Protocol A3921234

A PHASE 3, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED STUDY OF THE EFFICACY AND SAFETY OF TOFACITINIB (CP-690,550) IN CHINESE SUBJECTS WITH ACTIVE PSORIATIC ARTHRITIS AND AN INADEQUATE RESPONSE TO AT LEAST ONE CONVENTIONAL SYNTHETIC DMARD

Statistical Analysis Plan (SAP)

Version: 2

Date: 22-FEB-2021

[redacted]

[redacted]

TABLE OF CONTENTS

[redacted]

[redacted]

[redacted]

[redacted]

LIST OF TABLES

LIST OF FIGURES

APPENDICES

[redacted]

[redacted]

1. VERSION HISTORY

Table 1 Summary of Major Changes in SAP Amendments

SAP Version	Associated Protocol Amendment	Change	Rationale
1 (17-APR-2018)	Protocol Amendment 1 (20-NOV-2017)	None	Not Applicable
2 (22-FEB-2021)	Protocol Amendment 4 (04-JUN-2020)	Sections 2.2 and 5.3: Protocol languages were updated	To be consistent with protocol amendment 4.
		Section 3.4.3: Added details for prior DMARDs use summaries.	To clarify how prior DMARDs use will be summarized.
		Section 4.2.1: Added one additional criterion for subject exclusion from Per-Protocol analysis set.	To exclude subjects with history of or current autoimmune rheumatic disease or rheumatic inflammatory disease other than PsA.
		Sections 5.2 and 5.2.1.1: Updated method for reporting response rate when the rate is 0 or 100%.	To clarify the reporting method.
		Section 6.1.1.2 and Appendix 1: Sensitivity analysis in the per-protocol analysis set is renamed a supportive analysis.	Sensitivity analysis is reserved for a different statistical method but otherwise everything else (including the analysis set and missing data handling method) stays the same.
		Sections 6.2.1 and 6.2.2: Line graph will	For TLF reduction.

		only be produced for selected endpoints.	
		Section 6.4: Updated specifications for subgroup analysis by baseline csDMARD use.	To clarify the categories in subgroup analysis by baseline csDMARD use.
		Section 6.6.1: Analysis periods for AE display were updated.	For TLF reduction.
		Section 6.7: A new section was added for additional analyses to address COVID-19 pandemic impacts	To add additional analyses addressing COVID-19 pandemic impacts.
		Appendix 2: Added clarifications for definitions of visit windows in reporting.	To clarify visit window definition in special cases.
		Appendix 3: Added details in efficacy endpoint calculations and updated AEs of special interest and list of DMARDs, corrected the waist circumference threshold.	To make sure data derivation will be programmed correctly.
		Appendix 4: Re-formatted the descriptions of tipping point analysis into a step-by-step format and pre-specified the number of multiple imputations and the	To make sure tipping point analysis will be programmed correctly.

[redacted]

[redacted]

		values of MNAR adjustment.	
		Other updates in Sections 3.2.1, 3.4, 4.2, 5.3, 6.2.1, Appendix 1 and Appendix 3.	Minor corrections and clarifications for consistency.

2. INTRODUCTION

Psoriatic arthritis (PsA) is a chronic inflammatory autoimmune disease characterized by joint inflammation and destruction, psoriatic skin lesions, enthesitis, dactylitis, spondylitis, progressive disability and adverse effects on quality of life. Tofacitinib (CP-690,550) is a potent and selective inhibitor of the Janus Kinase (JAK) family of kinases. While tofacitinib shows nanomolar inhibitory potency against all JAK family kinases in enzymatic assays, it shows functional specificity for JAK1 and JAK1/3 over JAK2 in cell-based assays. The broad effects of JAK1/3 inhibition on multiple cytokine pathways provide the rationale for developing tofacitinib as treatment for PsA.

This is a 6-month Phase 3 study designed to evaluate the efficacy and safety of tofacitinib 5 mg BID as a treatment for PsA in Chinese subjects with active PsA who have had an inadequate response in their PsA to at least one conventional synthetic disease-modifying anti-rheumatic drug (csDMARD).

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study A3921234. This document may modify the plans outlined in the protocol; however, any major modifications of the primary endpoint definition or its analysis will also be reflected in a protocol amendment.

2.1. Study Objectives

2.1.1. Primary Objectives

- To compare efficacy of tofacitinib 5 mg BID versus placebo for treatment of rheumatological signs and symptoms of PsA in Chinese subjects with active PsA who have had an inadequate response to at least one csDMARD.
- To compare the safety and tolerability of tofacitinib 5 mg BID versus placebo in Chinese subjects with active PsA who have had an inadequate response in PsA to at least one csDMARD.

[redacted]

[redacted]

2.1.2. Secondary Objectives

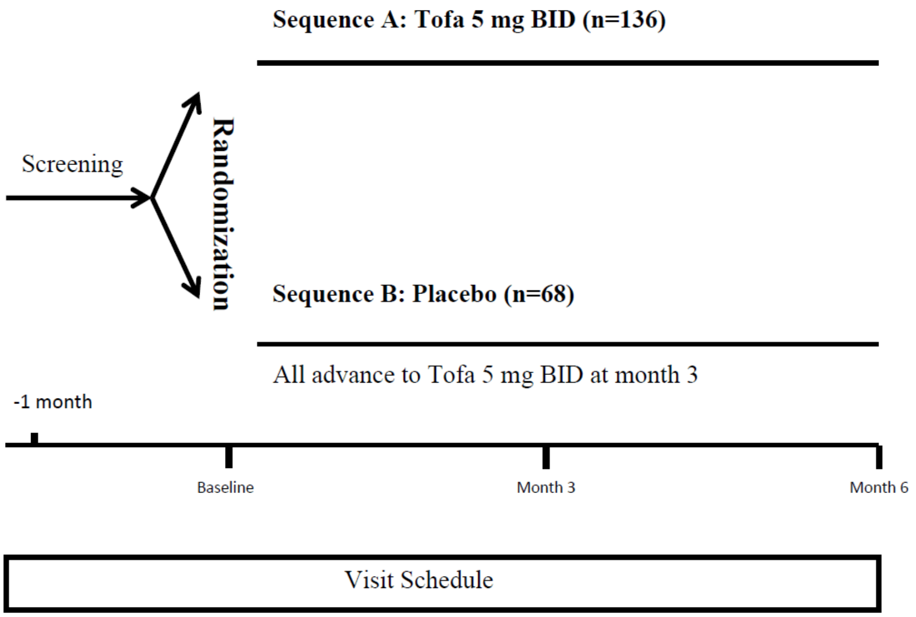
- To compare physical function status after administration of tofacitinib 5 mg BID versus placebo in Chinese subjects with active PsA who have had an inadequate response in PsA to at least one csDMARD.
- To compare the effects of tofacitinib 5 mg BID versus placebo on all health outcome measures in Chinese subjects with active PsA who have had an inadequate response to at least one csDMARD.
- To compare the efficacy of tofacitinib 5 mg BID versus placebo for the treatment of dermatological signs and symptoms of PsA in Chinese subjects who have had an inadequate response to at least one csDMARD.

2.2. Study Design

This is a Phase 3, randomized, 6-month, double-blind, placebo-controlled, parallel-group study designed to evaluate the efficacy and safety of tofacitinib in adult, Chinese subjects with active PsA who had an inadequate response in their PsA to at least one csDMARD. All subjects in the study will be evaluated for risk factors for venous thromboembolism. A total of approximately 204 subjects will be randomized in a 2:1 ratio to one of the following two parallel treatment sequences. The enrollment will be monitored to cap the proportion of subjects with baseline swollen joint count ≤5 at approximately 38% (i.e., approximately 78 subjects).

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

Figure 1 Study Design


Sequence A: Tofa 5 mg BID (n=136)
Screening
Randomization
Sequence B: Placebo (n=68)
All advance to Tofa 5 mg BID at month 3
-1 month
Baseline
Month 3
Month 6
Visit Schedule

[redacted]

*Primary study endpoints of ACR50 will be obtained at Month 3. All subjects randomized to placebo will receive tofacitinib 5 mg BID in a blinded manner after Month 3.

During the study, subjects are required to remain on a stable dose of one csDMARD, e.g., methotrexate or sulfasalazine, and should remain on that dose throughout the study.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

Visual analog scale (VAS) data will need to be rescaled prior to any calculation and analysis. VAS is recorded on CRF in terms of length at mark (X in mm) and overall length of line (Y in mm). The rescaled VAS for use in analysis will be: $Z=X/Y \times 100$ mm or $X/Y \times 10$ cm, depending on endpoints or their use in defining other endpoints.

3.1. Primary Endpoint(s)

- ACR50 responder rate at Month 3.

3.2. Secondary Endpoint(s)

3.2.1. Secondary Efficacy Endpoints

- ACR20 and ACR70 responder rates at all timepoints;
- ACR50 responder rates at all timepoints other than Month 3 (Note that ACR50 responder rate at Month 3 is already mentioned as the primary endpoint);
- Change from baseline (Δ) in ACR response criteria components (HAQ-DI, C-reactive protein [CRP], Patient’s Assessment of Arthritis Pain, Patient’s Global Assessment of Arthritis, Physician’s Global Assessment of Arthritis, swollen joint count, tender/painful joint count) at all timepoints;
- HAQ-DI response defined as a decrease from baseline ≥ 0.30 for subjects with baseline HAQ-DI ≥ 0.30 at all timepoints; HAQ-DI response defined as a decrease from baseline ≥ 0.35 for subjects with baseline HAQ-DI ≥ 0.35 at all timepoints;
- Psoriatic Arthritis Response Criteria (PsARC) at all timepoints;
- Physician’s Global Assessment of Psoriasis (PGA-PsO) at Month 1, Month 3 and Month 6;
    - PGA-PsO response of clear or almost clear and a ≥2-step improvement from baseline for subjects with baseline PGA-PsO ≥ 2;
    - ΔPGA-PsO for subjects with baseline PGA-PsO > 0;
- Psoriasis Area and Severity Index 75 (PASI75) response at Month 1, Month 3 and Month 6 for subjects with baseline psoriatic body surface area (BSA) ≥ 3% and baseline PASI > 0;

[redacted]

[redacted]

- Dactylitis severity score (DSS) at Month 1, Month 3 and Month 6 for subjects with baseline DSS > 0 (see Appendix 3 Section 12);
    - ΔDSS;
    - Resolution of dactylitis defined as a subject achieving DSS = 0;
- Enthesitis score [using Leeds enthesitis index (LEI)] at Month 1, Month 3 and Month 6 for subjects with baseline LEI > 0 (see Appendix 3 Section 11);
    - ΔLEI;
    - Resolution of enthesitis defined as a subject LEI = 0;

3.2.2. Secondary Physical Function and Health Outcome Measures

Assessed at baseline, Months 1, 3 and 6:

- Short-Form-36 Health Survey (ΔSF-36) Version 2, Acute;

  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 (PCS and MCS, respectively) (see Appendix 3 Section 6).
- EuroQol 5-Dimension Health State Profile – 3 Level Version (ΔEQ-5D-3L).

  This instrument yields 6 endpoints: scores from the five domains (ΔEQ-5D) and an EQ-VAS score of the subject's health state today (ΔEQ-VAS) (see Appendix 3 Section 7).

3.3. Other Endpoints

3.3.1. Other Efficacy Endpoints

- ΔDAS28-3(CRP) at all timepoints;
- Physician's Global Assessment of Psoriatic Arthritis (ΔPGA-PsA) at Month 1, Month 3 and Month 6;
- Percent change from baseline (%Δ) in PASI and PASI clinical sign component scores at Month 1, Month 3 and Month 6 for subjects with baseline BSA ≥ 3% and baseline PASI > 0;
- %ΔBSA at Month 1, Month 3 and Month 6 for subjects with baseline BSA > 0%;
- Nail Psoriasis Severity Index (ΔNAPSI) Score at Month 1, Month 3 and Month 6 for subjects with baseline NAPSI > 0.

