



Protocol A3921120


**A PHASE 3, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, STUDY
OF THE EFFICACY AND SAFETY OF TOFACITINIB IN SUBJECTS WITH
ACTIVE ANKYLOSING SPONDYLITIS (AS)**

**Statistical Analysis Plan
(SAP)**

Version: 4

Date: 17 Dec 2019

TABLE OF CONTENTS

1. VERSION HISTORY	6
2. INTRODUCTION	14
2.1. Study Objectives and Estimands	14
2.1.1. Estimands for ASAS20 and ASAS40 at Week 16.....	15
2.1.2. Estimands for Continuous Secondary Endpoints	16
2.2. Study Design	17
3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS	19
3.1. Primary Endpoint	19
3.2. Secondary Endpoints.....	19
3.2.1. Key Secondary Endpoint	19
3.2.2. Other Secondary Endpoints	19
3.3. Tertiary/Exploratory Endpoints.....	22
3.4. Baseline Variables.....	22
3.4.1. Stratification Factor	22
3.5. Safety Endpoints	22
	23
4. ANALYSIS SETS	23
4.1. Full Analysis Set	23
4.2. Per Protocol Analysis Set.....	23
4.2.1. Protocol Deviations	24
4.2.2. Deviations Assessed Prior to Randomization.....	24
4.2.3. Deviations Assessed Post-Randomization.....	25
4.3. Safety Analysis Set.....	25
4.4. Other Analysis Sets	25
4.4.1. Endpoint Specific Analysis Sets.....	25
5. GENERAL METHODOLOGY AND CONVENTIONS.....	26
5.1. Hypotheses and Decision Rules	27
5.2. General Methods	28
5.2.1. Analyses for Binary Endpoints	29
5.2.1.1. Cochran-Mantel Haenszel (CMH) Approach.....	29

5.2.1.2. Generalized Marginal Model for Repeated Measures (GMMRM)	30
5.2.1.3. Tipping Point Analysis	31
5.2.2. Analyses for Continuous (Repeated) Data	32
5.2.3. Analyses for Continuous (Single Visit) Data	32
5.3. Methods to Manage Missing Data	32
6. ANALYSES AND SUMMARIES	33
6.1. Primary Endpoint: ASAS20 at Week 16	34
6.1.1. Primary Analysis	34
6.1.2. Supportive Analyses	34
6.2. Key Secondary Endpoint: ASAS40 at Week 16	37
6.3. Other Secondary Endpoints and Tertiary/Exploratory Endpoints	37
6.4. Subset (Subgroup) Analyses	40
6.4.1. Subgroup Analyses for Primary Endpoint of ASAS20 at Week 16	40
6.4.2. Subgroup Analyses for Key Secondary Endpoint of ASAS40 at Week 16	42
6.5. Baseline and Other Summaries and Analyses	42
6.5.1. Baseline Summaries	42
6.5.2. Prior and Baseline Treatments for Ankylosing Spondylitis	45
6.5.3. Concomitant Medications	45
6.6. Safety Summaries and Analyses	46
6.6.1. Adverse Events	46
6.6.2. Laboratory Data	46
7. INTERIM ANALYSES	47
8. REFERENCES	48
9. APPENDICES	50
9.1. Data Set Description: On-Drug vs. On-Study Data	50
9.2. Endpoint Definitions and Data Derivation Details	50
9.2.1. ASAS Improvement Criteria	50
9.2.2. Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) & BASDAI50	51
9.2.3. Bath Ankylosing Spondylitis Metrology Index (BASMI)	52
9.2.4. Ankylosing Spondylitis Disease Activity Score (ASDAS _{CRP})	53

9.2.5. Swollen Joint Count (44).....	54
9.2.6. Maastricht Ankylosing Spondylitis Enthesitis Score (MASES)	55
9.2.7. 36-Item Short Form Survey Version 2 (SF-36v2) Acute	55
9.2.8. EuroQol-5D Health Questionnaire – 3 Level Version (EQ-5D-3L)	61
9.2.9. Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F).....	62
9.2.10. Work Productivity and Activity Impairment Questionnaire: Spondyloarthritis (WPAI).....	63
9.3. Tipping Point Analysis for Longitudinal Binary Data	64
9.4. Definition and Use of Visit Windows in Reporting.....	66
9.5. Range of Values for Continuous Efficacy Endpoints	68
9.6. Lists of DMARDs and NSAIDs.....	70
9.7. Metabolic Syndrome and Diabetes Mellitus at Baseline	72
9.7.1. Criteria for Clinical Diagnosis of Metabolic Syndrome.....	72
9.7.2. Definition of Diabetes Mellitus at Baseline.....	74
9.8. Summary of Efficacy Analyses.....	75
9.9. Abbreviations Used in SAP.....	82
9.10. Definition of NSAID-Inadequate Response (NSAID-IR)	84
9.11. Derivation of the Stratification Factor	85

LIST OF TABLES

Table 1.	Summary of Changes.....	6
Table 2.	Endpoint Specific Analysis Sets.....	26
Table 3.	Mock Table of ASAS20 Response at Week 16 and Reasons for Study Drug Discontinuation Prior to Week 16 (Estimand 1, FAS, MR=NR, On-Drug Data).....	36
Table 4.	Mock Table of ASAS20 Response at Week 16 and Reasons for Study Discontinuation Prior to Week 16 (Estimand 2, FAS, MR=NR, On-Study Data).....	37
Table 5.	Visit Windowing.....	67
Table 6.	Numerical Characteristics of Select Continuous Efficacy Endpoints	68
Table 7.	List of DMARDs and NSAIDs.....	70
Table 8.	Summary of Efficacy Analyses	75
Table 9.	Abbreviations Used in SAP.....	82

LIST OF FIGURES

Figure 1. Schematic of Study Design.....18

1. VERSION HISTORY**Table 1. Summary of Changes**

Version, Date	Associated Protocol Amendment	Rationale	Specific Changes
1 (05 March 2018)	Original (22 February 2018)	First version.	None.
2 (21 September 2018)	1 (06 September 2018)	Revision to align with study protocol amendment 1 which was revised based on regulatory input on Type I error controlled secondary endpoints and mandate to follow up subjects after discontinuation of investigational products.	<ol style="list-style-type: none"> 1. Updated Section 2.1 to have study objectives aligned with protocol's study objectives. 2. Updated the study schematic and clarified the investigators, subjects and sponsor will remain blinded to the first 16 weeks of treatment through the entire duration of the trial until database release in Section 2.2. 3. Updated ASAS40 at Week 16 as key secondary endpoint, re-ordered the other secondary endpoints to have same order of hierarchy in Type I error controlled hypothesis testing in Section 3.2. 4. Clarified the derivation of prior treatment history as a randomization factor in Section 3.4.1. 5. Added prior exposure to non-TNFi bDMARD as a protocol deviation for exclusion from Per

			<p>Protocol Analysis Set in Section 4.2.2.</p> <ol style="list-style-type: none"> 6. Updated the terminology of “study treatment” to “investigational product” to be consistent with protocol in Section 4.2.3. 7. Removed the statements regarding EMA’s consideration of ASAS40 as primary endpoint in Section 5.1. 8. For global type I error controlled endpoints in Section 5.1, changed SF-36v2 Physical Functioning domain to Physical Component Summary and added ASQoL to the hierarchy of testing. Also clarified the step-down testing procedure. 9. Added descriptions of analyses of ASAS20 at Week 16 for subjects who discontinued investigational product or study early in Section 6.1.2. 10. Clarified/Added EQ-5D-3L Utility Score (UK) to the list of endpoints in Sections 3.2.2, 6.3 and 9.8 (Table 8) as an endpoint, provided method of its derivation in Section 9.2.8 and data range in Section 9.5.
--	--	--	---

			<ol style="list-style-type: none"> 11. Updated country list in defining subgroup of geographic region in Section 6.4; added waist circumference to the list of baseline characteristics in Section 6.5. 12. Updated list of references in Section 8. 13. Clarified the names of CRF forms referred in Section 9.1 (On/Off-Drug Data) and Section 9.7 (Baseline Metabolic Syndrome). 14. Clarified the use of CRF-calculated BASDAI and inflammation scores at Screening and Baseline visits in Section 9.2.2. 15. Provided clarification for windowing observations for subjects who discontinue from study prior to Week 48 in Section 9.4. 16. Added abbreviations to Table 9 in Section 9.9. 17. Corrected minor typographical errors throughout the SAP.
3 (23 May 2019)	2 (10 April 2019)	Revision to incorporate suggestions from US FDA and to align with changes in study protocol amendment 2.	<ol style="list-style-type: none"> 1. Clarified terminology in study objectives and added definitions of estimands in Section 2.1 according to suggestions from FDA. 2. Updated study design in Section 2.2 to describe

			<p>the Week 16 and Week 48 analyses and the expanded inclusion of subjects with prior bDMARD use without IR per protocol amendment 2.</p> <ol style="list-style-type: none"> 3. Updated the stratum's label of the stratification factor from "TNFi-IR" to "TNFi-IR or bDMARD Use (Non-IR)" throughout the document. Also updated the study schematic in Section 2.2. 4. Re-ordered the endpoints in Section 3 to align with the updated sequence of endpoints under Type I error control testing per protocol amendment 2. 5. Removed the exclusion criterion of the per protocol analysis set: "Subject who were exposed to prior non-TNFi biological DMARD" in Section 4.2.2 per expanded inclusion under protocol amendment 2. 6. Provided descriptions and scopes of Week 16 and Week 48 analyses for efficacy and safety in Section 5 and Section 6 per protocol amendment 2. 7. Updated Global Type I Error Control testing in Section 5.1: moved ΔASQoL up and added
--	--	--	---

			<p>ΔFACIT-F Total Score per protocol amendment 2.</p> <ol style="list-style-type: none"> 8. Clarified the application of statistical adjustment for binary endpoints when 0%/100% response rate observed in treatments by stratum in Section 5.2.1.1. 9. Updated the MMRM model for continuous endpoints including interaction terms of stratification factor and baseline value with visit in the model in Section 5.2.2. 10. Added the ANCOVA model in Section 5.2 to analyze data of a single visit for some continuous endpoints in the Week 16 analysis. 11. Clarified the methods of handling missing data under the estimands for both binary and continuous endpoints in Section 5.3. 12. Updated the supportive analyses for the primary endpoint of ASAS20 at Week 16, key secondary endpoint of ASAS40 at Week 16 and other secondary binary and continuous endpoints in Sections 6.1, 6.2 and 6.3 and aligned these analyses with the
--	--	--	--

			<p>estimand framework described in Section 2.1.</p> <p>13. Added subgroup analyses of 3-category prior treatment history and baseline AS disease activity (with literature reference); updated subgroup definition of geographic region in Section 6.4. Updated a few baseline disease characteristics in Section 6.5.</p> <p>14. Updated the definition of treatment-emergent adverse event to remove the restriction of worsening severity from baseline in Section 6.6.1 per Sponsor's new data standard.</p> <p>15. Clarified the derivation of Modified Schober's test score using CRF value for BASMI in Section 9.2.3.</p> <p>16. Re-formatted the descriptions of Tipping Point Analysis in Section 9.3 into a step-by-step format and pre-specified the number of multiple imputations and the values of MNAR adjustment (δ) according to suggestion from FDA.</p> <p>17. Updated the list of DMARDs and NSAIDs in Section 9.6, noting that</p>
--	--	--	--

			<p>this is not an exhaustive list.</p> <p>18. Updated the summary table of efficacy analyses in Section 9.8 and updated abbreviation in Section 9.9.</p> <p>19. Added definition of NSAID inadequate response (NSAID-IR) in Appendix Section 9.10 and derivation of stratification factor in Appendix Section 9.11.</p>
4 (17 Dec 2019)	2 (10 April 2019)	Revision to provide additional clarification after study's blinded data review.	<p>1. Corrected typo to remove Week 2 for endpoint of MASES as it is not assessed at Week 2 per protocol (Section 3.2.2, Section 6.3, Section 9.8 Table 8);</p> <p>2. Clarified Tier-2 adverse events in Section 3.5;</p> <p>3. Clarified the handling of a case when a subject has an assessment prior to or on the date of the investigational product discontinuation in the same visit window as the discontinuation (see Section 9.4, Table 5), this assessment will be considered as on-drug and used in the analysis for that visit window (Section 5.3);</p> <p>4. Removed adverse event analyses for the incremental period: Week 16 to Week 48 as</p>

			<p>analyses for Baseline to Week 16 and Baseline to Week 48 are sufficient (Section 6.6.1);</p> <p>5. Clarified the handling of LLOQ for hsCRP is applicable to all efficacy and safety analyses, not just ASDAS_{CRP} (Section 9.2.4);</p> <p>6. Clarified the treatment switch date as the start date from the first entry of dosing log for the period between Week 16 to Week 24 (Section 9.4, Table 5);</p> <p>7. Removed Metamizole Sodium from the NSAID list as it is confirmed as an analgesic (Section 9.6, Table 7);</p> <p>8. Corrected typographical errors throughout the SAP.</p>
--	--	--	--

2. INTRODUCTION

Ankylosing Spondylitis (AS) is prevalent among spondyloarthropathies, a group of arthritic conditions affecting the spine. This under recognized disease is often not diagnosed for many years and typically presents in people between 20 and 40 years of age leading to progressive disability and adverse effects on health-related quality of life. Tofacitinib inhibits signaling of cytokines that are integral to lymphocyte activation, proliferation, and function and may thus result in suppression of multiple aspects of the immune response. This forms the basis of the rationale to investigate the effect of tofacitinib in active AS. This study is a follow-up to the A3921119 phase 2b study of tofacitinib in active AS.

2.1. Study Objectives and Estimands

Primary Objective

1. To compare the efficacy of tofacitinib 5 mg BID versus placebo on the ASAS20 response rate at Week 16 in subjects with active AS that have had an inadequate response to previous treatment.


Key Secondary Objective

1. To compare the efficacy of tofacitinib 5 mg BID versus placebo on the ASAS40 response rate at Week 16 in subjects with active AS that have had an inadequate response to previous treatment.

Other Secondary Objectives

1. To compare the safety and tolerability of tofacitinib 5 mg BID versus placebo in subjects with active AS that have had an inadequate response to previous treatment.
2. To compare the efficacy (including health-related quality of life, function, pain, and fatigue) of tofacitinib 5 mg BID versus placebo at all time points in subjects with active AS that have had an inadequate response to previous treatment.

Tertiary/Exploratory Objectives

1. 
2. To evaluate the effect of tofacitinib 5 mg BID on lymphocyte subsets using FACS analysis.
3. To measure the effect of tofacitinib 5 mg BID on healthcare resource utilization at all collected time points.

2.1.1. Estimands for ASAS20 and ASAS40 at Week 16

Only discontinuation of the investigational product will be considered as an intercurrent event to define the estimands for this study. There are three estimands for the primary endpoint of ASAS20 at Week 16.

Estimand 1:

The first estimand of ASAS20 at Week 16 is a composite estimand that accounts for both treatment adherence and response. A responder is defined as having a response without discontinuation of the investigational product prior to Week 16. This is also called the Primary Estimand defined according to the primary objective. It includes the following four attributes:

- Population: Subjects who have active AS;
- Variable: ASAS20 response at Week 16. A subject after an intercurrent event of discontinuation of investigational product will be considered a non-responder for the visit of interest;
- Intercurrent event: The intercurrent event of discontinuation of the investigational product prior to Week 16 is captured through the definition of the composite variable. Data collected after discontinuation of the investigational product prior to Week 16 (ie, off-drug data) are not included to derive the response, ie, only On-Drug data are used (see [Section 9.1](#));
- Population-level summary: The difference in ASAS20 response rates between tofacitinib 5 mg BID and placebo at Week 16.

Estimand 2:

The second estimand of ASAS20 at Week 16 is supportive to Estimand 1 and is a treatment policy estimand. It estimates the effect regardless of treatment adherence. It includes the following four attributes:

- Population: Subjects who have active AS;
- Variable: ASAS20 response at Week 16;
- Intercurrent event: The intercurrent event of discontinuation of the investigational product prior to Week 16 is not considered here for data exclusion. Data collected after discontinuation of the investigational product prior to Week 16 (ie, off-drug data) are also included to derive the response, ie, On-Study data are used (see [Section 9.1](#));
- Population-level summary: The difference in ASAS20 response rates between tofacitinib 5 mg BID and placebo at Week 16.

The difference between Estimand 1 and Estimand 2 is that Estimand 2 disregards treatment adherence and includes the additional data collected after the intercurrent event of discontinuation of the investigational product, ie, On-Study data. (see [Section 9.1](#)).

Estimand 3:

The third estimand of ASAS20 at Week 16 is supportive to Estimand 1 and is a hypothetical estimand. It estimates the treatment effect as if the intercurrent event of discontinuation of investigational product prior to Week 16 has not occurred. This includes the following four attributes:

- Population: Subjects who have active AS;
- Variable: ASAS20 response at Week 16;
- Intercurrent event: Only data collected before the intercurrent event of discontinuation of the investigational product prior to Week 16 are included to derive the response, ie, On-Drug data (see [Section 9.1](#));
- Population-level summary: The odds ratio of ASAS20 response between tofacitinib 5 mg BID and placebo at Week 16.

The main difference between Estimand 1 and Estimand 3 is that Estimand 3 assumes the intercurrent event of discontinuation of investigational product prior to Week 16 has not occurred, while Estimand 1 considers the response after discontinuation of investigational product as non-response via the composite strategy. Also, the population-level summary in Estimand 3 is an odds ratio instead of difference in response rates as in Estimand 1.

The same three estimand definitions for ASAS20 at Week 16 described above are also applicable to ASAS40 at Week 16 with ASAS20 substituted by ASAS40 in the definitions. Specifically, Estimand 1 for ASAS40 at Week 16 is called the Key Secondary Estimand, defined according to the key secondary objective. Estimand 1 will be used for other binary secondary endpoints by appropriately substituting the endpoint.

2.1.2. Estimands for Continuous Secondary Endpoints

Only discontinuation of the investigational product will be considered as an intercurrent event to define the estimands for this study. The endpoint of change from baseline (Δ) in Patient Global Assessment of Disease (PGA) will be used as an example in defining the estimands. Estimand 4 will be used for other continuous secondary endpoints by appropriately substituting the endpoint. Estimand 5 will be used only for the Type I error controlled continuous secondary endpoints as supportive analyses.

Estimand 4:

This estimand of Δ PGA at Week 16 is a hypothetical estimand. It estimates the treatment effect as if the intercurrent event of discontinuation of investigational product prior to Week 16 has not occurred. This includes the following four attributes:

- Population: Subjects who have active AS;
- Variable: Δ PGA at Week 16;
- Intercurrent event: Only data collected before the intercurrent event of discontinuation of the investigational product prior to Week 16 are included, ie, On-Drug data (see [Section 9.1](#));
- Population-level summary: The mean difference in Δ PGA between tofacitinib 5 mg BID and placebo at Week 16.

Estimand 5:

This estimand of Δ PGA at Week 16 is supportive to Estimand 4 and is a treatment policy estimand. It estimates the effect regardless of treatment adherence. It includes the following four attributes:

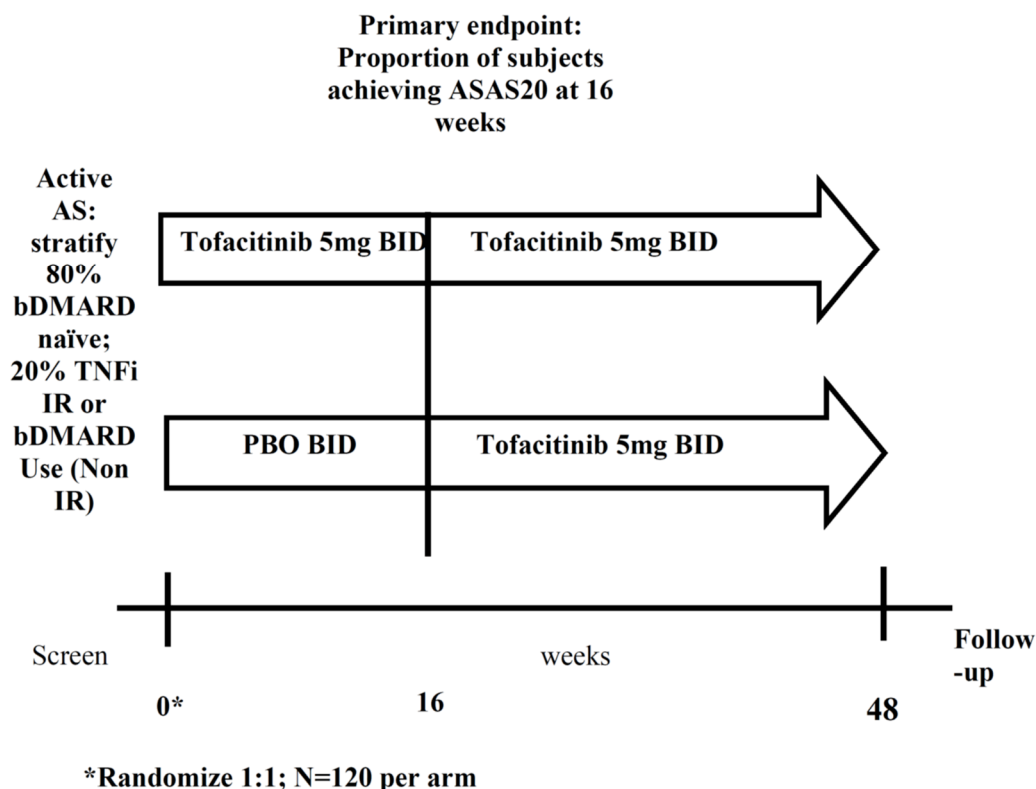
- Population: Subjects who have active AS;
- Variable: Δ PGA at Week 16;
- Intercurrent event: The intercurrent event of discontinuation of the investigational product prior to Week 16 is not considered here for data exclusion. Data collected after the discontinuation of the investigational product prior to Week 16 (ie, off-drug data) are also included, ie, On-Study data are used (see [Section 9.1](#));
- Population-level summary: The mean difference in Δ PGA between tofacitinib 5 mg BID and placebo at Week 16.

