STATISTICAL ANALYSIS PLAN

Study: PA0011

Product: Bimekizumab

A MULTICENTER, PHASE 3, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED STUDY EVALUATING THE EFFICACY AND SAFETY OF BIMEKIZUMAB IN THE TREATMENT OF SUBJECTS WITH ACTIVE PSORIATIC ARTHRITIS

SAP/Amendment Number	Date
Final SAP	24 May 2019
Amendment 1	02 Jul 2021
Amendment 2	09 Nov 2021

Confidentiality Statement

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACR	American College of Rheumatology
ACR20,50,70	American College of Rheumatology 20, 50, 70% response criteria
ADAb	anti-drug antibody
ADR	adverse drug reaction
AE	adverse event
ANCOVA	analysis of covariance
ATC	anatomical therapeutic chemical classification
BASDAI	Bath Ankylosing Spondylitis Disease Activity Index
BKZ	bimekizumab
BLQ	below the limit of quantification
BMI	body mass index
BP	blood pressure
BSA	body surface area
CASPAR	Classification Criteria for Psoriatic Arthritis
CCP	cyclic citrullinated peptide
cDMARD	conventional disease modifying antirheumatic drug
CI	confidence Interval
COVID-19	Coronavirus Disease 2019
CRF	case report form
CRP	C-reactive protein
CV	coefficient of variation
DAPSA	disease activity index for psoriatic arthritis
DAS28[CRP]	disease activity score-28 based on C-reactive protein

DILI	drug-induced liver injury
DMC	data monitoring committee
DMARDs	disease modifying antirheumatic drugs
ECG	electrocardiogram
eC-SSRS	electronic Columbia-Suicide Severity Rating Scale
EQ-5D-3L	Euro-Quality of Life 5-Dimensions 3 Level version
ET	early termination
EAER	exposure adjusted event rate
EAIR	exposure adjusted incidence rate
FACIT	Functional Assessment of Chronic Illness Therapy
FAS	full analysis Set
FDA	Food and Drug Administration
HAQ-DI	Health Assessment Questionnaire - Disability Index
HLA-B27	human leukocyte antigen B27
HLT	high level term
hs-CRP	high sensitivity C-reactive protein
ICH	International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use
IE	intercurrent event
IGA	Investigator's Global Assessment
IMP	investigational medicinal product
IPD	important protocol deviation
IWRS	interactive web response system
KM	Kaplan-Meier

LCL	lower confidence limit
LDI	Leeds Dactylitis Index
LEI	Leeds Enthesitis Index
LLOQ	lower limit of quantification
LN	natural logarithm
LSM	least square means
MACE	major adverse cardiac events
MAR	missing at random
MCMC	Markov-Chain Monte Carlo
MCS	Mental Component Summary
MDA	minimal disease activity
MedDRA	medical dictionary for regulatory activities
MI	multiple imputation
mNAPSI	modified Nail Psoriasis Severity Index
MTX	methotrexate
NAb	neutralizing anti-bimekizumab antibodies
NRI	nonresponder imputation
NRS	numeric rating scale
NSAID	nonsteroidal anti-inflammatory drug
OC	observed cases
OLE	open label extension
OR	odds ratio
PASDAS	psoriatic arthritis disease activity score
PASI	psoriasis area and severity index

PASI75, PASI90, PASI100	psoriasis area and severity index 75%, 90%, 100%
PCS	Physical Component Summary
PGA-Arthritis	Patient's Global Assessment of arthritis
PGA-PsA	Patient's Global Assessment of Psoriatic Arthritis
PhGA-Arthritis	Physician's Global Assessment of arthritis
PhGA-PsA	Physician's Global Assessment of Psoriatic Arthritis
PHQ-9	Patient Health Questionnaire-9
PK-PPS	pharmacokinetics per-protocol set
PPS	per-protocol set
PsA	psoriatic arthritis
PsAID-12	Psoriatic Arthritis Impact of Disease-12
PsARC	psoriatic arthritis response criteria
PsAQoL	Psoriatic Arthritis Quality of Life
PSO	psoriasis
PT	preferred term
PtAAP	Patient's Assessment of Arthritis Pain
Q4W	every 4 weeks
RS	randomized set
SAE	serious adverse event
SAP	statistical analysis plan
SC	subcutaneously
SD	standard deviation
SF-36	Short-Form 36-item Health Survey
SFU	safety follow-up

SOC	system organ class
SPARCC	Spondyloarthritis Research Consortium of Canada
SJC	swollen joint count
SMQ	standardized MedDRA query
SS	safety set
TB	tuberculosis
TEAE	treatment emergent adverse event
TESAE	treatment emergent serious adverse event
TFLs	tables, figures, and listings
TJC	tender joint count
TNF	tumor necrosis factor
TNFα	tumor necrosis factor alpha
TNFα-IR	TNFα -inadequate responders
UCL	upper confidence limit
ULN	upper limit of normal
VAS	visual analog scale
WHO-DD	World Health Organization Drug Dictionary
VLDA	very low disease activity
WPAI-SHP	Work Productivity and Activity Impairment questionnaire - Specific Health Problem

1 INTRODUCTION

This statistical analysis plan (SAP) provides the necessary information to perform the final statistical analysis as well as the interim analysis for study PA0011. It also defines the summary tables, figures, and listings (TFLs) to be generated for the final clinical study report.

The SAP is based on Protocol Amendment 2 (1 APR 2021). All references to study protocol hereafter refer to this version of the protocol, and, unless otherwise specified, the study will be analyzed as described in this version of the protocol. If a future protocol amendment necessitates a substantial change to the statistical analysis of the study data, or if analysis definitions must be modified or updated, this SAP will be amended accordingly.

The content of this SAP is compatible with the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH)/Food and Drug Administration (FDA) E9 Guidance documents (ICH E9, 1998).

2 PROTOCOL SUMMARY

2.1 Study objectives

2.1.1 Primary objective

The primary objective is to demonstrate the clinical efficacy of bimekizumab (BKZ) administered subcutaneously (sc) every 4 weeks (Q4W) for 16 weeks compared with placebo in the treatment of prior tumor necrosis factor alpha-inadequate responders (TNFα-IR) subjects with active psoriatic arthritis (PsA), as assessed by the American College of Rheumatology 50% (ACR50) response.

2.1.2 Secondary objectives

The secondary objectives are as follows:

- To assess the efficacy of bimekizumab compared with placebo
- To assess the safety and tolerability of bimekizumab
- To assess the impact of bimekizumab on patient-reported quality of life
- To assess the impact of bimekizumab on skin psoriasis (PSO) in the subgroup of affected subjects at Baseline.
- To assess the impact of bimekizumab on functional improvement

2.1.3 Other objectives

Other objectives are as follows:

- To assess the immunogenicity of bimekizumab
- To assess the impact of bimekizumab on extra-articular disease manifestations (dactylitis, enthesitis)
- To assess nail PSO in the subgroup of affected subjects at Baseline
- To explore the exposure-response relationship of bimekizumab
- To assess the impact of bimekizumab treatment on axial disease

- To assess the effect of bimekizumab on gene and protein expression, and explore the relationship between genomic, genetic, and proteomic biomarkers and disease biology, drug treatment and inflammatory and immune responses (from consenting subjects who agree to participate in the biomarkers sub-study)
- To assess the impact of bimekizumab on social life and work productivity

2.2 Study variables

2.2.1 Efficacy variables

2.2.1.1 Primary efficacy variable

The primary efficacy variable is the ACR50 response at Week 16.

2.2.2 Secondary variables

2.2.2.1 Secondary efficacy variables

The secondary efficacy variables are as follows:

- Change from Baseline in Health Assessment Questionnaire - Disability Index (HAQ-DI) at Week 16
- Reduction of 90% from Baseline in Psoriasis Area and Severity Index (PASI90 response) at Week 4 and Week 16 in the subgroup of subjects with PSO involving at least 3% body surface area (BSA) at Baseline
- Change from Baseline in the Short-Form 36-item Health Survey (SF-36) Physical Component Summary (PCS) score at Week 16
- Minimal Disease Activity (MDA) response at Week 16
- ACR20 response at Week 16
- ACR70 response at Week 16
- Proportion of subjects with an Investigator Global Assessment (IGA) score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline at Week 4 and Week 16 in the subset of subjects with psoriatic skin lesions at Baseline
- Change from Baseline in the Patient's Assessment of Arthritis Pain (PtAAP) at Week 16
- Change from Baseline in Psoriatic Arthritis Impact of Disease-12 (PsAID-12) total score at Week 16

2.2.2.2 Secondary safety variables

Secondary safety variables to be assessed are as follows:

- Incidence of treatment-emergent adverse events (TEAEs)
- Incidence of treatment-emergent serious adverse events (TESAEs)
- TEAEs leading to withdrawal from Investigational Medicinal Product (IMP)

2.2.3 Other variables

2.2.3.1 Other efficacy variables

The following other efficacy variables will be assessed (all time points not specified in Section 2.2.1.1 or Section 2.2.2.1 are exploratory):

- Time to ACR20, ACR50, and ACR70 response from Baseline (Day 1)
- ACR20, ACR50 and ACR70 response
- PASI75, PASI90 and PASI100 response in the subgroup of subjects with PSO involving at least 3% BSA at Baseline
- Composite endpoint composed of ACR50 and PASI90 response in subjects with PSO involving at least 3% BSA at Baseline
- Composite endpoint composed of ACR50 and PASI100 response in subjects with PSO involving at least 3% BSA at Baseline
- Proportion of Psoriatic Arthritis Response Criteria (PsARC) responders
- Psoriatic Arthritis Disease Activity Score (PASDAS) categories
- Change from Baseline in the PASDAS
- MDA response
- Very Low Disease Activity (VLDA) response
- Proportion of subjects with an IGA score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline in the subset of subjects with psoriatic skin lesions at Baseline
- Disease Activity Index for Psoriatic Arthritis (DAPSA) score categories,
- Change from Baseline in DAPSA score
- Change from Baseline in the Disease Activity Score-28 based on C-reactive protein (DAS28 [CRP])
- Change from Baseline in all individual ACR core components:
  - Swollen Joint Count (SJC)
  - Tender Joint Count (TJC)
  - HAQ-DI
  - PtAAP
  - Physician's Global Assessment for Psoriatric Arthritis (PhGA-PsA)
  - Patient's Global Assessment for Psoriatric Arthritis (PGA-PsA)
  - High sensitivity C-reactive protein (hs-CRP)
- Enthesitis-free state based on the Leeds Enthesitis Index (LEI) in the subgroup of subjects with enthesitis at Baseline (ie, LEI>0 at Baseline)

- Change from Baseline in the LEI in the subgroup of subjects with enthesitis at Baseline (ie, LEI>0 at Baseline)
- Enthesitis-free state based on the Spondyloarthritis Research Consortium of Canada (SPARCC) index in the subgroup of subjects with enthesitis at Baseline (ie, SPARCC>0 at Baseline)
- Change from Baseline in SPARCC index in the subgroup of subjects with enthesitis at Baseline (ie, SPARCC>0 at Baseline)
- Dactylitis-free state based on the Leeds Dactylitis Index (LDI) in the subgroup of subjects with dactylitis at Baseline (LDI>0)
- Change from Baseline in the LDI in the subgroup of subjects with dactylitis at Baseline (LDI>0)
- Proportion of subjects with a decrease of HAQ-DI from Baseline of at least 0.35 in the subgroup of subjects with Baseline HAQ-DI≥0.35
- Change from Baseline in the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) in the subgroup of subjects with axial involvement defined by a score of ≥4 at Baseline
- Change from Baseline in the modified Nail Psoriasis Severity Index (mNAPSI) score in the subgroup of subjects with psoriatic nail disease at Baseline (mNAPSI score>0)
- Change from Baseline in PsAID-12 total score
- Change from Baseline in the individual domain scores of PsAID-12
- Proportion of subjects achieving PsAID-12 total score ≤4
- Proportion of PsAID-12 responders (decrease from Baseline in PsAID-12 total score ≥3) in subjects with PsAID-12 total score ≥3 at Baseline
- Change from Baseline in the Psoriatic Arthritis Quality of Life (PsAQoL) total score
- Change from Baseline in the SF-36 PCS and SF-36 Mental Component Summary (MCS), as well as the 8 domain scores of the SF-36 (Physical Functioning, Role Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role Emotional, and Mental Health)
- Change from Baseline in Functional Assessment of Chronic Illness Therapy (FACIT)-Fatigue subscale score
- Proportion of FACIT-Fatigue subscale responders (subjects with a minimum clinically-important difference for FACIT-Fatigue subscale score, defined as an increase of ≥4) in subjects with FACIT-Fatigue subscale score ≤48 at Baseline
- Change from Baseline in Work Productivity and Activity Impairment Questionnaire–Specific Health Problem (WPAI-SHP) v2.0 adapted to PsA scores
- Responses to the EuroQol-5D-3-Level (EQ-5D-3L) dimensions
- Change from Baseline in EQ-5D-3L Visual Analog Scale (VAS) scores
- Change from Baseline in Physician's Global Assessment of Arthritis (PhGA-Arthritis)

- Change from Baseline in Patient's Global Assessment of Arthritis (PGA-Arthritis)

2.2.3.2 Other safety variables

Other safety variables to be assessed are:

- Change from Baseline in vital signs (blood pressure [BP], temperature, and pulse rate)
- Standard 12-lead electrocardiogram (ECG) results
- Change from Baseline in clinical laboratory values (hematology, biochemistry and urinalysis)
- Change from Baseline in Patient Health Questionnaire-9 (PHQ-9)

Physical examination findings considered clinically significant changes since the physical examination at the Screening Visit will be recorded as AEs.