  This is a score ranging from 0-8 based on 1 targeted nail.

[redacted]

[redacted]

Health Outcome Measures Assessed at baseline, Months 3 and 6:

- Work Productivity and Activity Impairment – Psoriatic Arthritis Questionnaire (ΔWPAI-PsA).

3.4. Baseline Variables

3.4.1. Demographic Characteristics

- Baseline age (2 categorizations: 18-44, 45-64, ≥ 65 years; 18-44, 45-64, 65-74, 75-84, ≥ 85 years; and continuous in years);
- Sex (female, male);
- Baseline body weight (< 60, ≥ 60 to ≤ 100, > 100 kg; and continuous in kg),
- Baseline height (cm, continuous);
- Baseline Body Mass Index (BMI: < 18.5, 18.5 to < 25, 25 to < 30, 30 to < 40, and ≥ 40 kg/m^2; and continuous in kg/m^2);
- Screening smoking status (never smoked, former smoker, current smoker); smoking duration (years) for former and current smokers.
- Alcohol use [never, former, current; continuous (units/week) for subjects who consumed any alcohol].

3.4.2. Baseline Disease Characteristics

- PsA duration (< 2, ≥ 2 years; and continuous in years);
- Baseline PsA Subtype: (< 5 joints, ≥ 5 joints);
- Screening presence of distal interphalangeal joints involvement (yes, no);
- Screening presence of arthritis mutilans (yes, no);
- Baseline swollen joint count (66) (continuous);
- Baseline tender/painful joint count (68) (continuous);
- Baseline HAQ-DI (continuous);
- Baseline PGA-PsO (0, 1, 2, 3, 4);
- Baseline PGA-PsO (continuous) for those subjects with PGA-PsO > 0 at baseline;
- Baseline PASI (0, > 0 to ≤ 20, > 20);
- Baseline PASI (continuous) for those subjects with BSA ≥ 3% and PASI > 0 at baseline;
- Baseline total psoriatic BSA (0, > 0 to < 3, ≥ 3%);

[redacted]

[redacted]

- Baseline total psoriatic BSA (continuous) for those with BSA > 0% at baseline,
- Number (%) of subjects with baseline PGA-PsO ≥ 3, PASI ≥ 12 and BSA ≥ 10%;
- Baseline presence of dactylitis (yes, no). Yes is defined for those subjects with baseline DSS > 0;
- Baseline DSS (continuous) for those subjects with DSS > 0 at baseline;
- Baseline presence of enthesitis measured by LEI (yes, no). Yes is defined for those subjects with baseline LEI > 0;
- Baseline enthesitis index measured by LEI (continuous) for those with LEI > 0 at baseline;
- Baseline SF-36 (8 domains, physical component score [PCS] and mental component score [MCS]);
- Baseline EQ-5D-3L (5 domains) and EQ-5D VAS;
- Baseline DAS28-3(CRP);
- Baseline CRP (≤ 2.87, > 2.87 mg/L; and continuous in mg/L);
- Baseline PGA-PsA;
- Baseline presence of nail psoriasis measured by NAPSI (yes, no). Yes is defined for those subjects with baseline NAPSI > 0;
- Baseline NAPSI for those with NAPSI>0 at baseline;
- Baseline diabetes mellitus (yes, no) (see Appendix 3 Section 14 for definition);
- Baseline metabolic syndrome (yes, no) (see Appendix 3 Section 15 for definition);
- Baseline cardiovascular risks;
- Baseline rheumatoid factor positive (yes, no);
- Baseline cyclic citrullinated peptide antibody positive (yes, no);

3.4.3. Prior and Baseline Treatments for Psoriatic Arthritis

- Prior oral corticosteroids use (oral only: yes, no);
- Baseline oral corticosteroid use (oral only: yes, no), Day 1 predose;
- Prior NSAIDs use (yes, no);
- Baseline NSAID use (yes, no), Day 1 predose;
- Prior Other DMARDs use (yes, no). Other DMARDs are DMARDs other than bDMARDs and csDMARDs for primary diagnosis;

[redacted]

[redacted]

- Number of prior DMARDs experience (1 csDMARD, 2 csDMARDs, ≥ 3 csDMARDs, ≥ 1 bDMARD). This categorization is based on bDMARDs and csDMARDs experience, regardless of other DMARDs use. Subjects who were treated with any bDMARD or both csDMARDs and bDMARDs will be counted in the "≥ 1 bDMARD" category;
- Prior csDMARDs experience (yes for csDMARDs only, no). This categorization is based on bDMARDs and csDMARDs experience, regardless of other DMARDs use.

3.5. Safety Endpoints

- Incidence and severity of adverse events (AEs);
- AEs of special interest (see Appendix 3 Section 13);
- Clinical laboratory tests (e.g., clinical chemistry, hematology);
- Vital sign measurements (blood pressure, pulse rate and temperature);
- Physical examinations;
- Electro-cardiogram (ECG) measurements.

3.5.1. Adverse Events

An adverse event is considered treatment emergent relative to a given treatment if:

- the event occurs for the first time during the effective duration of treatment and was not seen prior to the start of treatment (for example, during the baseline or run-in period), or
- the event was seen prior to the start of treatment but increased in severity during treatment.

Any event occurring after the first day of the treatment will be attributed to the corresponding treatment period regardless whether the event occurs on- or off-treatment.

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. Different analyses will be performed for different tiers (See Section 6.6.1).

Tier-1 events: These are pre-specified events of clinical importance. The AEs of special interest (see Appendix 3 Section 13) are equivalent to Tier-1 events.

Tier-2 events: These are events that are not tier-1 but are "common". A MedDRA PT is defined as a tier-2 event if occurring in at least 4 subjects in any treatment group.

Tier-3 events: These are events that are neither tier-1 nor tier-2 events.

[redacted]

[redacted]

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.

3.5.2. Laboratory Data

Blood and urine samples will be collected at the time points identified in the protocol for clinical safety laboratory tests and exploratory biomarkers (Table 2). Unscheduled clinical laboratory tests may be performed at any time during the study to assess any perceived safety concerns.

Table 2 Clinical Laboratory Testing

Laboratory Testing Profile	Tests Included
Laboratory Tests Required at Screening Only	QuantiFERON®-TB Gold In-Tube, hepatitis C virus antibody (HCV Ab), hepatitis C virus RNA (HCV RNA)[a], hepatitis B surface antigen (HBsAg), hepatitis B core antibody (HBcAb), hepatitis B surface antibody (HBsAb)[b], HIV-1/HIV-2 antibody screen, Prothrombin time (PT/INR)[c]; FSH (optional for post-menopausal women only)
Hematology	Hemoglobin, hematocrit, RBC, RBC morphology, reticulocyte (abs); White blood cell (WBC) count and differential, [neutrophils (%, abs), lymphocytes (%, abs), monocytes (%, abs), eosinophils (%, abs), basophils (%, abs)], platelets Hemaglobin A1c (HbA1c)
Chemistry Panel	Urea nitrogen, creatinine, glucose, calcium, sodium, potassium, bicarbonate, chloride, total protein, total bilirubin, direct bilirubin, indirect bilirubin, alanine transaminase (ALT), aspartate transaminase (AST), alkaline phosphatase, gamma-glutamyl transferase (GGT), albumin, creatine kinase (CK) Rheumatoid Factor (RF), Cyclic Citrullinated Peptide Antibody (CCP)
Lipid Panel	Fasting total cholesterol, HDL, LDL, triglyceride; apolipoprotein A-1, B and other lipoprotein tests potentially including particle size measurements
Urinalysis	Specific gravity, pH, protein, glucose, ketones, blood, leukocyte esterase. Urine HCG pregnancy testing for women of childbearing potential. Microscopy and/or culture to be performed if clinically indicated or if urinalysis results positive (blood, protein or leukocyte esterase/WBC)
Acute Phase Reactants	C-reactive protein (CRP, tested centrally)

[a] Only subjects who are HCV Ab positive should be reflex tested for HCV RNA.

[b] Only subjects who are HBsAg$^-$ and HBcAb$^+$ should be reflex tested for HBsAb.

[c] All subjects will be screened for normal prothrombin time (PT/INR). PT should also be evaluated to rule out acute hepatic injury in cases of hepatic enzyme elevations.

4. ANALYSIS SETS

Below is a description of the analysis sets defined for this study. Note that subjects who are screened but not randomized will appear on the subject evaluation table and that subjects who are randomized but not treated will appear on the subject evaluation table; this will the extent of how much data for these subjects will be reported. If a subject is treated but not randomized, then the subject will be excluded from any analyses. A narrative will be provided for this subject in the clinical study report (CSR).

Data for all subjects will be assessed to determine if subjects meet the criteria for inclusion in each analysis population prior to unblinding and releasing the database and classifications will be documented per standard operating procedures.

4.1. Full Analysis Set

The Full Analysis Set (FAS) will include all subjects randomized and who have received at least one dose of randomized study drug (tofacitinib or placebo). Subjects will be analyzed in the treatment groups as they are randomized regardless of what treatment they received. FAS is the primary patient population for the primary endpoint.

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 have 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 since the missingness mechanism is likely to be missing completely at random (MCAR) (Rubin, 1987). 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 (since subjects not responding are more likely to drop out prior to a post-baseline measure than those who are responding). 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 be a subset of FAS and will exclude all subjects who had a protocol deviation thought to have a material impact on the primary efficacy analysis. The PP analysis set will be used in the PP efficacy analysis for the primary endpoint of ACR50 at Month 3.

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 4.2.1 and Section 4.2.2 below.

[redacted]

[redacted]

4.2.1. Protocol Deviations Assessed Prior to Randomization

- Subjects who failed to meet the Classification Criteria for Psoriatic Arthritis (CASPAR) diagnostic criteria for PsA at screening. The PsA diagnosis at screening will be recorded on the PRIMARY DIAGNOSIS CRF where PSORIATIC ARTHRITIS is pre-filled. Subjects who did not have the CRF filled out will be considered as having this exclusion criteria met.

- Subjects who had < 3 tender/painful joints on motion out of 68 joints assessed or < 3 swollen joints out of 66 joints assessed at either screening or baseline (i.e., Day 1 Pre-dose). The “JOINT SWELLING AND TENDERNESS 66/68 – JOINT COUNT ASSESSMENT” CRF will be used to check this condition.

- Subjects who failed to meet the criterion of “an inadequate response (IR) to at least one csDMARD due to lack of efficacy (LOE) and/or toxicity/lack of tolerance (due to treatment-related AE)”. csDMARD-IR is defined as those csDMARDs that were discontinued due to AE, LOE, or both AE/LOE on or prior to Day 1, or a csDMARD that was started prior to Day 1, but was not discontinued and continued beyond Day 1 regardless of csDMARD dose change.

- Subjects who met the exclusion criterion of “history of or current autoimmune rheumatic disease or rheumatic inflammatory disease other than PsA”. The “SIGNIFICANT MEDICAL HISTORY” CRF and anti-CCP tested at baseline will be used to check this condition. Subjects who were considered to have current autoimmune rheumatic disease or rheumatic inflammatory disease other than PsA by the investigator and/or the Sponsor will also be excluded.

4.2.2. Protocol Deviations Assessed Post-Randomization

- Subject who was randomized but took or received incorrect treatment other than the randomized treatment for the entire duration from baseline through Month 3. This will be checked manually since there is no CRF for this condition.

- Subject who had < 80% compliance with tablet on 2 consecutive visits for the period from baseline through Month 3. The study treatment compliances will be calculated based on “ORAL DOSING” CRF. Both randomized treatments (i.e., tofacitinib and placebo) will be taken into account.

- Subject who received rescue medication for PsA on the preceding day or the same day of the primary endpoint assessment at Month 3. The rescue medication is recorded on “PRIOR AND CONCOMITANT MEDICATIONS : OTHER” CRF in the “RESCUE MEDICATIONS” category.

A list of DMARDs is given in Appendix 3 Section 17. The list will be reviewed and updated as new versions of MedDRA dictionary become effective and any updates will be maintained in a programming plan.