The difference between Estimand 5 and Estimand 4 is that Estimand 5 disregards treatment adherence and includes the additional data collected after the intercurrent event of discontinuation of the investigational product prior to Week 16, ie, On-Study data are used. (see [Section 9.1](#)).

2.2. Study Design

This is a phase 3, multicenter, randomized, double-blind, placebo-controlled efficacy and safety study designed to compare tofacitinib to placebo in subjects with active AS. An estimate of approximately 480 AS subjects will be screened globally in order that approximately 240 eligible subjects (120 per arm) will be randomized in a 1:1 ratio to tofacitinib 5 mg BID (twice daily) or matching placebo BID. [Figure 1](#) shows the schematic of the study design.

Figure 1. Schematic of Study Design



Abbreviations: AS = Ankylosing Spondylitis; ASAS20 = Assessment of SpondyloArthritis International Society 20; bDMARD = biological disease-modifying antirheumatic drug; BID = twice daily; IR = inadequate responder; N = number of subjects; PBO = placebo; TNFi = tumor necrosis factor inhibitor.

The duration of participation for eligible subjects will be approximately 56 weeks. This will include a screening period of approximately 30 days, a 16-week double-blind treatment period, a 32-week open-label treatment period and a 28-day follow-up period. During the 16-week treatment period subjects will visit the clinic every two weeks (± 3 days) until the Week 4 visit and then every 4 weeks (± 7 days) until the completion of Week 16. At the Week 16 visit all subjects will be assigned to open-label tofacitinib and will visit the clinic every two months (± 7 days) until Week 48. Subjects will then return to the clinic for a Follow-up visit approximately 28 days after the Week 48 visit. There will be a total of 2 planned analyses conducted when all applicable subjects have completed their Week 16 and Week 48 (including follow-up) visits, respectively. The first analysis will be conducted when all applicable subjects have completed their Week 16 visit. The investigators, subjects, and sponsor study team will remain blinded to the first 16 weeks of treatment assignment through the entire duration of trial until database release. Randomization will be stratified by prior treatment history: (1) bDMARD-naïve and (2) Tumor Necrosis Factor inhibitor-inadequate responder (TNFi-IR) or bDMARD Use (Non-IR). Subjects who had prior bDMARD use (Non-IR) will be eligible to participate in the study after washout and will be included in the "TNFi-IR or bDMARD Use (Non-IR)" stratum. Approximately 80% of the subjects will be bDMARD-naïve (designed as Stratum "bDMARD-naïve"), and the other approximately

20% of subjects who had an inadequate response to at least one but not more than 2 TNF inhibitors or who had prior bDMARD use (Non-IR) (designed as Stratum "TNFi-IR or bDMARD Use [Non-IR]"). All subjects will have an inadequate response to at least 2 NSAIDs. An inadequate response to NSAID or TNFi treatment is defined as having a treatment related adverse event or lack of response to NSAID or TNFi treatment that was administered in accordance with its labeling recommendations.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoint

- ASAS20 response at Week 16 (see [Section 9.2.1](#)).

3.2. Secondary Endpoints

3.2.1. Key Secondary Endpoint

- ASAS40 response at Week 16 (see [Section 9.2.1](#)).

3.2.2. Other Secondary Endpoints

- ASAS20 response at all other time points (Weeks 2, 4, 8, 12, 24, 32, 40, and 48) (see [Section 9.2.1](#));
- ASAS40 response at all other time points (Weeks 2, 4, 8, 12, 24, 32, 40, and 48) (see [Section 9.2.1](#));
- Change from Baseline (Δ) in Ankylosing Spondylitis Disease Activity Score using C-Reactive Protein (Δ ASDAS_{CRP}) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.4](#));
- Δ in High Sensitivity C-Reactive Protein (Δ hsCRP, mg/L) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ in Ankylosing Spondylitis Quality of Life (Δ ASQoL) at Weeks 16 & 48;

A total score is calculated by summing the 18 items. The total score ranges from 0 to 18, with higher values indicating more impaired health-related quality of life.

- Δ in 36-Item Short Form Health Survey Version 2 (Δ SF-36v2), Acute at Weeks 16 & 48;

This survey yields 10 endpoints: 8 general health domains: physical functioning, role limitations due to physical health, bodily pain, general health perceptions, vitality, social functioning, role limitations due to emotional problems, and mental health. These domains can also be summarized as physical and mental component summary scores (see [Section 9.2.7](#)).

- Δ in Bath Ankylosing Spondylitis Metrology Index (Δ BASMI) and its 5 components (Δ in Lateral Spinal Flexion, Tragus-to-Wall Distance, Lumbar Flexion [as measured by Modified Schober's test score], Maximal Intermalleolar Distance, and Cervical Rotation) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.3](#));
- Δ in Functional Assessment of Chronic Illness Therapy-Fatigue (Δ FACIT-F) (Total Score, Experience Domain Score, and Impact Domain Score) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);

This instrument has 13 items that assess fatigue. Instrument scoring yields a range from 0 to 52 for Total Score, with higher scores representing better subject status (less fatigue). For Experience Domain Score, range is 0-20; for Impact Domain Score, range is 0-32 (see [Section 9.2.9](#)).

- Δ in Patient Global Assessment of Disease (Δ PGA) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48). PGA is one of the 4 ASAS20/ASAS40 components;
- Δ in Patient Assessment of Spinal Pain (two numerical rating scales [NRS]: Δ in Nocturnal Spinal Pain and Total Back Pain, respectively) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48). Total Back Pain is one of the 4 ASAS20/ASAS40 components;
- Δ in Bath Ankylosing Spondylitis Functional Index (Δ BASFI) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);

This instrument consists of 10 questions. BASFI is the average of these 10 scores and it ranges from 0 to 10, with higher scores indicating greater functional limitation. If the number of missing scores is ≤ 3 , then BASFI is calculated based on the average of the remaining non-missing scores, otherwise if the number of missing scores is >3 , then BASFI is considered missing (Ramiro et al., 2014).¹ BASFI is one of the 4 ASAS20/ASAS40 components.

- Δ in Inflammation at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);

Inflammation is one of the 4 ASAS20/ASAS40 components, which is the average of the answers to questions 5 & 6 of BASDAI (see [Section 9.2.2](#)). If answer to one of the two questions is missing, the other non-missing answer will be used as inflammation. If both answers are missing, inflammation is considered missing (Ramiro et al., 2014).¹

- ASAS 5/6 response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.1](#));
- ASAS Partial Remission response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.1](#));

- Δ in Bath Ankylosing Spondylitis Disease Activity Index (Δ BASDAI) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.2](#));
- BASDAI50 response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.2](#));
- ASDAS_{CRP} Clinically Important Improvement response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.4](#));
- ASDAS_{CRP} Major Improvement response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.4](#));
- ASDAS_{CRP} Inactive Disease response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.4](#));
- Δ in Maastricht Ankylosing Spondylitis Enthesitis Score (Δ MASES) at all time points (Weeks 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.6](#));
- Δ in extra-articular involvement (Specific Medical History and peripheral articular involvement [as assessed by Swollen Joint Count (44) (Δ SJC[44])]) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) (see [Section 9.2.5](#));
- Δ in Spinal Mobility: Chest Expansion at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);

The chest expansion (cm) is measured as the difference between maximal inspiration and expiration. Two attempts will be performed and the better (ie, larger) of the two attempts will be utilized for data analysis.

- Δ in EuroQol-5D Health State Profile – 3 Level Version (Δ EQ-5D-3L) (5 Dimensions: Mobility, Self-Care, Usual Activities, Pain/Discomfort and Anxiety/Depression, “Your own health state today” [Δ EQ-VAS], and EQ-5D-3L Utility score based on United Kingdom (UK) weights) at Weeks 16 & 48 (see [Section 9.2.8](#));
- Δ in Work Productivity & Activity Impairment (Δ WPAI) Questionnaire: Spondyloarthritis (4 subscale scores in range of 0-100%) at Weeks 16 & 48 (see [Section 9.2.10](#)).

The 4 subscales are (1) percent work time missed due to health problem, (2) percent impairment while working due to health problem, (3) percent overall work impairment due to health problem, and (4) percent activity impairment due to health problem.

3.3. Tertiary/Exploratory Endpoints

- [REDACTED]
- Fluorescence Activated Cell Sorting (FACS) analysis of lymphocyte subsets;
- Ankylosing Spondylitis HealthCare Resource Utilization Questionnaire (AS-HCRU and Δ AS-HCRU Self-Rating of Job Performance at Work which is question 14 with range of 0-10) at Weeks 16 & 48.

3.4. Baseline Variables

Baseline value of an endpoint is the value obtained on Day 1 (ie, Baseline visit) or prior before the first dose of the investigational product. A subject's prior treatment history at randomization, defined by two categories: "bDMARD-naïve" and "TNFi-IR or bDMARD Use (Non-IR)" (see [Section 3.4.1](#)), which is used as a factor for stratified randomization, is described in more detail below and also in [Appendix Section 9.11](#).

3.4.1. Stratification Factor

Subjects will be stratified by prior treatment history (strata of "bDMARD-naïve" and "TNFi-IR or bDMARD Use [Non-IR]") at randomization. The stratum derived from data collected on prior medications taken and on the reasons for their discontinuation in the clinical database will be used for analysis. Specifically, prior medications taken before Day 1 collected on the Prior and Concomitant Medications DMARD CRF pages (bDMARDs subcategory) will be used to derive the strata. An inadequate response to a medication is defined when its discontinuation was due to either adverse event or lack of efficacy (see [Appendix Section 9.11](#) for additional details to derive this stratification factor). For the primary endpoint of ASAS20 at Week 16 and the key secondary endpoint of ASAS40 at Week 16 only, a sensitivity analysis may need to be performed using the stratum from the randomization system (if there are differences between that from the randomization system and that derived from the clinical database) (see [Section 6.1.2](#) and [Section 6.2](#)).

3.5. Safety Endpoints

- Incidence and severity of Adverse Events (AE);
- Clinical laboratory tests, vital signs, physical examination and 12-lead ECG parameters;
- Fluorescence Activated Cell Sorting (FACS) analysis of lymphocyte subsets.

In addition to standard safety displays, a 3-tier approach will be used to summarize AEs. Under this approach, AEs are classified into 1 of 3 tiers.

- Tier-1 events: These are pre-specified events of clinical importance and are maintained in a list in the product's Safety Surveillance Review Plan. This is a set of AEs of special interest that has been identified for the tofacitinib compound.

- Tier-2 events: These are events that are not Tier-1 but are “common” (ie, occurring ≥ 4 subjects in any treatment group). A Medical Dictionary for Regulatory Activities (MedDRA) Preferred Term (PT) is defined as a Tier-2 event if there are at least 4 subjects in any treatment group reporting an event.
- Tier-3 events: These are events that are neither Tier-1 nor Tier-2 events.

The 3 tiers are mutually exclusive. Tier-3 events will be included in standard safety displays and not separately displayed in specific Tier-3 tables.



4. ANALYSIS SETS

Below are descriptions of the Analysis Sets defined for this study.

4.1. Full Analysis Set

The full analysis set (FAS) will include all subjects who were randomized to the study and received at least one dose of the randomized investigational product (ie, tofacitinib or placebo). The primary efficacy population for this study is defined by the FAS of subjects. Subjects will be analyzed in the treatment groups as they are randomized. If a subject is treated but not randomized, then the subject will be excluded from the FAS. A narrative will be provided for this subject in the clinical study report (CSR).

Continuous and ordered-categorical (analyzed as continuous) endpoints for which change or percent change from Baseline is the measure to be analyzed, would require that a subject has a Baseline value and at least one post-baseline value to be included in the FAS for that endpoint. It is anticipated that there should be little bias due to excluding subjects with no Baseline value since the missingness mechanism is likely to be missing completely at random (MCAR) (Rubin, 1987).² Excluding a subject with no post-baseline would likely cause a more favorable estimate of the mean change from Baseline for that endpoint for the treatment group in which that subject was assigned. However, it is anticipated that very few subjects will be excluded for this reason minimizing any potential bias.

4.2. Per Protocol Analysis Set

The Per-Protocol (PP) analysis set will exclude all subjects from the FAS who had a protocol deviation (see [Section 4.2.1](#) below) thought to have a material impact on the primary efficacy analysis. The PP analysis set will be used in conjunction with Estimand 1 as a supportive analysis for the primary endpoint of ASAS20 at Week 16 and the key secondary endpoint of ASAS40 at Week 16 (see [Section 2.1.1](#)).

4.2.1. Protocol Deviations

Only protocol deviations that are thought to affect the primary efficacy endpoint of ASAS20 at Week 16 and the key secondary endpoint of ASAS40 at Week 16 will be considered in defining the PP analysis set. The following sections describe protocol deviations that will lead to subject exclusion from the PP analysis set. It is possible that unexpected deviations will arise, becoming known only after the study has been active for a long period of time; hence more deviations may be added to the list at a later date. As of this writing, the protocol deviations that will define the PP analysis set can all be found in [Section 4.2.2](#) and [Section 4.2.3](#) below.

The list of subjects along with the protocol deviation that excludes them from the PP analysis set will be finalized prior to the treatment unblinding for the Week 16 analysis and put into the Trial Master File (TMF).

4.2.2. Deviations Assessed Prior to Randomization

Exceptions to the inclusion or exclusion criteria are not expected to occur; but any subject who fails meeting any of the following protocol inclusion criteria will be excluded from the PP analysis set.

- Subjects who failed to meet the Modified New York Criteria for Ankylosing Spondylitis (1984) at screening. The AS diagnosis at screening is recorded on the Case Report Form (CRF) page for primary diagnosis. A subject who did not have this CRF page filled out will be considered having this exclusion criterion for PP analysis set met.
- Subjects who had BASDAI score <4 or back pain score (BASDAI Question 2) <4 assessed at either screening or Baseline (ie, Day 1) visit. The BASDAI CRF pages at screening and Baseline will be used to check this condition.
- Subject who failed to meet the criteria of at least 2 NSAIDs-IR (see definition of inadequate response to NSAID in [Appendix Section 9.10](#)).
- Subjects who failed to meet the criteria of:
 - Being naïve to prior bDMARD, or
 - Having an inadequate clinical response or intolerance due to treatment-related AE to at least 1, but not more than 2 approved TNFi's (ie, 1 or 2 TNFi-IR) or having bDMARD Use (Non-IR).

[Appendix Section 9.11](#) provides more details on how to check these criteria and to derive the stratification factor.

Full lists of DMARDs and NSAIDs can be found in [Section 9.6](#).

4.2.3. Deviations Assessed Post-Randomization

These specific classes of protocol deviations will be assessed post randomization up to Week 16 only:

- Subjects who were randomized but took or received incorrect treatment other than the randomized treatment for the entire duration from Baseline through Week 16. This will be checked manually since there is no CRF page for this condition.
- Subjects who had <80% compliance on the investigational product's tablets on 2 consecutive visits (ie, 2 dosing periods) for the period from Baseline through Week 16. The investigational product compliance (%) will be derived from the total number of doses actually taken divided by the total number of doses expected to take per dosing period as recorded in Oral Dosing CRF page. Both the randomized treatments (ie, tofacitinib and placebo) will be taken into account.
- Subjects who were dosed with rescue medication for AS on the preceding day or the same day of the primary endpoint assessment at the Week 16 study visit. The rescue medication is recorded on the Concomitant Medications – Rescue Medications CRF page.

Each subject's presence on the list of exclusion from the PP analysis set means that there was at least one deviation for that subject.

4.3. Safety Analysis Set

The safety analysis set (SAFETY) will include all subjects who were randomized and received at least one dose of the investigational product (ie, tofacitinib or placebo). Subjects will be analyzed in the treatment groups as they received. If a subject is treated but not randomized, then the subject will be excluded from the SAFETY analysis set. A narrative will be provided for this subject in the CSR.

4.4. Other Analysis Sets

4.4.1. Endpoint Specific Analysis Sets

Subjects will be excluded from the FAS for analyzing a specific endpoint if the criterion for inclusion is not met for the endpoint.

Table 2. Endpoint Specific Analysis Sets

Endpoint	Inclusion	Rationale
ASDAS _{CRP} Clinically Important Improvement (improvement [decrease] from Baseline in ASDAS _{CRP} ≥ 1.1 units)	Include subjects with Baseline ASDAS _{CRP} ≥ 1.1 units in FAS	Restrict Baseline ASDAS _{CRP} ≥ 1.1 units to allow room for response
ASDAS _{CRP} Major Improvement (improvement [decrease] from Baseline in ASDAS _{CRP} ≥ 2.0 units)	Include subjects with Baseline ASDAS _{CRP} ≥ 2.0 units in FAS	Restrict Baseline ASDAS _{CRP} ≥ 2.0 units to allow room for response
ASDAS _{CRP} Inactive Disease (ie, ASDAS _{CRP} < 1.3 units)	Include subjects with Baseline ASDAS _{CRP} ≥ 1.3 units in FAS	Restrict Baseline ASDAS _{CRP} ≥ 1.3 units to allow room for response
Δ in Swollen Joint Count (44)	Include subjects with Baseline swollen joint count (44) > 0 in FAS	Having swollen joints is not an enrollment criterion, it is expected only a subset of subjects will have swollen joints at Baseline.
Δ MASES	Include subjects with Baseline MASES > 0 in FAS	Enthesitis is not an enrollment criterion, it is expected only a subset of subjects will have enthesitis at Baseline.
Abbreviations: Δ = change from Baseline, ASDAS _{CRP} = Ankylosing Spondylitis Disease Activity Score - C-Reactive Protein, FAS = Full Analysis Set, MASES = Maastricht Ankylosing Spondylitis Enthesitis Score.		

5. GENERAL METHODOLOGY AND CONVENTIONS

There will be a total of 2 planned analyses conducted when all applicable subjects have completed their Week 16 and Week 48 (including follow-up) visits, respectively. The investigators, subjects and sponsor study team will remain blinded to the first 16 weeks of treatment assignment through the entire duration of the trial until database release at Week 48. The Week 16 analysis will include all efficacy data through Week 16 and the safety data through the data cutoff date (for those subjects who have passed the Week 16 visit). As the primary endpoint (ASAS20), the key secondary endpoint (ASAS40) and the other Type I error controlled secondary endpoints are at Week 16, there will be no additional adjustment made for Type I error rate at the final analysis at Week 48. The efficacy analysis results through Week 16 obtained from the Week 16 analysis will be considered final and definitive. CCI

The Week 48 analysis will be conducted when all the applicable subjects have completed their Week 48 (including follow-up) visit and the database is released. All the Week 48 analysis results will be secondary in nature. The Week 48 analysis will contain results for earlier visits prior to or at Week 16 including those for the primary endpoint and the key secondary endpoint; however, they will serve as a sensitivity analysis only to ensure there are

no major changes to the definitive results for the primary endpoint and the key secondary endpoint obtained at the Week 16 analysis.

5.1. Hypotheses and Decision Rules

This protocol is designed to establish the superiority of tofacitinib 5 mg BID to placebo for the treatment of active AS based on the primary endpoint of ASAS20 at Week 16 in subjects who have had an inadequate response to previous treatments.

All statistical tests will be conducted at the 2-sided 5% (or equivalently 1-sided 2.5%) significance level for comparing tofacitinib 5 mg BID to placebo. Type I error will be controlled at 2-sided 5% or equivalently 1-sided 2.5%.

For the primary endpoint of ASAS20 at Week 16, if the 2-sided p-value is $\leq 5\%$, the superiority of tofacitinib 5 mg BID to placebo will be declared for this primary endpoint and the primary objective of this study is met.

Hypothesis testing will continue to the global family of select set of secondary endpoints stated below [Global Type I Error Control (for the primary endpoint and a select set of secondary endpoints)]. For each of the endpoints within this family, superiority of tofacitinib 5 mg BID to placebo will be declared if statistical significance is achieved under the step-down testing procedure.

In addition, three other families of hypotheses, ASAS family endpoints (Type I Error Control for Endpoints in the ASAS Family), ASAS20's earlier time points (Type I Error Control for ASAS20 at Earlier Time Points) and ASAS40's earlier time points (Type I Error Control for ASAS40 at Earlier Time Points) will also be tested. For each of the endpoints or time points of ASAS20/ASAS40, superiority of tofacitinib 5 mg BID to placebo will be declared if statistical significance is achieved under its respective step-down testing procedure.