2.2.3.3 Pharmacokinetic variable

The pharmacokinetic (PK) variable is the plasma concentration of BKZ. Blood samples for these measurements will be obtained at all visits except Screening.

2.2.3.4 Pharmacogenomic variables

Additional blood samples will be collected from subjects who consent to participate in the sub study at specific time points and stored at -80°C for up to 20 years.
Genomic, genetic, epigenetic, proteins, and metabolite biomarkers may be measured to evaluate the relationship with response to treatment with BKZ, PsA disease biology, and inflammatory and immune response processes. The nature and format of these tentative sub study analyses will be determined when the results of the main study are made available.
The candidate exploratory variables are the blood or blood derivative (eg, serum) concentrations of cytokines and chemokines of relevance to IL-17A/F signaling pathway and PsA biology. Additional variables may include but will not be limited to serum complement concentrations.

2.2.3.5 Immunological variables

The immunological variables allow evaluation of immunogenicity. The immunological variables are:

- The anti- BKZ antibody status
- The treatment-emergent antibody status derived from antidrug antibody (ADAb) assays
- The neutralizing antidrug antibody (NAb) status

These 3 variables will be assessed at all visits except Screening.

2.3 Study design and conduct

2.3.1 Study description

PA0011 is a Phase 3, multicenter, randomized, double-blind, placebo-controlled study. To be eligible to participate in this study, subjects must be adults with a diagnosis of PsA based on the Classification Criteria for Psoriatic Arthritis (CASPAR) criteria and have disease with TJC ≥3 and SJC ≥3, as well as an inadequate response (lack of efficacy after at least 3 months of therapy

at an approved dose) or intolerance to treatment with 1 or 2 TNFα inhibitors for either PsA or PSO. Study entrance criteria are described in the protocol.

The study consists of the following periods (see Figure 7–1 for a schematic diagram).

- a Screening Period (≥14 days to ≤35 days),
- a 16-week placebo-controlled Double-Blind Treatment Period,
- and a Safety Follow-Up (SFU) Period, 20 weeks after the final dose of IMP (for subjects who complete the study and do not enter the open-label extension (OLE) study or for subjects who discontinue early, including those withdrawn from IMP).

The maximum study duration per subject will be up to 37 weeks. The end of the study is defined as the date of the last visit of the last subject in the study.

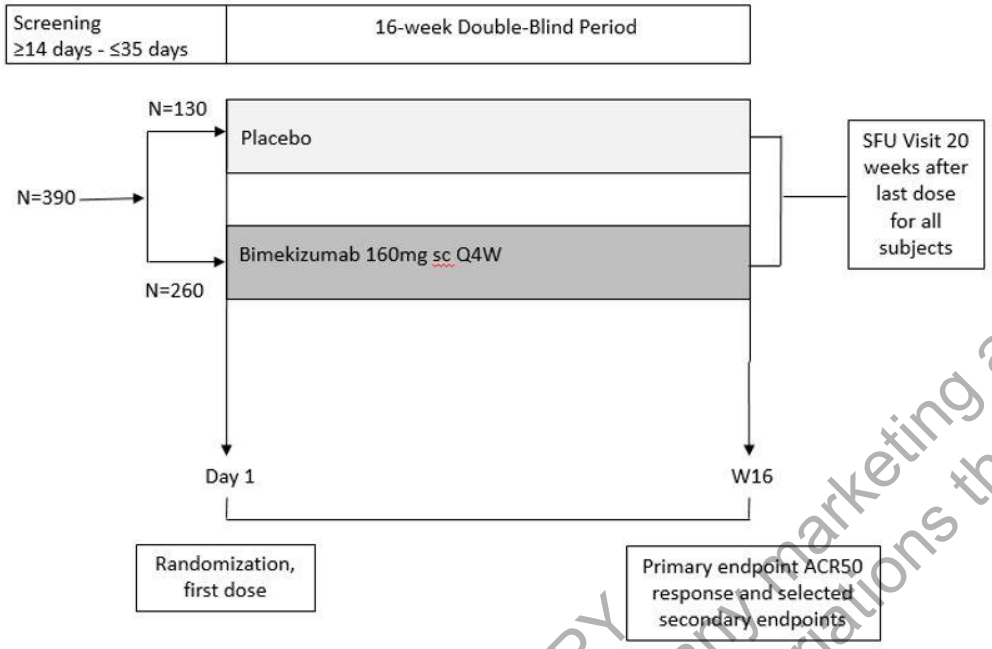
Approximately 390 subjects will be randomized in a 2:1 ratio, stratified by region (North America, Western Europe, Eastern Europe, and Asia) and prior TNFα inhibitor exposure (inadequate response to 1 prior TNFα inhibitor, inadequate response to 2 prior TNFα inhibitors, or intolerance to TNFα inhibitor) to receive one of the two following blinded treatment groups:

- Bimekizumab 160mg sc Q4W,
- Placebo

Subjects discontinuing IMP during the Double-Blind Treatment Period will be encouraged to return for all scheduled visits through Week 16 and the SFU Visit, as applicable.

At the completion of the Double-Blind Treatment Period, subjects completing the study (ie, performing all visits up to Week 16 and not having permanently discontinued IMP) and who did not meet any of the withdrawal criteria will be given the opportunity to enter an OLE study at Week 16.

Figure 7–1: Study schematic diagram



ACR50=American College of Rheumatology 50% response criteria; Q4W=every 4 weeks; SFU=Safety Follow-Up; sc=subcutaneous; W=week

Refer to the clinical study protocol for more detailed information on the study assessments and schedule of procedures.

2.3.2 Study periods

The following study periods are defined for the classification by study period:

Screening Period (≥14 days to ≤35 days)

Starts at time of the informed consent date (Screening Visit -Visit 1) and ends the day before the first dose administration of IMP (Baseline Visit – Visit 2) (ie, generally the day before Visit 2 date). This period should last between 14 to 35 days and will involve obtaining laboratory data and verifying that the doses of nonsteroidal anti-inflammatory drugs (NSAIDs) or permitted DMARDs, if used to treat PsA, are stable. The Screening Period will also enable washout of any medications not permitted for use during the study.

Double-Blind Treatment Period (16 weeks)

Starts on the day of first dose administration of study drug (Visit 2). This period should last 16 weeks maximum. During the Double-Blind Treatment Period, subjects will be randomized 2:1 (stratified by region and prior TNFα inhibitor exposure [inadequate response to 1 or 2 prior TNFα inhibitors, or intolerance to TNFα inhibitors]) to receive 1 of 2 blinded treatment regimens:

- Bimekizumab 160mg sc Q4W,
- Placebo

Visit windows of ±3 days are allowed from the first dose at all visits through Week 16. The visit window is relative to the Day 1 (Baseline) Visit.
Bimekizumab and placebo will be administered sc by unblinded study personnel at the clinical site.
The time between IMP doses should be ≥21 days and ≤35 days during the Double-Blind Treatment Period.
Subjects discontinuing IMP during the Double-Blind Treatment Period will return for all scheduled visits through Week 16 and the SFU Visit (20 weeks after the final dose of IMP) as applicable. Subjects withdrawing from the study will have an Early Termination (ET) Visit and a SFU Visit 20 weeks after the final dose of IMP, as applicable.

The end of the 16-week Double-Blind Treatment Period will be either the date of Week 16 visit (Visit 6) for subjects completing the treatment period, or the date of the early termination (ET) visit for subjects who discontinued during the Treatment Period.
If a subject does not have a Visit 6 or ET visit, then either the date of the last scheduled or unscheduled visit during the Treatment Period or the date of last known dose of study drug during the Treatment Period, whichever is later, will define the end date of the Treatment Period.

Safety Follow-Up

The SFU Period will be defined for all subjects who complete the study and do not enter the OLE study, or for subjects who discontinue early including those withdrawn from IMP. This period should last 20 weeks from the final dose.

- For subjects who complete the study and do not enter the OLE study, the SFU Period starts the day after Visit 6 date and ends on the day of the SFU Visit.
- For subjects who discontinue early, the SFU Period starts the day after the ET date and ends on the day of the SFU Visit

A subject will be considered to have completed a study period if they complete the last scheduled study visit for that period. Note that subjects who previously discontinued IMP and are continuing for all scheduled visits through Week 16 will also be considered as completed the Double-Blind Treatment Period.

2.4 Determination of sample size

Approximately 390 subjects will be randomly assigned in a 2:1 ratio to the following treatment groups: BKZ 160mg sc Q4W (260 subjects) and placebo (130 subjects).
The primary efficacy analysis is based on the primary comparison of BKZ versus placebo for ACR50 response at Week 16.
All sample size and power calculations were done at a significance level of 0.05.
All sample size and power calculations were performed using the software nQuery Advisor® 7.0.

2.4.1 Power calculation for primary endpoint

The sample size assumptions for BKZ vs placebo are based on the ACR50 response data in the subgroup of TNFα-IR patients from the Phase 2b BKZ study in a mixed prior TNFα therapy

population of subjects with moderate-to-severe PsA (PA0008). Sample size calculations for a TNFα-IR population are also based on ACR50 responses at Week 16 in the SPIRIT-P2 study.

Observed ACR50 at Week 12 results in the PA0008 TNFα-IR populations were in a small number of subjects; BKZ 160mg (n=7), 320mg (n=8), and 320mg (initial dose) plus 160mg (n=8) and ranged from 14.3% to 37.5%.

The ixekizumab Phase 3 study SPIRIT–P2 is conducted on a similar patient population to that in this study and showed a 35% (n=122) ACR50 response at Week 16. Therefore, taking into account the range of ACR50 responses at Week 12 in PA0008, the estimated ACR50 response at Week 16 in the BKZ 160mg sc Q4W group is conservatively assumed to be 26%.

For placebo, a similar approach as above is used. In the PA0008 TNFα-IR population, an ACR50 response of 11.1% (n=9) was observed at Week 12. The observed placebo ACR50 response at Week 16 was less than 10% in the SPIRIT-P2 study (n=118). Therefore, the estimated ACR50 response at Week 16 in the placebo group is assumed to be 10%.

The sample size for showing statistical superiority of BKZ vs placebo was calculated using a 2-sided 2-sample Chi-square test with continuity correction (Fleiss et al, 1980). Assuming 260 subjects in the BKZ group and 130 subjects in the placebo group, the test for detecting statistical superiority of BKZ 160mg vs placebo based on ACR50 response at Week 16 has 96% power to detect a true treatment difference of 16% (odds ratio (OR) of 3.16).

2.4.2 Power calculation for secondary endpoints

The assumptions for power calculations of the secondary endpoints included in the hierarchy and for which supporting data are available, in the TNFα-IR population are based on the results of PA0008 and the SPIRIT-P2 studies. All power calculations for binary endpoints were performed using a 2-sided 2-sample Chi-square test with continuity correction (Fleiss et al, 1980). All power calculations for continuous endpoints were performed using a 2-sided 2-group Satterthwaite t-test (Moser et al, 1989).

For the change from Baseline in HAQ-DI at Week 16, based on data from the TNFα-IR population in SPIRIT-P2 Week 12 as well as Week 12 and Week 16 data from the PA0008 and PsA001 studies respectively, estimates of the SD in the active treatment and placebo groups are 0.71. A two-group Satterthwaite t-test with a 0.050 2-sided significance level will have 85% power to detect a difference of 0.23 in the mean change from Baseline in HAQ-DI at Week 16 between BKZ treatment and placebo assuming a sample size of 390 subjects randomized 2:1.