4.3. Safety Analysis Set

The safety analysis set (SAFETY) will include all subjects who received at least one dose of the study drug (tofacitinib or placebo). Subjects will be classified according to the actual study treatment received.

4.4. Other Analysis Sets

4.4.1. Endpoint Specific Analysis Sets

Subjects will be excluded from FAS for analyzing a specific endpoint if the criterion for inclusion is not met for the endpoint as described in Table 3.

Table 3 Endpoint Specific Analysis Sets

Instrument/Endpoint	Inclusion	Rationale
HAQ-DI responder (decrease ΔHAQ-DI $\geq$ 0.30)	Include subjects with baseline HAQ-DI $\geq$ 0.30	Restrict baseline HAQ-DI to $\geq$ 0.30 to allow room for response.
HAQ-DI responder (decrease ΔHAQ-DI $\geq$ 0.35)	Include subjects with baseline HAQ-DI $\geq$ 0.35	Restrict baseline HAQ-DI to $\geq$ 0.35 to allow room for response.
ΔPGA-PsO	Include subjects with baseline PGA-PsO > 0	Baseline PGA-PsO = 0 means no PsO disease at baseline.
PGA-PsO responder (PGA-PsO = 0 or 1 and decrease ΔPGA-PsO $\geq$ 2)	Include subjects with baseline PGA-PsO $\geq$ 2	Restrict baseline PGA-PsO to $\geq$ 2 to allow room for response.
PASI75	Include subjects with baseline BSA $\geq$ 3% and baseline PASI > 0	PASI is not assessed if baseline BSA < 3% per the study protocol; PASI = 0 means absence of psoriasis as measured by PASI.
ΔDSS	Include subjects with baseline DSS > 0	Baseline DSS = 0 means absence of dactylitis. Dactylitis is not an enrollment criterion so it is expected that only a subset of the subjects will have dactylitis at baseline.
Resolution of dactylitis (DSS = 0)	Include subjects with baseline DSS > 0	Baseline DSS = 0 means absence of dactylitis. Dactylitis is not an enrollment criterion so it is expected that only a subset of the subjects will have dactylitis at baseline.

[redacted]

[redacted]

ΔLeeds Enthesitis Index (ΔLEI)	Include subjects with baseline LEI > 0	Baseline LEI = 0 means absence of enthesitis as measured by LEI. Enthesitis is not an enrollment criterion so it is expected that only a subset of the subjects will have enthesitis at baseline.
Resolution of enthesitis (LEI = 0)	Include subjects with baseline LEI > 0	Baseline LEI = 0 means absence of enthesitis as measured by LEI. Enthesitis is not an enrollment criterion so it is expected that only a subset of the subjects will have enthesitis at baseline.
%Δ in PASI and PASI clinical sign component scores	Include subjects with baseline BSA ≥ 3% and baseline PASI > 0	PASI is not assessed if baseline BSA < 3% per the study protocol; PASI = 0 means absence of psoriasis as measured by PASI.
%ΔBSA	Include subjects with baseline BSA > 0%	Percent change from baseline cannot be calculated if baseline BSA = 0%.
ΔNAPSI	Include subjects with baseline NAPSI > 0	Baseline NAPSI=0 means absence of nail psoriasis. Nail psoriasis is not an enrollment criterion so it is expected that only a subset of the subjects will have nail psoriasis at baseline.

5. GENERAL METHODOLOGY AND CONVENTIONS

5.1. Hypotheses and Decision Rules

This protocol is designed to establish the superiority of tofacitinib 5 mg BID to placebo for treatment of rheumatological signs and symptoms of active PsA in subjects who have had an inadequate response in PsA to at least one csDMARD.

The null hypothesis is that there is no difference between tofacitinib 5 mg BID and placebo, and the alternative hypothesis is that there is a difference between tofacitinib 5 mg BID and placebo.

Tofacitinib 5 mg BID will be considered superior to placebo with respect to ACR50 response rate at Month 3 if the difference between tofacitinib 5 mg BID and placebo is statistically significant at the 2-sided 5% level, i.e., 2-sided p-value ≤ 0.05.

For endpoints other than the primary (ACR50 response rate at Month 3), this study is not statistically powered to test the difference between tofacitinib 5 mg BID and placebo. Thus, any reported confidence intervals (CIs) and p-values should only be considered nominal.

5.2. General Methods

As subjects randomized to treatment group of placebo → tofacitinib 5 mg BID (where → means switching to) will advance from placebo to open-label treatment of tofacitinib 5 mg BID at Month 3, the treatment label used for reporting visits up to Month 3 will be "Placebo" and treatment label for reporting visits after Month 3 through Month 6 will be "Placebo → Tofacitinib 5 mg BID".

For analyses through Month 3, the following treatment comparison will be made at each time point, where applicable:

- Tofacitinib 5 mg BID vs. Placebo.

For analyses after Month 3 through Month 6, 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 ACR50 will be between tofacitinib 5 mg BID and placebo at Month 3.

Descriptive Summaries:

In general, the data for all continuous endpoints will be summarized by time point and treatment sequence in tables containing descriptive statistics (n, mean, standard deviation (s.d.), standard error (SE) of the mean, minimum, 1st, 2nd (median) and 3rd quartiles and maximum) for baseline and change or percent change from baseline for those endpoints measured at baseline. The data for all binary endpoints will be summarized by treatment group and by time point in tables showing descriptive statistics: N, n (i.e., 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 will not be reported. 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.

[redacted]

[redacted]

5.2.1. Analyses for Binary Data

5.2.1.1. Normal Approximation

The normal approximation for the difference in binomial proportions (such as ACR50 response rate) can be used to test the superiority of tofacitinib 5 mg BID to placebo at post-baseline timepoints. The normal approximation to the test statistic for the difference in binomial random variables is calculated as

$$Z = \frac{\hat{p}_t - \hat{p}_c}{\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{N_t} + \frac{\hat{p}_c(1-\hat{p}_c)}{N_c}}},$$

where $\hat{p}$ refers to the relative frequency, N to the number of observations, the subscript *c* refers to the comparator group (e.g., placebo group) and the subscript *t* refers to the test group (e.g., tofacitinib 5 mg BID group) so that test statistics and p-values can be calculated for the contrast comparing the two treatment groups.

Two-sided 95% confidence intervals will be formed by

$$\hat{p}_t - \hat{p}_c \pm z_{0.975}\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{N_t} + \frac{\hat{p}_c(1-\hat{p}_c)}{N_c}}$$

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 distribution function.

If there is no (i.e., 0) response or 100% response in any one of the two treatment groups for the comparison, e.g., tofacitinib 5 mg BID vs. placebo, when calculating the proportions above, 0.5 will be added to the number of responses (i.e., numerator) and 1 will be added to the denominator in each treatment corresponding to the pair of comparison for calculating the treatment difference, standard error (i.e., $\sqrt{var(\hat{p}_t - \hat{p}_c)}$), 95% CI and 2-sided p-value (Agresti, 2002). Estimated response rates and standard errors will be reported without adjustment for both groups. (Note that this adjusted method also applies to the case where the response rate is 0% in one treatment group and 100% in the other.)

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

The final results will be expressed in percentages, i.e., (proportions × 100)%.

[redacted]

[redacted]

5.2.1.2. Generalized Marginal Model for Repeated Measures (GMMRM)

As a supportive analysis of ACR50, scheduled to be collected multiple times during the first 3 months of the placebo-controlled period (Week 2, Months 1, 2 and 3) of the study, a generalized marginal linear model for repeated measures (GMMRM) will be used. This model will have fixed effects for time (discrete), treatment, and time by treatment interaction; the dependent variable will be logit of the probability of 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 AR(1) variance-covariance matrix for both treatment sequences 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 indicate good performance as measured by type I error, confidence interval coverage and bias even in the presence of data that is MAR (Liu and Zhan, 2011).

From this model, one can obtain estimate of the probability of response for each treatment sequence at each timepoint as well inferential comparisons (odds ratio, CI and p-value) between treatment sequences.

In the event that a response rate is 0 or 100% at a particular visit for a treatment sequence that may lead to convergence issues, data from that visit for both treatment sequences 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 binary data measured during the placebo-controlled period at Week 2, Months 1, 2 and 3 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 ACR50 at Month 3. 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).

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 is used for the analysis, 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, and a latent subject-level random effect. The logit link is used to model the ACR50 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 data without imputation are included in this generalized linear mixed effect model to estimate the model parameters.

[redacted]

[redacted]

Imputation of missing ACR50 response is performed only at Month 3 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 ACR50 response at Month 3 for subjects with missing response in the tofacitinib 5 mg BID group and placebo group independently (i.e., 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 ACR50 response at Month 3 for subjects with missing response in the tofacitinib 5 mg BID group or placebo group independently to assess when the conclusion might change (i.e., 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) will be used to combine the results of multiple imputed samples for inference. More detailed descriptions are provided in Appendix 4.

5.2.2. Analyses for Continuous Data

Continuous endpoints will be analyzed by Mixed Model Repeated Measures (MMRM) that includes the fixed effects of treatment, visit, treatment by visit interaction, and baseline value. An unstructured variance covariance matrix for within-subject correlation (an alternative variance covariance matrix will be used if the model does not converge under the unstructured variance covariance matrix). Comparison of the two treatment sequences at each post-baseline timepoint will be generated from this model (providing 2-sided p-values and 95% CI).

5.2.3. Analyses of Tier-2 Adverse Events

Number and percentage of subjects with AEs over the first 3 months will be provided for each treatment group. Tier-2 events will be analyzed using asymptotic methods proposed by Miettinen and Nurminen (1985). Risk ratios (tofacitinib 5mg BID compared to placebo) and 2-sided 95% CI will be reported. P-values will not be reported for Tier-2 events.

5.3. Methods to Manage Missing Data

As stated in the protocol, in the primary analysis of the primary endpoint, ACR50 response rate at Month 3, missing values for any reason, e.g., due to a subject dropping from the study or due to the impact of COVID-19 pandemic, will be handled by setting the ACR50 value to nonresponsive. Note that this can be viewed as a composite endpoint in the sense that a response requires the subject completes a visit of interest, e.g., Month 3, and achieves a response per the defined ACR50 response criteria (otherwise it is considered a nonresponse). This method of handling missing response is known as missing response as non-response (MR=NR). This MR=NR approach will be used for all "response-type" endpoints at all timepoints except for the supportive analyses using GMMRM (see Section 5.2.1.2) and Tipping Point Analysis (see Section 5.2.1.3) for ACR50 up to Month 3. Note that if a subject

[redacted]

[redacted]

discontinues from the study at a visit of interest, say Month 3, but the value of the endpoint is not missing at Month 3, the value of the endpoint will be used for Month 3.

For a composite “response-type” endpoints (such as ACR50), if values in any of the components at a timepoint are missing, the component variables that are not missing will be used to determine the response status. If one could not determine the response status in the presence of missing components at the timepoint, then the composite response-type endpoint status is considered as missing and is handled by MR=NR for that timepoint for all of the inferential analyses except for the supportive analyses using GMMRM (see Section 5.2.1.2) and Tipping Point Analysis (see Section 5.2.1.3) for ACR50 up to Month 3.

Repeated measures data for continuous and ordered-categorical (analyzed as continuous) endpoints will be analyzed with MMRM. This model will yield unbiased estimates and valid inferences in the presence of data that is MCAR or missing at random (MAR). Given that subjects who drop out are scheduled to have an end of study evaluation, the more stringent condition of MAR (Rubin, 1987) is believed to be a reasonable assumption. Hence missing values will not be imputed in the routine analysis using the repeated measures model.

For the SF-36, EQ-5D and WPAI-PsA instruments, rules suggested by the developers of these 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.

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

For the demographic and baseline characteristics, if a value is missing, a missing category will be added to summaries for the categorical variables.

Missing values for safety endpoints will not be imputed.