Global Type I Error Control (for the primary endpoint and a select set of secondary endpoints): The family-wise Type I error rate will be controlled at the 2-sided 5% (or equivalently 1-sided 2.5%) significance level using a step-down testing procedure for the primary endpoint of ASAS20 at Week 16, the key secondary endpoint of ASAS40 at Week 16, and a select set of secondary endpoints at Week 16 tested in the sequence below: ASAS20, ASAS40, Δ ASDAS_{CRP}, Δ hsCRP, Δ ASQoL, Δ SF-36v2 Physical Component Summary (PCS), Δ BASMI, and Δ FACIT-F Total Score. When an endpoint fails to declare statistical significance, this endpoint and the remaining endpoints lower in the hierarchy will be considered non-significant. The rationale for the selection and ordering of the select set of secondary endpoints are clinical importance, precedence and likelihood of statistical success based on the results of the A3921119 study.

Type I Error Control for Endpoints in the ASAS Family: Δ PGA, Δ in total back pain, Δ BASFI, and Δ in inflammation (average of questions 5 and 6 of BASDAI) are the 4 ASAS components used in deriving ASAS20, thus they are considered belonging to the ASAS family of endpoints. A step-down testing procedure will be applied to them and tested in the sequence below: ASAS20, Δ PGA, Δ in total back pain, Δ BASFI, and Δ in inflammation (average of questions 5 and 6 of BASDAI) at Week 16. When an endpoint fails to declare

statistical significance, this endpoint and the remaining endpoints lower in the hierarchy will be considered non-significant. Though this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the family of ASAS endpoints.

Type I Error Control for ASAS20 at Earlier Time Points: In order to be more rigorous about establishing the onset of efficacy as measured by ASAS20 at the earliest time point at which there is statistical separation between tofacitinib 5 mg BID and placebo, a step-down approach with the ASAS20 from Week 16 to earlier time points (order of testing: Weeks 16, 12, 8, 4, and 2) will also be used for each time point. Though this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the family of ASAS20 time points.

Type I Error Control for ASAS40 at Earlier Time Points: In order to be more rigorous about establishing the onset of efficacy as measured by ASAS40 at the earliest time point at which there is statistical separation between tofacitinib 5 mg BID and placebo, a step-down approach with the ASAS40 from Week 16 to earlier time points (order of testing: Weeks 16, 12, 8, 4, and 2) will also be used for each time point. Though this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the family of ASAS40 time points.

5.2. General Methods

As subjects randomized to treatment group of placebo → tofacitinib 5 mg BID (ie, → means switching to) will advance from placebo to open-label treatment of tofacitinib 5 mg BID at Week 16, the treatment label used for reporting visits up to Week 16 will be “Placebo” and treatment label for reporting visits after Week 16 through Week 48 will be “Placebo → Tofacitinib 5 mg BID”. For analyses through Week 16, the following treatment comparison will be made at each time point, where applicable:

- Tofacitinib 5 mg BID vs. Placebo.

For analyses after Week 16 through Week 48, the following treatment comparison will be made at each time point,

- Tofacitinib 5 mg BID vs. Placebo → Tofacitinib 5 mg BID.

The primary efficacy comparison for the primary endpoint of ASAS20 will be between tofacitinib 5 mg BID and placebo at Week 16.

Descriptive Summaries

In general, the data for all continuous endpoints will be summarized by treatment group (ie, 2 groups) and by time point in tables containing descriptive statistics (N [which is the number of subjects evaluable for the endpoint at the time point], mean, standard deviation, standard error of the mean, minimum, 1st, 2nd (ie, median) and 3rd quartiles and maximum) for actual and change from Baseline (or percent change from Baseline) values for those endpoints measured at Baseline. In case when N=1 (ie, only one subject is

available/evaluable for summary), standard deviation and standard error will be reported as “NA” (not applicable) while the remaining statistical parameters will have the same value as the mean. The data for all response-type endpoints will be summarized by treatment group and by time point in tables showing descriptive statistics: N, n (ie, number of responders), response rate (%), standard error of the response rate, and 95% confidence interval (CI) based on normal approximation. If \hat{p} is the estimated response rate, then the 95% CI is calculated as:

$$\hat{p} \pm z_{0.975} \sqrt{\frac{\hat{p}(1 - \hat{p})}{N}}$$

where $z_{0.975}$ is the 97.5th percentile of the standard normal distribution and N is the number of subjects evaluable for the endpoint at the time point. If the lower bound is calculated to be negative, it will be set to 0%; if the upper bound is calculated to be larger than 100%, it will be set to 100%. In case when response rate is 0 or 100%, standard error will be reported as “NA” and the 95% CI bounds will be the same as the response rate. The displays described above for continuous and response-type endpoints will only use available data with no imputation. Therefore, the calculation of response rates will use the number of evaluable subjects as denominators.

5.2.1. Analyses for Binary Endpoints

5.2.1.1. Cochran-Mantel Haenszel (CMH) Approach

For a single time point (eg, ASAS20 response rate at Week 16), difference in response proportion between tofacitinib 5 mg BID and placebo groups will be estimated using the Cochran-Mantel Haenszel (CMH) approach adjusting for stratification factor of prior treatment history (strata of "bDMARD-naïve" and "TNFi-IR or bDMARD Use [Non-IR]") (Cochran, 1954; Mantel and Haenszel, 1959).^{3,4} Large sample approximation will be used for testing the superiority of tofacitinib 5 mg BID to placebo at Week 16 and prior time points and for forming 95% CI's and calculating p-values. This approach will be referred to as CMH.

Explicitly, let \hat{p}_{Ai} and \hat{p}_{Bi} represent the estimated response rates for tofacitinib 5 mg BID (A) and placebo (B), respectively, of the i^{th} stratum. Using the CMH weights, the weighted difference in proportions between tofacitinib 5 mg BID and placebo groups, d, is expressed by:

$$d = \sum_i w_i (\hat{p}_{Ai} - \hat{p}_{Bi})$$

where $w_i = \frac{(n_{Ai}n_{Bi})/(n_{Ai}+n_{Bi})}{\sum_i (n_{Ai}n_{Bi})/(n_{Ai}+n_{Bi})}$, is the CMH weight for the i^{th} stratum, with n_{Ai} and n_{Bi}

equal the numbers of subjects in the tofacitinib 5 mg BID and placebo groups, respectively, per stratum. The weights are normalized such that the sum across strata adds up to 1.

The variance of d is:

$$\text{var}(d) = \sum_i w_i^2 \text{var}(\hat{p}_{Ai} - \hat{p}_{Bi}),$$

$$\text{with } \text{var}(\hat{p}_{Ai} - \hat{p}_{Bi}) = \frac{\hat{p}_{Ai}(1 - \hat{p}_{Ai})}{n_{Ai}} + \frac{\hat{p}_{Bi}(1 - \hat{p}_{Bi})}{n_{Bi}}.$$

Two-sided 95% CI will be estimated using the normal approximation to the binomial distribution via the following method,

$$d \pm [z_{1-\alpha/2} \sqrt{\text{var}(d)}].$$

Two-sided p-value for the test of the 0 difference between tofacitinib 5 mg BID and placebo groups will be calculated as:

$$p = 2(1 - \Phi(|Z|)),$$

where $\Phi(\cdot)$ is the Gaussian cumulative density function and $Z = \frac{d}{\sqrt{\text{var}(d)}}$ is the Normal

Z-test statistic.

If there is no (ie, 0) response or 100% response in any one or both of the two treatment groups for the comparison in a stratum, eg, tofacitinib 5 mg BID vs. placebo, when calculating the proportions above, 0.5 will be added to the number of responses (ie, numerator) and 1 will be added to the denominator in each treatment in that stratum corresponding to the pair of comparison for calculating the treatment difference, standard error (ie, $\sqrt{\text{var}(d)}$), 95% CI and 2-sided p-value (Agresti, 2002).⁵

When response rate of 0% or 100% is observed in both treatments in comparison and in both strata, no formal comparison will be performed. Estimated response rate of 0% or 100% will be reported as observed. Standard error will be reported as 0.

The final results will be expressed in percentages, ie, (proportions x 100)%.

5.2.1.2. Generalized Marginal Model for Repeated Measures (GMMRM)

As a supportive analysis of ASAS20 (or ASAS40) for Estimand 3 (Section 2.1.1) scheduled to be collected multiple times during the first 16 weeks of placebo-controlled period (Weeks 2, 4, 8, 12 and 16) of the study, a generalized marginal linear model for repeated measures (GMMRM) will be used. This model will have fixed effects for treatment, visit (discrete), and treatment-by-visit interaction; the dependent variable will be the logit of the probability of ASAS20 (or ASAS40) response. Note that the model includes fixed effects of treatment, visit, and treatment-by-visit interaction only to ensure the estimated response rates and the treatment difference have a population average or marginal interpretation. A common autoregressive of order 1 [AR(1)] variance-covariance matrix for the two treatment groups will be used to model the variability among observations within a subject. If AR(1)

matrix fails to converge, compound symmetry will be attempted. The parameters of the model will be estimated with pseudo-likelihood type methods. Simulations investigating generalized linear models (GLM) for correlated binary data fit with generalized linear mixed effect model indicate good performance as measured by Type I error, CI coverage and bias even in the presence of data that is MAR (Liu and Zhan, 2011).⁶ From this model, one can obtain estimates of response rates for each treatment group at each time point (at Week 16 and prior) as well as inferential comparisons between treatment groups (odds ratio [OR], 95% CI, and p-value). In the unlikely event that a response rate is 0 or 100% at a particular visit for a treatment group that may lead to convergence issues, data from that visit for both treatment groups will be removed from the model fitting in order to overcome any numerical difficulty.

5.2.1.3. Tipping Point Analysis

A method to analyze the longitudinal data of a binary endpoint measured during the placebo-controlled period (eg, ASAS20 [or ASAS40] response rates at Weeks 2, 4, 8, 12 and 16) under the missing not at random (MNAR) assumption is called the tipping point analysis. This analysis is used as a supportive analysis for the primary endpoint of ASAS20 and the key secondary endpoint of ASAS40 both at Week 16. It assesses the robustness of the binary data to potential deviations from the missing at random (MAR) assumption for both tofacitinib 5 mg BID and placebo groups and it is based on multiple imputation (Yan, Lee and Li, 2009; Ratitch, O’Kelly and Tosiello, 2013).^{7,8}

In this tipping point analysis, a single saturated generalized linear mixed effect model is used as the imputation model. The normal approximation of the difference in binomial proportions using the CMH method adjusting for prior treatment history at randomization ("bDMARD-naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]") is used as the analysis model, which is the same as the method of analyzing response-type endpoint at a single time point described in [Section 5.2.1.1](#). The generalized linear mixed effect model includes the fixed effects of treatment, visit, treatment-by-visit interaction, prior treatment history at randomization ("bDMARD-naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]"), and a latent subject-level random effect. The logit link is used to model the ASAS20 (or ASAS40) response rate as dependent variable for all visits. Estimation of the model parameters is performed under the Bayesian framework using Markov Chain Monte Carlo (MCMC) methods. Only the available on-drug data (see [Appendix Section 9.1](#)) without imputation are included in this generalized linear mixed effect model to estimate the model parameters.

Imputation of missing ASAS20 (or ASAS40) response is performed only at Week 16 based on the predictive distribution of the generalized linear mixed effect model. Under missing not at random (MNAR) assumption, a fixed MNAR quantity (favorable or unfavorable) will be applied to the probability of the ASAS20 (or ASAS40) response at Week 16 for subjects with missing response in the tofacitinib 5 mg BID group and placebo group independently (ie, the analysis will be two-dimensional applied to both tofacitinib 5 mg BID and placebo). A series of fixed quantities will be applied to the probability of the ASAS20 (or ASAS40) response at Week 16 for subjects with missing response in the tofacitinib 5 mg BID group or placebo group independently to assess when the conclusion might change (ie, tipping). A scenario included in the tipping point analysis framework is an analysis under MAR, if there is no fixed quantity applied to the mean of the missing response for subjects in the tofacitinib

5 mg BID group and placebo group. The Rubin's rules (Rubin, 1987)² will be used to combine the results of multiple imputed samples for inference. More detailed step-by-step descriptions are provided in the [Appendix \(Section 9.3\)](#).

5.2.2. Analyses for Continuous (Repeated) Data

Repeated measures data for continuous or order-categorical (analyzed as continuous) endpoints will be analyzed as change from baseline as appropriate with a mixed model for repeated measures (MMRM) that includes fixed effects of treatment group, visit, treatment-group by visit interaction, stratification factor (ie, prior treatment history: "bDMARD-naïve" vs "TNFi-IR or bDMARD Use [Non-IR]") at randomization, stratification-factor by visit interaction, baseline value, and baseline-value by visit interaction. A common unstructured variance-covariance matrix will be used, provided the model converges, otherwise an alternative covariance structure, eg, heterogeneous compound symmetry (CSH), will be attempted. The Kenward-Roger degrees of freedom approximation will be used. This model is referred to as MMRM. Comparison of tofacitinib 5 mg BID to placebo (providing least squares means [LSM] of the treatments, LSM of the treatment difference, 2-sided p-value and 95% CI) at each time point during the first 16 weeks will be generated using this MMRM. Comparisons of tofacitinib 5 mg BID and placebo → tofacitinib 5 mg BID for visits after Week 16 through Week 48 are also reported from the same model. For each endpoint, the number of subjects by treatment group in the FAS, the number of subjects by treatment group included in the MMRM and the number of subjects by treatment group evaluable at each of the visits are also to be reported. If the Baseline is missing or if there are no post-baseline measurements, the subject will be excluded from this analysis (see [Section 4.1](#)).

5.2.3. Analyses for Continuous (Single Visit) Data

For continuous or order-categorical (analyzed as continuous) endpoints, when the analysis includes only a single post-baseline visit, these endpoints will be analyzed as change from baseline with an analysis of covariance (ANCOVA) model that includes fixed effects of treatment group, stratification factor (ie, prior treatment history: "bDMARD-naïve" vs "TNFi-IR or bDMARD Use [Non-IR]") at randomization, and baseline value. This model is referred to as ANCOVA. Comparison of tofacitinib 5 mg BID to placebo (providing LSM of the treatments, LSM of the treatment difference, 2-sided p-value and 95% CI) at this visit will be generated using this ANCOVA. For this visit, the number of subjects by treatment group in the FAS and the number of subjects by treatment group included in the ANCOVA at this visit are also to be reported. If the Baseline is missing or if there is no post-baseline measurement at this visit, the subject will be excluded from this analysis (see [Section 4.1](#)).

5.3. Methods to Manage Missing Data

As described in [Section 2.1.1](#), Estimands 1, 2 and 3 will be used for analyses of ASAS20 at Week 16. After accounting for the intercurrent event of discontinuation of investigational product ([Section 2.1.1](#)), missing ASAS20 values at Week 16 will be handled by setting the ASAS20 values to nonresponsive (in conjunction with Estimand 1 and Estimand 2). This method of handling missing response is known as missing response as non-response (MR=NR). Note that for Estimand 1 which employs the composite strategy ([Section 2.1.1](#)), this can be viewed as a composite endpoint in the sense that a response requires the subject

completes a visit of interest while receiving investigational product eg, Week 16, and achieves a response per the defined ASAS20 response criteria (otherwise it is considered a nonresponse). Note that when a subject has an ASAS20 assessment prior to or on the same date as the discontinuation of the investigational product within the Week 16 visit window, this ASAS20 assessment will be used to evaluate ASAS20 response as on-drug observation for Week 16 (this same rule will be applied to other visits and other applicable binary endpoints). For Estimand 3, missing response will not be explicitly imputed but will be implicitly handled in the GMMRM analysis. The above methods will also be applied to ASAS40 response at Week 16 as well as time points prior to Week 16 for both ASAS20 and ASAS40. The MR=NR in conjunction with Estimand 1 using on-drug data will be applied to all other response-type secondary endpoints (not Type I error controlled) including ASAS 5/6, ASAS partial remission, ASDAS_{CRP} clinically important improvement, ASDAS_{CRP} major improvement, ASDAS_{CRP} inactive disease, and BASDAI50 at all time points collected.

Repeated measures data for continuous and ordered-categorical (analyzed as continuous) endpoints will be analyzed with a MMRM model. This model will yield unbiased estimates and valid inferences in the presence of data that is MCAR or MAR (Rubin, 1987).² Hence missing values will not be explicitly imputed in the routine analysis using the repeated measures model. Continuous and ordered-categorical (analyzed as continuous) endpoints with only a single post-baseline visit will be analyzed with an ANCOVA model. Missing values will not be imputed when using the ANCOVA model.

Tipping point analyses will be conducted to address impact of missing values on the conclusions for ASAS20 and ASAS40 at Week 16 ([Section 5.2.1.3](#)).

For the SF-36v2, ASQoL, EQ-5D-3L (including EQ-VAS and EQ-5D-3L Utility Score [UK]), WPAI, AS-HCRU, and FACIT-F instruments, rules suggested by the developers of these instruments will be followed in calculating scores when individual question/items may be missing. If these rules are not enough for calculating a score, then the endpoint will be considered to have a missing value, and this missing value will be addressed as specified in the paragraph above. For BASDAI, inflammation based on BASDAI's questions 5 and 6, and BASFI, methods of handling missing scores follow the those suggested by Ramiro et al. (2014)¹ (see [Sections 3.2.2](#) and [9.2.2](#)). Specific methods of handling missing components for BASMI, missing joints for SJC(44), and missing sites for MASES can be found in [Appendix](#) (see [Sections 9.2.3](#), [9.2.5](#), and [9.2.6](#) respectively).

In general, missing values in any of the endpoints will not be imputed when summarizing these endpoints using descriptive statistics.

In addition, missing values for safety endpoints will not be imputed.

6. ANALYSES AND SUMMARIES

All efficacy analysis will use FAS ([Section 4.1](#)) or FAS with some subjects excluded from analyses for certain endpoints ([Section 4.4.1](#)) except where indicated otherwise. For each endpoint, comparison ([Section 5.2](#)) of tofacitinib 5 mg BID group to the placebo group at Week 16, which is the primary interest of the treatment comparison will be provided (point estimate, SE, p-value, and 95% CI). Treatment comparisons at other visits will be provided

as well. The Week 16 analysis will include efficacy data through Week 16 only, while safety data through Week 16 and beyond for those subjects who have passed the Week 16 visit at data cut. The Week 48 analysis will include efficacy and safety data through the Week 48 visit including follow-up visit ([Section 5](#)).

6.1. Primary Endpoint: ASAS20 at Week 16

6.1.1. Primary Analysis

The ASAS20 response rate at Week 16 is the primary efficacy endpoint in this trial. The analysis of primary endpoint will be based on the FAS, which includes all randomized subjects who take at least 1 dose of investigational product (ie, tofacitinib or placebo). For the primary analysis, the normal approximation for the difference in binomial proportions adjusting for the stratification factor (ie, prior treatment history: "bDMARD-naïve" vs "TNFi-IR or bDMARD Use [Non-IR]") at randomization (prior treatment history will be derived from the clinical database and used in the analysis; see [Section 3.4.1](#) and [Section 9.11](#)) via the CMH approach will be used to test the superiority of tofacitinib 5 mg BID to placebo and to generate 95% CI for the difference (see [Section 5.2.1.1](#)). Missing values for any reason will be handled by setting the ASAS20 value to nonresponsive, MR=NR (see [Section 5.3](#)). This is the primary analysis for the ASAS20 response rate at Week 16. This analysis uses the on-drug data only (see [Appendix Section 9.1](#)). This corresponds to Estimand 1 or the Primary Estimand (see [Section 2.1.1](#)). Analysis of ASAS20 at other time points will also be analyzed using the same method and the on-drug data on the FAS.

6.1.2. Supportive Analyses

Supportive Analysis 1: As a supportive analysis based on Estimand 2 of ASAS20 at Week 16 (see [Section 2.1.1](#)), ASAS20 at Week 16 will be analyzed using the on-study data (see [Appendix Section 9.1](#)). Analysis of ASAS20 at other time points will also be analyzed using the same method as in the primary analysis ([Section 6.1.1](#)) but using the on-study data on the FAS.

Supportive Analysis 2: The same analysis as the primary analysis for ASAS20 response rate at Week 16 (see [Section 5.2.1.1](#)) will be conducted on the PP analysis set (see [Section 4.2](#)) as a supportive analysis based on Estimand 1 of ASAS20 at Week 16 (see [Section 2.1.1](#)) since the FAS may include instances of protocol deviations which may have material impact on the primary efficacy endpoint. Missing values will be handled by setting the ASAS20 value to non-responsive, ie, MR=NR (see [Section 5.3](#)). This analysis uses the on-drug data only (see [Appendix Section 9.1](#)).

Supportive Analysis 3: As a supportive analysis based on Estimand 3 of ASAS20 at Week 16 (see [Section 2.1.1](#)), GMMRM without imputation for missing responses will be applied to the available ASAS20 data collected at Weeks 2, 4, 8, 12, and 16. The FAS will be used for this analysis (see [Section 5.2.1.2](#)). Missing ASAS20 response status will not be imputed and will be treated as missing. This analysis uses the on-drug data only (see [Appendix Section 9.1](#)).

Supportive Analysis 4: Another supportive analysis on the ASAS20 response rate at Week 16 is performed to assess the robustness of the data to the departures from the MAR assumption.