For PASI90 response at Week 16, a PASI90 response at Week 12 of 38% (N=122), from the SPIRIT-P2 study is assumed. The PASI90 response at Week 12 in the BKZ 160mg group is 50.0% (n=8) in the subgroup analysis of the TNFα-IR patients. Placebo PASI90 response at Week 16 is based on SPIRIT-P2 data (6.0% at Week 12). However, for this endpoint the placebo response is rounded to 10%. With those assumptions, the endpoint has 99% power to detect a treatment difference of 28% (OR of 5.52) at an assumed 50% of subjects with BSA ≥3% of the planned sample size.

For the change from Baseline in PCS score of SF-36 at Week 16, based on data from the TNFα-IR population in SPIRIT-P2 where estimates of the SD in the active treatment and placebo groups at Week 12 are ixekizumab Q4W (SD=11.95) and placebo (SD=12.15). A 2-group Satterthwaite t-test with a 0.050 2-sided significance level will have 89% power to detect a

difference in means of 4.4 change from Baseline in SF-36 PCS score at Week 16 between the bimekizumab and placebo treatment groups with a sample size of 390 subjects randomized 2:1.

For MDA at Week 16, it is assumed that 5% of the placebo subjects achieve MDA at Week 16 (SPIRIT-P2 Week 12). Given the planned sample size and using a 2-sided 2-sample Chi-square test with continuity correction this study has 92% power to detect an absolute difference in MDA response at Week 16 of 12% (OR of 3.892) with a total sample size of 390 subjects.

3 DATA ANALYSIS CONSIDERATIONS

3.1 General presentation of summaries and analyses

Statistical analysis and generation of tables, figures, subject data listings, and statistical output will be performed using SAS® Version 9.3 or higher unless otherwise specified. All tables and listings will use Courier New font size 9.

Descriptive statistics will be displayed to provide an overview of the study results.

For categorical parameters, the number and percentage of subjects in each category will be presented. Unless otherwise noted, the denominator for percentages will be based on the number of subjects included in the respective analysis set. Subjects with missing data can generally be accounted for using one of the 2 following approaches:

- Percentages will be summarized based on all subjects in the analysis set and a "Missing" category (corresponding to subjects with missing data for the variable being summarized) will be included as the last row in the list of categories being summarized (this approach will be used when reporting demographics and Baseline characteristics, some ADAb and NAb tables and some shift tables from Baseline for laboratory data).
- Percentages will be based only on those subjects with observed data for the variable being summarized. As the denominator may be different from the number of subjects in the analysis set being considered, the denominator should be displayed in the table. The general format for displaying this will be "n/Nsub (%)".

Unless otherwise noted, all percentages will be displayed to one decimal place. No percentage will be displayed for zero counts, and no decimal will be presented when the percentage is 100%.

Unless stated otherwise, continuous endpoints will be expressed as a change from Baseline which will be calculated as the value at a specific timepoint minus the value at Baseline.

For continuous parameters, descriptive statistics will include number of subjects with available measurements (n), arithmetic mean, standard deviation (SD), median, minimum, and maximum (unless otherwise stated).

For hs-CRP variable, the summary statistics should contain arithmetic mean, geometric mean, geometric coefficient of variation (CV), median, first and third quartile (Q1 and Q3), minimum and maximum (for the value and ratio to Baseline). The geometric CV (%) will be calculated using the following formula:

$$CV = \sqrt{e^{SD_{ln}^2} - 1}$$

where SD_{ln} represents the standard deviation of the ln-transformed hs-CRP values.

For BKZ PK concentrations, geometric mean, geometric CV, 95% confidence intervals (CIs) for geometric mean (assuming log-normally distributed data) will be calculated if at least ⅔ of the values of interest are above the lower limit of quantification (LLOQ). If this is not the case, only median, minimum and maximum will be presented.

Decimal places for descriptive statistics will always apply the following rules:

- "n" will be an integer.
- Mean, SD, SE, and median (Q1 and Q3 where applicable) will use 1 additional decimal place compared to the original data (for a derived score, number of decimals of the original data considered is the one obtained where deriving the score in the absence of missing data).
- Coefficient of variance (CV [%]) will be presented with 1 decimal place.
- Minimum and maximum will have the same number of decimal places as the original raw value.

If the number of decimal places reported in the original data is varied then either the maximum raw number of reported decimal places or 3 will be used, whichever is the lowest, as a guide for the descriptive statistics.

Table 7–1: List of decimals for continuous efficacy endpoints

Continuous efficacy endpoint	Decimal place used for minimum and maximum	Decimal places used for Mean, SD (or SE) and median
BASDAI	1	2
DAPSA	1	2
DAS28[CRP]	2	3
EQ-5D-3L dimension scores	0	1
EQ-5D-3L (VAS)	0	1
FACIT-Fatigue subscale score	1	2
HAQ-DI	3	4
hs-CRP	2	3
LDI	1	2
LEI	1	2
mNAPSI	0	1
PASDAS	1	2
PASI	1	2
PGA-Arthritis /PhGA-Arthritis	0	1
PGA-PsA / PhGA-PsA	0	1
PsAID-12	1	2
PsAQoL	1	2
PsARC	1	2
PtAAP	0	1
SF-36	2	3
SPARCC	1	2

TJC / SJC	1	2
WPAI-SHP	1	2

BASDAI=Bath Ankylosing Spondylitis Disease Activity Index; DAPSA=disease activity index for psoriatic arthritis; DAS28[CRP]=disease activity score-28 based on C-reactive protein; EQ-5D-3L=Euro-Quality of Life 5-Dimensions 3 Level version; FACIT=Functional Assessment of Chronic Illness Therapy; HAQ-DI=Health Assessment Questionnaire – Disability Index; hs-CRP=High sensitivity C-reactive protein; LDI=Leeds Dactylitis Index; LEI=Leeds Enthesitis Index; mNAPSI=Modified Nail Psoriasis Severity Index; PASDAS= Psoriatic Arthritis Disease Activity Score; PASI=Psoriasis Area and Severity Index; PGA-Arthritis=Patient's Global Assessment of Arthritis; PGA-PsA=Patient's Global Assessment of Psoriatic Arthritis; PhGA-Arthritis=Physician's Global Assessment of Arthritis; PhGA-PsA=Physician's Global Assessment of Psoriatic Arthritis; PHQ-9=Patient Health Questionnaire-9; PSAID-12=Psoriatic Arthritis impact of disease-12; PsAQoL=Psoriatic Arthritis Quality of Life; PsARC= Psoriatic Arthritis Response Criteria; PtAAP=Patient's assessment of Arthritis Pain; SF-36=Short-Form 36-item Health Survey; SJC= Swollen Joint Count; SPARCC= Spondyloarthritis Research Consortium of Canada ; TJC=Tender Joint Count; WPAI-SHP=Work Productivity and Activity Impairment questionnaire – Specific Health Problem

Unless stated otherwise, statistical tests of efficacy variables will be performed 2-sided and p-values will be rounded to 3 decimal places. P-values less than 0.001 will be presented as "<0.001" and p-values greater than 0.999 will be presented as ">0.999." Statistical comparison will be performed at the 0.05 level of significance. Least square means (LSM), odds-ratios (OR) and corresponding confidence intervals (CI) will be presented using 3 decimals.

UCB uses SAS in a 64-bit Windows environment, and it is well-documented that in this environment the maximum accuracy of any numeric value is 15 significant digits. However, SAS by default does not limit the accuracy of numeric values to 15 significant digits which, in certain instances, may result in inaccurate representation of the data and cause errors when used in subsequent calculations, particularly when comparing a value to a chosen threshold. This, in turn, could potentially result in a change in classification of a subject from a responder to a nonresponder (and vice versa) if these values occur on a threshold used in the evaluation of response (or a critical laboratory value for example).

Therefore, in order to avoid issues caused by inaccurate floating point representation of numeric values, temporary variables are created (ie, for absolute values, change and percentage change from Baseline) during programming which are rounded to 12 decimal places prior to comparison to a specific threshold in the derivation of a response parameter. This does not imply inherent rounding on the analysis variables for absolute value, change or percentage of change which are retained unrounded in the final analysis dataset. Thus, rounding is applied exclusively during the derivation of new response parameters or critical value variables, and the rounded values are created on a temporary basis only.

The SAS® outputs supportive of any inferential statistics that are part of the hierarchical testing procedure (ie, all inferential statistics associated with the endpoints in Table 7–4 will be provided as a separate PDF document in addition to TFLs. These outputs will be included in the 'Documentation of Statistical Methods' section of the clinical study report.

The order of treatment groups to be presented in tables from left to right will be:

- Placebo
- Bimekizumab 160mg Q4W

Selected tables may also include columns for all subjects (regardless of study treatment).

The abbreviation for bimekizumab is BKZ and will be used in tables and listings headers. In the TLFs subjects randomized to bimekizumab 160mg will be labeled as "BKZ 160mg Q4W".

Per protocol, visit windows of ±3 days from the first dose at all visits through Week 16 are permissible. For the SFU Visit, the visit should occur no more than 3 days prior to the scheduled visit date and within 7 days after the scheduled visit date. Only data collected at the scheduled visit will be included in the analysis. The only deviation may concern unscheduled vendor data as explained in Section 10.5.

All by-visit summaries will contain nominal (ie, scheduled) visits only. Unscheduled visits will not be mapped to scheduled visits except for some assessments that may occur within 3-day time window of a scheduled visit. In that case, the assessment will be mapped to the corresponding scheduled visit and will be used for the analysis unless the data are present for that scheduled visit.

Assessments collected at unscheduled visits will only be listed, except in the case of determining treatment-emergent markedly abnormal criteria for laboratory and vital sign parameters where all post-Baseline (scheduled and unscheduled) values will be summarized. For PK, immunological and immunogenicity analyses, SFU data will be a part of the data reported in some tables (Section 9).

For by-visit tables summarizing efficacy data, the SFU visit will not be included, but SFU efficacy data will be listed. By-visit tables summarizing safety will however include the SFU visit.

A complete set of data listings containing all documented data and all derived data (eg, change from Baseline) will be generated.

Repeated visits will not be summarized but listed.

Unless otherwise stated, listings will be sorted by treatment, subject number within each treatment group (not randomization number), variable (if applicable) and visit (if applicable, including timing relative to dosing if applicable). For listings including nonrandomized subjects, the nonrandomized subjects will be shown first in the listing, ordered by subject number. All listings will include repeat and unscheduled measurements; such measurements will appear in chronological order together with the scheduled visits, ie, a repeated measurement will appear directly after the visit and time relative to dosing for which the repeat measurement was performed. In all the listings dates will be presented in the format 'YYYY-MM-DD' and times will be presented in 24h clock format as 'hh:mm'.

3.2 General study level definitions

3.2.1 Relative day

The relative day will be included in some listings. The way that relative day is calculated depends on when the given event occurs relative to the date of first study drug administration.

- If the event starts (stops) on or after the first dose of study drug administration, but prior to the study drug stop date, relative day is calculated as start (stop) date of the event minus first dose date of study drug dose + 1.

Relative day 1 is the date of first study drug administration.

- If the event starts (stops) after the last dose of study drug administration, the relative day is calculated as start (stop) date of the event minus most recent study drug dose date. The relative day in this situation should be preceded by a '+'

- If the event starts (stops) before the first dose of study drug administration, the relative day is calculated as start (stop) date of the event minus first dose date of study drug dose. The relative day in this situation should be preceded by a '-'.

Relative day will only be computed for fully completed dates and will be left blank for partial dates.

For AEs, relative days for start and stop dates will be calculated as the number of days since the first injection of the medication. For non-treatment emergent AEs, relative day of onset will be negative if the event started before the date of first dose. A complete date must be established in order to correctly identify the AEs. Section 4.2.3 describes imputation rules in case of missing data for AEs.

3.2.2 Mapping of assessments performed at Early Termination Visit

If the early termination visit occurs at a scheduled visit as outlined in the schedule of assessments, then no mapping is necessary, and any early withdrawal assessments should correspond to that scheduled visit. Early termination visit assessments that do not have a scheduled visit with a matching date will be assigned a visit based on the protocol defined visit windows (±3 days from first dose to Week 16, and -3 to +7 days from the final dose for the SFU Visit). For early discontinuation visits that fall between protocol defined visit windows and not within them, the later of the 2 visits will be assigned.