6. ANALYSES AND SUMMARIES

6.1. Primary Endpoint(s)

6.1.1. ACR50 Response Rate at Month 3

6.1.1.1. Primary Analysis

Endpoint: ACR50 response rate at Month 3

- Analysis time points: Month 3
- Analysis population: FAS

[redacted]

[redacted]

- Method of imputation for missing data: MR=NR (specified in Section 5.3)
- Analysis methodology: The normal approximation to the difference in binomial proportions (specified in Section 5.2.1.1) will be used to test the difference between tofacitinib 5 mg BID and placebo and to generate 95% CI and p-value for the difference in response rates.

Reporting results:

- Response rate without imputation: N, frequency and percentage of responders, SE of the response rate and 95% CI at Month 3 will be presented by treatment group.
- Response rate with imputation: N, frequency and percentage of responders, SE of the response rate at Month 3 will be presented by treatment group.
- Response rate difference with imputation: Difference between treatment groups, and corresponding SE, 95% CI and p-value at Month 3 will be presented.

Figures:

- Line graph of ACR50 response rate (using imputed data) with ± SE will be plotted over time for both treatment sequences.

Note that ACR50 response rate at other timepoints will be analyzed by the same approach.

6.1.1.2. Supportive Analyses

Supportive Analysis 1: The analysis of ACR50 response rate at Month 3 described in Section 6.1.1.1 will also be conducted on the PP analysis set since the FAS may include instances of non-compliance which may diminish the potential efficacy of tofacitinib. This will serve as a supportive analysis of the primary endpoint.

Endpoint: ACR50 response rate at Month 3

- Analysis time points: Month 3
- Analysis population: PP analysis set
- Method of imputation for missing data: MR=NR (specified in Section 5.3)
- Analysis methodology: Normal approximation (specified in Section 5.2.1.1)

Reporting results:

- Response rate with imputation: N, frequency and percentage of responders, SE of the response rate at Month 3 will be presented by treatment group.
- Response rate difference with imputation: Difference between treatment groups, and corresponding SE, 95% CI and p-value at Month 3 will be presented.

[redacted]

[redacted]

Supportive Analysis 2: As a supportive analysis for the ACR50 response rate at Month 3, GMMRM without imputation for missing data will be applied to the available ACR50 data collected at post-baseline visits at Week 2, Months 1, 2 and 3.

Endpoint: ACR50 response rate

- Analysis time points: Week 2, Months 1, 2 and 3
- Analysis population: FAS
- Method of imputation for missing data: No imputation
- Analysis methodology: GMMRM (specified in Section 5.2.1.2)

Reporting results:

- Response rate (model-based): N, frequency and percentage of responders, and SE of the response rate up to Month 3 will be presented by treatment sequence.
- Odds ratio: Odds ratio between the two treatment sequences, and corresponding 95% CI and p-value at each analysis timepoint will be presented.

Supportive Analysis 3: Another supportive analysis on the ACR50 response rate at Month 3 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 Appendix 4.

Endpoint: ACR50 response rate

- Analysis time points: Week 2, Months 1, 2 and 3
- Analysis population: FAS
- Method of imputation for missing data: Multiple imputation
- Analysis methodology: Tipping point analysis (specified in Section 5.2.1.3 and Appendix 4)

Reporting results:

- Response rate (model-based): N, frequency and percentage of responders, and SE of the response rate at Month 3 will be presented by treatment sequence, for each pair of MNAR quantities (favorable or unfavorable) applied to the probability of the ACR50 response at Month 3 for subjects with missing response in the tofacitinib 5 mg BID group and placebo group.
- Response rate difference: Difference between treatment groups, and corresponding SE, 95% CI and p-value at Month 3 will be presented, for each pair of MNAR quantities (favorable or unfavorable) applied to the probability of the ACR50

[redacted]

[redacted]

response at Month 3 for subjects with missing response in the tofacitinib 5 mg BID group and placebo group.

6.2. Secondary Endpoint(s)

6.2.1. Binary Endpoints

The analysis and summary methods described in this section apply to all the endpoints below.

Endpoints: ACR20 and ACR70 response rates

- Analysis time points: All post-baseline timepoints through Month 6
- Analysis population: FAS

Endpoint: ACR50 response rate

- Analysis time points: All post-baseline timepoints prior to and after Month 3
- Analysis population: FAS

Endpoint: ΔHAQ-DI (decrease) ≥ 0.30 response rate

- Analysis time points: All post-baseline timepoints through Month 6
- Analysis population: Subjects in FAS with baseline HAQ-DI ≥ 0.30

Endpoint: ΔHAQ-DI (decrease) ≥ 0.35 response rate

- Analysis time points: All post-baseline timepoints through Month 6
- Analysis population: Subjects in FAS with baseline HAQ-DI ≥ 0.35

Endpoint: PsARC response rate

- Analysis time points: All post-baseline timepoints through Month 6
- Analysis population: FAS

Endpoint: PGA-PsO response (clear or almost clear and a 2-step improvement from baseline) rate

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline PGA-PsO ≥ 2

Endpoint: PASI75 response rate

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline BSA ≥ 3% and baseline PASI > 0

Endpoint: Resolution of dactylitis (achieving DSS = 0)

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline DSS > 0

[redacted]

Endpoint: Resolution of enthesitis (achieving LEI = 0)

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline LEI > 0

Method of imputation for missing data: MR=NR (specified in Section 5.3)

Analysis methodology: Normal approximation (specified in Section 5.2.1.1)

Reporting results:

- Response rate without imputation: N, frequency and percentage of responders, SE of the response rate, and 95% CI at each analysis timepoint will be presented by treatment group.
- Response rate with imputation: N, frequency and percentage of responders, SE of the response rate at each analysis timepoint will be presented by treatment group.
- Response rate difference with imputation: Difference between treatment groups, and corresponding SE, 95% CI and p-value at each analysis timepoint will be presented.

Figures:

- For endpoints assessed at more than 3 post-baseline timepoints, line graph of response rates (using imputed data) with ± SE will be plotted over time for both treatment groups.

6.2.2. Continuous Endpoints

The analysis and summary methods described in this section apply to all the endpoints below.

Endpoints: Change from baseline (Δ) in ACR response criteria components (HAQ-DI, C reactive protein [CRP], Patient's Assessment of Arthritis Pain, Patient's Global Assessment of Arthritis, Physician's Global Assessment of Arthritis, swollen joint count, tender/painful joint count)

- Analysis time points: All post-baseline timepoints
- Analysis population: FAS

Endpoint: ΔPGA-PsO

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline PGA-PsO > 0

Endpoint: ΔDSS

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline DSS > 0

[redacted]

[redacted]

Endpoint: ΔLEI

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline LEI > 0

Endpoint: ΔSF-36v2

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: FAS

Endpoint: ΔEQ-5D-3L and EQ-VAS

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: FAS

Method of imputation for missing data: No imputation

Analysis methodology: MMRM (specified in Section 5.2.2)

Reporting results:

- Raw data: N, mean, s.d., SE of the mean, minimum, 1st, 2nd (median) and 3rd quartiles and maximum at baseline and post-baseline visits will be presented by treatment sequence.
- Change from baseline: N, mean, s.d., SE of the mean, minimum, 1st, 2nd (median) and 3rd quartiles and maximum will be presented for each analysis time point by treatment sequence. The LS means, SE for the LS means, and LS mean of the difference between the two treatment sequences with corresponding SE and 95% CI will be presented for each analysis timepoint.

Figures:

- For endpoints assessed at more than 3 post-baseline timepoints, line graph of LS mean change from baseline with ± SE will be plotted over time for both treatment sequences.

6.3. Other Endpoint(s)

The following continuous endpoints will be analyzed and reported by methods described in Section 6.2.2.

Endpoint: ΔDAS28-3(CRP)

- Analysis time points: All post-baseline timepoints
- Analysis population: FAS

Endpoint: ΔPGA-PsA

[redacted]

[redacted]

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: FAS

Endpoint: %Δ in PASI and PASI clinical sign component scores

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline BSA ≥ 3% and baseline PASI > 0

Endpoint: %ΔBSA

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline BSA > 0%

Endpoint: ΔNAPSI

- Analysis time points: Month 1, Month 3, Month 6
- Analysis population: Subjects in FAS with baseline NAPSI > 0

Endpoint: ΔWPAI-PsA

- Analysis time points: Month 3, Month 6
- Analysis population: FAS

6.4. Subset Analyses

The primary purpose of the subgroup analyses is to check for consistency of results of the primary endpoint of ACR50 response rate at Month 3 across subgroups, i.e., make sure overall results are not being driven by some subset of subjects. There is no intention to have any specific inference within subgroups.

The subgroup comparison will be made on FAS unless stated otherwise. If sample size for a category of a subgroup factor is too small, the analysis for that subgroup may not be performed.

Endpoint: ACR50 response rate at Month 3

- Analysis time points: Month 3
- Method of imputation for missing data: MR=NR (specified in Section 5.3)
- Analysis methodology: Normal approximation (specified in Section 5.2.1.1)

Subgroup Analysis will be performed by:

- Gender (female, male),
- Baseline weight (< 60, ≥ 60 to ≤ 100, > 100 kg),
- Baseline Body Mass Index (BMI) (< 25, ≥ 25 to < 30, ≥ 30 to < 40, and ≥ 40 kg/m^2);

[redacted]

[redacted]

- Age (< 65, ≥ 65 years),
- PsA disease duration (< 2 years, ≥ 2 years),
- Baseline PsA subtype (< 5 joints, ≥ 5 joints) formed based on counting the number of joints at baseline that are either swollen (66) or tender/painful (68), i.e., if a joint is swollen only, tender/painful only, or both swollen and tender/painful, it is only counted as one joint,
- Screening PsA subtype (arthritis mutilans: yes, no),
- Screening PsA subtype (spondylitis: yes, no),
- Baseline CRP (≤ 2.87, > 2.87 mg/L),
- Baseline PASI (≤ 20, > 20) for subjects with baseline PASI > 0 and BSA ≥ 3%,
- Baseline smoking status (never smoked, former smoker, current smoker),
- Prior bDMARD experience (bDMARD-naive vs bDMARD-experienced),
- Number of prior DMARDs-IR (1 csDMARD-IR, 2 csDMARDs-IR, ≥ 3 csDMARDs IR, ≥ 1 bDMARD-experienced). Only prior csDMARDs-IR and bDMARD use will be used to form the categories (see Appendix 3 Section 16 for prior DMARDs-IR definition). Subjects who were treated with any bDMARD or both csDMARDs-IR and bDMARDs will be counted in the "≥ 1 bDMARD-experienced" category,
- Baseline oral corticosteroid use (oral only: yes, no), Day 1 predose,
- Baseline csDMARD use (MTX only, Sulfasalazine only, Other, None), Day 1 predose.

Reporting results:

- Response rate without imputation: N, frequency and percentage of responders, SE of the response rate, and 95% CI at Month 3 will be presented by treatment group for each defined category of each subgroup.
- Response rate with imputation: N, frequency and percentage of responders, SE of the response rate at Month 3 will be presented by treatment group.
- Response rate difference with imputation: Difference between treatment groups, and corresponding SE and 95% CI (no p-value) at Month 3 will be presented for each defined category of each subgroup.

Figures:

- Forest plots of the differences between treatment groups (using imputed data) with corresponding 95% CI will be presented.

[redacted]

[redacted]

6.5. Baseline and Other Summaries and Analyses

6.5.1. Baseline Summaries

Baseline characteristics will include but may not be limited to the demographic and baseline disease characteristics listed in Section 3.4.1 and Section 3.4.2 and will be summarized descriptively. For continuous variables, the summary will include N, mean, s.d. and range; 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 both treatment groups combined.

6.5.2. Prior and Baseline Treatments for Psoriatic Arthritis

Prior and baseline treatments for PsA will be characterized by the variables listed in Section 3.4.3. These variables will be summarized descriptively by frequencies and percentages for each treatment group and for both treatment groups combined.