This analysis is called tipping point analysis and its methodology is briefly described in [Section 5.2.1.3](#) and further detailed in the [Appendix \(Section 9.3\)](#). The FAS will be used and the analysis will include both tofacitinib 5 mg BID and placebo groups. The available on-drug data of ASAS20 without imputation from the following five time points: Weeks 2, 4, 8, 12, and 16 are included in the imputation model to estimate the parameters. The missing ASAS20 at Week 16 with respect to the on-drug data will be subject to multiple imputation. The tipping point analysis includes MAR for both treatment groups as a special case. Both parameter estimation and multiple imputation use only the on-drug data (see [Appendix Section 9.1](#)).

Supportive Analysis 5: Similar to Supportive Analysis 4 above, the tipping point analysis is repeated on the On-Study data (see [Appendix Section 9.1](#) regarding the scope of on-study data). The same available on-drug data of ASAS20 without imputation from the following five time points: Weeks 2, 4, 8, 12, and 16 are included in the imputation model to estimate the parameters using the on-drug data only. However, the missing ASAS20 at Week 16 with respect to the on-study (both available on-drug and off-drug taken into account together) will be subject to multiple imputation. Therefore, parameter estimation uses only on-drug data, but multiple imputation for missing ASAS20 response accounts for on-study data which includes both on-drug and off-drug data (see [Appendix Section 9.1](#)).

Supportive Analysis 6: MMRM without imputation for missing data described in [Section 5.2.2](#) will be used to analyze the change from Baseline for the 4 ASAS20/ASAS40 component endpoints [Δ PGA, Δ in total back pain, Δ BASFI, and Δ in inflammation (average of questions 5 and 6 of BASDAI)] on the FAS using the on-study data (including both on-drug and off-drug data, see [Appendix Section 9.1](#)). These analyses of the 4 ASAS20/ASAS40 component endpoints at Week 16 correspond to Estimand 5 (see [Section 2.1.2](#)). The model will include the time points of Weeks 2, 4, 8, 12, and 16 for the Week 16 analysis only. Only comparisons between tofacitinib 5 mg BID with placebo for visits up to Week 16 will be reported. Prior treatment history will be derived from the clinical database and used in the analysis (see [Section 3.4.1](#)).

Note that MMRM analyses of the 4 ASAS20/ASAS40 component endpoints [Δ PGA, Δ in total back pain, Δ BASFI, and Δ in inflammation (average of questions 5 and 6 of BASDAI)] on the FAS using the on-drug data are detailed in [Section 6.3](#). These endpoints at Week 16 correspond to Estimand 4 (see [Section 2.1.2](#)). These are secondary endpoints/analyses themselves but can be viewed supportive to the primary endpoint as well.

In addition, a summary of subjects will be produced for those who complete the Week 16 visit by their ASAS20 response status at Week 16 and those who discontinue from the investigational product prior to Week 16 visit by their reason (ie, status) of discontinuation using the Disposition – End of Treatment CRF page and the on-drug data ([Appendix Section 9.1](#)). A mock table is given below.

Table 3. Mock Table of ASAS20 Response at Week 16 and Reasons for Study Drug Discontinuation Prior to Week 16 (Estimand 1, FAS, MR=NR, On-Drug Data)

ASAS20 Response Outcome at Week 16 – Estimand 1, FAS, MR=NR, On-Drug Data		
Status - n (%)	Tofacitinib 5 mg BID (N=xxx)	Placebo (N=xxx)
ASAS20 Responders	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders – Who Completed the Week 16 Visit with Observed On-Drug Data	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders – Who Completed the Week 16 Visit with Missing On-Drug Data	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders – Who Discontinued Investigational Product Prior to Week 16 Visit	xx (xx.xx)	xx (xx.xx)
Reasons for Discontinuation of Investigational Product		
Lack of Efficacy	xx (xx.xx)	xx (xx.xx)
Adverse Event	xx (xx.xx)	xx (xx.xx)
Death	xx (xx.xx)	xx (xx.xx)
Lost to Follow-up	xx (xx.xx)	xx (xx.xx)
Withdrawal by Subject	xx (xx.xx)	xx (xx.xx)
Non-Compliance with Study Drug	xx (xx.xx)	xx (xx.xx)
Protocol Deviation	xx (xx.xx)	xx (xx.xx)
Pregnancy	xx (xx.xx)	xx (xx.xx)
Physician Decision	xx (xx.xx)	xx (xx.xx)
Medication Error without Associated Adverse Event	xx (xx.xx)	xx (xx.xx)
Study Terminated by Sponsor	xx (xx.xx)	xx (xx.xx)
Other	xx (xx.xx)	xx (xx.xx)

Another summary of subjects will also be produced for those who complete the Week 16 visit by their ASAS20 response status at Week 16 and those who permanently discontinue from the study prior to Week 16 visit by their reason (ie, status) of discontinuation using Disposition – Follow-Up CRF page and on-study data ([Appendix Section 9.1](#)).

Table 4. Mock Table of ASAS20 Response at Week 16 and Reasons for Study Discontinuation Prior to Week 16 (Estimand 2, FAS, MR=NR, On-Study Data)

ASAS20 Response Outcome at Week 16 – Estimand 2, FAS, MR=NR, On-Study Data		
Status - n (%)	Tofa 5 mg BID (N=xxx)	Placebo (N=xxx)
ASAS20 Responders	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders – Who Completed the Week 16 Visit with Observed On-Study Data	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders – Who Completed the Week 16 Visit with Missing On-Study Data	xx (xx.xx)	xx (xx.xx)
ASAS20 Non-Responders – Missing On-Study Data due to Discontinuation from Study Participation Prior to Week 16 Visit	xx (xx.xx)	xx (xx.xx)
Reasons for Discontinuation from Study		
Death	xx (xx.xx)	xx (xx.xx)
Lost to Follow-up	xx (xx.xx)	xx (xx.xx)
Withdrawal by Subject	xx (xx.xx)	xx (xx.xx)
Study Terminated by Sponsor	xx (xx.xx)	xx (xx.xx)
Other	xx (xx.xx)	xx (xx.xx)

6.2. Key Secondary Endpoint: ASAS40 at Week 16

ASAS40 at Week 16 will be analyzed using the same methods as those for the primary endpoint ASAS20 detailed in [Section 6.1](#). Analysis of ASAS40 at other time points will also be analyzed using the same method as for ASAS20 on the FAS. These include the analyses for Estimand 1 (also known as the Key Secondary Estimand for Key Secondary Analysis), Estimand 2 (Supportive Analysis 1), Estimand 1 on the Per Protocol Analysis Set (Supportive Analysis 2), Estimand 3 (Supportive Analysis 3) (see [Section 2.1.1](#)), and supportive tipping point analyses (Supportive Analysis 4 and Supportive Analysis 5).

6.3. Other Secondary Endpoints and Tertiary/Exploratory Endpoints

Analyses of all endpoints listed in this section will use the FAS ([Section 4.1](#)) or FAS with certain subjects excluded for some endpoints ([Section 4.4.1](#)) using the on-drug data (main) and are also repeated using the on-study data for the following Type I error controlled secondary endpoints only: Δ ASDAS_{CRP}, Δ hsCRP, Δ ASQoL, Δ SF-36v2 (PCS score as well as the non-Type I error controlled endpoints of MCS and 8 norm-based domain scores), Δ BASMI (as well as the 5 non-Type I error controlled components), Δ FACIT-F (Total score as well as the non-Type I error controlled endpoints of Experience Domain and Impact Domain Scores), Δ PGA, Δ in Total Back Pain, Δ BASFI and Δ Inflammation (supportive) ([Appendix Section 9.1](#)).

For the binary endpoints listed below, the observed response rates and SEs of these observed response rates, without imputation for missing data, will be provided descriptively for the two treatment groups at the time points indicated per [Section 5.2](#). For treatment comparisons at all time points, the normal approximation for the difference in binomial proportions

adjusting for the stratification factor (ie, prior treatment history: "bDMARD naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]") at randomization (prior treatment history will be derived from the clinical database and used in the analysis; see [Section 3.4.1](#)) via the CMH approach will be used for both testing and forming 95% CI's for the treatment difference between tofacitinib 5 mg BID and placebo as described in [Section 5.2.1.1](#). Setting missing responses to non-responses (ie, MR=NR) will be used for missing data.

- ASAS20 response at all other time points (Weeks 2, 4, 8, 12, 24, 32, 40, and 48);
- ASAS40 response at all other time points (Weeks 2, 4, 8, 12, 24, 32, 40, and 48) (see [Section 6.2](#));
- ASAS 5/6 response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- ASAS partial remission response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- ASDAS_{CRP} Clinically Important Improvement response, ASDAS_{CRP} Major Improvement response and ASDAS_{CRP} Inactive Disease response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48); and
- BASDAI50 response at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48).

When using the on-drug data, the treatment comparisons of these binary endpoints correspond to Estimand 1 ([Section 2.1.1](#)). Only ASAS20 and ASAS40 (the only two binary endpoints under Type I Error Control at earlier time points) will be repeated using the on-study data ([Section 9.1](#)), and the treatment comparisons of these two binary endpoints correspond to Estimand 2 ([Section 2.1.1](#)).

MMRM without imputation for missing data described in [Section 5.2.2](#) will be used to analyze the change from Baseline for the following endpoints. The model will include up to Week 16 (Weeks 2, 4, 8, 12, and 16) for the Week 16 analysis while all the time points of the entire study duration (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48) for the Week 48 analysis as indicated below for each of the endpoints. Prior treatment history will be derived from the clinical database and used in the analysis (see [Section 3.4.1](#) and [Section 9.11](#)). When the on-drug data is used ([Section 9.1](#)), the treatment comparisons of these endpoints correspond to Estimand 4 ([Section 2.1.2](#)) for both Week 16 and Week 48 analyses. These same analyses will be repeated only for the following Type I error controlled secondary endpoints: Δ ASDAS_{CRP}, Δ hsCRP, Δ BASMI (as well as the 5 non-Type I error controlled components), Δ FACIT-F (Total score as well as the non-Type I error controlled endpoints of Experience Domain and Impact Domain Scores), Δ PGA, Δ in Total Back Pain, Δ BASFI and Δ Inflammation (supportive) ([Section 6.1.2](#)) using the on-study data ([Section 9.1](#)) for the Week 16 analysis only, and the treatment comparisons of these endpoints correspond to the Estimand 5 ([Section 2.1.2](#)).

- Δ ASDAS_{CRP} at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);

- Δ hsCRP (mg/L) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ BASDAI at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ BASMI and its 5 components (Δ in Lateral Spinal Flexion, Tragus-to-Wall Distance, Lumbar Flexion [as measured by Modified Schober's test score], Maximal Intermalleolar Distance, and Cervical Rotation) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ BASFI at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ MASES at all time points (Weeks 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ SJC(44) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ in Spinal Mobility: Chest Expansion at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ PGA at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ in Patient Assessment of Spinal Pain (two numerical rating scales [NRS]: Δ in Nocturnal Spinal Pain and Total Back Pain) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ in Inflammation (defined as the average of questions 5 & 6 of BASDAI) at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48);
- Δ FACIT-F Total, Experience Domain, and Impact Domain Scores at all time points (Weeks 2, 4, 8, 12, 16, 24, 32, 40, and 48).

For the Week 16 analysis, the ANCOVA model will include Week 16 as indicated below for each of the endpoints. When the on-drug data is used ([Section 9.1](#)), the treatment comparisons of these endpoints correspond to Estimand 4 ([Section 2.1.2](#)). These same analyses will be repeated only for the following Type I error controlled secondary endpoints: Δ SF-36v2 (PCS, as well as non-Type I error controlled MCS and 8 norm-based domains) and Δ ASQoL using the on-study data ([Section 9.1](#)) for the Week 16 analysis only, and the treatment comparisons of these endpoints correspond to the Estimand 5 ([Section 2.1.2](#)). The Week 48 analysis will use the MMRM model. The MMRM model will include all the time points of the entire study duration (Weeks 16 & 48) using the on-drug data only corresponding to Estimand 4 ([Section 9.1](#)) as indicated below for each of the endpoints. Prior treatment history will be derived from the clinical database and used in the analysis (see [Section 3.4.1](#) and [Section 9.11](#)).

- Δ SF-36v2, Acute (8 domains (norm-based), PCS, and MCS scores) at Weeks 16 & 48;

- Δ EQ-5D-3L (5 dimensions), Δ EQ-VAS, and Δ EQ-5D-3L Utility Score (UK) at Weeks 16 & 48;
- Δ ASQoL at Weeks 16 & 48;
- Δ WPAI Questionnaire: Spondyloarthritis (4 subscale scores in range of 0-100%) at Weeks 16 & 48;
- Δ AS-HCRU Self-Rating of Job Performance at Work at Weeks 16 & 48.

However, AS-HCRU will only be summarized descriptively for Baseline, Weeks 16 & 48.

6.4. Subset (Subgroup) Analyses

6.4.1. Subgroup Analyses for Primary Endpoint of ASAS20 at Week 16

All subgroup comparisons for the primary endpoint ASAS20 at Week 16 will be made on the FAS with missing value handled by MR=NR using the on-drug data corresponding to Estimand 1 for the following comparison only:

- Tofacitinib 5 mg BID vs. Placebo.

Subgroup analysis will be performed for each of the following subgroup variables and their categories in parentheses:

- Prior treatment history - 2 categories ("bDMARD-naïve", "TNFi-IR or bDMARD Use [Non-IR]") where prior treatment history is derived from the clinical database (see [Appendix Section 9.11](#));
- Prior treatment history – 3 categories ("bDMARD-naïve", "TNFi-IR", "bDMARD Use [Non-IR]") where prior treatment history – 3 categories is derived from the clinical database. Subjects under "bDMARD Use (Non-IR)" category are not TNFi-IR and have prior bDMARD Use without IR. This subgroup analysis is only done when there is sufficient number of subjects in the "bDMARD Use (Non-IR)" category (see [Appendix Section 9.11](#));
- Geographic region – 4 regions ("North America (US and Canada)" [United States, Canada], "European Union" [Bulgaria, Czech Republic, France, Hungary, Poland], "Rest of World" [Ukraine, Russia, Australia, Turkey], "Asia" [China, South Korea]);
- Gender (Female, Male);
- Race (White, Asian, Other), when a subject selects more than one race category, this subject will be grouped to the "Other";
- Age at Baseline (<65, \geq 65 years);
- Baseline weight (<60, \geq 60 to \leq 100, >100 kg);

- Baseline Body Mass Index (BMI) (<25, ≥25 to <30, ≥30 to <40, and ≥40 kg/m²);
- AS disease symptom duration (<5, ≥5 years);
- Baseline AS disease activity (ASDAS_{CRP} <1.3 [inactive disease], ≥1.3 to <2.1 [low disease activity], ≥2.1 to ≤3.5 [high disease activity], > 3.5 [very high disease activity] units, and continuous) (van der Heijde et al. 2016; Machodo et al., 2018);⁹
- Baseline hsCRP (≤2.87, >2.87 mg/L), where 2.87 mg/L (the upper limit of normal) is defined in this study as elevated;
- Baseline smoking status (Never Smoked, Former Smoker, Current Smoker);
- Day 1 Concomitant csDMARD Use (Yes, No);
- HLA-B27 (Positive, Negative).

Estimates of the treatment difference between tofacitinib 5 mg BID and placebo at Week 16 along with the 95% CI without p-value, will be presented for each category of a subgroup variable. The normal approximation using CMH approach adjusting for prior treatment history ("bDMARD-naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]") (see [Sections 3.4.1](#) and [5.2.1.1](#) as well as [Section 9.11](#)) will be used for each category of a subgroup variable, except for the subgroup of prior treatment history.

In addition to Week 16, subgroup analysis of ASAS20 by prior treatment history (2 category definition: "bDMARD-naïve", "TNFi-IR or bDMARD Use [Non-IR]") and (3-category definition: "bDMARD-naïve", "TNFi-IR", and "bDMARD Use [Non-IR]" provided there is sufficient number of subjects in the category of bDMARD Use [Non-IR]) will be conducted at each visit prior to Week 16. ASAS20 will be analyzed separately by prior treatment history (both 2-category and 3-category definitions) using the normal approximation approach to the difference in binomial proportions on the FAS ([Section 4.1](#)) with missing value considered as non-responsive (ie, MR=NR). For each category of "bDMARD-naïve" and "TNFi-IR or bDMARD Use (Non-IR)" of the 2-category definition, as an example, the 95% CI of the treatment difference is calculated as:

$$(\hat{p}_A - \hat{p}_B) \pm z_{0.975} \sqrt{\frac{\hat{p}_A(1 - \hat{p}_A)}{n_A} + \frac{\hat{p}_B(1 - \hat{p}_B)}{n_B}}$$

where \hat{p}_A and \hat{p}_B represent the estimated response rates for tofacitinib 5 mg BID and placebo respectively, and n_A and n_B are their respective sample sizes. $z_{0.975}$ is the 97.5th percentile of the standard normal distribution. No p-value will be produced. The method of handling 0 or 100% response rate in either or both treatment groups can be found in [Section 5.2.1.1](#).

The primary purpose of subgroup analyses is to check for consistency of the primary endpoint results across subgroup categories, ie, making sure overall results are not driven by some subset of subjects (ie, particular categories of a subgroup). Graphical display (eg, forest plots) of the treatment differences in ASAS20 at Week 16 between tofacitinib 5 mg BID and placebo will be presented. There is no intention to have any specific inference within subgroup variables, therefore only 95% CI's are produced without corresponding p-values.

6.4.2. Subgroup Analyses for Key Secondary Endpoint of ASAS40 at Week 16

As ASAS40 at Week 16 is the key secondary endpoint, all of the subgroup analyses described for ASAS20 in [Section 6.4.1](#) will also be performed on ASAS40 at Week 16. The analysis by prior treatment history described for ASAS20 in [Section 6.4.1](#) will also be performed on ASAS40. Only the on-drug data corresponding to Estimand 1 will be used (see [Appendix Section 9.1](#)). The other Type 1 controlled secondary endpoints will not be included in subgroup analyses.

6.5. Baseline and Other Summaries and Analyses

6.5.1. Baseline Summaries

Baseline characteristics will include but may not be limited to the ones listed below and will be summarized descriptively. For continuous variables, the summary will include N (which is the number of subjects evaluable for the baseline characteristics), mean, median, SD and range (ie, minimum and maximum); for binary and categorical variables, the summary will include frequencies and percentages. A missing category will be included for those subjects with missing value. In addition to displays by treatment groups, the summaries will also be provided for all the treatment groups combined.

Demographic characteristics:

- Baseline age (2 categorizations: <18, ≥18 to ≤44, ≥45 to ≤64, ≥65 years; <18, ≥18 to ≤44, ≥45 to ≤64, ≥65 to ≤74, ≥75 to ≤84, ≥85 years; and continuous in years);
- Gender (Female, Male);
- Race (White, Black, Asian, Other);
- Geographic region (see [Section 6.4.1](#) for definition);
- Ethnicity (Hispanic/Latino, non-Hispanic/Latino);
- Baseline body weight (<60, ≥60 to ≤100, >100 kg; and continuous in kg);
- Baseline height (continuous in cm);
- Baseline waist circumference (continuous in cm);

- Baseline BMI (<18.5, ≥18.5 to <25, ≥25 to <30, ≥30 to <40, and ≥40 kg/m²; and continuous kg/m²);
- Baseline smoking status (Never Smoked, Former Smoker, Current Smoker);
- Duration of smoking started (continuous in years) for subjects who are Current Smokers or Former Smokers;
- Duration of smoking stopped (continuous in years) for subjects who are Former Smokers only;
- Baseline pack-year (continuous in pack-year) for subjects who are Current Smokers or Former Smokers;
- Baseline Alcohol use (Yes, No; continuous in units/week for subjects who have Current Alcohol use at Baseline); Yes is defined for subjects who have Current Alcohol use at Baseline, else No.

Baseline disease characteristics:

- Prior treatment history at randomization ("bDMARD-naïve", "TNFi-IR or bDMARD Use [Non-IR]") where prior treatment history is derived from the clinical database (see [Section 9.11](#)). Under the "TNFi-IR or bDMARD Use (Non-IR)" category, two subcategories: "TNFi-IR" and "bDMARD Use (Non-IR)" will be summarized. Subjects under "bDMARD Use (Non-IR)" are not TNFi-IR and have prior bDMARD Use without IR (see [Section 9.11](#));
- AS disease symptom duration (<5, ≥5 years; and continuous in years);
- AS disease duration since diagnosis (<5, ≥5 years; continuous in years);
- Family history of Spondyloarthritis (Yes, No);
- History of Uveitis (Yes, No);
- Current symptom of Uveitis (Yes, No) for subjects with history of uveitis;
- History of Psoriasis (Yes, No);
- Current symptom of Psoriasis (Yes, No) for subjects with history of psoriasis;
- History of Inflammatory Bowel Disease (IBD) (Yes, No);
- Current symptom of IBD (Yes, No) for subjects with history of IBD;
- HLA-B27 (Positive, Negative);
- Baseline hsCRP (≤2.87, >2.87 mg/L [ie, Elevated hsCRP]; and continuous in mg/L);

- Baseline metabolic syndrome (Yes, No) (see [Section 9.7.1](#) for definition);
- Baseline diabetes mellitus (Yes, No) (see [Section 9.7.2](#) for definition);
- Baseline PGA (continuous);
- Baseline Patient Assessment of Pain – Total Back Pain (continuous);
- Baseline Patient Assessment of Pain – Nocturnal Spinal Pain (continuous);
- Baseline BASFI (continuous);
- Baseline BASDAI Total score (continuous);
- Baseline Inflammation (Average of questions 5 and 6 of BASDAI) (continuous);
- Baseline BASMI and its 5 components (continuous);
- Baseline Spinal Mobility – Chest Expansion (continuous in cm)
- Baseline AS disease activity (ASDASCRP <1.3 [inactive disease], ≥1.3 to <2.1 [low disease activity], ≥2.1 to ≤3.5 [high disease activity], >3.5 [very high disease activity] units, and continuous) (van der Heijde et al. 2016; Machodo et al., 2018);^{9,10}
- Baseline presence of enthesitis based on MASES (Yes, No). Yes is defined for those subjects with Baseline MASES >0;
- Baseline MASES (continuous) for those subjects with Baseline MASES >0;
- Baseline presence of swollen joints (Yes, No). Yes is defined for those subjects with Baseline SJC(44) >0;
- Baseline SJC(44) (continuous) for those subject with Baseline SJC(44) >0;
- Baseline SF-36v2 – 8 domain scale (ie, norm-based), PCS, MCS scores (continuous);
- Baseline EQ-5D-3L – 5 dimension scores (continuous), EQ-VAS (Your own health state today) (continuous in mm) and EQ-5D-3L Utility score (UK) (continuous);
- Baseline FACIT-F – Total score, Impact domain and Experience domain scores (continuous);
- Baseline WPAI – 4 subscores (continuous in %);
- Baseline ASQoL Total score (continuous).