If there is an existing scheduled site visit in the window, then the assessments at the ET visit will be mapped to the next scheduled site visit.

The ADAb assessments and PK are exceptions to this rule:

- ADAb levels, PK from an ET visit will be mapped to the next visit where antibody levels / PK are measured.

3.3 Definition of Baseline values

A Baseline value for a subject is defined as the latest measurement for that subject up to and including the day of administration of first study medication (ie, generally Visit 2 by default), unless otherwise stated.

By default, for randomized subjects for which no start date of treatment is available, the Baseline value will be considered as the last available value prior to the randomization date.

If a Baseline assessment is taken on the same day as first administration of study medication, it is eligible to be used as the Baseline value, even in the case that the time of the assessment is recorded as taking place after the time of first study medication administration. This is considered acceptable as this measurement is still the best representation of the Baseline value of the given assessment since it is highly unlikely that the study medication could have an impact on any measurement in such a short period of time. However, such cases should be rare as study center personnel are instructed to do all assessments at the Baseline visit prior to administering study medication. One exception to this rule is plasma concentration. If Baseline plasma concentration is measured at a time after the first administration of study medication, then it should not be eligible to be considered as a Baseline plasma concentration. Such cases should be discussed with the quantitative clinical pharmacologist.

If a Baseline measurement is missing or not collected, and a Screening value is available, the Screening value will be utilized as Baseline instead. If no measurement is available prior to receiving study medication, then the Baseline value is treated as missing.

Baseline values for component scores should be computed using components from the same visit where the relevant measurements were recorded prior to dosing. For example, if the Screening Visit has all of the components, and the Baseline visit is missing one or more components, the Baseline value for the component score should be calculated using the Screening Visit values.

When the time of first dose is derived, it should be based on the first injection of study treatment, regardless of whether or not it is an active treatment.

Chest x-ray performed within 3 months before the Screening Visit is considered as Baseline measurement.

3.4 Study treatment discontinuation and intercurrent event

The concept of intercurrent event is one of the estimand attributes.

Unless otherwise noted, treatment discontinuation due to any reason will be considered as an intercurrent event (IE). Since there is no treatment discontinuation date collected, the treatment discontinuation date is defined as:

- Treatment end date + 31 days.

3.5 Protocol deviations

Important protocol deviations (IPD) are deviations from the protocol which potentially could have a meaningful impact on study conduct or on the primary efficacy, key safety, or PK outcomes for an individual subject. The criteria for identifying IPD will be defined within the appropriate protocol-specific document. All protocol deviations (PDs) will be reviewed as part of the ongoing data cleaning process and data evaluation. The IPD will be identified and documented prior to unblinding to confirm exclusion from Per-Protocol Set (PPS) and Pharmacokinetics Per-Protocol Set (PK-PPS).

Also, an IPD will not always necessitate the removal of a subject from the PPS. Depending on the type of protocol deviation, the subject may remain in the PPS. The IPD document developed by the clinical study team should clearly state which deviations will result in a removal from the PPS.

In addition to IPDs resulting in the exclusion from the PPS, subjects who reduce the dose or dosing frequency of certain medications due to intolerance, AE, side-effects or receive new prohibited medication for AE will be removed from the PPS. While this is not an IPD, as it is allowed per protocol for safety reasons, these subjects will be removed from the PPS as this non-PD could have an effect on the primary efficacy outcome in the same way as flagged IPDs resulting in exclusion from the PPS.

A specific category of PD, as a consequence of the Coronavirus Disease 2019 (COVID-19) pandemic, called COVID-19 related PD will be assessed based especially on the information collected on a dedicated eCRF page (and other available sources).

3.6 Analysis sets

The analysis sets are described below.

3.6.1 Enrolled Set

The Enrolled Set (ES) consists of all subjects who have given informed consent.

Subject dispositions will be summarized on the ES.

3.6.2 Randomized Set

The Randomized Set (RS) consists of all enrolled subjects who have been randomized.

Demographic tables, primary, secondary, and other efficacy variables will be presented on the RS.

3.6.3 Safety Set

The Safety Set (SS) consists of all subjects who received at least 1 dose of the IMP.

Demographic tables, study treatment compliance and exposure and safety variables will be presented on the SS. The anti-BKZ antibody will also be analyzed on the SS for subjects receiving BKZ.

3.6.4 Full Analysis Set

The Full Analysis Set (FAS) consists of all randomized subjects who received at least 1 dose of the IMP and have a valid measurement of all components of the primary efficacy variable at Baseline.

Supportive analysis of the primary efficacy variable will be performed on the FAS.

3.6.5 Per Protocol Set

The Per-Protocol Set (PPS) consists of all subjects in the RS who had no IPD or non-PD related to prohibited medications (along with subjects with a reduced frequency/dosing of a drug) affecting the primary efficacy variable. The deviations will be predefined and subjects with deviations will be evaluated during ongoing data cleaning meetings prior to unblinding of the data. See Section 3.5 for further details. Exclusion from the FAS will be considered as an IPD that also result in exclusion from the PPS.

Supportive analysis of the primary efficacy variable will be performed on the PPS.

3.6.6 Pharmacokinetic Per-Protocol Set

The Pharmacokinetic Per Protocol Set (PK-PPS) consists of all randomized subjects who received at least 1 dose of the IMP and provided at least 1 quantifiable plasma concentration post-dose without important protocol deviations that would affect the concentration.

PK variables will be analyzed for all subjects in the PK-PPS.

3.6.7 COVID-19-free Set

The COVID-19-free Set will consist of all subjects in the RS who had no COVID-19 impact up to the primary efficacy endpoint. This will be defined as subjects:

- not having a COVID-19 related IPD
- not having an impact based on the COVID-19 eCRF.
- not having an AE related COVID-19 (Section 10.2.5)

- not discontinuing due to COVID-19

The disposition data, the primary efficacy endpoint and the secondary efficacy endpoints included in the testing hierarchy will be analyzed on the COVID-19-free Set.

3.7 Treatment assignment and treatment groups

It is expected that subjects receive treatment as randomized and hence safety analyses will be based on the SS, as randomized. However, if after unblinding it is determined that subjects randomized to placebo received BKZ at any time, then for safety analyses these subjects will be reallocated to the BKZ treatment group. Subjects randomized to BKZ will only be reallocated to the placebo treatment group if they never received BKZ.

Efficacy analyses will be performed according to randomization and not actual treatment received. Data collected after study treatment discontinuation will be listed only and not reported in table (except in the context of the treatment policy analysis performed on the primary efficacy endpoint in Section 8.1.4.5).

3.8 Center pooling strategy

Geographic regions have been categorized as North America, Western Europe, Eastern Europe, and Asia. Below table displays the geographic regions with corresponding countries.

Table 7–2: Region definitions

Region	Countries
North America	Canada, United States
Western Europe	Germany, Italy, United Kingdom
Eastern Europe	Czech Republic, Poland, Russia, Hungary
Asia	Australia, Japan

The above regions will be the ones used when considering the region as factor in the efficacy analyses.

3.9 Coding dictionaries

Medical history and AEs will be coded using the Medical Dictionary for Regulatory Activities (MedDRA®) version 19.0. Medications will be coded using the World Health Organization Drug Dictionary (WHO-DD) version MAR2021. Medical procedures will not be coded.

To maintain consistency across studies in the bimekizumab development program, the dictionary versions will be held constant to support consistency of coding in the regulatory submissions.

3.10 Changes to protocol-defined analyses

The following changes from the protocol will be considered:

- The subgroup analysis on BASDAI will be performed on the categories: <4 vs. ≥4 rather than ≤4 vs. >4.
- Proportion of subjects with a decrease of HAQ-DI from Baseline of at least 0.35 (HAQ-DI responders) in those subjects with HAQ-DI ≥0.35 instead of >0.35.

- Proportion of PsAID-12 responders (decrease from Baseline in PsAID-12 total score ≥3) in subjects with PsAID-12 total score ≥3 at Baseline instead of >3 at Baseline.
- An additional supportive analysis for primary endpoint based on the analysis of the individual components of ACR will be performed using the Reference-Based imputation method.
- The main analyses of the secondary continuous variables included in the testing hierarchy will be performed using the Reference-Based imputation method.
- Additional subgroup analyses will be performed for HAQ-DI responders at Week 16.
- Additional subgroups based on the combination of concomitant methotrexate (MTX) and baseline conventional disease modifying antirheumatic drug (cDMARD).

3.11 Changes related to COVID-19

The impact of the COVID-19 pandemic on study procedures/conduct as well as the efficacy and safety endpoints will be investigated, and additional analysis outputs will be provided as appropriate. These additional analyses were not planned as part of the original protocol as the pandemic was not ongoing at the time of protocol finalization. These additional analyses will include analyses by period of the COVID-19 pandemic (pre/during/post) as defined below:

- Pre COVID-19 pandemic period: Period prior to COVID-19 pandemic start date defined as 11-Mar-2020.
- COVID-19 pandemic period: Period from 11-Mar-2020 though the COVID-19 pandemic end date which is currently not defined at the time of approval of the SAP.
- Post COVID-19 pandemic period: Period after the declaration of the end of the pandemic

The additional analyses are described in the following sections:

- Subject disposition (Section 5.2)
- Protocol deviations (Section 5.3)
- Efficacy analyses related to the hierarchy endpoints (Section 8.1.4.8 and Section 8.2.3)
- Adverse events (Section 10.2.5)

4 STATISTICAL/ANALYTICAL ISSUES

4.1 Adjustments for covariates

The primary efficacy analysis will investigate the treatment effect, adjusting on the 2 randomization stratification variables:

- Region
  - North America, Western Europe, Eastern Europe, Asia

  or
  - North America, Eastern Europe, Western Europe + Asia if the percentage of randomized subjects is less than 10% in either of the Asia or Western Europe regions

- Prior TNFα inhibitor exposure (inadequate response to 1 prior TNFα inhibitor, inadequate response to 2 prior TNFα inhibitors, intolerance to TNFα inhibitors)

When adjusting on these 2 variables, a statistical model might not converge (ie, in case the likelihood maximization algorithm failed to converge). In that case, the statistical model will be run after dropping stratification variables in their order of appearance, successively region (one first and the other after) and prior TNFα inhibitor exposure.

In the event that a subject is stratified in the incorrect stratum (ie, the stratum recorded in the Interactive voice or web Response System (IXRS) differs from the actual stratum the subject belongs to), the actual stratum will be used for the analysis.

The secondary analyses will be adjusted on the same categorical factors as the primary analysis. For continuous variables, Baseline value (of the variable of interest) will also be included as covariate where appropriate.

4.2 Handling of dropouts or missing data

4.2.1 Strategy for handling missing data for efficacy analyses

Different approaches will be used to handle missing data including how the intercurrent events will be considered (Section 3.4).

In this section, 3 terms will be defined for binary endpoints:

- Non-composite binary endpoint: Binary endpoints derived based on 1 continuous measurement (eg, PASI90).
- Composite binary endpoint: Binary endpoints derived based on several continuous measurements (eg, ACR50).
- Composite continuous endpoint: Continuous endpoint derived based on several continuous measurements (eg, PASDAS).