6.5.3. Concomitant DMARD Treatments

Type of concomitant DMARDs use (methotrexate, sulfasalazine, other) will be summarized by frequencies and percentages. Subjects who were treated with >1 DMARD(s) will be counted in the 'Other' category. Concomitant DMARDs use will also be summarized by counting each csDMARD regardless of other DMARD use. Frequencies and percentages will be provided for each csDMARD.

These summaries will be presented at two time points: one for Day 1 and the other for Day 1 through Month 3.

6.6. Safety Summaries and Analyses

All 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). All the safety data analyses will be performed on the safety analysis set.

6.6.1. Adverse Events

Data related to adverse events will be summarized descriptively through appropriate data tabulations, descriptive statistics, and graphical presentations:

- Incidence and severity of adverse events;
- Incidence of AEs of special interest;

[redacted]

[redacted]

- Serious infections will be summarized separately;
- Any safety events that trigger withdrawal of a subject;
- Special attention will be given to the events of anemia.

Adverse events will be reported as treatment-emergent all causality and treatment-emergent treatment-related. A treatment-emergent adverse event is any event that has: onset after the start of the first dose of study treatment, or onset prior to the first dose of study treatment and worsens in severity after the first dose of study treatment. The relationship of adverse event to study treatment is assigned by the investigator.

Adverse events will be displayed by two analysis periods: Baseline to Month 3 and Baseline to Month 6. Adverse events will be displayed from Baseline to Month 6 by treatment sequence groups and from Baseline to Month 3 by treatment groups.

In addition to the AE analyses described above, the AEs (all causality) are classified in 3 tiers and will be analyzed per the 3-tier approach. Below provides a brief summary.

Analysis of Tier-1 AEs (AEs of special interest) is covered above.

Tier-2 AEs will be analyzed per Section 5.2.3. No multiplicity adjustment will be used. Both tabular form and graphic (e.g., forest plots) display will be provided. They will be sorted by System Organ Class (SOC) and preferred terms (PT) within SOC, alphabetically, as well as by magnitude in risk ratio.

Tier-3 AEs will not be tabled separately but will be included in the standard AE analyses along with all other AEs.

It should be recognized that most studies are not designed to reliably demonstrate a causal relationship between the use of a pharmaceutical product and an adverse event or a group of adverse events. Except for select events in unique situations, studies do not employ formal adjudication procedures for the purpose of event classification. As such, safety analysis is generally considered as an exploratory analysis with the potential to generate hypotheses for further investigation. The 3-tier approach facilitates this exploratory analysis.

6.6.2. Other Safety Summaries

For laboratory tests, 12-lead ECG parameters and vital signs, summaries include categorical tables (e.g., normal, high, low), and descriptive statistics for change from baseline by treatment group and visit. For the fluorescence-activated cell sorting (FACS), the endpoints to be summarized, which are subsets of lymphocytes and are measured at baseline, Month 3, and Month 6/early termination, are as follows: *CD3+ (%, abs), CD3+CD4+ (%, abs), CD3+CD8+ (%, abs), CD19+ (%, abs), CD16+/CD56+ (%, abs)*. Raw values as well as percent change from baseline will be summarized with descriptive statistics by treatment group and visit.

[redacted]

[redacted]

Demographic parameters will be summarized (as number and percent or mean, standard deviation, and range) by treatment group and gender following CaPS.

The number and percent of subjects reporting specific past and present medical histories at screening will be summarized by medical history reporting term (from the Medical Dictionary for Regulatory Activities (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 mediation usage by medication type will be tabulated by treatment group using the WHO-Drug dictionary following CaPS.

6.7. Additional Analyses to Address COVID-19 Pandemic Impacts

6.7.1. Additional Supportive Analysis of the Primary Endpoint

The analysis of ACR50 response rate at Month 3 described in Section 6.1.1.1 will also be conducted by excluding subjects with ACR50 response at Month 3 missing due to COVID-19 pandemic impact or with ACR50 components collected remotely during COVID-19 pandemic. This will serve as an additional supportive analysis of the primary endpoint.

Endpoint: ACR50 response rate at Month 3

- Analysis time points: Month 3
- Analysis population: FAS excluding subjects impacted by COVID-19 pandemic (i.e., excluding subjects with missing or remote Month 3 visit due to COVID-19 pandemic, or any PD due to COVID-19 pandemic that had an impact on the primary endpoint)
- Method of imputation for missing data: MR=NR (specified in Section 5.3)
- Analysis methodology: Normal approximation (specified in Section 5.2.1.1).

Reporting results:

- Response rate with imputation: N, frequency and percentage of responders, SE of the response rate at Month 3 will be presented by treatment group.
- Response rate difference with imputation: Difference between treatment groups, and corresponding SE, 95% CI and p-value at Month 3 will be presented.
- A listing of subjects who were impacted by COVID-19 pandemic and excluded from this additional supportive analysis.

[redacted]

[redacted]

6.7.2. Other Summaries and Listings Depicting COVID-19 Pandemic Impacts

- An anchor date will be used as a start date for COVID-19 pandemic related periods in analyses. For China, the date COVID-19 was identified as the causative agent of outbreak in Wuhan by the China Center for Disease Control and Prevention (January 9, 2020) will be used as the anchor date for COVID-19 pandemic.
- A summary table showing visits attended both before and after the anchor date will be produced. Visit attendance will be based on swollen joint count data.
- A summary table of number of subjects with major impact by COVID-19 pandemic, which includes:
    - Subjects who were excluded from the additional supportive analysis in Section 6.7.1.
    - Subjects who withdraw from the study due to COVID-19 pandemic before Month 3.
- Protocol deviations related to COVID-19 pandemic will be summarized and listed separately. Both important and non-important PDs related to COVID-19 pandemic will be reported.
- A separate summary table solely for subject discontinuations related to COVID-19 pandemic, if any, will be provided.
- COVID-19 related AEs, if any, will be reported separately.
- A summary table showing AEs reported both before and after the anchor date will be produced by treatment group. This summary will consider windows of equal duration before and after the anchor date.
- Data collected remotely during COVID-19 pandemic for PROs and laboratory tests will be marked and footnoted in corresponding listings.

7. INTERIM ANALYSES

There is no interim analysis planned for this study.

7.1. Introduction

Not applicable.

7.2. Interim Analyses and Summaries

Not applicable.

[redacted]

[redacted]

8. REFERENCES

1. Agresti A. (2002). Categorical Data Analysis, 2nd Edition, New York: John Wiley & Sons.

2. Alberti KG, Eckel RH, Grundy SM, et al. 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, 2009, 120: 1640-1645.

3. Bruce B, Fries JF. The health assessment questionnaire (HAQ). Clinical and Experimental Rheumatology; 2005, 23 (Supp. 39), S14-S18.

4. Gottlieb A, Korman NJ, Gordon KB, et al. Guidelines of care for the management of psoriasis and psoriatic arthritis: Section 2. Psoriatic arthritis: overview and guidelines of care for treatment with an emphasis on the biologics. J Am Acad Dermatol 2008, 58: 851-64.

5. Liu, GR and Zhan X Comparisons of methods for analysis of repeated binary responses with missing data. Journal of Biopharmaceutical Statistics, 2011, 21: 371-392.

6. Miettinen O and Nurminen M. Comparative analysis of two rates. Statistics in Medicine, 1985, 4: 213-226.Mumtaz A, Gallagher P, Kirby B, et al. Development of a preliminary composite disease activity index in psoriatic arthritis. Ann Rheum Dis, 2011, 70: 272-277.

7. Rubin DB (1987). Multiple imputation for nonresponse in surveys. New York: Wiley.Schoels M, Aletaha D, Funovits J, Kavanaugh A, Baker D, Smolen JS. Application of the DAREA/DAPSA score for assessment of disease activity in psoriatic arthritis. Ann Rheum Dis, 2010, 69: 1441-1447.

8. Yan X, Lee S, Li N. (2009). Missing data handling methods in medical device clinical trials. Journal of Biopharmaceutical Statistics. 19: 1085-1098.

9. 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.

10. Madsen, O.R. Is DAS28-CRP with three and four variables interchangeable in individual patients selected for biological treatment in daily clinical practice? Clin Rheumatol; 2011, 30: 1577–1582.

[redacted]

[redacted]

9. APPENDICES

Appendix 1. SUMMARY OF EFFICACY ANALYSES

Endpoint	Analysis Set	Analysis Timepoint	Statistical Method	Model	Missing Data Imputation	Interpretation
ACR50 Response Rate	FAS	Month 3	Normal approximation		MR=NR	Primary
ACR50 Response Rate	PP	Month 3	Normal approximation		MR=NR	Supportive Analysis
ACR50 Response Rate	FAS excluding subjects impacted by COVID-19 pandemic	Month 3	Normal approximation		MR=NR	Supportive Analysis
ACR50 Response Rate	FAS	Week 2, Months 1, 2, and 3	GMMRM	Fixed effects: treatment, visit, treatment*visit	None	Supportive Analysis
ACR50 Response Rate	FAS	Week 2, Months 1, 2, and 3	Tipping point analysis	Imputation model: Fixed effects: treatment, visit, treatment*visit	Multiple Imputation	Supportive Analysis
ACR50 Response Rate	Each defined category of each subgroup on FAS (see Section 6.4)	Month 3	Normal approximation		MR=NR	Subset Analysis
ACR50 Response Rate	FAS	Week 2, Months 1, 2, 4 and 6	Normal approximation		MR=NR	Secondary

[redacted]

[redacted]

ACR20 Response Rate	FAS	Week 2, Months 1, 2, 3, 4, and 6	Normal approximation		MR=NR	Secondary
ACR70 Response Rate	FAS	Week 2, Months 1, 2, 3, 4 and 6	Normal approximation		MR=NR	Secondary
ΔHAQ-DI (decrease) ≥ 0.30 response rate	Subjects in FAS with baseline HAQ-DI ≥ 0.30	Week 2, Months 1, 2, 3, 4 and 6	Normal approximation		MR=NR	Secondary
ΔHAQ-DI (decrease) ≥ 0.35 response rate	Subjects in FAS with baseline HAQ-DI ≥ 0.35	Week 2, Months 1, 2, 3, 4 and 6	Normal approximation		MR=NR	Secondary
PsARC response rate	FAS	Week 2, Months 1, 2, 3, 4 and 6	Normal approximation		MR=NR	Secondary
PGA-PsO response rate	Subjects in FAS with baseline PGA-PsO ≥ 2	Months 1, 3, 6	Normal approximation		MR=NR	Secondary
PASI75 response rate	Subjects in FAS with baseline BSA ≥ 3% and baseline PASI > 0	Months 1, 3, 6	Normal approximation		MR=NR	Secondary
Resolution of dactylitis	Subjects in FAS with baseline DSS > 0	Months 1, 3, 6	Normal approximation		MR=NR	Secondary
Resolution of enthesitis	Subjects in FAS with baseline LEI > 0	Months 1, 3, 6	Normal approximation		MR=NR	Secondary

[redacted]

[redacted]

ΔHAQ-DI	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
ΔCRP	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
Δ in Patient's Assessment of Arthritis Pain	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
Δ in Patient's Global Assessment of Arthritis	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
Δ in Physician's Global Assessment of Arthritis	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
Δ in swollen joint count	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
Δ in tender/painful joint count	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary

[redacted]

[redacted]

ΔPGA-PsO	Subjects in FAS with baseline PGA-PsO > 0	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
ΔDSS	Subjects in FAS with baseline DSS > 0	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
ΔLEI	Subjects in FAS with baseline LEI > 0	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
ΔSF-36v2 (10 endpoints)	FAS	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
ΔEQ-5D (5 endpoints) and EQ-VAS	FAS	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Secondary
ΔDAS28-3(CRP)	FAS	Week 2, Months 1, 2, 3, 4 and 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Other
ΔPGA-PsA	FAS	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Other

[redacted]

[redacted]

%Δ in PASI and PASI clinical sign component scores	Subjects in FAS with baseline BSA ≥ 3% and baseline PASI > 0	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Other
%ΔBSA	Subjects in FAS with baseline BSA > 0%	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Other
ΔNAPSI	Subjects in FAS with baseline NAPSI > 0	Months 1, 3, 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Other
ΔWPAI-PsA	FAS	Month 3 & Month 6	MMRM	Fixed effects: treatment, visit, treatment*visit, baseline value	None	Other

[redacted]

[redacted]

Appendix 2. DEFINITION AND USE OF VISIT WINDOWS IN REPORTING

For reporting purposes, the following visit windows will be used for efficacy/PRO variables and for any safety displays that display by week/month. All observations regardless of on- or off-treatment will be used. For an endpoint that consists of multiple components, e.g., ACR50 and DAS28-3(CRP), the components should be collapsed first in the same visit window to derive the endpoint.