6.5.2. Prior and Baseline Treatments for Ankylosing Spondylitis

Prior use denotes any relevant drug use prior to Day 1. If there are missing values for any of the characteristics, add a “Missing” category to capture these subjects unless noted otherwise.

- Prior oral corticosteroids use (oral only: Yes, No, Missing is considered as No);
- Number of prior NSAIDs-IR (2 NSAID-IR, ≥ 3 NSAID-IR) for all subjects (see [Appendix Section 9.10](#) for definition of NSAID-IR);
- Number of prior TNFi-IR (1 TNFi-IR, 2 TNFi-IR) for subjects who have prior inadequate response to TNFi (ie, with prior treatment history of "TNFi-IR" derived from the clinical database, see [Appendix Section 9.11](#) for definition of TNFi-IR);
- Number of prior bDMARD Use (1 bDMARD Use, ≥ 2 bDMARDs Use) for subjects who are not grouped to "TNFi-IR" stratum but have prior bDMARD Use without IR (ie, bDMARD Use [Non-IR] derived from clinical database, see [Appendix Section 9.11](#)).

6.5.3. Concomitant Medications

Concomitant treatments are those treatments that are taken on Day 1 or after. These treatments can be summarized based on the periods of assessment and the summary will include only those subjects who received any concomitant medications post-baseline.

- Taken on Day 1 (at Baseline) only;
- Taken on Day 1 up to Week 16;
- Taken on Day 1 up to Week 48.

Concomitant medications for primary diagnosis:

- Concomitant csDMARDs (for any csDMARDs use [n, %] and for each csDMARD use, eg, Methotrexate, Sulfasalazine, etc. Full but non-exhaustive list of csDMARDs can be found in [Section 9.6 Table 7. List of DMARDs and NSAIDs](#)); Note that csDMARD is defined as non-bDMARD.
- Concomitant NSAIDs (for any NSAIDs use [n, %] and for each NSAID use. Full but non-exhaustive list of NSAIDs can be found in [Section 9.6 Table 7. List of DMARDs and NSAIDs](#));
- Corticosteroids (for any Corticosteroids use [n, %] and for each corticosteroid use, eg, Glucocorticoids);
- Pain Management/Analgesics (for any Pain Management/Analgesics use [n, %] and for each pain management/analgesics use).

6.6. Safety Summaries and Analyses

All the safety data will be summarized descriptively through appropriate data tabulations, descriptive statistics, and graphical presentations. AEs, vital signs, 12-lead ECG parameters, and laboratory tests will be summarized according to the Clinical Data Interchange Standards Consortium (CDISC) and Pfizer Standards (CaPS).

6.6.1. Adverse Events

AEs will be reported as treatment-emergent all causality and treatment-emergent treatment-related. A treatment-emergent AE is any event that starts on or after the first dosing day. The algorithm will not consider any events that start prior to the first dose date.

The relationship to study treatment is assessed by the investigator.

AEs will be displayed by two analysis periods: (1) Baseline to Week 16; and (2) Baseline to Week 48 (end of study including follow-up).

Tier-1 events, ie, AEs of special interest will be summarized descriptively.

Tier-2 events will be analyzed using asymptotic methods proposed by Miettinen and Nurminen (1985).¹¹ Risk ratios (tofacitinib 5 mg BID compared to placebo) and 2-sided 95% CI will be reported. P-values will not be reported for Tier-2 events. This analysis is applicable for the period of Baseline to Week 16 only.

6.6.2. Laboratory Data

For laboratory tests, 12-lead ECG parameters and vital signs, these summaries include categorical tables (eg, normal, high, low), and descriptive statistics for change or percent change (for fasting lipids) from Baseline by treatment group and visit. For the FACS, the endpoints to be summarized, which are subsets of lymphocytes and are measured at Baseline, Week 4, Week 16, Week 32, and Week 48, are as follows: CD3+(%, abs), CD3+CD4+(%, abs), CD3+CD8+(%, abs), CD19+(%, abs), CD56+/CD16+(%, abs). Raw values as well as percent change from Baseline will be summarized with descriptive statistics by treatment group and visit.

The number and percent of subjects reporting specific past and present medical histories at screening will be summarized by medical history reporting term (from MedDRA) and treatment group following CaPS.

Clinical findings on any physical examination during the study will be tabulated by treatment group following CaPS.

Previous and concomitant medication usage by medication type will be tabulated by treatment group using the WHO-Drug dictionary following CaPS.

7. INTERIM ANALYSES

No formal interim analysis will be conducted for this study. The analysis of the primary endpoint, key secondary endpoint and other Type I error controlled secondary endpoints conducted at Week 16 is the final analysis (ie, Week 16 analysis) for these endpoints. The E-DMC will be responsible for ongoing monitoring of the safety of subjects in the study according to the charter. The recommendations made by the E-DMC to alter the conduct of the study will be forwarded to Pfizer for final decision. Information about the E-DMC can be found in the E-DMC Charter, which outlines the operating procedures of the committee, including specific description of the scope of their responsibilities, including a plan where communication timelines are defined.

The final analysis (ie, Week 48 analysis) will be performed at the official database release.

8. REFERENCES

1. Ramiro S, van Tubergen A, van der Heijde D, van den Bosch F, Dougados M, Landewé R. (2014). How to deal with missing items in BASDAI and BASFI. *Rheumatology*. 53: 374-376.
2. Rubin DB. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
3. Cochran WG. (1954). Some methods of strengthening the common χ^2 tests. *Biometrics*. 10: 417-451.
4. Mantel N, Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.* 22:719-748.
5. Agresti A. (2002). *Categorical Data Analysis*, 2nd Edition, New York: John Wiley & Sons.
6. Liu GR, Zhan X. (2011). Comparisons of methods for analysis of repeated binary responses with missing data. *Journal of Biopharmaceutical Statistics*. 21: 371-392.
7. Yan X, Lee S, Li N. (2009). Missing data handling methods in medical device clinical trials. *Journal of Biopharmaceutical Statistics*. 19: 1085-1098.
8. Ratitch B, O'Kelly M, Tosiello R. (2013). Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharmaceutical Statistics*. 12: 337-347.
9. van de Heijde D, Joshi A, Pangan AL, Chen N, Betts K, Mittal M, Bao Y. (2016). ASAS40 and ASDAS clinical responses in the ABILITY-1 clinical trial translate to meaningful improvements in physical function, health-related quality of life and work productivity in patients with non-radiographic axial spondyloarthritis. *Rheumatology*. 55: 80-88.
10. Machado PM, Landewé R, van der Heijde D, et al. (2018). Ankylosing Spondylitis Disease Activity Score (ASDAS): 2018 update of the nomenclature for disease activity states. *Ann. Rheum. Dis.* 77: 1539-1540.
11. Miettinen O, Nurminen M. (1985). Comparative analysis of two rates. *Statistics in Medicine*. 4: 213-226.
12. Sieper J, Rudwaleit M, Baraliakos X, Brandt J, Braun J, Burgos-Vargas R, Dougados M, Hermann KG, Landewe R, Maksymowych W, van der Heijde D. (2009). The assessment of spondyloarthritis international society (ASAS) handbook: a guide to assess spondyloarthritis. *Ann. Rheum. Dis.* 68 (Suppl. II): ii1-ii44.
13. Machado P, Landewe R, Lie E, et al. (2011). Ankylosing Spondylitis Disease Activity Score (ASDAS): defining cut-off values for disease activity states and improvement scores. *Ann. Rheum. Dis.* 70(1):47-53.

14. van der Heijde D, Landewe R, Feldtkeller E. (2008). Proposal of a linear definition of the Bath Ankylosing Spondylitis Metrology Index (BASMI) and comparison with the 2-step and 10-step definitions. *Ann. Rheum. Dis.* 67:489–493.
15. Lukas C, Landewe R, Sieper J, Dougados M, Davis J, Braun J, van der Linden S, van der Heijde D, for the Assessment of SpondyloArthritis International Society. (2009). Development of an ASAS-endorsed disease activity score (ASDAS) in patients with ankylosing spondylitis. *Ann. Rheum. Dis.* 68: 18-24.
16. Ware Jr JE, Kosinski M, Bjorner JB, Turner-Bowker DM, Gandek B, Maruish ME. (2007). *User’s Manual for the SF-36v2[®] Health Survey (2nd ed.)*. Lincoln, RI: QualityMetric Incorporated.
17. Dolan P. (1997). Modeling valuations for EuroQol Health States. *Medical Care.* 35(11): 1095-1108.
18. Szende A, Oppe M, Devlin N. (2007). *EQ-5D value sets: inventory, comparative review and user guide*. 2nd ed., Dordrecht, The Netherlands: Springer.
19. Cella D, Lai JS, Chang CH, Peterman A, Slavin M. (2002). Fatigue in cancer patients compared with fatigue in the general United States population. *Cancer*, 94 (2): 528-538.
20. Cella D, Lai JS, Stone A. (2011). Self-reported fatigue: one dimension or more? Lessons from the Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F) questionnaire. *Support Care Cancer*, 19: 1441-1450.
21. Reilly MC, Gooch KL, Wong RL, Kupper H, van der Heijde D. (2010). Validity, reliability and responsiveness of the work productivity and activity impairment questionnaire in ankylosing spondylitis. *Rheumatology (Oxford)*, 49: 812-819.
22. Alberti KG, Eckel RH, Grundy SM, Zimmet PZ, Cleeman JI, Donato KA, Fruchart JC, James PT, Loria CM, Smith SC. (2009). Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation*, 120: 1640-1645.

9. APPENDICES

9.1. Data Set Description: On-Drug vs. On-Study Data

If for any reason, when subjects have prematurely and permanently discontinued the investigational product (ie, tofacitinib or placebo) during the course of the study, every effort will be made by the study sites and investigators to engage these subjects to follow through the remaining planned study visits and to collect both efficacy and safety data through the end of the study while they are off-drug. It is important that subjects are still blinded to the treatment they receive through the end of the study. For a subject who prematurely discontinues the investigational product (ie, tofacitinib or placebo), the date of the permanent drug discontinuation as recorded on the Disposition – End of Treatment CRF page will be used to separate the efficacy data collected in visits on or before this date as this subject's on-drug data and data collected in visits after this date as this subject's off-drug data.

Because of this data differentiation, there will be two different datasets included for analyses. The first dataset, referred to as On-Drug Data, includes only data that are collected in visits on or before the permanent drug discontinuation. The second dataset, referred to as On-Study Data, includes both the data collected in visits on or before as well as visits after the permanent drug discontinuation. Data that are collected during the on-drug period will be flagged as on-drug data and data collected during the off-drug period will be flagged as off-drug data. Therefore, both on-drug and off-drug data constitute the on-study data. All of the analyses specified in this SAP will use the On-Drug Data unless otherwise specified when using the On-Study Data. The On-Study data will only be used for specific supportive analyses of the primary endpoint (ASAS20), key secondary endpoint (ASAS40) and their 4 components (see [Section 2.1](#), [Section 6.1.2](#) and [Section 6.2](#)) as well as other Type I error controlled secondary endpoints.

All safety data will be analyzed regardless of investigational product discontinuation.

9.2. Endpoint Definitions and Data Derivation Details

9.2.1. ASAS Improvement Criteria

The Assessment of SpondyloArthritis International Society (ASAS) has developed improvement criteria for clinical trials in AS which include the ASAS20, ASAS40, ASAS 5/6 (Sieper et al., 2009)¹² assessments and partial remission (Machado et al., 2011).¹³ These composite scores are derived from several of the Patient Reported Outcome (PRO) measures or disease activity assessments. The composite score will be calculated by the sponsor. In order for this calculation to be completed, the investigator is responsible to ensure completeness of the PROs and appropriately conduct the assessments that comprise this endpoint.

ASAS20 and ASAS40 assess 4 domains: the “Patient Global Assessment of Disease” (PGA, from the CRF of Patient Global Assessment of Disease, range 0-10, with higher scores indicating greater disease activity), Spinal Pain (from the CRF of Total Back Pain, range 0-10, with higher scores indicating greater pain), Function (which is the mean of the 10 questions from the BASFI CRF, range 0-10, with higher scores indicating greater

functional limitation) and Inflammation (which is the mean of questions 5 and 6 from the BASDAI CRF, range 0-10, with higher scores indicating greater inflammation).

ASAS20 improvement (response) is defined as $\geq 20\%$ improvement (decrease) from Baseline and ≥ 1 unit improvement (decrease) from Baseline in at least 3 domains on a scale of 0-10 and no worsening from Baseline of $\geq 20\%$ and ≥ 1 unit in the remaining domain.

ASAS40 improvement (response) criteria are classified as $\geq 40\%$ improvement (decrease) from Baseline and ≥ 2 units improvement (decrease) from Baseline in at least 3 domains on a scale of 0-10 and no worsening from Baseline at all in the remaining domain.

ASAS partial remission is based on the same 4 domains noted above. Partial remission is defined as a response if a score of 2 or less (on a scale of 0-10) for each of the 4 domains. ASAS20, ASAS40, and ASAS partial remission will be considered missing if at least one of the 4 domain scores is missing.

ASAS 5/6 assesses 6 domains: the 4 domains as noted in the ASAS20 and 40 (ie, PGA, Total Back Pain, BASFI, and Inflammation), and additionally two more domains, hsCRP (mg/L) and Spinal mobility, specifically lateral flexion (from the BASMI). Computation details for lateral flexion (cm) can be found in [Section 9.2.3](#) for BASMI. ASAS 5/6 response is defined as $\geq 20\%$ improvement (decrease) from Baseline in at least 5 domains.

In order to avoid numerical difficulty, if the Baseline value of any domain is equal to 0, the following algorithm will be used in evaluating the percent change from Baseline:

1. If change from Baseline is also equal to 0, then percent change from Baseline is set to be 0%;
2. If change from Baseline is > 0 , then percent change from Baseline is set to be 999999%.

These percentages will be used to derive the ASAS20, ASAS40, and ASAS 5/6 responses and will be reported in listings. Change from Baseline cannot be < 0 if Baseline is equal to 0 since the range of values in these domains is 0-10.

9.2.2. Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) & BASDAI50

The BASDAI score is obtained by computing the mean score for the 2 questions related to morning stiffness (questions 5 and 6) and then adding that value to the sum of the scores for the first 4 questions and then dividing the total by 5. This can be written as:

$$\text{BASDAI} = (Q1 + Q2 + Q3 + Q4 + (Q5 + Q6)/2) / 5.$$

The total score will range from 0 to 10, with higher scores indicating greater disease activity.

If answer to only one question is missing, BASDAI score will be calculated using the remaining 5 non-missing answers; otherwise if answers to two or more questions are missing, BASDAI will be considered missing (Ramiro et al., 2014).¹ Specifically, if the answer to either question 5 or 6 is missing, then the score for the one available question can be used. If both are missing, then the BASDAI is considered missing. Regarding the other 4 questions, if only 1 is missing, then the other 3 non-missing answers are used; else if more than 1 of these 4 answers is missing, then the BASDAI is considered missing. Explicitly, the calculation of BASDAI score for any one missing answer is detailed in the table below:

Missing Q5 or Q6	Missing any Q1-Q4	Formula
No	Yes	$BASDAI = [\text{sum of the 3 non-missing of (Q1, Q2, Q3, Q4)} + (Q5 + Q6) / 2] / 4$
Yes	No	$BASDAI = [Q1 + Q2 + Q3 + Q4 + (\text{non-missing of Q5 or Q6})] / 5$

At a given visit, BASDAI50 response is defined as an improvement (decrease) from Baseline in BASDAI score $\geq 50\%$.

As one of the 4 domains to derive ASAS improvement criteria, the mean score of the 2 questions related to morning stiffness (questions 5 and 6) is used as a measure of inflammation. If one of the 2 scores is missing, the non-missing one is used for inflammation; else if both scores are missing, inflammation is set as missing (Ramiro et al., 2014).¹ This score will range from 0 to 10, with higher scores indicating greater inflammation.

The BASDAI score and the inflammation score (mean of questions 5 and 6) at all visits will be derived using the algorithm described above to account for missing values. The BASDAI score and the inflammation score obtained from the CRF page at Screening and Baseline visits are calculated only for the purpose of determining subject's eligibility and protocol deviation status (see [Section 4.2.2](#)) and will not be used for statistical analyses for these two visits. BASDAI score and inflammation score at Screening and Baseline visits will be re-calculated using the above algorithm accounting for missing answers above and will be used for all BASDAI and BASDAI50 statistical analyses.

9.2.3. Bath Ankylosing Spondylitis Metrology Index (BASMI)

The BASMI consists of 5 components: cervical rotation (left and right, degree), intermalleolar distance (cm), lumbar flexion as measured by Modified Schober's test score (cm) (note: This score is obtained by subtracting 15 cm from the value entered on the CRF page which intentionally includes the 15 cm.), tragus-to-wall distance (left and right, cm) and lateral flexion (left and right, neutral and flexion, cm). Cervical rotation, tragus-to-wall distance and lateral flexion have multiple measurements while intermalleolar distance and Modified Schober's test score have single measurement. Each measurement is done twice (ie, two trials) at a visit.

The calculation of BASMI follows the algorithm published by van der Heijde et al. (2008).⁹ The algorithm is reproduced in the table below. In order to obtain the lateral flexion for each side (left or right) from the measures recorded on the CRFs, one needs to compute the following difference: [(lateral flexion-neutral) – (lateral flexion-flexion)]. This is done for each of the two trials. For tragus-to-wall distance, the better (ie, smaller) of two trials, while for cervical rotation and lateral flexion, the better (ie, larger) of two trials will be chosen for left and right separately. Then the mean of left and right of those chosen (denoted as A) will be used to compute scores in range of 0-10 (denoted as S). For Modified Schober’s test score and intermalleolar distance, the better (ie, larger) of the two trials (denoted as A) will be used to compute scores in range of 0-10 (denoted as S). It is important to note that the better values of the measurements do not necessarily all come from the same trial, meaning better one of the measurements is separately chosen between the two trials.

Each of the score obtained for cervical rotation, intermalleolar distance, Modified Schober’s test score, tragus-to-wall distance and lateral flexion (A in the table below), will be mapped via linear functions to a score in range of 0-10 (S in the table below). To obtain the BASMI score, one takes the mean of the five component scores (ie, five S) that range between 0-10, with higher scores indicating more severe the patient’s limitation of movement due to their AS. If the number of missing component scores ≤ 2 , BASMI score is computed using the mean of the remaining non-missing component scores, otherwise if the number of missing component scores > 2 , BASMI is considered as missing.

Components	S=0 if	0 <S <10 if	S=10 if
Lateral lumbar flexion (cm)	$A \geq 21.1$	$S = (21.1-A)/2.1$	$A \leq 0.1$
Tragus-to-wall distance (cm)	$A \leq 8$	$S=(A-8)/3$	$A \geq 38$
Lumbar flexion (modified Schober) (cm)	$A \geq 7.4$	$S=(7.4-A)/0.7$	$A \leq 0.4$
Intermalleolar distance (cm)	$A \geq 124.5$	$S=(124.5-A)/10$	$A \leq 24.5$
Cervical rotation angle (degree, °)	$A \geq 89.3$	$S=(89.3-A)/8.5$	$A \leq 4.3$

9.2.4. Ankylosing Spondylitis Disease Activity Score (ASDAS_{CRP})

The ASDAS_{CRP} endpoint is derived from several patient reported outcomes and hsCRP (mg/L) which will be calculated by the sponsor. Following is the formula used for calculating the ASDAS_{CRP} (Sieper et al., 2009; Lukas et al., 2009).^{12,15}

$$ASDAS_{CRP} = 0.121 \times \text{Back Pain} + 0.058 \times \text{Duration of Morning Stiffness} + 0.110 \times \text{Patient Global} + 0.073 \times \text{Peripheral Pain/Swelling} + 0.579 \times \text{Ln}(\text{hsCRP}+1).$$

Question 2 of the BASDAI provides the data for Back Pain, Question 6 of the BASDAI provides the data for Duration of Morning Stiffness, the score from the Patient Global Assessment of Disease (PGA) is utilized for the Patient Global and Question 3 of the BASDAI contributes the data for Peripheral Pain/Swelling. If any one of the five components is missing, ASDAS_{CRP} is considered missing. The range of the ASDAS_{CRP} is ≥ 0 , with higher scores indicating greater disease activity.

hsCRP will be set to 0.199 mg/L in the formula above and in all efficacy and safety analyses, where 0.2 mg/L is the lower limit of quantification (LLOQ) of the assay.