The table below summarizes all continuous endpoints (excluding composite continuous endpoints) along with their observed ranges and indicate whether they are represented by an integer value. These endpoints will be analyzed and / or used in the derivation of composite or non-composite endpoints:

Table 7–1: List of non-composite continuous efficacy endpoints

Efficacy endpoint	Minimum	Maximum	Integer Value
BASDAI	0	10	
EQ-5D-3L dimension scores	1	3	Yes
EQ-5D-3L (VAS)	0	100	Yes
FACIT-Fatigue subscale score	0	52	
HAQ-DI	0	3	
hs-CRP	LLOQ/2 where LLOQ=0.10mg/L	No maximum	

Efficacy endpoint	Minimum	Maximum	Integer Value
IGA	0	4	Yes
LDI	0	No maximum	
LEI	0	6	
Tender dactylitis count	0	20	Yes
mNAPSI	0	13	Yes
PASI	0	72	
PGA-Arthritis /PhGA-Arthritis	0	100	Yes
PGA-PsA / PhGA-PsA	0	100	Yes
PsAID-12	0	10	
PsAQoL	0	20	
PtAAP	0	100	Yes
SF-36[a] Physical Functioning (PF) Role Physical (RP) Bodily Pain (BP) General Health (GH) Vitaly (VT) Social Functioning (SF) Role Emotional (RE) Mental Health (MH) Physical component summary (PCS) Mental component summary (MCS)	 19.26 21.23 21.68 18.95 22.89 17.23 14.39 11.63 5.02 -3.33	 57.54 57.16 62.00 66.50 70.42 57.34 56.17 63.95 79.78 80.09	
SPARCC	0	16	
TJC / SJC	0	68/66	
WPAI-SHP	0	100	

BASDAI=Bath Ankylosing Spondylitis Disease Activity Index; EQ-5D-3L=Euro-Quality of Life 5-Dimensions 3 Level version; FACIT=Functional Assessment of Chronic Illness Therapy; HAQ-DI=Health Assessment Questionnaire – Disability Index; hs-CRP=High sensitivity C-reactive protein; IGA=Investigator Global assessment; LDI=Leeds Dactylitis Index; LEI=Leeds Enthesitis Index; LLOQ=lower limit of quantification; mNAPSI=Modified Nail Psoriasis Severity Index; PASI=Psoriasis Area and Severity Index; PGA-Arthritis=Patient's global assessment of Arthritis; PGA-PsA=Patient's Global assessment of Psoriatic Arthritis; PhGA-Arthritis=Physician's Global Assessment of Arthritis; PhGA-PsA=Physician's Global Assessment of Psoriatic Arthritis; PHQ-9=Patient Health Questionnaire-9; PSAID-12=Psoriatic Arthritis Impact of Disease-12; PsAQoL=Psoriatic Arthritis Quality of Life; PtAAP=Patient's Assessment of Arthritis Pain; SF-36=Short-Form 36-item Health Survey; SJC= Swollen Joint Count; SPARCC= Spondyloarthritis Research Consortium of Canada ; TJC=Tender Joint Count; WPAI-SHP=Work Productivity and Activity Impairment questionnaire – Specific Health Problem
[a] minimum and maximum of the norm-based scores – Normal range for SF-36 scores is 0-100

4.2.1.1 Primary analysis of binary endpoints that are part of the testing hierarchy (primary and secondary endpoints)

All binary endpoints (composite & non-composite) are based on continuous component variables.

The primary analysis of binary efficacy endpoints that are part of the testing hierarchy is considered under a framework estimand in which missing data (due to study treatment discontinuation) is considered indicative of failed treatment and imputed to nonresponse. Further data missing while on treatment, or data observed while not on treatment, will also be imputed to nonresponse.

This composite estimand approach to handling missing data which is the primary analysis method for binary efficacy endpoints is similar to the non-responder imputation (NRI) method.

4.2.1.2 Primary analysis of continuous endpoints that are part of the testing hierarchy (secondary endpoints)

For continuous efficacy endpoints that are part of the testing hierarchy, the primary analysis method is based on the hypothetical estimand approach as follows:

- If subjects have missing data regardless whether the missing data is preceded by an IE, then missing data will be imputed based on the MI-MCMC/Reference-Based imputation method (Section 4.2.2.3).
- If subjects have non-missing data after IE, then such data will be set to missing prior to running MI.

4.2.1.3 Supportive analyses of primary efficacy endpoint

For the primary endpoint, which is a composite binary endpoint, several supportive analyses assuming different missing data mechanisms will be conducted:

- The modified composite estimand approach: the IE is changed from all treatment discontinuation to discontinuation of treatment due to AE or lack of efficacy. The same as the primary method will be used with the difference that only data after discontinuation due to AE or lack of efficacy is set to nonresponse. All other imputed data will be used.
- For the analysis of individual components of ACR, under the hypothetical estimand approach using the Reference-Based imputation method (Section 4.2.2.3).
- The tipping point approach implemented within the MI framework, but only if the primary endpoint analysis result is statistically significant at alpha=0.05. (Section 4.2.2.4).
- The treatment policy strategy (Section 4.2.2.5).
- The observed case (OC) analysis that will only include the observed data. For subjects who had treatment discontinuation, only observed data prior to treatment discontinuation date will be analyzed. For visits with missing data without treatment discontinuation, the data will remain missing. For observed data after treatment discontinuation, the corresponding visit data and subsequent visit data will be treated as missing.

4.2.1.4 Supportive analyses of endpoints that are part of the testing hierarchy (secondary endpoints)

The following supportive analysis methods will be performed for the secondary and the other efficacy variables that are part of the testing hierarchy:

For binary endpoints:

- The modified composite estimand approach: the IE is changed from all treatment discontinuation to discontinuation of treatment due to AE or lack of efficacy. For composite or non-composite binary endpoints, the standard MI approach) will be implemented (similarly as in Section 4.2.1.3) on the raw continuous score(s) (eg, PASI for the PASI90 endpoint) before deriving the binary endpoint based on the imputed score(s).
- The OC analysis.

For continuous endpoints:

- The hypothetical estimand approach:
  - If subjects have missing data regardless whether the missing data is preceded by an IE, then missing data will be imputed based on the MI-MCMC/Monotone regression method (Section 4.2.2.2).
  - If subjects have non-missing data after IE, then such data will be set to missing prior to running MI.
- The OC analysis.

4.2.1.5 Analyses of secondary endpoints that are not part of the testing hierarchy and other efficacy endpoints

The other efficacy endpoints including the secondary efficacy endpoints that are not part of the testing hierarchy will be analyzed as the secondary efficacy endpoints (with the exclusion of the MI-MCMC/reference-based imputation method for continuous variables) but with no designated priority.

For a combined binary endpoint based on the combination of a non-composite and a composite binary endpoint (as with the ACR50 and PASI90 combined endpoint), each binary endpoint will be derived independently, as described above, before deriving the combined endpoint.

4.2.1.6 Summary table of missing data handling approaches for efficacy analyses

The below table summarizes which missing data handling approaches will be used for each type of efficacy endpoint.

Table 7–2: Missing data handling approaches for efficacy endpoints

Efficacy Endpoint Priority	Variable Type	Missing data handling approach						
		Composite estimand (NRI)	Modified Composite estimand (MI)	Hypothetical estimand (MI)	Hypothetical estimand Reference-Based (MI)	OC	Tipping Point analysis (MI)	Treatment Policy (MI)
Primary endpoints	Composite Binary	P	S[a]			S	S[a]	S[a]
	Binary	P	S[a]			S		

Efficacy Endpoint Priority	Variable Type	Missing data handling approach						
		Composite estimand (NRI)	Modified Composite estimand (MI)	Hypothetical estimand (MI)	Hypothetical estimand Reference-Based (MI)	OC	Tipping Point analysis (MI)	Treatment Policy (MI)
Secondary endpoints included in the testing hierarchy	Continuous			S	P	S		
Other Secondary endpoints (not included in the testing hierarchy	Binary	X	X[a]			X		
	Continuous			X		X		
Other endpoints	Binary	X	X[a,b]			X		
	Continuous			X[b]	X[d]	X		
	Categorical	X[c]	X[c]			X		

P=Primary method, S=Supportive method, X=Method to be used (no priority designated) MI=multiple imputation, NRI=non-responder imputation, OC=observed cases.
Note: Composite estimand (NRI) refers to the approach in which data preceded by the intercurrent event of study treatment discontinuation are imputed as nonresponse, and other missing data are also imputed as nonresponse.
Note: Modified Composite Estimand (MI) refers to the approach in which data preceded by the intercurrent event of study treatment discontinuation due to AE or lack of efficacy are imputed as nonresponse (or worst category for categorical variable), and other missing data are imputed via a MI model.
Note: Hypothetical Estimand (MI) refers to the approach where outcomes for subjects without an intercurrent event of study treatment discontinuation are as observed, and outcomes for subjects with the intercurrent event are imputed via a MI model.
Note: Referenced Based imputation refers to the approach in which it is assumed that the statistical behavior of the bimekizumab and placebo-treated subjects after discontinuing study medication becomes that of the placebo-treated subjects that remain in the study.
[a] Imputation method is applied on continuous data, and binary variable is derived from the continuous variable based on complete imputed data set.
[b] MI will be used to provide point estimates at each timepoint only.
[c] Missing data to be imputed to the worst category.
[d] For ACR components analysis at Week 16.

4.2.2 Methods for handling missing data for efficacy analyses

The sections below describe the method to be used for missing efficacy data handling.

Note: For a specific endpoint, if analyzed subjects in a one or several treatment groups have no missing data, then analysis of observed cases will be performed in the concerned treatment group(s).

4.2.2.1 Non-responder imputation

For binary endpoints, the non-responder imputation (NRI) analysis, also described above as the primary composite estimand, will consider the following subjects as “non-responders” for the timepoint of interest:

- Subjects with missing data at the timepoint of interest. In the case of ACR50, nonresponse is imputed if ACR50 cannot be derived (ie, derived as “missing”) based on available ACR components. (Section 8.1.1).

- Subjects remaining in the study at the timepoint of interest but discontinuing study treatment before the timepoint of interest (data after IE).
- Subjects with missing Baseline value (for composite or non-composite binary variables based on change(s) from Baseline of continuous endpoint(s)).

4.2.2.2 MI – MCMC/Monotone Regression

In instances where MI is used, the missing value is replaced by a set of plausible values, where each value is a Bayesian draw from the conditional distribution of the missing data given the observed data. Intermittent missing data will be imputed using the MCMC method, followed by regression for monotone missing data (called "MI-MCMC/Monotone regression" method in this SAP). The MI procedures planned for efficacy analyses are based on an assumption of data missing at random (MAR).

This section describes the algorithms to be implemented for the MI – MCMC/Monotone Regression procedures for non-composite binary endpoints, composite binary endpoints, and continuous endpoints. These descriptions focus on the MI procedure itself and do not specifically account for dealing with intercurrent events which is addressed in Section 4.2.1.

4.2.2.2.1 Non-Composite binary endpoints

For non-composite binary endpoints, the MI method will be applied as follows based on raw values:

- Step 1: Imputation of missing data using MI:

Create datasets, one for each treatment group, of subjects with observed values and missing values (needing estimation by MI). For the imputation step, missing values will be separated into 2 categories: intermittent missing values (ie, missing values for a given subject that has available data before and after the missing timepoint, including missing value at Baseline) and monotone missing values (ie, where all subject data is missing after a given time point). Datasets should be designed in a horizontal structure meaning that each subject should be present in a single observation, with a set of values, one for each scheduled visit where the endpoint is supposed to be collected according to the protocol (this excludes the unscheduled visits).

Datasets should also be sorted by subject number before proceeding with the MI process.

a. **For the intermittent missing values**, the missing value in each dataset will be filled in using the Markov-Chain Monte Carlo (MCMC) method with multiple chains, monotone missing data imputing pattern, and non-informative priors for all parameters. Unless specified differently, the first 200 iterations will not be used (the 'burn-in' option). A total of 100 sets of imputations will be performed. The seed used for these imputations will be 2017 and all other MI procedures described in this SAP will use this same seed as well. The procedure will be performed for each treatment group separately.

The resulting 100 imputed data sets will have a monotone missing pattern and will be imputed using a method for monotone missingness:

b. **For monotone missing data,** one monotone regression will be used to impute missing data. A separate regression model is estimated for each variable with missing values (ie, measurement at each time point). Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable. Since the data set has a

monotone missing data pattern, the process is repeated sequentially for variables with missing values. The procedure will be based on the 100 imputed datasets generated from the MCMC procedure and will be performed for each treatment group separately by imputation.

The SAS® PROC MI procedure will be used for the imputation.

In both cases, the imputation model will include the randomization strata as stratification variables (region and prior TNFα inhibitor exposure), the value at Baseline and at each post-Baseline visit (up to the week of interest). The imputation model based on the MCMC method will only allow joint multivariate normal variables. For the MCMC method (when imputing intermittent missing values), randomization strata will be re-coded as indicator variables, and will be always specified in the following order if they are retained in the model:

- North America region (1 for North America, 0 otherwise)
- Western Europe region (1 for Western Europe, 0 otherwise)
- Eastern Europe region (1 for Eastern Europe, 0 otherwise)
- Prior TNFα 1 (1 for inadequate response to 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior or more TNFα inhibitors)

  No variable needs to be created for the 3rd category of prior TNFα inhibitors (intolerance to TNFα inhibitors) or the 4th region (Asia) which is covered when all above variables are "0".