If two or more visits/observations fall into the same window, the visit/observation closest to the target day will be used in the analyses. If there is a tie, the later visit/observation will be used. However, for a subject who discontinues from the study early prior to the Month 6 visit and who does not have an observation within the Month 6 visit window, if two or more observations fall in the same visit window (i.e., a visit window prior to Month 6), 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.

Visit Label	**Target Day**	**Definition [Day Window]**
Screening		< Day 1
Baseline	Day 1	Last non-missing assessment on or before Day 1 and prior to first dose of study drug administration
Week 2	Day 15	Day 2 to Day 21
Month 1	Day 29	Day 22 to Day 42
Month 2	Day 57	Day 43 to Day 70
Month 3	Day 85	Day 71 to [Treatment Switch Day or Day 98, whichever is smaller]
Month 4	Day 113	[Treatment Switch Day + 1 or Day 99, whichever is smaller] to Day 140
Month 6	Day 169	≥ Day 141

Target day is equal to the target week × 7 days + 1 day. A month is 4 weeks or 28 days.
Treatment Switch Day = date of the first dose of treatment switch at month 3 (i.e., treatment switch date – date of the first dose + 1). The first dose of the treatment switch at Month 3 will be taken from the first dosing date from the Month 4 oral dosing CRF.
All the data collected on the treatment switch day will be counted in ≤ Month 3 except for the AE data. Any new treatment emergent AE occurred on the day of treatment switch will be counted in ≥ Month 4.

Note that the upper limit of the Month 3 visit window and the lower limit of the Month 4 visit window are defined for each of ALL the subjects in the study since, per the study protocol, subjects randomized to the placebo group will receive tofacitinib 5 mg BID in a blinded fashion for the remainder of the study. Further details for the Month 3/Month 4 visit windows are in order.

1. If Treatment Switch Day falls between Day 71 and Day 98, then the upper limit of the Month 3 visit window is the smaller of Treatment Switch Day or Day 98 and the lower limit of the Month 4 visit window is the smaller of Treatment Switch Day + 1 or Day 99;
2. If Treatment Switch Day is < 71, > 98, or missing, the upper limit of the Month 3 visit window will be Day 98 and lower limit of the Month 4 visit window will be Day 99.

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

	If Treatment Switch Day is							
	Missing		< 71		71-98		> 98	
Visit label	Month 3	Month 4	Month 3	Month 4	Month 3	Month 4	Month 3	Month 4
Day Window	71 to 98	99 to 140	71 to 98	99 to 140	from [71 to 71] up to [71 to 98]	from [72 to 140] up to [99 to 140]	71 to 98	99 to 140

The visit window definitions will be applied to all the endpoints, such as efficacy, PRO, safety etc. Note some of the endpoints may not be measured at every visit. For example, dactylitis will be assessed at baseline, Months 1, 3, and 6, and visit windows at Week 2, Months 2 and 4 will be undefined. In the event that an observation may fall into one of the undefined visit windows, that observation will not be used in analysis.

Appendix 3. ENDPOINT DEFINITION AND DATA DERIVATION DETAILS

1. ACR Assessments

The American College of Rheumatology's definition for calculating improvement in RA (ACR50) is calculated as a ≥ 50% improvement in tender and swollen joint counts and ≥ 50% improvement in 3 of the 5 remaining ACR-core set measures: patient and physician global assessments, pain, disability, and an acute-phase reactant. Similarly, ACR20 and ACR70 are calculated with the respective percent improvement. This efficacy measurement will be made at every study visit. The specific components of the ACR Assessments that will be used in this study are:

1) Tender/Painful Joint count (68);
2) Swollen Joint Count (66);
3) Patient's Assessment of Arthritis Pain (VAS);
4) Patient's Global Assessment of Arthritis (VAS);
5) Physician's Global Assessment of Arthritis (VAS);
6) C-Reactive Protein (CRP);
7) Health Assessment Questionnaire – Disability Index (HAQ-DI).

If the value in any of the components at a timepoint is missing, the component variables that are not missing will be used to determine the response status. As a general principle, if there are sufficient non-missing components to determine whether the ACR endpoint is a response or non-response, then ACR endpoint is not missing, else if the available non-missing components are not sufficient to determine the response status of ACR endpoint then it is considered missing.

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

- If change from baseline is also equal to 0, then percent change from baseline is set to be 0%;
- If change from baseline is > 0, then percent change from baseline is set to be 999999%.

These percentages will be used to derive the ACR endpoints. Change from baseline cannot be < 0 since none of the components should have negative value.

The swollen joint and tender/painful joint counts will be calculated per Section 9 and Section 10 of Appendix 3.

Lower limit of quantification (LLOQ) for CRP is 0.020 mg/dL (or 0.200 mg/L). Any CRP value below the LLOQ will be reported as "< 0.020" in database and will be set to 0.019

[redacted]

[redacted]

mg/dL. The unit for CRP will be mg/dL or mg/L depending on endpoint of interest. The normal range for the CRP is ≤ 2.87 mg/L.

The VAS data will need to be rescaled prior to any calculation (Section 3).

2. Health Assessment Questionnaire – Disability Index (HAQ-DI)

The HAQ-DI assesses the degree of difficulty a subject has experienced during the past week in 8 domains of daily living activities: dressing and grooming, arising, eating, walking, hygiene, reach, grip, and other activities. Each activity category consists of 2-3 items. For each question in the questionnaire, the level of difficulty is scored from 0 to 3 with 0 representing “no difficulty”, 1 as “some difficulty”, 2 as “much difficulty”, and 3 as “unable to do”. The domain score for each domain is the maximum (i.e., worst) of the scores from the items/questions within the domain. If this domain score is ≥ 2, no further adjustment is needed. If it is < 2 (i.e., 0 or 1) but no aids, devices or help from another person is indicated, then also no adjustment is needed. However, if it is < 2 and any aid, device or help from another person is indicated, then it is further adjusted upward to 2 as described below.

Any activity that requires assistance from another individual or requires the use of an assistive device adjusts to a minimum score of 2 to represent a more limited functional status.

Domain	If domain score is either 0 or 1, adjust to 2 when the following is satisfied.
Dressing and grooming	“Devices used for dressing (button hook, zipper pull, long-handled shoe horn, etc.)” or help from others on “Dressing and Grooming” is checked
Arising	“Special or built up chair” or help from others on “Arising” is checked.
Eating	“Built up or special utensils” or help from others on “Eating” is checked.
Walking	“Cane”, “Walker”, “Crutches”, “Wheelchair”, or help form others on “Walking” is checked
Hygiene	“Raised toilet seat”, “Bathtub bar”, “Long-handled appliances in bathroom”, “Bathtub seat” or help from others on “Hygiene” is checked.
Reach	“Long-handled appliances for reach” or help from others on “Reach” is checked
Grip	“Jar opener (for jars previously opened)” or help from others on “Gripping or opening things” is checked.
Activities	Help from others on “Errands and chores” is checked.

Note: For "Other, (specify)", whether checked or unchecked or specifying the other "aids or devices" in this category, is not to be used in the adjustment of the domain score.

For each domain, the domain score will be determined by non-missing scores from the questions. A domain score will only be missing if all scores within the domain are missing. The HAQ-DI score is the average of all the 8 domain scores. If > 2 domain scores are missing, HAQ-DI score is considered missing, else the HAQ-DI score is computed as the average of the non-missing domain scores (Bruce & Fries, 2005). A higher score represents a more limited physical functional status/ability.

3. PsA Response Criteria (PsARC)

The PsARC will be calculated at all study visits in addition to the ACR response criteria. The PsARC consists of 4 measurements:

1) Tender joint count (68);
2) Swollen joint count (66);
3) Physician's Global Assessment of Arthritis (VAS);
4) Patient's Global Assessment of Arthritis (VAS).

The same tender/painful joint count and swollen joint count used for ACR response criteria will be applied to the PsARC. In order to be a 'PsARC responder', subjects must achieve improvement in 2 of 4 measures, one of which must be joint pain or swelling, without worsening in any measure.

Specifically, the PsARC response is defined as improvement in two of the following 4 criteria, one of which must be joint pain or swelling, without worsening in any measure: (1) ≥ 20% improvement in Physician's Global Assessment of Arthritis (VAS); (2) ≥ 20% improvement in Patient's Global Assessment of Arthritis (VAS); (3) ≥ 30% improvement in tender joint count (68); and (4) ≥ 30% improvement in swollen joint count (66) (Gottlieb et al, 2008).

Methods for calculating percent change from baseline when baseline value of any component is equal to 0 can be found in Appendix 3 Section 1.

If values in any of the components at a timepoint were missing, the component variables that were not missing were used to determine the response status. It should be noted that since PsARC response requires no worsening in any measure, a subject with any missing components cannot be a responder at that timepoint. If one could not determine the response status in the presence of missing components at the timepoint, then the PsARC response will be considered as missing. Handling of missing values for a subject at timepoint in analyses is found in Section 5.3.

[redacted]

[redacted]

The swollen joint and tender/painful joint counts will be calculated per Section 9 and Section 10 in Appendix 3.

The VAS data will need to be rescaled prior to any calculation (Section 3).

4. Physician Global Assessment (PGA) of Psoriasis (PGA-PsO)

The Physician Global Assessment of Psoriasis is scored on a 5-point scale, reflecting a global consideration of the erythema, induration and scaling across all psoriatic lesions. Average erythema, induration and scaling are rated separately over the whole body according to a 5-point severity scale, scored as 0 (none), 1, 2, 3, or 4 (most severe), The severity rating scores are summed and the average taken – the total average is rounded to the nearest whole number score to determine the PGA. See study protocol for additional details.

PGA-PsO response is defined for a subject who achieves a PGA-PsO score of clear (0) or almost clear (1) and a ≥ 2-step improvement from baseline.

5. Psoriasis Area and Severity Index (PASI)

The Psoriasis Area and Severity Index (PASI) quantifies the severity of a subject's psoriasis based on both lesion severity and the percentage of body surface area (BSA) affected. PASI is a composite scoring by the investigator of degree of erythema, induration, and scaling (each scored separately) for each of four body regions, with adjustment for the percent of BSA involved for each body region and for the proportion of the body region to the whole body. Note: PASI should only be performed if ≥ 3% of subject's BSA is affected at baseline.

- Lesion severity: The basic characteristics of psoriatic lesions – erythema, induration and scaling – provide a means for assessing the severity of lesions. Assessment of these three main signs is performed separately for four areas of the body: head and neck, upper limbs, trunk (including axillae and groin), and lower limbs (including buttocks). Average erythema, induration and scaling are rated for each body area according to a 5-point scale: 0, no involvement; 1, slight; 2, moderate; 3, marked; 4, very marked.

- Body surface area involvement (%BSA): The extent (%) to which each of the four areas of the body is affected by psoriasis is assigned a numerical score according to the following area scoring criteria: 0, no involvement; 1, > 0 to 9%; 2, 10 to 29%; 3, 30 to 49%; 4, 50 to 69%; 5, 70 to 89%; 6, 90 to 100%.

In each area, the sum of the severity rating scores for erythema, induration and scaling is multiplied by the score representing the percentage of this area involved by psoriasis, multiplied by a weighting factor (head 0.1; upper limbs 0.2; trunk 0.3; lower limbs 0.4). The sum of the numbers obtained for each of the four body areas is the PASI.