The ASDAS_{CRP} clinically important improvement, major improvement and inactive disease will be calculated from the ASDAS_{CRP} data (van de Heijde et al. 2016).⁹

- ASDAS_{CRP} clinically important improvement is defined as a response if improvement (decrease) from Baseline in ASDAS_{CRP} ≥ 1.1 units.
- ASDAS_{CRP} major improvement is defined as a response if improvement (decrease) from Baseline in ASDAS_{CRP} ≥ 2.0 units.
- ASDAS_{CRP} inactive disease is defined as a response if actual ASDAS_{CRP} < 1.3 units.

9.2.5. Swollen Joint Count (44)

Swollen joints are recorded in the CRF page of joint swelling. This page is to be completed if answer to the question, “Does this subject exhibit any joint swelling in the physical exam?” is “Yes.” For each of the 44 joints, swelling can be recorded as “Present”, “Absent”, “Not Done”, “Not Applicable” or this joint assessment is simply missing. Before calculating the swollen joint count, the following pre-processing steps for each joint should be performed:

- If a joint receives an intra-articular injection (either at Baseline or post-baseline visits), set the joint status to “Present” on or after the date of injection regardless of the assessment of the joint recorded on the CRF page.
- If there is no associated injection record available, a joint recorded as “Not Done” is set to missing.
- A joint recorded as “Not Applicable” means that measurement cannot be made and this joint is set to missing.

After this pre-processing, a joint assessment that is set to missing is to be imputed by the average of the non-missing values of the other joints, with the restriction that the number of missing joint assessments cannot exceed 50% of the expected total number of joints evaluated for the swollen joint count (ie, 22, 50% of 44). In that case, the swollen joint count is set to missing.

Explicitly after pre-processing, let:

- N_1 =number of joints recorded as “Present”;
- N_2 =number of joints recorded as “Absent”;
- N_3 =number of joints with missing values.

The total number of joints assessed is $N_1 + N_2 + N_3$ (ie, 44). The imputed swollen joint count is therefore $N_1 + N_3 \times (N_1 / (N_1 + N_2))$ if $N_1 + N_2 > 22$; otherwise the swollen joint count is set to missing. It can be noted that the swollen joint count may not be a whole number. Note that if there is no missing joint assessment, the swollen joint count will be the number of joints recorded as “Present” (ie, N_1).

9.2.6. Maastricht Ankylosing Spondylitis Enthesitis Score (MASES)

Thirteen sites are assessed for tenderness on the MASES CRF page. For each site, tenderness can be recorded as “Present”, “Absent”, “Not Done”, or this site assessment is simply missing. Before calculating MASES, the following pre-processing step for each site will be performed:

- If a site is recorded as “Not Done”, set the site to missing.

After this pre-processing, a site assessment that is set to missing is to be imputed by the average of the non-missing values of the other sites, with the restriction that the number of missing site assessments cannot exceed 50% of the expected total number of sites evaluated for the MASES (ie, 6.5, 50% of 13). In that case, the MASES is set to missing.

Explicitly after pre-processing, let:

- N_1 =number of sites recorded as “Present”;
- N_2 =number of sites recorded as “Absent”;
- N_3 =number of sites with missing values.

The total number of sites assessed is $N_1 + N_2 + N_3$ (ie, 13). The imputed MASES is therefore $N_1 + N_3 \times (N_1 / (N_1 + N_2))$ if $N_1 + N_2 > 6.5$; otherwise the MASES is set to missing. It can be noted that the MASES may not be a whole number. Note that if there is no missing site assessment, the MASES will be the number of sites recorded as “Present” (ie, N_1).

9.2.7. 36-Item Short Form Survey Version 2 (SF-36v2) Acute

The SF-36 v.2 (Acute) is a 36-item generic health status measure. It measures 8 general health domains: physical functioning, role limitation due to physical health, bodily pain, general health perceptions, vitality, social functioning, role limitations due to emotional problems and mental health. These domains can also be summarized as physical and mental component summary scores (Ware Jr et al., 2007).¹⁶

For each of the 8 domains, the raw score is first transformed into a score within the range of 0-100. This score will then be standardized into a Z-score using the domain’s mean and standard deviation based on the 1998 general U.S. population (ie, norm). This Z-score is finally re-scaled to a scale score (or norm-based score) with mean of 50 and standard deviation of 10. The scale scores (ie, norm-based scores) of the 8 domains are used for analyses. These 8 domains are as follows:

- a. Physical Functioning (PF). This score is based on the responses to the 10 items that compose Question 3 and reflects the degree to which various physical activities have been limited in the previous week by the subject's health.
- b. Role-Physical (RP). This score is based on the responses to the four items that compose Question 4 and reflects the relative amount of time that the subject has had problems with work or other regular daily activities as a result of their physical health during the previous week.
- c. Bodily Pain (BP). This score is based on the responses to Questions 7 and 8 and reflects bodily pain and its effects on normal work during the previous week.
- d. General Health (GH). This score is based on responses to Question 1 and the four items in Question 11 and reflects the subject's perception of their general health during the previous week.
- e. Vitality (VT). This score is based on responses to Question 9 items a, e and g, and reflects the subject's physical energy level relative to time during the previous week.
- f. Social Functioning (SF). This score is based on responses to Questions 6 and reflects how physical health or emotional problems have interfered with social activities during the previous week.
- g. Role-Emotional (RE). This score is based on responses to the three items in Question 5 and reflects the amount of time during the previous week that emotional problems have interfered with work or regular daily activities.
- h. Mental Health (MH). This score is based on responses to Question 9 items b, c, d, f, and h and reflects various mental/emotional states relative to time during the previous week.

The summary component scores are:

- Physical Component Summary (PCS).
- Mental Component Summary (MCS).

In addition, there is another subscale in the SF-36: Health Transition (TR). This score is based on the response to Question 2 and is a rating of current general health compared to one week previous.

The summary component summary (PCS, MCS) scores are based on a normalized sum of the 8 scale scores (ie, norm-based scores): PF, RP, BP, GH, VT, SF, RE, and MH. All domains and summary components are scored such that a higher score indicates a higher functioning or health level.

VARIABLE	DERIVATION
SF-36 PF scale score	<p>raw score = sum (items 3A, 3B, 3C, 3D, 3E, 3F, 3G, 3H, 3I, 3J)</p> <p>$PF = (\text{raw score} - 10) * 5$</p> <p>$PF_Z = (PF - 82.62455)/24.43176$</p> <p>PF scale score = (PF_Z*10) + 50</p> <p>When calculating the raw score, if 5 or more of the items are non-missing then replace any missing values as follows:</p> <p>Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores.</p> <p>Otherwise, if less than 5 of the items are non-missing then PF scale score is missing.</p> <p>The response scale for each activity ranges from 1 to 3 where 1=limited a lot, 2=limited a little, and 3=not limited at all. A higher PF scale score indicates better physical functioning.</p>
SF-36 RP scale score	<p>raw score = sum (items 4A, 4B, 4C, and 4D)</p> <p>$RP = [(\text{raw score} - 4)/16] * 100$</p> <p>$RP_Z = (RP - 82.65109)/26.19282$</p> <p>RP scale score = (RP_Z * 10) + 50</p> <p>When calculating the raw score, if 2 or more of the items are non-missing then replace any missing values as follows:</p> <p>Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores.</p> <p>Otherwise, if less than 2 of the items are non-missing then RP scale score is missing.</p> <p>The response scale for each item ranges from 1 to 5 where 1=All of the time, 2=Most of the time, 3=Some of the time, 4=A little of the time, and 5=None of the time. A higher RP scale score indicates better role-physical functioning.</p>

<p>SF-36 BP scale score</p>	<p>raw score = sum (reversed item 7 and reversed item 8) $BP = (raw\ score - 2) * 10$ $BP_Z = (BP - 73.86999)/24.00884$ $BP\ scale\ score = (BP_Z * 10) + 50$</p> <p>Reverse direction of Item 7 as follows if =1, set to 6 if =2, set to 5.4 if =3, set to 4.2 if =4, set to 3.1 if =5, set to 2.2 if =6, set to 1</p> <p>Reverse direction of item 8 as follows: if =1 and original value of item 7=1, set to 6 if =1 and original value of item 7>=2, set to 5 if =2, set to 4 if =3, set to 3 if =4, set to 2 if =5, set to 1</p> <p>If item 7 is answered and item 8 is missing, set 8 = reversed 7 as defined above. If 8 is answered and 7 is missing, set 7 as reverse item 8 as follows: if =1, set to 6 if =2, set to 4.75 if =3, set to 3.5 if =4, set to 2.25 if =5, set to 1</p> <p>If 1 or more questions were answered, calculate BP scale score as defined above. If neither question was answered then BP scale score is missing.</p> <p>The scale for Question 7, amount of bodily pain, ranges from 1 to 6 where 1=None, 2=Very mild, 3=mild, 4=Moderate, 5=Severe, and 6=Very severe.</p> <p>The scale for Question 8, the degree to which pain interfered with normal work, ranges from 1 to 5 where 1=Not at all, 2=A little bit, 3=Moderately, 4=Quite a bit, and 5=Extremely. A higher BP scale score indicates lack of bodily pain.</p>
<p>SF-36 GH scale score</p>	<p>raw score = sum (reversed item 1, item 11A, reversed 11B, 11C and reversed 11D) $GH = (raw\ score - 5) * 5$ $GH_Z = (GH - 70.78372)/21.28902$ $GH\ scale\ score = (GH_Z * 10) + 50$</p> <p>Reverse direction of Item 1 as follows: if =1, set to 5 if =2, set to 4.4 if =3, set to 3.4 if =4, set to 2 if =5, set to 1</p> <p>Reverse direction of item 11B and 11D by subtracting score from 6.</p> <p>When calculating the raw score, if 3 or more of the items are non-missing then replace any missing values as follows:</p>

	<p>Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores.</p> <p>Otherwise, if less than 3 of the items are non-missing then GH scale score is missing.</p> <p>Responses for Question 1, an assessment of self-perceived health status, range from 1 to 5 where 1=Excellent, 2=Very good, 3=Good, 4=Fair, and 5=Poor.</p> <p>Responses for the items in Question 11 range from 1 to 5 where 1=Definitely true, 2=Mostly true, 3=Don't know, 4=Mostly false, and 5=Definitely false and reflect the subject's perception of their relative health and expectations of their future health status.</p> <p>A higher GH scale score indicates better general health perceptions.</p>
SF-36 VT scale score	<p>raw score = sum (reversed item 9a, reversed 9e, 9g and 9i)</p> $VT = [(raw\ score - 4)/16] * 100$ $VT_Z = (VT - 58.41968)/20.87823$ $VT\ scale\ score = (VT_Z * 10) + 50$ <p>Reverse direction of Items 9a and 9e by subtracting score from 6.</p> <p>When calculating the raw score, if 2 or more of the items are non-missing then replace any missing values as follows:</p> <p>Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores.</p> <p>Otherwise, if less than 2 of the items are non-missing then VT scale score is missing.</p> <p>The scale for these items ranges from 1 to 5 where 1=All of the time, 2=Most of the time, 3=Some of the time, 4=A little of the time, and 5=None of the time. A higher VT scale score indicates more vitality.</p>
SF-36 SF scale score	<p>raw score = sum (reversed 6 and 10)</p> $SF = [(raw\ score - 2)/8] * 100$ $SF_Z = (SF - 85.11568) / 23.24464$ $SF\ scale\ score = (SF_Z * 10) + 50$ <p>Reverse direction of score for item 6 by subtracting score from 6.</p> <p>When calculating the raw score, if 1 of the items is missing then substitute the missing score with the score on the non-missing item. If both items are missing then SF scale score is missing.</p> <p>Responses to Question 6, an assessment of the extent to which health/emotional problems interfered with social activities, range from 1 to 5 where 1=Not at all, 2=Slightly, 3=Moderately, 4=Quite a bit, and 5=Extremely. Responses to Question 10 reflect the amount of time that health/emotional problems interfered with social activities and range from 1 to 5 where 1=All of the time, 2=Most of the time, 3=Some of the time, 4=A little of the time, and 5=None of the time. A higher SF scale score indicates better social functioning.</p>

<p>SF-36 RE scale score</p>	<p>raw score = sum (items 5A, 5B, and 5C) $RE = [(raw\ score - 3) / 12] * 100$ $RE_Z = (RE - 87.50009) / 22.01216$ $RE\ scale\ score = (RE_Z * 10) + 50$</p> <p>When calculating the raw score, if 2 or more of the items are non-missing then replace any missing values as follows:</p> <p>Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores.</p> <p>Otherwise, if less than 2 of the items are non-missing then RE scale score is missing.</p> <p>Responses to the items in Question 5 range from 1 to 5 where 1=All of the time, 2=Most of the time, 3=Some of the time, 4=A little of the time, and 5=None of the time. A higher RE scale score indicates better role-emotional functioning.</p>
<p>SF-36 MH scale score</p>	<p>raw score = sum (items 9B, 9C, reversed 9D, 9F and reversed 9H) $MH = (raw\ score - 5) * 5$ $MH_Z = (MH - 75.76034) / 18.04746$ $MH\ scale\ score = (MH_Z * 10) + 50$</p> <p>Reverse direction of scores for 9D and 9H, by subtracting score from 6.</p> <p>If 3 or more of the items are non-missing then replace any missing values as follows:</p> <p>Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores.</p> <p>Otherwise, if less than 3 of the items are non-missing then MH scale score is missing.</p> <p>The scale for these items ranges from 1 to 5 where 1=All of the time, 2=Most of the time, 3=Some of the time, 4=A little of the time, and 5=None of the time. A higher MH scale score indicates better mental health.</p>

SF-36 PCS score	<p>PCS score includes the 8 scales for GH, PF, RP, RE, SF, MH, BP, and VT.</p> <p>PF1= (PF-82.62455)/24.43176; RP1=(RP-82.65109)/26.19282; BP1=(BP-73.86999)/24.00884; GH1 = (GH-70.78372)/21.28902; VT1= (VT-58.41968)/20.87823; SF1=(SF-85.11568)/23.24464; RE1= (RE-87.50009)/22.01216; MH1=(MH-75.76034)/18.04746;</p> <p>Raw Score = ((GH1*.24954)+(PF1* .42402)+(RP1*.35119) + (RE1*-.19206)+(SF1*-.00753)+(MH1*-.22069)+(BP1*.31754) + (VT1*.02877))</p> <p>PCS Summary Scale Score = (raw score *10) + 50</p> <p>Raw Score is missing if one of the component scale scores is missing.</p>
SF-36 MCS score	<p>MCS score includes the 8 scales for GH, PF, RP, RE, SF, MH, BP, and VT.</p> <p>PF1= (PF-82.62455)/24.43176; RP1=(RP-82.65109)/26.19282; BP1=(BP-73.86999)/24.00884; GH1 = (GH-70.78372)/21.28902; VT1= (VT-58.41968)/20.87823; SF1=(SF-85.11568)/23.24464; RE1= (RE-87.50009)/22.01216; MH1=(MH-75.76034)/18.04746;</p> <p>Raw Score =((GH1*-.01571)+(PF1*-.22999)+(RP1*-.12329) + (RE1*.43407)+(SF1*.26876)+(MH1*.48581)+(BP1*- 0.09731) + (VT1*.23534))</p> <p>MCS Summary Scale Score = (raw score*10)+50</p> <p>Raw score is missing if one of the component scale scores in missing.</p>

9.2.8. EuroQol-5D Health Questionnaire – 3 Level Version (EQ-5D-3L)

There are 5 health dimensions on the EQ-5D-3L Health Questionnaire CRF page: Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression. Each dimension has three possible levels of response (no problem, some problems, and extreme problems). If a subject reports no problem in a dimension, the score is coded as 1; if some problems, the score is coded as 2; and if extreme problems, the score is coded as 3. An ambiguous value such as when two boxes are checked, it will be treated as missing.

“Your own health state today” (EQ-VAS) records the subject’s self-rated health, a score ranging from 0 to 100 mm is recorded, with higher scores representing better health state today. The raw VAS data will need to be rescaled to a 100 mm scale prior to any calculation.

In addition, EQ-5D-3L Utility score based on United Kingdom (UK) sample in 1993 will be derived (Dolan, 1997; Szende, Oppe, and Delvin, 2007).^{17,18} The time trade-off (TTO) value set is used for scoring. The following table provides the scoring algorithm for a subject’s health state coded in 5 digits in the order of Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression. If any dimension is coded as 1, nothing (ie, 0) will be subtracted. Therefore, if a subject has a health state of 11111, the Utility score will be

1 (ie, maximum Utility score). An example of a health state of 21232 has Utility score of 0.088, as illustrated below. If any of the 5 dimension scores is missing, the Utility score is set to missing.

UK TTO value set		Example: the value for health state 21232	
Full health (11111)	1	Full health	= 1
At least one 2 or 3	- 0.081	Minus a constant	- 0.081
At least one 3	- 0.269	Minus a constant	- 0.269
Mobility = 2	- 0.069	Minus Mobility level 2	- 0.069
Mobility = 3	- 0.314		
Self-care = 2	- 0.104	Minus Self-care level 1	- 0.000
Self-care = 3	- 0.214		
Usual Activities = 2	- 0.036	Minus Usual Activities level 2	- 0.036
Usual Activities = 3	- 0.094		
Pain/Discomfort = 2	- 0.123		
Pain/Discomfort = 3	- 0.386	Minus Pain/Discomfort level 3	- 0.386
Anxiety/Depression = 2	- 0.071	Minus Anxiety/Depression level 2	- 0.071
Anxiety/Depression = 3	- 0.236		
		State 21232	= 0.088

9.2.9. Functional Assessment of Chronic Illness Therapy – Fatigue (FACIT-F)

FACIT-F is a 13-item questionnaire (each item score ranges from 0 to 4). There will be 3 endpoints derived: FACIT-F total score, FACIT-F experience domain score and FACIT-F impact domain score (Cella et al., 2002; Cella, Lai, and Stone, 2011).^{19,20} FACIT-F total score (range 0-52) is calculated by summing the 13 items. FACIT-F experience domain score (range 0-20) is calculated by summing 5 items of Q1 I feel fatigued, Q2 I feel weak all over, Q3 I feel listless (“washed out”), Q4 I feel tired and Q7 I have energy, while FACIT-F impact domain score (range 0-32) is calculated by summing the remaining 8 items. Note that all the scores except for Q7 & Q8 should be reversed (ie, 0 →4, 1→3, 3→1, 4→0) prior to

summing for both the total score and the domain scores such that higher scores represent better functioning (ie- less fatigue).

All responses are added with equal weight to obtain the total score. In cases where some answers are missing, a total score is prorated from the score of the answered items, so long as more than 50% of the items (ie, at least 7 of 13 for FACIT-F total score, at least 3 of 5 for FACIT-F experience domain score, and at least 5 of 8 for FACIT-F impact domain score) are answered.

9.2.10. Work Productivity and Activity Impairment Questionnaire: Spondyloarthritis (WPAI)

The WPAI assesses work productivity and impairment (Reilly et al., 2010).²¹ It is a 6-item questionnaire used to assess the degree to which Ankylosing Spondylitis affected work productivity and regular activities over the past 7 days. The questions are as follows:

Q1 = currently employed;

Q2 = hours missed due to health problems;

Q3 = hours missed due to other reasons;

Q4 = hours actually worked;

Q5 = degree health affected productivity while working (0-10 scale, with higher numbers indicating less productivity);

Q6 = degree health affected regular activities (0-10 scale, with higher numbers indicating greater impairment of regular activities).

Subscale scores that are calculated from these questions are:

- Percent work time missed due to health problem: $Q2/(Q2+Q4)$ for those who were currently employed;
- Percent impairment while working due to health problem: $Q5/10$ for those who were currently employed and actually worked in the past 7 days;
- Percent overall work impairment due to health problem: $Q2/(Q2+Q4)+[(1-Q2/(Q2+Q4))\times(Q5/10)]$ for those who were currently employed;
- Percent activity impairment due to health problem: $Q6/10$ for all respondents.

Each subscale score is expressed as an impairment percentage (range: 0-100%) where higher numbers indicate greater impairment and less productivity.

9.3. Tipping Point Analysis for Longitudinal Binary Data

A method to analyze the longitudinal data of a binary endpoint measured during the placebo-controlled period (eg, ASAS20 [or ASAS40] response rates at Weeks 2, 4, 8, 12 and 16) under the MNAR assumption is called the tipping point analysis. This analysis is used as a supportive analysis for the primary endpoint of ASAS20 and the key secondary endpoint of ASAS40 both at Week 16. It assesses the robustness of the binary data to potential deviations from the MAR assumption for both tofacitinib 5 mg BID and placebo groups and it is based on multiple imputation (Yan, Lee and Li 2009; Ratitch, O’Kelly and Tosiello, 2013).^{7,8}

In this tipping point analysis, a single saturated generalized linear mixed effect model is used as the imputation model. The normal approximation of the difference in binomial proportions using the CMH method adjusting for prior treatment history at randomization ("bDMARD-naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]") (prior treatment history will be derived from the clinical database and used in the analysis, see [Section 3.4.1](#)) is used as the analysis model, which is the same as the method of analyzing binary endpoint at a single time point described in [Section 5.2.1.1](#). This tipping point analysis will be implemented in the following steps.