The randomization strata (region and prior TNFα inhibitor exposure) may be dropped from the imputation model to facilitate model convergence if required (this will be the case for all imputation models) as described below:

- If the MI fails to converge when adjusting on the stratification variables specified above, or if the percentage of randomized subjects is less than 10% in either of the Asia or Western Europe regions, then the Western Europe and Asia region will be combined and the binary variables for region that will be left in the MI will then be for North America and Eastern Europe regions.
- If after doing that, the MI still fails to converge then all region factors will be removed from the MI.
- If the MI still does not converge after dropping the regions, the remaining stratification variables (Prior TNFα 2,1) will also be removed.
- If a variable is dropped in order to allow convergence for one model in the study, that variable does not have to be dropped from other models in the study of the model converges without dropping the variable. That is, model convergence should be evaluated for each efficacy table independently. Furthermore, this means that a table based on a single timepoint may have different variables in the imputation model than a separate by-visit table for the same variable.

The post-Baseline values will need to be specified in chronological order after the Baseline value in the imputation model so that the SAS® PROC MI imputes variables from left to right (ie, the Week 4 value will be first imputed using regression based on Baseline value, and then the Week 8 value will be imputed using regression based on Baseline and Week 4 values, etc.). The

resulting datasets for each treatment arm will be combined into 1 complete dataset containing 100 times the number of subjects analyzed.

Note: When imputing missing hs-CRP values, missing values due to values below LLOQ (0.10 mg/L) will be replaced by the midpoint value between 0 and LLOQ prior to running MI.

- Step 2: Standard rounding rules will be applied to the imputed values after each SAS® PROC MI. If an imputed value falls outside of the range for the given variable (as listed in Table 7–1), the value will be updated to be within the predefined range. For example, the imputed value for PASI will be updated to 0 or 72 in the case of an imputed value less than 0 or greater than 72, respectively. For endpoints that can take only integer values (eg,: IGA as listed in Table 7–1), the imputed values will be rounded to the closest integer after each SAS® PROC MI.

  Note: For hs-CRP, the lower limit used the midpoint value between 0 and LLOQ. For some parameters, there is not necessarily an upper limit to the imputed value (ie, LDI)

- Step 3: On the dataset obtained from Step 2, the binary responder variable will be derived. In practice, the value at the week of interest (eg, Week 16) in the imputed data sets will be used to categorize the subject as a responder or not. If subjects have an IE, then the endpoint at all subsequent visits will be set to "nonresponse" and the following will be performed:
  - When date of visit is available: by comparing IE date vs. date of visit
  - When the date of visit is missing in case of fully imputed data at a visit: by comparing the next visit number after the visit where the last study medication occurs vs. the visit number where data are imputed

- Step 4: At each timepoint the (unadjusted) proportion of responders will be calculated by treatment group from the imputed datasets using SAS® PROC FREQ.

- Step 5: (For primary and secondary efficacy endpoints only): For each value of the imputation number from 1 to 100, the adjusted proportion of responders will be analyzed using a logistic regression model with a fixed effect for treatment. The suitability of including region and prior TNFα inhibitor exposure as fixed effects will be assessed using goodness-of-fit tests (Deviance and Pearson's and Hosmer-Lemeshow) and added, if appropriate, if it allows model convergence. Covariates kept in the modelling should also be the same as the ones previously used for the MI. Comparisons will be made using 2-sided Wald test level of significance of 5%. The results obtained from the 100 logistic regression analyses (ie, the adjusted proportion of responders for each treatment group, the OR and the difference of proportions for the BKZ-placebo comparison and corresponding 95% CI) will be combined for overall inference using Rubin's rules, which account for the uncertainty associated with the imputed values (Rubin, 1987). This will be done using SAS® PROC MIANALYZE.

4.2.2.2.2 Composite binary endpoints

For composite binary endpoints, the MI method will be applied as follows:

- Step1: The MCMC/monotone regression approach described above in Step 1 for non-composite binary endpoints will be performed separately for each individual component variable.

- Step 2: The same rounding rules described in Step 2 above for non-composite binary endpoints will be performed for each component.
- Step 3: Based on the multiple imputed data sets obtained for each component, the binary response will be derived as follows:
  - The dataset obtained for each component will be merged by imputation number and subject number
  - On the dataset obtained, the binary endpoint will be derived for each subject/visit based on the component values.
  - If subjects have an IE, then the endpoint at all subsequent visits (from the day after the intercurrent event date, whether the data were observed or not) will be set to "non-response".
- Steps 4 and 5: Same as Steps 4 and 5 above for non-composite binary endpoints (Section 4.2.2.2.1).

For the primary endpoint, the 'burn-in' option in SAS® PROC MI may be set higher than the default value. The goodness-of-fit tests to assess convergence of the MCMC will be provided in a SAS® output.

4.2.2.2.3 Calculation of adjusted responder rates, odds ratio and CIs for binary endpoints.

Estimates of the adjusted responder rates for each treatment group and the associated standard errors are obtained from the logistic regression in Step 5 on the logit scale and as such are assumed to follow a normal distribution. These estimates will be combined using Rubin's rules and the combined estimates and associated SEs will be used to construct 95% CIs on the logit scale. The combined estimates and 95% CIs on the logit scale will be back transformed using the inverse logit link function to obtain the adjusted responder rates (%) and associated 95% CIs.

Some key points to consider relative to the calculation of the odds ratios and corresponding confidence intervals. As the estimates of the odds ratios from the logistic regression model in step 5 follow a lognormal distribution, a log transformation is needed to normalize these 100 odd-ratio estimates since the procedures for combining results from multiple imputed datasets are based on the assumption that the statistics estimated from each imputed dataset are normally distributed. Therefore, the log of the odds ratio estimates from the logistic regression model are used when combining into a single inference (step 3). Additionally, the standard errors (SE) for the odds ratios are transformed as follows:

$$SE = \frac{\log(UCL) - \log(LCL)}{2Z_{\alpha/2}}$$

where UCL and LCL are the upper and lower confidence limit, respectively, for the CI of the odds ratio from the logistic regression model, and $Z_{\alpha/2}$ is the relevant critical value from the standard normal distribution (1.96 for a 95% CI). The estimates of the log odds ratio for BKZ relative to placebo and the corresponding upper and lower confidence limits will be provided. The odds ratio will be then estimated by exponentiating the estimate of the log odds ratio. The odds ratio and the confidence limits of the odds ratio will be estimated as follows:

$$OR = \exp(Log\ odds\ ratio\ estimate\ )$$

$$LCL = OR * \exp(-SE * Z_{\alpha/2}\ )$$

$$UCL = OR * \exp(SE * Z_{\alpha/2}\ )$$

where OR is the back-transformed estimate of the odds ratio just described, SE is the standard error of the log odds ratio and $Z_{\alpha/2}$ is the relevant critical value from the standard normal distribution (1.96 for a 95% CI). These calculations will be done such that odds ratios and corresponding CIs are calculated for the odds ratio of BKZ versus placebo.

In addition to calculating the odds ratio, associated CIs, and p-values for the comparisons of BKZ vs. placebo, the estimated proportion of responders (ie, estimated responder rate) and the difference in the proportion of responders between BKZ and placebo will be estimated, along with 2-sided 95% CIs. The creation of the estimates of this difference will be completed using the process detailed below:

1. Use the logistic regression model to calculate:

   Least squares mean (LSM) estimates of the log odds of BKZ (G_B) and placebo (G_P), as well as their corresponding standard errors (S_B and S_P, respectively).

   Standard error of the LSM estimate of the log odds ratio (S_R)

2. Compute estimates for predicted proportions using the following transformations:

   $P_B = \exp(G_B)/(1 + \exp(G_B))$

   $P_P = \exp(G_P)/(1 + \exp(G_P))$

   The difference in proportions is then given by:

   $D = P_B - P_P$

3. Estimate the standard error of D by:

   $S_D = \text{sqrt}[P_B^2(1\text{-}P_B)^2S_B^2 + P_P^2(1\text{-}P_P)^2S_P^2 + P_B(1\text{-}P_B)P_P(1\text{-}P_P)S_R^2 - P_B(1\text{-}P_B)P_P(1\text{-}P_P)(S_B^2 + S_P^2])$

The MI/MCMC monotone regression method, as previously outlined, will be used to account for missing values. The calculation steps described above will be based on the results provided from the logistic regression model of the multiple imputed datasets. An estimate for the difference in proportions, D, and corresponding standard error, SD, will be computed for each of these datasets. The results from these analyses will be combined into a single estimate of the difference in predicted proportion of response and a 2-sided 95% CI interval using SAS PROC MIANALYZE.

4.2.2.2.4 Continuous efficacy endpoints

For secondary and other continuous efficacy endpoints, the same methods as the ones described above for the non-composite binary endpoints will apply based on the values at the timepoint(s) of interest, with the following differences:

- Step 1: The same imputation procedure as the described in step 1 for the non-composite binary endpoints will be used.

- Step 2: The same rounding rules described in step 2 above for non-composite binary endpoints will be performed.
- Step 3: On the dataset obtained with the imputation number from 1 to 100; change from Baseline will be derived. Simple means and standard errors will be calculated using Rubin's rules (via SAS® PROC MIANALYZE) for each timepoint(s) of interest. For calculation of other descriptive statistics such as the median, min and max, Rubin's rules do not apply. MI estimates will be computed by simply averaging the estimates from the multiple repetitions of the imputation algorithm. For CRP data (that will be presented using the geometric mean and corresponding 95% CI, arithmetic mean, median, Q1, Q3, minimum and maximum). The change from Baseline will be expressed as the ratio to Baseline (value at the visit divided by value at Baseline) in the by-visit summaries. The following approach will be applied:
  - Following the MI procedure, the ratio to Baseline will be calculated for any of the imputed values
  - The natural logarithm of the absolute values and of the ratios to Baseline will be calculated
  - The logged values will be summarized by treatment, visit and imputation
  - The datasets will be combined using PROC MIANALYZE in order to get the mean from the absolute values and ratios to Baseline across imputations
  - The estimates of the mean will be back transformed to obtain the geometric mean on the original scale
  - For the median, Q1, Q3, minimum and maximum the procedure outlined above for the other endpoints will be followed
- Step 4: (for primary and secondary efficacy variables): For each value of the imputation number from 1 to 100, the change from Baseline will be analyzed using an ANCOVA model with treatment group, region, and prior TNFα inhibitor exposure as fixed effects and the Baseline value as covariate. The results obtained from the 100 ANCOVA analyses (ie, the LSM for the treatment difference and the 95% CI for the contrasts) will be combined using Rubin's rules (via SAS® PROC MIANALYZE).

For composite continuous endpoints, the MI will be applied on each component the same way as for composite binary endpoint. The datasets obtained for each component will be merged by imputation number and subject number. On the dataset obtained, the endpoint will be derived for each subject/visit based on the component values. The endpoint obtained will be analyzed as above.

4.2.2.2.5 Ordinal efficacy endpoints

For ordinal efficacy endpoints, the MI method will be applied as for the continuous efficacy endpoints with the following exceptions:

- After the imputation of the intermittent and monotone missing data , the rounding rules will be performed to align the data with the possible responses for the respective endpoint (eg, If the variable can take the value 1, 2 or 3, the imputed values will be rounded to the closest possible value).

- If the ordinal efficacy endpoint is used to derive a binary endpoint, the same analysis strategy as described above for the binary endpoints will be used.
- If the ordinal efficacy endpoint is not used to derive a binary endpoint, the unadjusted proportions of subject by categories will be presented.
- If subjects have an IE, then the ordinal endpoint at all subsequent visits (from the day after the IE date, whether the data were observed or not) will be set to the value used for the worst category.

4.2.2.3 MI – MCMC/Reference-Based imputation

For the primary analysis of the secondary continuous efficacy variable included in the testing hierarchy and the individual components of the ACR response, the Reference-Based multiple imputation method will be used. In this imputation method, the missing data will be imputed based on data from the placebo group only (Mallinckrodt, 2013).

Reference-Based MI assumes that the statistical behavior of the BKZ and placebo-treated subjects after discontinuing study medication becomes that of the placebo-treated subjects. All time points after study treatment discontinuation for BKZ and placebo groups will be considered missing. MIs are used to replace missing outcomes for BKZ- and placebo-treated subjects who discontinued using multiple draws from the posterior predictive distribution estimated from the placebo arm.

Note: SAS 9.4 may be used for this analysis where equivalent steps to those outlined below will be used to perform the imputation.