PASI = 0.1Ah (Eh + Ih + Sh) + 0.2Au (Eu + Iu + Su) + 0.3At (Et + It + St) + 0.4Al (El + Il + Sl)

[redacted]

[redacted]

Where A = area of involvement score; E = erythema; I =induration; S = scaling; h = head; u = upper limbs; t = trunk; l = lower limbs.

The PASI score can vary in increments of 0.1 units from 0.0 to 72.0, with higher scores representing increasing severity of psoriasis.

Each PASI clinical signs component score can be computed as:

- PASI_E = 0.1Ah (Eh) + 0.2Au (Eu) + 0.3At (Et) + 0.4Al (El)
- PASI_I = 0.1Ah (Ih) + 0.2Au (Iu) + 0.3At (It) + 0.4Al (Il)
- PASI_S = 0.1Ah (Sh) + 0.2Au (Su) + 0.3At (St) + 0.4Sl (Sl)

Any missing component will result in PASI as missing.

6. Short Form 36 (SF-36, version 2, acute)

The SF-36v2 (Acute) is a 36-item generic health status measure. It measures 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.

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.

[redacted]

[redacted]

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 scores, PCS and MCS, are based on a normalized sum of the 8 scale 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.

Data Derivation Details to Obtain Scale Scores for SF-36

VARIABLE	DERIVATION
SF-36 PF scale score	raw score = sum (items 3A, 3B, 3C, 3D, 3E, 3F, 3G, 3H, 3I, 3J) PF = (raw score -10) * 5 PF_Z = (PF – 82.62455) / 24.43176 **PF scale score = (PF_Z*10) + 50** When calculating the raw score, if 5 or more of the items are non-missing then replace any missing values as follows: Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores. Otherwise, if less than 5 of the items are non-missing then PF scale score is missing. 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.

SF-36 RP scale score	raw score = sum (items 4A, 4B, 4C, and 4D) RP = [(raw score -4)/16] * 100 RP_Z = (RP – 82.65109) / 26.19282 **RP scale score = (RP_Z * 10) + 50** When calculating the raw score, if 2 or more of the items are non-missing then replace any missing values as follows: Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores. Otherwise, if less than 2 of the items are non-missing then RP scale score is missing. 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.

[redacted]

[redacted]

SF-36 BP scale score	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
	
	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
	
	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
	
	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
	
	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.
	
	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.
	
	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.

SF-36 GH scale score	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** 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 Reverse direction of item 11B and 11D by subtracting score from 6. When calculating the raw score, if 3 or more of the items are non-missing then replace any missing values as follows: Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores. Otherwise, if less than 3 of the items are non-missing then GH scale score is missing. 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. 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. A higher GH scale score indicates better general health perceptions.

SF-36 VT scale score	raw score = sum (reversed item 9a, reversed 9e, 9g and 9i) VT = [(raw score -4)/16] * 100 VT_Z = (VT – 58.41968) / 20.87823 **VT scale score = (VT_Z * 10) + 50** Reverse direction of Items 9a and 9e by subtracting score from 6. When calculating the raw score, if 2 or more of the items are non-missing then replace any missing values as follows: Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores. Otherwise, if less than 2 of the items are non-missing then VT scale score is missing. 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.
SF-36 SF scale score	raw score = sum (reversed 6 and 10) SF = [(raw score -2) / 8] * 100 SF_Z = (SF – 85.11568) / 23.24464 **SF scale score = (SF_Z * 10) + 50** Reverse direction of score for item 6 by subtracting score from 6. 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. 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.

SF-36 RE scale score	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** When calculating the raw score, if 2 or more of the items are non-missing then replace any missing values as follows: Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores. Otherwise, if less than 2 of the items are non-missing then RE scale score is missing. 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.
SF-36 MH scale score	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** Reverse direction of scores for 9D and 9H, by subtracting score from 6. If 3 or more of the items are non-missing then replace any missing values as follows: Calculate the mean of the answered questions, using the number of answered questions as the denominator. Substitute the mean for any missing scores. Otherwise, if less than 3 of the items are non-missing then MH scale score is missing. 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.
SF-36 TR scale score	raw score = item 2 **TR scale score = raw score** The scale for this item ranges from 1 to 5 where 1=Much better now than one week ago, 2=Somewhat better now than one week ago, 3=About the same as one week ago, 4=Somewhat worse now than one week ago, and 5=Much worse now than one week ago. A higher TR scale score indicates worse general health currently relative to one week previous.

[redacted]

[redacted]

SF-36 PCS score	PCS score includes the 8 scales for GH, PF, RP, RE, SF, MH, BP, and VT.
	
	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;
	
	Raw Score = ((GH1*.24954)+(PF1* .42402)+(RP1*.35119)
	+ (RE1*-.19206)+(SF1*-.00753)+(MH1*-.22069)+(BP1*.31754)
	+ (VT1*.02877))
	PCS Summary Scale Score = (raw score *10) + 50
	
	Raw Score is missing if one of the component scale scores is missing.

SF-36 MCS score	MCS score includes the 8 scales for GH, PF, RP, RE, SF, MH, BP, and VT.
	
	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;
	
	Raw Score =((GH1*-.01571)+(PF1*-.22999)+(RP1*-.12329)
	+ (RE1*.43407)+(SF1*.26876)+(MH1*.48581)+(BP1*- 0.09731)
	+ (VT1*.23534))
	MCS Summary Scale Score = (raw score*10)+50
	
	Raw score is missing if one of the component scale scores is missing.

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

On the EQ-5D (subject version, 3 categories of response per question), 5 dimensions of health (mobility, self-care, usual activities, pain/discomfort and anxiety/depression) are assessed. The status of each dimension has three possible responses (no problem in the relevant health dimension to severe problems). The EQ-VAS records the patient's self-rated health on a vertical visual analogue scale where the endpoint is labelled 'Best imaginable health state' and 'Worst imaginable health state'. Based on the patient's mark on the VAS form a score ranging from 0 to 100 mm is recorded.

8. DAS28-3(CRP)

The Disease Activity Score (DAS) is a derived measurement with differential weighting given to each component. DAS28-3(CRP) will be calculated from measurements made at all study visits (Madsen et al., 2011).

The components of the DAS28-3 arthritis assessment are:

1) Tender/Painful Joint Count (28);
2) Swollen Joint Count (28);
3) C-Reactive Protein (CRP in mg/L).

DAS28-3(CRP) = [0.56*sqrt(TJC28) + 0.28*sqrt(SJC28) + 0.36*ln(CRP+1)] *1.10 + 1.15

where sqrt() refers to the square root, and ln() refers to the natural logarithm.

The swollen joint count (28) and tender/painful joint count (28) will be calculated similarly as for the swollen joint count (66) and tender/painful joint count (68) per Section 9 and Section 10 in Appendix 3.

Any missing component will result in DAS28-3(CRP) as missing. A higher score represents a more severe disease activity.

9. Swollen Joint Count (66)

Swollen joints are recorded in the CRF page of joint swelling. For each of the 66 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 per Pfizer’s standard:

- 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 (i.e., 33, 50% of 66). In that case, the swollen joint count is set to missing.

[redacted]

[redacted]

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 $N_1 + N_2 + N_3$ (i.e., 66). The calculated swollen joint count is therefore $N_1 + N_3 \times (N_1 / (N_1 + N_2))$ if $N_1 + N_2 > 33$; otherwise the swollen joint count is set to missing. The calculated swollen joint count will be rounded to a whole number. For example, a value of 24.5 is rounded to 25, a value of 24.49 is rounded to 24. Note that if there is no missing joint assessment, the swollen joint count will be the number of joints recorded as “Present” (i.e., N_1).

10. Tender/Painful Joint Count (68)

Tender/painful joints are recorded in the CRF page of joint tender/painful. For each of the 68 joints, tender/painful can be recorded as “Present”, “Absent”, “Not Done”, “Not Applicable” or this joint assessment is simply missing. The tender/painful joint count will be calculated per Pfizer’s standard similarly as for the swollen joint count (see Appendix 3 Section 9).

11. Leeds Enthesitis Index (LEI) and Resolution of Enthesitis

Six sites, including (right and left): lateral_epicondyle humerus, medial femoral condyle and Achilles tendon insertion, are assessed for enthesitis. Leeds Enthesitis Index (LEI) is the number of sites with presence of enthesitis.

Each of the 6 sites is assessed as “Present”, “Absent”, “Not Done”, “Not Applicable” or simply missing. Before calculating the LEI, the pre-processing steps for each site will be performed per Pfizer’s standard as follows.

- A site recorded as “Not Done” is set to missing.
- A site recorded as “Not Applicable” means that measurement cannot be made and this site is set 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 LEI (i.e., 3, 50% of 6). In that case, the LEI is set to missing.

Explicitly after pre-processing, let

- N_1=number of sites recorded as “Present”;

[redacted]

[redacted]

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

The total number of sites assessed is $N=N_1+ N_2+ N_3$ (i.e., 6). The calculated LEI is therefore $N_1+ N_3\times(N_1/(N_1+ N_2))$ if $N_1+ N_2 > 3$; otherwise the LEI is set to missing. The calculated LEI will be rounded to a whole number. For example, a value of 4.5 is rounded to 5, a value of 4.49 is rounded to 4. Note that if there is no missing site assessment, the LEI will be the number of sites recorded as “Present” (i.e., N_1).

Resolution of enthesitis is defined as a subject with LEI=0.

12. Dactylitis Severity Score (DSS) and Resolution of Dactylitis

Each of the 20 digits in hands and feet with dactylitis will be evaluated. Dactylitis severity will be scored based upon digit tenderness using a scale of 0-3, (0-NONE, 1-MILD, 2-MODERATE, 3-SEVERE, 8-NOT APPLICABLE, 9-NOT DONE), in each digit of the hands and feet. Dactylitis severity score (DSS) is the sum of the severity scores of the 20 digits (range: 0-60, with a higher score indicating more severe dactylitis).

Before calculating the DSS, the pre-processing steps for each digit will be performed per Pfizer’s standard as follows.

- A digit recorded as “Not Done” is set to missing.
- A digit recorded as “Not Applicable” means that measurement cannot be made and this digit is set to missing.

Let N_1 be the number of digits with missing values, and N_2 and S be the number of the digits and the sum of the digit scores assessed as 0-NONE, 1-MILD, 2-MODERATE, OR 3-SEVERE, respectively, the calculated DSS is: $S + N_1 \times \frac{S}{N_2}$, if $N_1<N_2$; otherwise the DSS is set as missing.

The calculated DSS is rounded to the whole number. For example, a value of 24.5 is rounded to 25, a value of 24.49 is rounded to 24.

Resolution of dactylitis is defined as a subject with DSS=0.

13. AEs of Special Interest

[redacted]

[redacted]

• Serious infections (SIs)
• Herpes zoster (HZ)
• Adjudicated opportunistic infections (OIs) excluding tuberculosis (TB)[a]
• Hematologic events (anaemia, neutropenia, lymphopenia, and thrombocytopenia)
• Adjudicated malignancies[a]
• Adjudicated major adverse cardiovascular events (MACE)[a,b]
• Adjudicated deep vein thrombosis (DVT)[a]
• Adjudicated pulmonary embolism (PE)[a]
• Adjudicated venous thromboembolism (VTE)[a] (i.e., adjudicated DVT[a] or adjudicated PE[a])
• Arterial thromboembolism (ATE)[c]
• Thromboembolism (TE) (i.e., adjudicated DVT[a], adjudicated PE[a], or ATE[c])
• Adjudicated drug-induced liver injury (DILI)[a] and other hepatic events[d]
• Renal events[e]
• Adjudicated gastrointestinal (GI) perforations[a]
• Adjudicated interstitial lung disease (ILD)[f]
Abbreviations: CMQ = Customized MedDRA Queries; DILI = drug-induced liver injury; GI = gastrointestinal; HZ = herpes zoster; ILD = interstitial lung disease; MACE = major adverse cardiovascular events; MedDRA = Medical Dictionary for Regulatory Activities; OI = opportunistic infection; SI = serious infection; SMQ = Standardised MedDRA Queries; TB = tuberculosis. a. Adjudicated by Adjudication Committees independent and external to Pfizer. b. Components: cardiovascular death, non-fatal myocardial infarction, and non-fatal stroke. c. CMQs to be included will be specified in the programming plan. d. Including SMQs: Hepatic failure, fibrosis, and cirrhosis, and other liver damage-related conditions; and Hepatitis, non-infectious. e. Including acute renal failure SMQ. f. Adjudicated by an internal review committee.