Step 1: Specification of the imputation model.

The generalized linear mixed effect model includes the fixed effects of treatment, visit (discrete), treatment-by-visit interaction, prior treatment history at randomization ("bDMARD-naïve" vs "TNFi-IR or bDMARD Use [Non-IR]"), and a latent subject-level random effect. The logit link is used to model the ASAS20 (or ASAS40) response rate as dependent variable for all visits. Estimation of the model parameters is performed under the Bayesian framework using MCMC methods. Only the available on-drug data (see [Appendix Section 9.1](#)) without imputation are included in this generalized linear mixed effect model to estimate the model parameters. If the binary response is denoted by Y_i for subject i , then the saturated generalized linear mixed effect model for two treatment groups and the five post-baseline visits is parameterized as:

$$\begin{aligned} \text{logit}(\pi_i) = \text{logit}(P(Y_i = 1)) = & \beta_0 + \beta_{\text{dose}}\text{dose} + \beta_{\text{time1}}\text{time1} + \beta_{\text{time2}}\text{time2} + \beta_{\text{time3}}\text{time3} + \\ & \beta_{\text{time4}}\text{time4} + \beta_{\text{int1}}\text{dose} * \text{time1} + \beta_{\text{int2}}\text{dose} * \text{time2} + \\ & \beta_{\text{int3}}\text{dose} * \text{time3} + \beta_{\text{int4}}\text{dose} * \text{time4} + \beta_{\text{strata}}\text{strata} + \phi_i \cdot \end{aligned}$$

In this model, using reference coding, *dose* is a dummy variable representing treatment assignment. When it is 0, it represents the placebo group; 1 represents tofacitinib 5 mg BID. Similarly, *time1*, *time2*, *time3*, and *time4* are dummy variables. When all of them are 0, it represents Week 2; *time1* = 1 and all others being 0 represent Week 4; *time2* = 1 and all others being 0 represent Week 8; *time3* = 1 and all others being 0 represent Week 12, and *time4* = 1 and all others being 0 represent Week 16. *strata* is a dummy variable representing prior treatment history with two strata: "bDMARD-naïve" and "TNFi-IR or bDMARD Use (Non-IR)". When it is 0, it represents "bDMARD-naïve"; 1 represents "TNFi-IR or bDMARD Use (Non-IR)". Interaction terms between the treatment and time point are also

included in the model, leading to a saturated model. ϕ_i 's represent the subject-specific random effects.

Step 2: Estimating the imputation model.

Estimation of the model parameters is performed under the Bayesian framework using MCMC methods. All of the β 's are the respective effect parameters and they are all assigned the same non-informative prior Normal distribution, $N(\mu_\beta = 0, \sigma_\beta^2 = 9)$. The variance of 9 on the logit scale ensures this prior is non-informative on the probability scale over its support of $[0, 1]$. ϕ_i 's are the subject-specific random effects and they are all assigned a common prior Normal distribution, $N(0, \sigma^2)$. σ^2 is the common variance and it is further assigned a weakly informative Inverse-Gamma distribution with shape=1 and scale=1. In this prior distribution, the prior 90th percentile for σ^2 is approximately 9.

Step 3: Imputation of missing response at Week 16.

A single imputation of missing ASAS20 (or ASAS40) response at Week 16 is performed based on the predictive distribution of the generalized linear mixed effect model. Inferences at earlier time points are not of interest, thus imputation of missing ASAS20 (or ASAS40) at earlier time points will not be performed. The following model parameters pertinent to the assessment of treatment difference at Week 16: β_0^* , β_{dose}^* , β_{time4}^* , β_{int4}^* , β_{strata}^* , and ϕ_i^* can be sampled from their posterior distributions. Multiply imputed datasets can be generated using multiple samples from the MCMC of these parameters. The number of imputations (R) is specified as 100.

Step 4: MNAR parameter (δ) and predictive distribution.

For example, for a subject i randomized to tofacitinib 5 mg BID with prior treatment history of "TNFi-IR or bDMARD Use (Non-IR)" with missing ASAS20 (or ASAS40) response at Week 16, the probability of response of this subject π_i^* (derived from sampled parameters) is subtracted by a fixed MNAR quantity (favorable or unfavorable) of δ ($-1 < \delta < 1$) to become π_i^p ,

$$\pi_i^p = \pi_i^* - \delta$$

where $\pi_i^* = \frac{\exp(\beta_0^* + \beta_{dose}^* + \beta_{time4}^* + \beta_{int4}^* + \beta_{strata}^* + \phi_i^*)}{1 + \exp(\beta_0^* + \beta_{dose}^* + \beta_{time4}^* + \beta_{int4}^* + \beta_{strata}^* + \phi_i^*)}$. In the situation when $\pi_i^* < 0$, π_i^p will be set to 0; similarly when $\pi_i^* > 1$, π_i^p will be set to 1. For a given δ , the missing ASAS20 (or ASAS40) response at Week 16 for this subject in the tofacitinib 5mg BID group with prior treatment history of "TNFi-IR or bDMARD Use (Non-IR)" will be imputed based on Bernoulli model with the probability of π_i^p . This is similarly done for a subject in the placebo group. A series of fixed MNAR quantity δ including 0 (ie, MAR) will be applied to both treatment groups. The same series of δ values is used for both treatment groups: -0.9, -0.7, -0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5, 0.7, 0.9. The δ values may be adjusted as appropriate. The value of 0 represents MAR assumption, while positive values represent unfavorable scenarios (ie, penalizing the subject's probability of response) and negative values represent favorable scenarios. Once all subjects with missing ASAS20 (or ASAS40)

responses at Week 16 are imputed, this step results in a single complete imputed data set for Week 16.

Step 5: Analysis of complete imputed data set.

Analysis of an imputed data set will produce an estimate as well as standard error of the treatment difference applying the normal approximation of the difference in binomial proportions using the CMH approach adjusting for prior treatment history at randomization ("bDMARD-naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]"). For a given value of MNAR quantity, this is repeated for R=100 times to generate R=100 complete imputed data sets and these R=100 sets of estimates are combined using the Rubin's rules (Rubin, 1987).² The analysis method is the same as that used for the primary analysis.

Step 6: Assessment of tipping point

Steps 4-5 can then be repeated for different values of MNAR quantities δ to evaluate their impact on the treatment difference for subjects with missing ASAS20 (or ASAS40) responses at Week 16 between tofacitinib 5 mg BID and placebo. The specific implementation (ie, seed for random number generation, burn-in and thinning of the MCMC samples) of this tipping point analysis will be pre-specified in the statistical programming plan. Note that there is no need to repeat Steps 1-3 for different values of δ .

9.4. Definition and Use of Visit Windows in Reporting

For reporting purposes, the following visit windows will be used for efficacy, patient-reported outcomes and safety display that show by scheduled visits. If two or more observations fall into the same visit window, the observation closest to the target day will be used in the analyses. If there is a tie, the later observation will be used. However, for a subject who discontinues from the study early prior to the Week 48 visit and who does not have an observation within the Week 48 visit window, if two or more observations fall in the same visit window (ie, a visit window prior to Week 48), then the latest observation (rather than the observation closest to the target day) within this visit window will be used in the analysis for that visit window.

The visit windows will be applied to on-drug data and on-study data separately (see [Section 9.1](#)).

Table 5. Visit Windowing

Visit Label	Target Day ^a	Window Definition
Screening		< Day 1
Baseline	Day 1	Last non-missing assessment on or before Day 1 and prior to first dose of investigational product (ie, tofacitinib or placebo)
Week 2	Day 15	Day 2 to Day 21
Week 4	Day 29	Day 22 to Day 42
Week 8	Day 57	Day 43 to Day 70
Week 12	Day 85	Day 71 to Day 98
Week 16 ^c	Day 113	Day 99 to [Treatment Switch Day ^b or Day 140, whichever is smaller]
Week 24	Day 169	[Treatment Switch Day ^b +1 or Day 141, whichever is smaller] to Day 196
Week 32	Day 225	Day 197 to Day 252
Week 40	Day 281	Day 253 to Day 308
Week 48	Day 337	≥ Day 309

- a. Target day is equal to the target week×7 days + 1 day.
- b. Treatment switch day = day of the first dose of treatment switch at Week 16 (ie, treatment switch day = treatment switch date – date of the first dose +1), where treatment switch date is taken from the Oral Dosing CRF, as the start date in the first entry of the dosing log for the period of Week 16 to Week 24.
- c. All the data collected on treatment switch day will be counted in ≤ Week 16 except for the AE data. Any new treatment emergent AE occurred on the day of treatment switch will be counted in ≥ Week 24.

Note that the upper limit of the Week 16 visit window and the lower limit of the Week 24 visit window are defined uniquely for each of the subjects in the study regardless of which treatment group the subjects are randomized to, since per the study protocol, subjects randomized to the placebo → tofacitinib 5 mg BID treatment group will advance to the predetermined treatment (tofacitinib 5 mg BID) in a blinded fashion for the remainder of the study. Further details for the Week 16/Week 24 visit windows are in order.

1. If treatment switch day falls between Day 99 and Day 140, then the upper limit of the Week 16 visit window is the smaller of treatment switch day or Day 140 and the lower limit of the Week 24 visit window is the smaller of treatment switch day+1 or Day 141;
2. If treatment switch day < 99, >140, or missing, then the upper limit of the Week 16 visit window will be set to Day 140 and the lower limit of the Week 24 visit will be set to Day 141.

Effects of Bullets 1 and 2 above are further illustrated as follows.

Visit label	If treatment switch day is							
	Missing		<99		99-140		>140	
	Week 16	Week 24	Week 16	Week 24	Week 16	Week 24	Week 16	Week 24
Day window	99 to 140	141-196	99 to 140	141-196	from [99 to 99] up to [99 to 140]	from [100 to 196] up to [141 to 196]	99 to 140	141-196

9.5. Range of Values for Continuous Efficacy Endpoints

The following table displays the unit, theoretical range of values, and the direction of improvement for each of the select continuous efficacy endpoints.

Table 6. Numerical Characteristics of Select Continuous Efficacy Endpoints

Endpoint	Unit	Theoretical Range of Values	Direction of Improvement from Baseline
Patient Global Assessment of Disease	None	0-10	Decrease from Baseline
Patient Assessment of Spinal Pain (Total Back Pain, Nocturnal Spinal Pain)	None	All: 0-10	Decrease from Baseline
BASFI	None	0-10	Decrease from Baseline
BASDAI	None	0-10	Decrease from Baseline
Inflammation Score (ie, Average of Q5 and Q6 of BASDAI)	None	0-10	Decrease from Baseline
hsCRP	mg/L	≥0	Decrease from Baseline
BASMI score and its 5 component scores (A, S) (A is the unmapped component score, S is the mapped component score [range 0-10] via linear method, see Section 9.2.3)	BASMI, 5 components (S): None Lateral flexion, Tragus-to-wall distance, lumbar flexion, and intermalleolar distance (A): cm Cervical rotation angle (A): degree (°)	BASMI, 5 components (S): 0-10 5 components (A): ≥0	BASMI, 5 components (S), Tragus-to-wall distance (A): Decrease from Baseline Lateral flexion, lumbar flexion, intermalleolar distance, and cervical rotation (A): Increase from Baseline.
Spinal Mobility – Chest Expansion	cm	≥0	Increase from Baseline (ie, higher

			score represents more spinal mobility)
ASDAS _{CRP}	None	≥0	Decrease from Baseline
MASES	None	0-13	Decrease from Baseline
Swollen Joint Count (44)	None	0-44	Decrease from Baseline
SF-36v2, 8 domain scale (ie, norm-based), PCS, and MCS scores	None	All: Real values (Mean=50, SD=10)	Increase from Baseline
EQ-5D-3L, 5 dimension scores	None	All: 1, 2, 3	Decrease from Baseline
EQ-VAS	mm	0-100	Increase from Baseline
EQ-5D-3L, Utility Score (UK)	None	-0.594 - 1	Increase from Baseline
FACIT-F (Total, Impact domain, Experience domain scores)	None	Total: 0-52 Impact domain: 0-32 Experience domain: 0-20	Increase from Baseline (ie, higher score represents less fatigue)
ASQoL	None	0-18	Decrease from Baseline
WPAI 4 subscale scores	%	All: 0-100	Decrease from Baseline
AS-HCRU Self-Rating of Job Performance	None	0-10	Decrease from Baseline
<p>Abbreviations: % = percent; ASDAS_{CRP} = Ankylosing Spondylitis Disease Activity Score using C-Reactive Protein; AS-HCRU = Ankylosing Spondylitis – HealthCare Resource Utilization Questionnaire; ASQoL = Ankylosing Spondylitis Quality of Life; BASDAI = Bath Ankylosing Spondylitis Disease Activity Index; BASFI = Bath Ankylosing Spondylitis Functional Index; BASMI = Bath Ankylosing Spondylitis Metrology Index; cm = centimeter; EQ-5D-3L = EuroQol Health State Profile – 5 Dimensions – 3 Levels; EQ-VAS = EuroQol Your own health state today-Visual Analog Scale; FACIT-F = Functional Assessment of Chronic Illness Therapy - Fatigue; hsCRP = high sensitivity C-Reactive Protein; MASES = Maastricht Ankylosing Spondylitis Enthesitis Score; MCS = mental component summary; mg/L = milligrams per liter; PCS = physical component summary; SD = standard deviation; SF-36v2 = 36-Item Short Form Survey Version 2 Acute; UK = United Kingdom; WPAI = Work Productivity & Activity Impairment.</p>			

9.6. Lists of DMARDs and NSAIDs

The following table enlists the generic names of biological DMARDs (bDMARDs) including both TNFi bDMARDs and non-TNFi bDMARDs, csDMARDs (also known as non-bDMARDs), and NSAIDs. This list is not an exhaustive list and additional terms may be present in the clinical database.

Table 7. List of DMARDs and NSAIDs

WHO Drug Preferred Term	bDMARDs		csDMARDs ¹	NSAIDs
	TNFi	Non-TNFi		
ABATACEPT		x		
ACECLOFENAC				x
ACEMETACIN				x
ADALIMUMAB ²	x			
ANAKINRA		x		
APREMILAST ³			x	
AZATHIOPRINE			x	
CELECOXIB				x
CERTOLIZUMAB ²	x			
CERTOLIZUMAB PEGOL ²	x			
CHLOROQUINE			x	
CICLOSPORIN			x	
DEXKETOPROFEN				x
DEXKETOPROFEN TRAMADOL				x
DEXKETOPROFEN TROMETAMOL				x
DICLOFENAC				x
DICLOFENAC DIETHYLAMINE				x
DICLOFENAC POTASSIUM				x
DICLOFENAC SODIUM				x
DICLOFENAC SODIUM/MISOPROSTOL				x
ETANERCEPT ²	x			
ETODOLAC				x
ETORICOXIB				x
FLURBIPROFEN				x
GOLIMUMAB ²	x			
GUSELKUMAB		x		
HYDROXYCHLOROQUINE			x	

WHODrug Preferred Term	bDMARDs		csDMARDs ¹	NSAIDs
	TNFi	Non-TNFi		
HYDROXYCHLOROQUINE PHOSPHATE			x	
IBUPROFEN				x
IBUPROFEN W/PARACETAMOL				x
INDOMETACIN				x
INFLIXIMAB ²	x			
IXEKIZUMAB		x		
KETOPROFEN				x
KETOROLAC				x
LEFLUNOMIDE			x	
LORNOXICAM				x
MELOXICAM				x
METHOTREXATE			x	
METHOTREXATE SODIUM			x	
NABUMETONE				x
NAPROXEN				x
NAPROXEN SODIUM				x
NIFLUMIC ACID				x
NIMESULIDE				x
PELUBIPROFEN				x
PHENYL BUTAZONE				x
PIROXICAM				x
PIROXICAM BETADEX				x
RITUXIMAB		x		
ROFECOXIB				x
SALSALATE				x
SECUKINUMAB		x		
SULFASALAZINE			x	
SULINDAC				x
TALNIFLUMATE				x
TENOXICAM				x
TIAPROFENIC ACID				x
TOCILIZUMAB		x		
USTEKINUMAB		x		

WHODrug Preferred Term	bDMARDs		csDMARDs ¹	NSAIDs
	TNFi	Non-TNFi		
Abbreviations: TNFi = tumor necrosis factor inhibitor; bDMARD = biological disease-modifying antirheumatic drug; csDMARD = conventional synthetic disease-modifying antirheumatic drug.				
<ol style="list-style-type: none"> csDMARD is also defined as non-biological DMARD (ie, non-bDMARD). TNFi approved for Ankylosing Spondylitis by the US FDA. Apremilast is a targeted synthetic DMARD (tsDMARD), but is considered as a csDMARD in the analysis. A footnote, "Apremilast is considered as a csDMARD", will be added to any display where applicable. 				

9.7. Metabolic Syndrome and Diabetes Mellitus at Baseline

9.7.1. Criteria for Clinical Diagnosis of Metabolic Syndrome

Criteria for clinical diagnosis of the metabolic syndrome will be based on the 2009 Joint Statement Harmonizing the Metabolic Syndrome (Alberti et al., 2009).²² There are 5 measures or risk factors as detailed in Table 1 of the 2009 Joint statement and excerpted below:

Measure	Categorical Cut Points
Elevated waist circumference*	Population- and country-specific definitions
Elevated triglycerides (drug treatment for elevated triglycerides is an alternate indicator†)	≥150 mg/dL (1.7 mmol/L)
Reduced HDL-C (drug treatment for reduced HDL-C is an alternate indicator†)	<40 mg/dL (1.0 mmol/L) in males; <50 mg/dL (1.3 mmol/L) in females
Elevated blood pressure (antihypertensive drug treatment in a patient with a history of hypertension is an alternate indicator)	Systolic ≥130 and/or diastolic ≥85 mm Hg
Elevated fasting glucose‡ (drug treatment of elevated glucose is an alternate indicator)	≥100 mg/dL

HDL-C indicates high-density lipoprotein cholesterol.

*It is recommended that the IDF cut points be used for non-Europeans and either the IDF or AHA/NHLBI cut points used for people of European origin until more data are available.

†The most commonly used drugs for elevated triglycerides and reduced HDL-C are fibrates and nicotinic acid. A patient taking 1 of these drugs can be presumed to have high triglycerides and low HDL-C. High-dose ω-3 fatty acids presumes high triglycerides.

‡Most patients with type 2 diabetes mellitus will have the metabolic syndrome by the proposed criteria.

Any 3 abnormal findings out of the 5 measures will qualify a subject for the metabolic syndrome. A missing measure is considered as its criteria not satisfied and thus not counted as an abnormal finding. The baseline (ie, Day -1 per CaPS) value will be used in the derivation. If the Day -1 value is not available, a value prior to Day -1 will be used.

- Any drug treatment taken on Day -1 recorded on the Prior and Concomitant Medications – Anti-Diabetic Agents CRF will be considered having the drug treatment of elevated glucose criterion satisfied.
- Any drug treatment taken on Day -1 recorded on the Prior and Concomitant Medications – Anti-Hypertension Agents CRF will be considered having the drug treatment of elevated blood pressure criterion satisfied.
- Any drug treatment taken on Day -1 recorded on the Prior and Concomitant Medications – Lipid Lowering Agents CRF will be considered having the drug treatment of elevated triglycerides criterion satisfied.
- For the HDL-C criterion, only the lab categorical cutpoints will be considered and drug treatment will not be considered.
- The waist circumference thresholds will be population and country specific per Table 2 of the 2009 Joint Statement excerpted below. Explicitly, the US specific threshold will be used for the subjects enrolled from US, the Canada threshold for the subjects enrolled from Canada, the European threshold for the subjects enrolled from Western Europe, Eastern Europe and Russia, the Asian (including Japanese) threshold for the subjects enrolled from Asia, and Ethnic Central and South American threshold for the subjects enrolled from Central and South America.

Population	Organization (Reference)	Recommended Waist Circumference Threshold for Abdominal Obesity	
		Men	Women
Europid	IDF (4)	≥94 cm	≥80 cm
Caucasian	WHO (7)	≥94 cm (increased risk)	≥80 cm (increased risk)
		≥102 cm (still higher risk)	≥88 cm (still higher risk)
United States	AHA/NHLBI (ATP III)* (5)	≥102 cm	≥88 cm
Canada	Health Canada (8,9)	≥102 cm	≥88 cm
European	European Cardiovascular Societies (10)	≥102 cm	≥88 cm
Asian (including Japanese)	IDF (4)	≥90 cm	≥80 cm
Asian	WHO (11)	≥90 cm	≥80 cm
Japanese	Japanese Obesity Society (12)	≥85 cm	≥90 cm
China	Cooperative Task Force (13)	≥85 cm	≥80 cm
Middle East, Mediterranean	IDF (4)	≥94 cm	≥80 cm
Sub-Saharan African	IDF (4)	≥94 cm	≥80 cm
Ethnic Central and South American	IDF (4)	≥90 cm	≥80 cm

9.7.2. Definition of Diabetes Mellitus at Baseline

A subject is considered having diabetes mellitus at baseline, if any of the conditions below is met. The baseline (ie, Day -1 per CaPS) value will be used in the derivation. If the Day -1 value is not available, a value prior to Day -1 will be used.