In the case of continuous endpoints, the procedure will be implemented as follows on the raw values:

1. Data will be processed sequentially, one timepoint (visit) at a time, by repeatedly calling SAS® PROC MI to impute missing outcome data at visits t=1,…, Week 16 (Week 16 being the time point of interest) using data from the placebo-treated subjects arm only.

a. *Initialization*. Set t=1 (Baseline visit). Create a dataset combining all records from all subjects with columns for covariates (prior TNFα exposure and region) and outcome at Baseline. Impute missing values at Baseline using prior TNFα exposure at Baseline and region.

b. *Iteration*. Set t=t+1. Create a dataset combining records from BKZ- (subjects with missing data at visit t) and all placebo-treated subjects with columns for covariates (prior TNFα exposure and region) and outcomes at visits 1 to t. In this dataset, outcomes for BKZ-randomized subjects are missing at visit t and observed or previously imputed at visits 1 to t-1. Outcomes for placebo-treated subjects are observed or missing at visit t and observed or previously imputed at visits 1 to t-1. The outcomes should be sorted in chronological order in the model.

c. *Imputation*. Impute missing values for visit t using previous outcomes for visits 1 to t-1, prior use of TNFα inhibitor, and region. Note that only placebo data will be used to estimate the imputation model since no outcome is available for BKZ- treated subjects at visit t. As a consequence, the input dataset should include all subjects from placebo but only subjects from the BKZ arm that have values at timepoint t missing.

d. Repeat steps 1b-1c for all timepoints, 100 times with different seed values (seeds ranging from 201 to 300) to create 100 imputed complete datasets. Subjects whose missing values were imputed in the last PROC MI call will be included in the input dataset for the next PROC MI call. Standard rounding rules will be applied to the imputed values. If the MI procedure yields a value outside of the pre-defined range for the given variable, the value will be updated to be within the predefined range of values for the endpoint of interest. For example, the imputed value for TJC will be updated to 0 or 68 in the case of an imputed value less than 0 or greater than 68, respectively.

2. The analysis will be done on the change from Baseline, and the model will be an ANCOVA model with treatment group, region, and prior TNFα inhibitor exposure classification as fixed effects and the Baseline value as covariate.

   For generation of summary statistics, the 100 imputed datasets will be combined, and simple means and standard errors will be calculated using Rubin's rules (via SAS® PROC MIANALYZE). For calculation of other descriptive statistics (median, Q1, Q3, minimum and maximum), Rubin's rules do not apply. Multiple imputation estimates will be computed by calculating arithmetic means of the estimates from the multiple repetitions of the imputation algorithm. Thus, for median, Q1, Q3, minimum and maximum the following approach will apply:

   - The data will be summarized by treatment, visit and imputation and the summary statistics will be computed.
   - Results will be summarized by taking the mean value of each summary statistic at each visit across all imputations.
   - The number of decimal places will remain the same as the original for display purposes (ie, if the mean was presented to 1 decimal place, the mean of the means will also be presented to 1 decimal place).

3. The results obtained from the 100 ANCOVA analyses in Step 4 (i.e. the Least Square Means for the treatment difference and the 95% CI for the contrasts) will be combined with Rubin's rules.

4.2.2.4 Tipping Point analysis

Tipping point analyses will also be implemented as a supportive analysis for the primary efficacy endpoint.

The objective of the tipping point analyses is to identify assumptions about the missing data under which the conclusions from the main analysis change (ie, under which there is no longer evidence of a treatment effect. These tipping point analyses will only be performed if the primary efficacy analysis results in a statistically significant treatment effect ($p<0.05$).

As a first step, the worst-case scenario will be evaluated. Specifically, all BKZ-randomized subjects with a missing ACR50 at Week 16 (or non-missing ACR50 after study treatment discontinuation) will be imputed as non-responders, while all placebo-randomized subjects with a missing ACR50 at Week 16 will be imputed as responders. While there is little justification for such an approach, it makes the most putative assumption possible against a BKZ treatment effect. After applying this imputation approach, a logistic regression model consistent with the

one described for the primary analysis will be applied. If the p-value for the odds ratio of BKZ versus placebo is less than 0.05, then no further tipping point analyses are needed.

If this analysis based on the worst-case scenario results in a p-value greater than 0.05, then additional tipping point analyses will be performed. In practice, it implies different delta adjustments will be made to the assumed responses on the missing data (where missing values include observations after the intercurrent event date and any other missing values) in each treatment group independently with various degrees of plausibility with the goal to find for each treatment group the "tipping point" that will significantly reverse the primary result which yielded a p-value less than 0.05. These delta adjustments will be done on each component of ACR50 as follows:

- Step 1: The same MCMC method described in Section 4.2.2.2 will be implemented for non-monotone (intermittent) missing pattern values, using the same imputation model. This will be based on 100 imputations.
- Step 2: Based on the dataset obtained in Step 1, a monotone regression model will be applied (using the same imputation model as in Step 1) and as described in Section 4.2.2.2 . This will be based on 1 imputation.
- Step 3: Imputed values will be shifted using a delta adjustment independently in each treatment group.

Once defined, the same delta adjustment value will be applied on the imputed values for the Week 16 visit only. The selected scenario will assume that subjects randomized to BKZ and who have missing data have a lower probability of response compared to subjects randomized to placebo with missing data.

- For ACR components for which high scores are associated with a less favorable outcome, it will mean that:
  - A positive shift is applied to the imputed value for subjects randomized to BKZ in order to increase the imputed value.
  - A negative shift is applied to the imputed value for subjects randomized to placebo in order to decrease the imputed value.

A set of possible values will be first pre-defined for the delta parameter as follows:

- For all ACR components, the value of the initial delta parameter will be equivalent to a specific percentage of the possible range of each component (ie, 5%). For hs-CRP, the range will be based on the log-transformed values at Week 16. The delta parameters for each endpoint are listed in the Table 7–3 .

Table 7–3: Tipping point analysis: Delta parameter for each ACR component

ACR component[1]	Range	Delta
SJC	0-66	3.3
TJC	0-68	3.4
HAQ-DI	0-3	0.15
PtAAP, PhGA-PsA, PGA-PsA	0-100	5

ACR component[1]	Range	Delta
hs-CRP[2]	Observed range of the $\log_e$ transformed values for all subjects at Week 16	5% of the observed range

[1]The shifted imputed value should not exceed the range for the ACR component.
[2]The delta adjustment will be applied on the $\log_e$ transformed imputed and observed values which will then be exponentiated prior to deriving endpoint.

- Step 4: Standard rounding rules will be applied to the imputed values. If the SAS® PROC MI yields a value outside of the range for the given component, the value will be updated after the imputation has been performed to be within the predefined range.
- Step 5: Repeat steps 1 to 4 for each ACR component. The composite binary endpoint (ACR50) will then be derived based on the multiple shifted imputed data sets obtained for each component.
- Step 6: Additionally, as the primary endpoint is derived using NRI, subjects randomized to BKZ with missing data (where missing values include observations after the IE date and any other missing values) should be set to non-response, after applying the delta adjustment outlined in Step 3 above. This ensures subjects randomized to BKZ do not have a higher probability of response in the tipping point analyses compared to the primary analysis (ie, a subject randomized to BKZ who is non-responder in the primary analysis cannot become a responder in the tipping point analyses).
- Step 7: In the data obtained, for each value of the imputation number, ACR50 will be analyzed using a logistic regression model with factors of treatment group, region, and prior TNFα inhibitor exposure at Baseline as fixed effects.
- Step 8: The results obtained from the 100 logistic regression analyses in Step 6 will be combined for overall inference using Rubin's rules, and the results obtained for each shift parameter will be presented in a single table.
- Step 9: Step 3 to Step 8 will be repeated so that, at each iteration, missing values are adjusted with a larger shift than at the previous iteration. Depending on the results obtained, delta parameters with more granularity (eg, 2 times, 3 times the initial delta) will be investigated. The process will go on until the p-value for the odds ratio between BKZ and placebo is no longer statistically significant (ie, ≥ 0.05). The odds ratio, 95% CI, and p-values obtained for each value of delta will be combined in one single table.

4.2.2.5 Treatment Policy Strategy

Another supportive analysis will be performed on the primary efficacy endpoint to address intercurrent events.

The treatment policy strategy will include all available data observed at the week of interest (Week 16) regardless of the occurrence of intercurrent events. This means the analysis includes on- and off-treatment values collected after subjects prematurely discontinued study treatment but agreed to remain on the study and continued to attend visits and provide assessments at those visits. Those observed values will be analyzed according to the subject's randomized treatment. Subjects for whom efficacy data cannot be obtained at the week of interest, despite attempts to

retain them in the study, will have their data imputed using MI-MCMC/monotone regression (Section 4.2.2.2).

4.2.3 Handling of missing data for AE

For analyses of AEs, a complete date must be established to correctly identify the AE as occurring during treatment or not. For purposes of imputing missing components of partially reported start and stop dates for AEs, the algorithms listed below will be followed. In the event of ambiguity or incomplete data which makes it impossible to determine whether the AE was treatment emergent, the AE will be considered treatment emergent.

Although the algorithms for treatment-emergence depend on the onset date, imputation rules are provided for stop date as well, as these may be needed for certain statistical analyses, such as an analyses of AE prevalence or AE duration.

Start and stop dates of AEs will be displayed as reported in the subject data listings (ie, no imputed values will be displayed in data listings).

Imputation of Partial AE Start Dates

- If only the month and year are specified:
  - if the month and year of first dose of IMP is not the same as the month and year of the start date, then use the 1st of the month,
  - if the month and year of first dose of IMP is the same as the month and year of the start date, then use the date of first dose of IMP,
- If only the year is specified:
  - if the year of first dose is not the same as the year of the start date, then use January 1 of the year of the start date.
  - if the year of first dose is the same as the year of the start date, then use the date of first dose.
- If the date is completely unknown:
  - use the date of first dose of IMP if the stop date is unknown or not prior to the onset date.

Imputation of Partial Stop Dates:

- if only the month and year are specified, then use the last day of the month
- if only the year is specified, then use December 31st of that year
- if the stop date is completely unknown, then do not impute the stop date

If the imputed stop date is prior to the imputed start date:

- For missing start day and start month:
  - if the year of start date is the same as the year of first dose and the imputed stop date is on or after the date of first dose then set the start date to the date of first dose
  - otherwise set to the 1st January of the year of the start date

- For missing start day only
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is on or after the date of first dose, then set the start date to the date of first dose
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is before the date of first dose, then set the start date to the 1st of that month

In the event of ambiguity or incomplete data which makes it impossible to determine whether an AE is treatment emergent, the AE will be considered as treatment emergent.

Other imputations

In addition, the following will apply for presenting AE in summary tables:

- if the intensity of an AE is unknown, it will be considered as severe.
- if the relationship to study drug is missing, it will be considered as related.

For missing seriousness, no imputation rule will be applied.

4.2.4 Handling of missing data for prior and concomitant medications

Any medications with incomplete start and end dates/times will be handled according to the following rules for classification as prior and concomitant and for the calculation of relative study days. In the event of ambiguity or incomplete data which makes it impossible to determine whether a medication was concomitant or not, the medication will be considered as concomitant. Such imputations will only be performed for these classifications and calculations; in the listings all data will be shown as recorded on the electronic case report form (eCRF).

Imputation of Partial Start Dates

- If only the month and year are specified:
  - if the month and year of first dose of IMP is not the same as the month and year of the start date, then use the 1st of the month.
  - if the month and year of first dose of IMP is the same as the month and year of the start, then use the date of first dose of IMP.
- If only the year is specified:
  - if the year of first dose of IMP is not the same as the year of the start date, then use the 1st of January of the year of the start date.
  - if the year of first dose of IMP is the same as the year of the start date, then use the date of first dose of IMP.
- If the start date is completely unknown:
  - if the stop date is unknown or not prior to the date of first dose of IMP, then use the date of first dose of IMP.
  - if the stop date is prior to the date of first dose of IMP, then set the start date to the 1st of January of the year of the end date.

Imputation of Partial Stop Dates:

- If only the month and year are specified, then use the last day of the month,
- If only the year is specified, then use December 31st of that year.
- If the stop date is completely unknown, do not impute the stop date. There will be no imputation of any other missing data.

If the imputed stop date is prior to the imputed start date:

- For missing start day and start month:
  - if the year of start date is the same as the year of first dose and the imputed stop date is on or after the date of first dose, then set the start date to the date of first dose
  - otherwise set to the 1st January of the year of the start date
- For missing start day only
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is on or after the date of first dose, then set the start date to the date of first dose
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is before the date of first dose, then set the start date to the 1st of that month

In the event of ambiguity or incomplete data which makes it impossible to determine whether a medication was concomitant, the medication will be considered as concomitant.