14. Definition of Diabetes Mellitus at Baseline

A subject is considered having diabetes mellitus at baseline, if any of the conditions below is met:

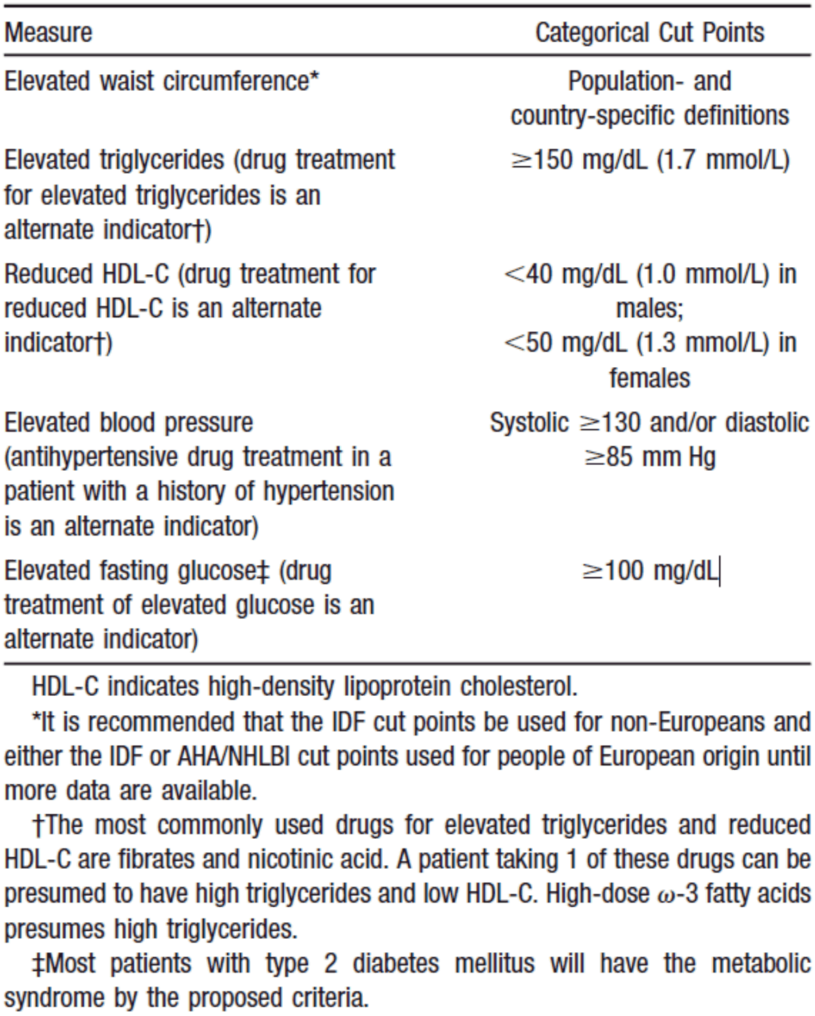
- subject had a diagnosis of diabetes mellitus recorded on the "SIGNIFICANT MEDICAL HISTORY" CRF at screening,
- subject received any drug treatment taken on Day 1 recorded on the "PRIOR AND CONCOMITANT MEDICATIONS: OTHER" CRF in the "ANTI-DIABETIC AGENTS" category,
- subject's HbA1c ≥ 6.5% at baseline or, if HbA1c is not available, baseline fasting plasma glucose ≥ 126 mg/dL.

[redacted]

[redacted]

15. Criteria for Clinical Diagnosis of the 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). 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.

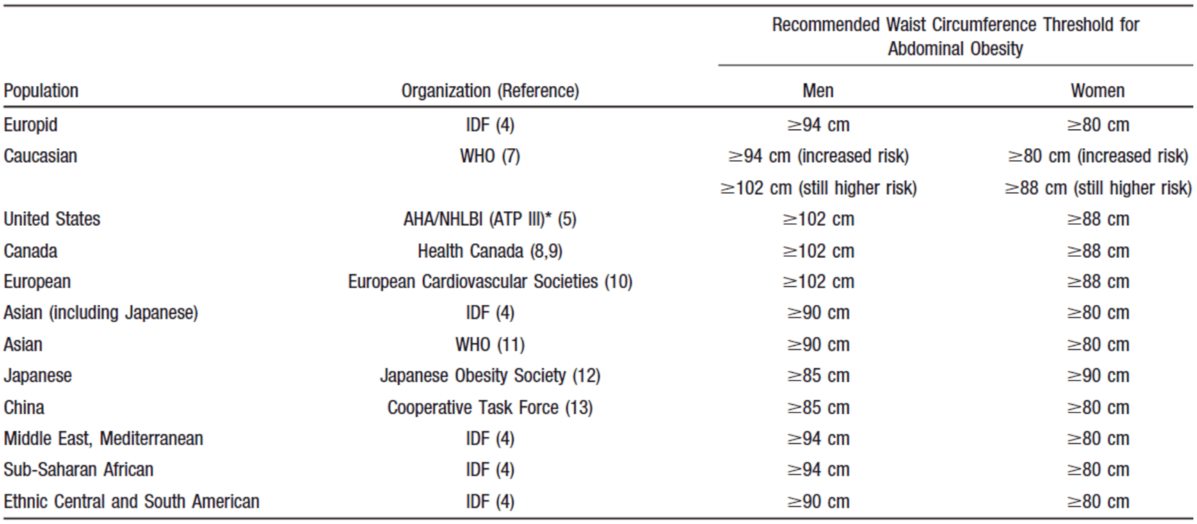
‡Most patients with type 2 diabetes mellitus will have the metabolic syndrome by the proposed criteria.

Any 3 abnormal findings out of the 5 measures will qualify a subject for the metabolic syndrome.
The baseline (i.e., Day 1) data will be used to derive the metabolic syndrome.

- Any drug treatment taken on Day 1 recorded on the "PRIOR AND CONCOMITANT MEDICATIONS: OTHER" CRF in the "ANTI-DIABETIC AGENTS" category 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: OTHER" CRF in the "ANTI-HYPERTENSION AGENTS"

category 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: OTHER" CRF in the "LIPID LOWERING AGENTS" category will be considered having the drug treatment of elevated triglycerides criterion satisfied.
- For the HDL-C criterion, only the lab categorical cut-points will be considered and the 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 China specific threshold will be used for this China-alone study.



		Recommended Waist Circumference Threshold for Abdominal Obesity	
Population	Organization (Reference)	Men	Women
Europid	IDF (4)	≥94 cm	≥80 cm
Caucasian	WHO (7)	≥94 cm (increased risk) ≥102 cm (still higher risk)	≥80 cm (increased 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

16. Prior DMARDs-IR

Prior DMARDs-IR count only those DMARDs that resulted in an inadequate response (IR), defined as those DMARDS that were discontinued due to AE, LOE or both AE/LOE on or prior to Day 1, or a DMARD that was started prior to Day 1, but was not discontinued and continued beyond Day 1 regardless of DMARD dose change. Any DMARD that was discontinued only due to "Other Reasons" on or before Day 1 will not be considered a DMARD IR. If the same one agent (e.g., methotrexate) had multiple dose entries, and if any of the doses had AE, LOE or AE/LOE, or the agent continued beyond Day 1 without being discontinued, it counts as one DMARD IR.

The prior and the concomitant DMARDs are recorded on the "PRIOR AND CONCOMITANT MEDICATIONS – DMARD" CRF.

17. List of DMARDS

[redacted]

WHO Drug Preferred Term	bDMARD		csDMARD
	TNFi	Non-TNFi	
ABATACEPT		x	
ADALIMUMAB	x		
ANAKINRA		x	
APREMILAST			x
AZATHIOPRINE			x
CERTOLIZUMAB	x		
CERTOLIZUMAB PEGOL	x		
CHLOROQUINE			x
CICLOSPORIN			x
ETANERCEPT	x		
GOLIMUMAB	x		
GUSELKUMAB		x	
HYDROXYCHLOROQUINE			x
HYDROXYCHLOROQUINE PHOSPHATE			x
INFLIXIMAB	x		
IXEKIZUMAB		x	
LEFLUNOMIDE			x
METHOTREXATE			x
METHOTREXATE SODIUM			x
RITUXIMAB		x	
SECUKINUMAB		x	
SULFASALAZINE			x
TOCILIZUMAB		x	
USTEKINUMAB		x	

TNFi= TNF inhibitor

bDMARD= biologic DMARD

csDMARD= conventional synthetic (i.e., traditional non-biologic) DMARD

A footnote “For analysis purposes, Apremilast (a targeted synthetic DMARD) was grouped with the csDMARDs.” will be added to any display where applicable.

[redacted]

[redacted]

Appendix 4. STATISTICAL METHODOLOGY DETAILS

Tipping Point Analysis for Longitudinal Binary Data

A method to analyze the longitudinal data of a binary endpoint measured during the placebo-controlled period at Week 2, Months 1, 2 and 3 under the MNAR assumption is called the tipping point analysis. This analysis is used as a supportive analysis for the primary endpoint of ACR50 at Month 3. 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).

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 is used for the analysis, 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.

<u>Step 1: Specification of the imputation model.</u>

The generalized linear mixed effect model includes the fixed effects of treatment, visit, treatment-by-visit interaction, and a latent subject-level random effect. The logit link is used to model the ACR50 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 data 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 four post-baseline visits is parameterized as

$$logit(\pi_i) = logit\big(P(Y_i = 1)\big) = \beta_0 + \beta_{dose} dose + \beta_{time1} time1 + \beta_{time2} time2 + \beta_{time3} time3 +$$

$$\beta_{int1} dose * time1 + \beta_{int2} dose * time2 + \beta_{int3} dose * time3 + \phi_i \cdot$$

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, $time1$, $time2$ and $time3$ are dummy variables. When all of them are 0, it represents Week 2; $time1 = 1$ and all others being 0 represent Month 1; $time2 = 1$ and all others being 0 represent Month 2; $time3 = 1$ and all others being 0 represent Month 3. Interaction terms between the treatment and time point are also included in the model, leading to a saturated model.

<u>Step 2: Estimating the imputation model.</u>

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

[redacted]

[redacted]

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 Month 3.

A single imputation of missing ACR50 response at Month 3 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 ACR50 at earlier time points will not be performed. The following model parameters pertinent to the assessment of treatment difference at Month 3: β_0^*, β_{dose}^*, β_{time3}^*, β_{int3}^*, 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 missing ACR50 response at Month 3, 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 $\pi_i^* = \frac{\exp(\beta_0^*+\beta_{dose}^*+\beta_{time3}^*+\beta_{int3}^*+\phi_i^*)}{1+\exp(\beta_0^*+\beta_{dose}^*+\beta_{time3}^*+\beta_{int3}^*+\phi_i^*)}$. In the situation when $\pi_i^p < 0$, π_i^p will be set to 0; similarly when $\pi_i^p > 1$, π_i^p will be set to 1. For a given δ, the missing ACR50 response at Month 3 for this subject in the tofacitinib 5mg BID group 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 with missing ACR50 response. A series of fixed MNAR quantity δ including 0 (i.e., 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 (i.e., penalizing the subject's probability of response) and negative values represent favorable scenarios. Once all subjects with missing ACR50 responses at Month 3 are imputed, this step results in a single complete imputed data set for Month 3.

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. 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

[redacted]

[redacted]

using the Rubin’s rules (Rubin, 1987). 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 ACR50 responses at Month 3 between tofacitinib 5 mg BID and placebo. The specific implementation (i.e., 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.

[redacted]

[redacted]