- Subject had a diagnosis of diabetes mellitus recorded on the Cardiovascular Risk Factors CRF or on the General Medical History CRF at screening;
- Subject received any drug treatment taken on Day -1 recorded on the Prior and Concomitant Medications – Anti-Diabetic Agents CRF;
- Subject's HbA1c $\geq 6.5\%$ at baseline or, if HbA1c is not available, baseline fasting plasma glucose ≥ 126 mg/dL.

9.8. Summary of Efficacy Analyses

The following table provides a summary of all of the efficacy analyses to be performed for the comparison of treatments across visits.

Table 8. Summary of Efficacy Analyses

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
Section references	Section 2.1	Section 6	Section 5		Section 9.1	Section 4	Section 5	Section 5	Section 5.3	Section 5.2	Section 5	
1. ASAS20 (Primary endpoint at Week 16) 2. ASAS40 (Key secondary endpoint at Week 16) (Type I Error Controlled)	Estimand 1 (ASAS20: Primary Estimand; ASAS40: Key Secondary Estimand)	ASAS20: Primary analysis; ASAS40: Key Secondary analysis	Week 16	Weeks 2, 4, 8, 12, 16	On-drug data	FAS	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Binary	ASAS20 at Week 16 primary, ASAS40 at Week 16 Key secondary, all others secondary
	Estimand 1 (ASAS20: Primary Estimand; ASAS40: Key Secondary Estimand)	ASAS20: Primary analysis; ASAS40: Key Secondary analysis	Week 48	Weeks 2, 4, 8, 12, 16, 24, 32, 40, 48	On-drug data	FAS	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	1. Tofacitinib 5 versus Placebo (Week 2 to Week 16) 2. Tofacitinib 5 versus Placebo → Tofacitinib 5 (After Week 16)	Binary	Supportive for ≤16 weeks and secondary for >16 weeks
	Estimand 2	Supportive analysis 1	Week 16	Weeks 2, 4, 8, 12, 16	On-study data	FAS	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Binary	Supportive
	Estimand 2	Supportive analysis 1	Week 48	Weeks 2, 4, 8, 12, 16, 24, 32, 40, 48	On-study data	FAS	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	1. Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Binary	Supportive

Protocol A3921120 (Tofacitinib (CP-690,550)) Statistical Analysis Plan

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
										2. Tofacitinib 5 versus Placebo → Tofacitinib 5 (After Week 16)		
	Estimand 1	Supportive analysis 2	Week 16	Week 16	On-drug data	PP	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	Tofacitinib 5 versus Placebo at Week 16	Binary	Supportive
	Estimand 3	Supportive analysis 3	Week 16	Weeks 2, 4, 8, 12, 16	On-drug data	FAS	GMMRM	Stratification factor not used in model	No imputation	Tofacitinib 5 versus Placebo (Week 2 to Week 16) reporting Odds Ratio	Binary	Supportive
	NA	Supportive analysis 4	Week 16	Estimation: Weeks 2, 4, 8, 12, 16; Report: Week 16	On-drug data for both parameter estimation and multiple imputation	FAS	Tipping Point Analysis	Prior treatment history as a covariate in the saturated generalized linear mixed effect model (ie, imputation model). Stratified by prior treatment history as CMH stratification factor in the analysis model.	Only missing data at Week 16 will be imputed using multiple imputation	Tofacitinib 5 versus Placebo at Week 16 for different values of MNAR adjustment δ	Binary	Supportive
	NA	Supportive analysis 5	Week 16	Estimation: Weeks 2, 4, 8, 12, 16; Report: Week 16	On-drug data for parameter estimation but on-study data for multiple imputation	FAS	Tipping Point Analysis	Prior treatment history as a covariate in the saturated generalized linear mixed effect model (ie, imputation model). Stratified by prior treatment history as CMH stratification factor in the analysis model.	Only missing data at Week 16 will be imputed using multiple imputation	Tofacitinib 5 versus Placebo at Week 16 for different values of MNAR adjustment δ	Binary	Supportive

Protocol A3921120 (Tofacitinib (CP-690,550)) Statistical Analysis Plan

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
	Estimand 1	Subgroup analysis of prior treatment history (2 categories): "bDMARD-naïve" and "TNFi-IR or bDMARD Use (Non-IR)"	Week 16	Weeks 2, 4, 8, 12, 16	On-drug data	FAS for each category of "bDMA RD-naïve" and "TNFi-IR or bDMARD Use (Non-IR)"	Normal Approximation	NA	MR=NR	Tofacitinib 5 versus Placebo (Week 2 to Week 16) Note: no p-values	Binary	Supportive
	Estimand 1	Subgroup analysis of prior treatment history (3 categories): "bDMARD-naïve", "TNFi-IR", and "bDMARD Use (Non-IR)"	Week 16	Weeks 2, 4, 8, 12, 16	On-drug data	FAS for each category of "bDMA RD-naïve", "TNFi-IR", and "bDMA RD Use (Non-IR)"	Normal Approximation	NA	MR=NR	Tofacitinib 5 versus Placebo (Week 2 to Week 16) Note: no p-values	Binary	Supportive
	Estimand 1	Other subgroup analyses	Week 16	Week 16 only	On-drug data	FAS for each category of a subgroup variable	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	Tofacitinib 5 versus Placebo at Week 16 Note: no p-values	Binary	Supportive
Four ASAS components : 1. ΔPGA 2. ΔPatient Assessment of Spinal Pain (Total Back Pain)	Estimand 4	Secondary analysis	Week 16	Weeks 2, 4, 8, 12, 16	On-drug data	FAS**	MMRM	MMRM*	No imputation	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Continuous	Secondary
	Estimand 4	Secondary analysis	Week 48	Weeks 2, 4, 8, 12, 16, 24, 32, 40, 48	On-drug data	FAS**	MMRM	MMRM*	No imputation	1. Tofacitinib 5 versus Placebo	Continuous	Supportive for ≤16 weeks and secondary for >16 weeks

Protocol A3921120 (Tofacitinib (CP-690,550)) Statistical Analysis Plan

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
3. ΔBASFI 4. ΔInflammation (All secondary endpoints, Type I Error Controlled)										(Week 2 to Week 16) 2. Tofacitinib 5 versus Placebo → Tofacitinib 5 (After Week 16)		
	Estimand 5	Supportive analysis 6	Week 16	Weeks 2, 4, 8, 12, 16	On-study data	FAS**	MMRM	MMRM*	No imputation	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Continuous	Supportive
1. ASAS 5/6 2. ASAS partial remission 3. ASDAS _{CRP} clinically important improvement 4. ASDAS _{CRP} major improvement 5. ASDAS _{CRP} inactive disease 6. BASDAI50 (All secondary endpoints)	Estimand 1	Secondary analysis	Week 16	Weeks 2, 4, 8, 12, 16	On-drug data	FAS/Endpoint-specific FAS	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Binary	Secondary
	Estimand 1	Secondary analysis	Week 48	Weeks 2, 4, 8, 12, 16, 24, 32, 40, 48	On-drug data	FAS/Endpoint-specific FAS	CMH Normal Approximation	Stratified by prior treatment history as CMH stratification factor	MR=NR	1. Tofacitinib 5 versus Placebo (Week 2 to Week 16) 2. Tofacitinib 5 versus Placebo → Tofacitinib 5 (After Week 16)	Binary	Supportive for ≤16 weeks and secondary for >16 weeks

Protocol A3921120 (Tofacitinib (CP-690,550)) Statistical Analysis Plan

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
1. Δ ASDAS _{CRP} 2. Δ hsCRP 3. Δ BASMI and Δ in 5 components 4. Δ FACIT-F (Total, Experience Domain, Impact Domain Scores) 5. Δ BASDAI 6. Δ MASES 7. Δ Swollen Joint Count (44) 8. Δ Spinal Mobility - Chest Expansion 9. Δ Patient Assessment of Spinal Pain (Nocturnal Spinal Pain) (All secondary endpoints)	Estimand 4	Secondary analysis	Week 16	Weeks 2, 4, 8, 12, 16 (note: MASES has no Week 2)	On-drug data	FAS/Endpoint-specific FAS**	MMRM	MMRM*	No imputation	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Continuous	Secondary
	Estimand 4	Secondary analysis	Week 48	Weeks 2, 4, 8, 12, 16, 24, 32, 40, 48 (note: MASES has no Week 2)	On-drug data	FAS/Endpoint-specific FAS**	MMRM	MMRM*	No imputation	1. Tofacitinib 5 versus Placebo (Week 2 to Week 16) 2. Tofacitinib 5 versus Placebo → Tofacitinib 5 (After Week 16)	Continuous	Supportive for ≤ 16 weeks and secondary for >16 weeks
	Estimand 5 (Only for the following four: Δ ASDAS _{CRP} , Δ hsCRP, Δ BASMI and 5 components, Δ FACIT-F (Total, Experience Domain, Impact Domain Scores) (Type I Error Controlled except FACIT-F Impact, Experience Domain Scores, and BASMI 5 components))	NA	NA	Week 16	Weeks 2, 4, 8, 12, 16	On-study data	FAS/Endpoint-specific FAS**	MMRM	MMRM*	No imputation	Tofacitinib 5 versus Placebo (Week 2 to Week 16)	Continuous

Protocol A3921120 (Tofacitinib (CP-690,550)) Statistical Analysis Plan

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
1. ΔASQoL 2. ΔSF-36v2 (PCS, MCS, 8 domains, norm-based) 3. ΔEQ-5D-3L (5 dimensions, EQ-VAS, Utility Score [UK]) 4. ΔWPAI (4 subscale scores) (All secondary endpoints)	Estimand 4	Secondary analysis	Week 16	Week 16	On-drug data	FAS**	ANCOVA	ANCOVA includes fixed effects of treatment, stratification factor (ie, prior treatment history) at randomization and baseline value	No imputation	Tofacitinib 5 versus Placebo at Week 16	Continuous	Secondary
	Estimand 4	Secondary analysis	Week 48	Weeks 16, 48	On-drug data	FAS**	MMRM	MMRM*	No imputation	1. Tofacitinib 5 versus Placebo at Week 16 2. Tofacitinib 5 versus Placebo → Tofacitinib 5 at Week 48	Continuous	Supportive for ≤16 weeks and secondary for >16 weeks
	Estimand 5 (Only for the following two: ΔASQoL, ΔSF-36v2 PCS, MCS, 8 domains, norm-based) (Type I Error Controlled except SF-36v2 MCS and 8 domain scores)	NA	NA	Week 16	Week 16	On-study data	FAS**	ANCOVA	ANCOVA includes fixed effects of treatment, stratification factor (ie, prior treatment history) at randomization and baseline value	No imputation	Tofacitinib 5 versus Placebo at Week 16	Continuous
1. ΔAS-HCRU Self-Rating of Job Performance at Work (Tertiary /Exploratory endpoint)	Estimand 4	Tertiary / Exploratory analysis	Week 16	Week 16	On-drug data	FAS**	ANCOVA	ANCOVA includes fixed effects of treatment, stratification factor (ie, prior treatment history) at randomization and baseline value	No imputation	Tofacitinib 5 versus Placebo at Week 16	Continuous	Exploratory
	Estimand 4	Tertiary / Exploratory analysis	Week 48	Weeks 16, 48	On-drug data	FAS**	MMRM	MMRM*	No imputation	1. Tofacitinib 5 versus Placebo at Week 16	Continuous	Supportive for ≤16 weeks and exploratory for >16 weeks

Endpoint	Week 16 Estimand	Week 16 Analysis	Week 16 or 48 Analysis Report	Visits to Report	Data Inclusion	Analysis Set	Analysis Method	Adjustment to Stratification at Randomization	Missing Data Handling	Treatment Comparison	Endpoint Type	Interpretation
										2. Tofacitinib 5 versus Placebo → Tofacitinib 5 at Week 48		

For complete list of abbreviations used for this table, please see [Section 9.9](#).
 In all the analyses, stratification factor (ie, prior treatment history: "bDMARD-naïve" vs. "TNFi-IR or bDMARD Use [Non-IR]") at randomization will be derived from the clinical database ([Section 3.4.1](#), [Section 9.11](#)).
 *MMRM includes fixed effects of treatment, visit, treatment-by-visit interaction, stratification factor (ie, prior treatment history) at randomization, stratification-factor-by-visit interaction, baseline value, and baseline-value-by-visit interaction.
 **Subjects are included in model if they have non-missing baseline values and at least one non-missing post-baseline value.

9.9. Abbreviations Used in SAP

The following table lists out the abbreviations used in this SAP except some mathematical notations and formulae.

Table 9. Abbreviations Used in SAP

Term or Abbreviation	Definition
Δ	Change from Baseline
AE	Adverse Event
ANCOVA	Analysis of Covariance
AR(1)	Autoregressive of order 1 type covariance structure
AS	Ankylosing Spondylitis
ASAS	Assessment of SpondyloArthritis International Society
ASAS20	Assessment of SpondyloArthritis International Society improvement (response) is defined as $\geq 20\%$ improvement (decrease) from Baseline and ≥ 1 unit improvement (decrease) from Baseline in at least 3 domains on a scale of 0-10 and no worsening from Baseline of $\geq 20\%$ and ≥ 1 unit in the remaining domain
ASAS40	Assessment of SpondyloArthritis International Society improvement (response) criteria are classified as $\geq 40\%$ improvement (decrease) from Baseline and ≥ 2 units improvement (decrease) from Baseline in at least 3 domains on a scale of 0-10 and no worsening from Baseline at all in the remaining domain.
AS-HCRU	Ankylosing Spondylitis – HealthCare Resource Utilization Questionnaire
ASDAS _{CRP}	Ankylosing Spondylitis Disease Activity Score using C-Reactive Protein
ASQoL	Ankylosing Spondylitis Quality of Life
BASDAI	Bath Ankylosing Spondylitis Disease Activity Index
BASDAI50	Bath Ankylosing Spondylitis Disease Activity Index improvement (response) is defined as decrease from Baseline in BASDAI score $\geq 50\%$
BASFI	Bath Ankylosing Spondylitis Functional Index
BASMI	Bath Ankylosing Spondylitis Metrology Index
bDMARD	biological disease-modifying antirheumatic drug
BID	twice daily
BMI	Body Mass Index
BP	Bodily Pain
CaPS	CDISC and Pfizer Standards
CDISC	Clinical Data Interchange Standards Consortium
■	■
cm	centimeter
CMH	Cochran–Mantel–Haenszel
CP-690,550	Tofacitinib
CRF	Case Report Form
CSR	Clinical Study Report
CI	Confidence Interval
csDMARD	Conventional synthetic disease-modifying antirheumatic drug
CSH	Heterogeneous Compound Symmetry type covariance structure
DMARD	Disease-modifying antirheumatic drug
ECG	Electrocardiogram
EMA	European Medicines Agency
EQ-5D-3L	EuroQol Health State Profile – 5 Dimensions – 3 Levels

EQ-VAS	EuroQol Your own health state today - Visual Analog Scale
FACIT-F	Functional Assessment of Chronic Illness Therapy-Fatigue
FACS	Fluorescence-activated cell sorting
FAS	Full Analysis Set
GMMRM	Generalized Marginal Model for Repeated Measures
GLM	Generalized Linear Model
GH	General Health
hsCRP	High Sensitive C-reactive protein
HLA-B27	Human leukocyte antigen - B27
IBD	Inflammatory Bowel Disease
IR	Inadequate Response, Inadequate Responder
kg	kilogram
kg/m ²	kilogram per meter squared
LLOQ	Lower limit of quantification
LSM	Least squares mean
Ln	Natural logarithm
MAR	Missing at random
MASES	Maastricht Ankylosing Spondylitis Enthesitis Score
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
MCS	Mental component summary
MedDRA	Medical Dictionary for Regulatory Activities
mg/L	milligram per liter
MH	Mental Health
MMRM	Mixed Model for Repeated Measures
MNAR	Missing not at random
MR=NR	Missing response as non-response
N	Number of subjects
NA	Not Applicable
NRS	Numerical Rating Scale
NSAID	Nonsteroidal anti-inflammatory drug
NSAID-IR	Inadequate response to nonsteroidal anti-inflammatory drug treatment
OR	Odds ratio
PBO	Placebo
PCS	Physical component summary
PF	Physical Functioning
PGA	Patient Global Assessment of Disease
█	█
PRO	Patient-reported outcome
PP	Per Protocol
PT	Preferred Term
RP	Role-Physical
RE	Role-Emotional
SAFETY	Safety Analysis Set
SAP	Statistical analysis plan
SD	Standard deviation
SE	Standard error
SF	Social Functioning
SF-36v2	36-Item Short-Form Health Survey Version 2 Acute
SI	Sacroiliac
SJC	Swollen Joint Count
TMF	Trial Master File
TNF	Tumor necrosis factor

TNFi	Tumor necrosis factor inhibitor
TNFi-IR	Inadequate response to tumor necrosis factor inhibitor treatment
Tofa	Tofacitinib
TR	Health Transition
tsDMARD	targeted synthetic disease-modifying antirheumatic drug
TTO	Time Trade-Off
UK	United Kingdom
VT	Vitality
WPAI	Work Productivity & Activity Impairment
vs.	versus
WHO-Drug	International classification of medicines created by World Health Organization Programme for International Drug Monitoring
Z _{0.975}	97.5 th percentile of the standard normal distribution

9.10. Definition of NSAID-Inadequate Response (NSAID-IR)

For prior use of a given NSAID, the criterion of an inadequate response (NSAID-IR) can be satisfied in any one of the following three possible conditions:

1. Discontinued due to adverse event or intolerability: NSAID was taken and was discontinued due to adverse event prior to Day 1 of the study. This will be checked using Prior and Concomitant Medications NSAID CRF page where the start date and end date were before Day 1 of the study and "Adverse Event" was checked under Reason for Discontinuation;
2. Discontinued due to lack of efficacy: NSAID was taken and was discontinued due to lack of efficacy prior to Day 1 of the study with at least a 4-week (ie, 28 days) trial prior to discontinuation. This will be checked using Prior and Concomitant Medications NSAID CRF page where the start date and end date were before Day 1 of the study, the NSAID treatment duration met the minimum duration: end date - start date + 1 \geq 28 days, and "Lack of Efficacy" was checked under Reason for Discontinuation; or
3. Ongoing but with lack of efficacy: NSAID was taken and was ongoing through Day 1 of the study with at least a 4-week (ie, 28 days) trial prior to Day 1 of the study, with active AS at Screening and Day 1 (ie, Baseline). This will be checked using (a) Prior and Concomitant Medications NSAID CRF page where the start date was before Day 1, the "Ongoing?" box was checked Yes, the NSAID treatment duration met the minimum duration: date of first dose of the study (ie Day 1) - start date + 1 \geq 28 days, (b) the date of diagnosis of Ankylosing Spondylitis was present on the Primary Diagnosis CRF page at Screening visit, and (c) BASDAI score of \geq 4 and back pain score (BASDAI question 2) of \geq 4 at both Screening and Baseline (ie Day 1) visits on the BASDAI CRF page.

Multiple occurrences of the above three conditions for the same NSAID whether in different doses or different dosing frequency are counted as only one unique NSAID-IR. Based on all NSAIDs entered into the Prior and Concomitant Medications NSAID CRF page, the Primary Diagnosis CRF page (for condition 3), and the BASDAI CRF page (for condition 3) for each

subject, the number of NSAIDs-IR per this subject will be counted. A subject has to meet at least 2 NSAIDs-IR to enroll.

9.11. Derivation of the Stratification Factor

The stratification factor will be derived from the Prior and Concomitant Medications DMARD CRF page, and specifically from the biologic DMARD (bDMARD) subcategory. Each of the strata will be derived as follows.

1. bDMARD-naïve: subjects who had no prior exposure to any bDMARD prior to Day 1 of the study. This will be checked using the DMARD CRF page where subjects should have no entry on the bDMARD subcategory.
2. TNFi-IR: subjects who had inadequate response to one or at most two different TNFi therapies. For prior use of a given TNFi, an inadequate response is defined in one of the following two possible conditions:
 - a. Discontinued due to adverse event or intolerability: TNFi was taken and was discontinued due to adverse event prior to Day 1 of the study. This will be checked using the Prior and Concomitant Medications DMARD CRF page (bDMARD subcategory) where the start date and end date were before Day 1 of the study and "Adverse Event" was checked under Reason for Discontinuation; or
 - b. Discontinued due to lack of efficacy: TNFi was taken and was discontinued due to lack of efficacy prior to Day 1 of the study. This will be checked using the Prior and Concomitant Medications DMARD CRF page (bDMARD subcategory) where the start date and end date were before Day 1 of the study and "Lack of Efficacy" was checked under Reason for Discontinuation.
3. bDMARD Use (Non-IR): subjects who had prior exposure to any bDMARD (TNFi or non-TNFi bDMARD) prior to Day 1 of the study documented on the DMARD CRF page (bDMARD subcategory).

Subjects in (1) above are grouped to the "bDMARD-naïve" stratum and subjects in (2) and (3) are grouped to the "TNFi-IR or bDMARD Use (Non-IR)" stratum.

Any subjects with three or more TNFi-IR or with non-TNFi bDMARD-IR due to protocol deviation will be included in the stratum of "TNFi-IR or bDMARD Use (Non-IR)" for analysis purpose.