4.3 Interim analyses and data monitoring

4.3.1 Interim analysis

One interim analysis will be performed after all randomized subjects have completed the Double-Blind Treatment Period at Week 16.

The purpose of the Week 16 interim analysis is to perform a comprehensive evaluation of all available double-blind data for the BKZ and placebo treatment arms to prepare a regulatory submission for a Marketing Authorization Application based on this analysis.

The interim analysis will evaluate the primary and secondary efficacy variables as well as all other efficacy, safety and PK variable results up to Week 16. In addition, disposition, demographics and other Baseline characteristics, protocol deviations and compliance will be described. All available data at the time of the snapshot will be included (ie, even SFU data available beyond the Week 16 timepoint). Events on-going at the time of the interim analysis (eg, concomitant medications, adverse events) will be analyzed as they are collected to date, and analyses based on these events will be up to date at time of the final analysis.

For the interim analysis, the database will be locked, and the treatment codes will be made available to UCB personnel. An interim report will be written. The investigators and subjects will remain blinded to the assigned BKZ dosing regimen until the final analysis is completed. A blinding maintenance plan will be written to evaluate the potential bias in data reporting and

describe the process of interim results generation and dissemination. The plan will be finalized prior to the lock of the database at the interim analysis.

No formal alterations to the further study conduct (eg, stopping rules, sample size re-estimation, or changes to eligibility criteria) are planned for the interim analysis.

No separate SAP for the interim analysis will be provided. The TFL shells for the interim and final analyses will be provided in the same document. The final analysis will consist of a re-run of all analyses, including new data occurring during the SFU Period that were not available for the interim Week 16 analysis.

4.3.2 Data Monitoring Committee

An independent Data Monitoring Committee (DMC) will be formed to monitor the ongoing safety of the study through periodic review. The general scope of DMC activities is presented in the protocol, and the composition and operation of the DMC will be defined in the DMC charter. The presentation and analysis of data for the DMC meetings will be described within a separate DMC SAP.

4.4 Multicenter studies

The data from all centers will be pooled for the purposes of the analysis. However, the effect of center (using a pooling of centers by region) on results will be evaluated as mentioned in Section 3.8.

4.5 Multiple comparisons/multiplicity

A fixed sequence testing procedure will be applied for the primary endpoint and the key secondary endpoints. The testing procedure will account for multiplicity and control the family-wise type I error rate at alpha=0.05 (2-sided).

For each test, on each binary efficacy endpoint, the null hypothesis is that the conditional odds ratio is equal to one (H_0: $OR_{T1T2} = 1$). The alternative hypothesis is that the conditional odds ratio is not equal to one (H_A: $OR_{T1T2} \neq 1$).

For each test, on each continuous efficacy endpoint, the null hypothesis is that there is no difference between treatment groups (H_0: $T_1 - T_2 = 0$). The alternative hypothesis is that there is a difference between treatment groups (H_A: $T1 - T2 \neq 0$).

T_1 refers to BKZ and T_2 to placebo.

According to this strategy, the statistical testing of an endpoint can be investigated only if the null hypothesis for the previous endpoint has been rejected (ie, if $p<0.05$). Table 7–4 shows the testing order for these endpoints.

Table 7–4: Sequential testing procedure of primary/secondary efficacy endpoints (fixed testing procedure procedure) (All efficacy endpoints at Week 16)

BKZ 160mg Q4W	
H_1	ACR50 Response superior to Placebo
H_2	Change from Baseline HAQ-DI superior to Placebo

H_3	PASI90 Response superior to Placebo in subjects with PSO BSA ≥3% at Baseline
H_4	Change from Baseline SF-36 PCS superior to Placebo
H_5	MDA superior to Placebo

ACR50=American College of Rheumatology 50% response crtiteria; BKZ=bimekizumab; H=hypothesis; HAQ-DI=Health Assessment Questionnaire - Disability Index; MDA=Minimal Disease Activity; PASI90=Psoriasis Area and Severity Index 90; PCS=Physical Component Summary; Q4W=every 4 weeks; SF-36=Short-Form 36-item Health Survey; W=week.

4.6 Use of an efficacy subset of subjects

The RS will be the primary analysis set for efficacy analyses, but analyses will also be repeated on the FAS and the PPS for the primary efficacy endpoint. The FAS analysis will evaluate whether there are differences in the efficacy analysis between randomized subjects and randomized subjects with a Baseline assessment; while the PPS analysis will evaluate the effect of IPD on the analysis.

4.7 Active-control studies intended to show equivalence

Efficacy outcomes are based on superiority comparisons of the BKZ 160mg sc Q4W to placebo, this section is then not applicable for this study.

4.8 Examination of subgroups

Subgroup analyses will be performed on the variables below. They are all assessed at Baseline, except concomitant cDMARDs, concomitant Methotrexate (MTX) and ADAb status which will be assessed during the 16-week period. Subgroup analyses will be performed on the ACR50 response, the PASI90 response and the HAQ-DI response (subjects with a decrease of HAQ-DI from Baseline of at least 0.35) at Week 16. The ADAb status will also be used for subgroup analysis for the PK endpoints.

The variables for subgroup analyses will be:

- Age (<45 years of age, ≥45 years of age),
- Gender (male, female),
- Disease duration (<1 year, ≥1 year),
- Region (eg, North America, Western Europe, Eastern Europe, Asia. Western Europe will be combined with Asia if any of these 2 regions have less than 10% of the randomized subjects),
- Race (White, Black, and Other),
- Body weight at Baseline (≤100kg, >100kg),
- Prior TNFα exposure (intolerance to TNFα inhibitor, inadequate response to at least 1 TNFα inhibitor, inadequate response to 2 prior or more TNFα inhibitors),
- hs-CRP at Baseline (<6mg/L, ≥6mg/L),
- Prior conventional disease-modifying antirheumatic drugs (cDMARDs) (0, 1, ≥2) (taken prior to Baseline). Such medications will be classified using an adjudicated spreadsheet.
- Concomitantly receiving cDMARDs versus no concomitant cDMARDs,
- Concomitantly receiving MTX versus no concomitant MTX,

- Concomitantly receiving MTX at Baseline vs. other cDMARDs at Baseline (MTX at Baseline, no MTX at Baseline and cDMARDs at Baseline, no MTX at Baseline and no cDMARDs at Baseline),
- PSO affected BSA at Baseline (<3%, ≥3% to 10%, >10%),
- BASDAI at Baseline (<4, ≥4),
- ADAb status (positive, negative) (See Section 9.2) for the BKZ 160mg Q4W group only,
- Human leukocyte antigen B27 (HLA-B27) (positive, negative).

Note: In the context of submission for Japanese health authorities, a subset of tables, listings and figures will be provided for the subgroup of subjects randomized in Japan.

5 STUDY POPULATION CHARACTERISTICS

5.1 Subject disposition

Subject disposition will be summarized for all subjects screened overall, by region and by site. In this summary, the dates of first subject screened and date of last subject last visit (end of study), number of screened subjects, and the number of subjects included in each analysis set (RS, SS, FAS, PPS, PK-PPS, COVID-19-free Set) will be presented.

The disposition of subjects into treatments groups and analysis sets will also be summarized on the RS.

Reasons for screen failures (as collected on the Study Termination Screen Failure CRF page) will be summarized for all screened subjects who failed to be randomized.

The number and percentage of randomized subjects who completed the study, completed the study not on randomized treatment, who discontinued IMP, overall and by reason, and who discontinued the study, with the primary reason for study discontinuation (as collected on the Study Termination CRF page), will be tabulated overall and by treatment group. The numbers and percentage of randomized subjects entering the SFU Period will additionally be presented. The numbers and percentages of these subjects who either complete the SFU visit or not will additionally be presented. The numbers and percentages of randomized subjects entering the Open Label Extension (OLE) study and those not entering the OLE study will also be presented.

A subject will be said to have completed the study if she/he had completed the last scheduled study visit, not including SFU visits. (The SFU visit is typically 140 days post last dose of IMP, which is roughly equivalent to 5 half-lives of BKZ).For subjects who previously discontinued IMP, the primary reason for discontinuation as collected on the Study Medication termination CRF page will be summarized in a separate table.

Finally, the number of randomized subjects who discontinued the study due to AEs by type of AEs (serious fatal, non-fatal, other) will be summarized.

Study disposition and termination details will be listed for all subject screened. Additional listings will be created on study discontinuation, subject analysis sets and visit dates (both actual visit and visit remapped for analysis). A listing will also be provided for subjects who did not meet the study eligibility criteria.

To assess subject disposition (entry and periods in the study) during the COVID-19 pandemic, subject disposition will be assessed by period of the COVID-19 pandemic (pre – during – post) (Section 3.6.7), by comparing the dates of visits (or events) to the dates of the COVID-19 pandemic period.

5.2 Impact of COVID-19

A listing of visits affected by COVID-19 will be presented based on the ES including the visit, date of visit, relationship to COVID-19, impact category, and a narrative (short description) of the event. These data will be summarized for non-randomized subjects and by treatment group and overall, for enrolled subjects.

A summary of study visits by COVID-19 pandemic period (pre – during – post) will be presented for subjects enrolled prior, during and after the pandemic.

In addition, in order to assess the potential impact of COVID-19 on the collection and reporting of efficacy data, a separate summary on the RS will be presented to display missing data as well as data collected via an alternative modality (eg, phone, video call) for efficacy endpoints included in the hierarchy (Table 7–4). For these displays, missing data will be presented only for visits affected by COVID-19, as reported on the dedicated eCRF page. Missing data at other visits and for other reasons will not be included.

For visits conducted remotely, it was not possible to assess some endpoints (eg, TJC, SJC, LDI, LEI, PASI, etc.) and therefore these assessments will be missing for those visits. In addition, for any missed visit or visit conducted remotely, the CRP assessment will also be missing. Such assessments will be considered missing as a result of the COVID-19 pandemic. For these visits, it will therefore not be possible to assess composite endpoints like ACR response. For this summary table, based on the RS, only visits at which efficacy assessments are scheduled will be included.

5.3 Protocol deviations

The definition of an IPD is given in Section 3.5.

A summary displaying the number and percentage of subjects with an important protocol deviation will be provided for the RS by treatment group. This will include a summary of subjects excluded from the PPS and PK-PPS. The summary will be provided overall and by type of deviation (inclusion criteria deviation, exclusion criteria deviation, withdrawal criteria deviation, prohibited concomitant medication, incorrect treatment or dose, treatment non-compliance, procedural non-compliance, COVID-19 related IPD) as well as the number and percentage of subjects excluded from the PPS due to reason other than PD. Criteria for exclusion of subjects from the PPS or PK-PPS will be defined in a separate document.

A listing of important protocol deviations will be provided for all subjects in the RS.

A listing of COVID-19 related PDs will be provided for all subjects in the RS.

A listing of subjects excluded from the PPS for reasons other than IPD will be provided.

6 DEMOGRAPHICS AND OTHER BASELINE CHARACTERISTICS

Unless otherwise specified, all summaries will be based on the SS and repeated using the RS (unless the SS and RS analysis sets are identical).

6.1 Demographics

The following demographic variables measured at Baseline will be summarized overall and by treatment group: age (years), gender, country, geographic region, country, race, ethnicity, height (cm), body weight (kg) and body mass index (BMI) (kg/m2).

Age, body weight and BMI will be summarized as continuous variables and as categorical variables based on the categories specified below.

For age, 3 sets of categories will be defined:

- ≤18, 19 to <65, ≥65 years (clinicaltrials.gov requirement),
- 18 to <65, 65 to <85, ≥85 years (EudraCT requirement),
- 18 to <45, ≥45 to <65, ≥65 years.

Body weight categories will be: ≤100, >100 kg.

Even if BMI is available in the database, BMI (in kg/m2) will be recalculated during analysis based on height (in m) and weight (in kg) values collected in the database, and the calculated values will be used in the statistical analysis since they are considered more accurate. The formula for BMI (kg/m2) calculation is:

$$\text{BMI} = \frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$$

BMI categories will be: <25, 25 to <30, ≥30 to <35, ≥35 kg/m2.

A separate frequency table will summarize the subject's lifestyle on the SS.

By-subject listings on demographics and female subjects' childbearing potential and birth control data will be provided for all screened subjects. Listing on lifestyle will be provided on the RS.

To understand the study population during the different periods of the COVID-19 pandemic, the reporting of Baseline demographics will be repeated by COVID-19 pandemic period (pre - during as determined by the Baseline visit date) on the SS.

6.2 Baseline characteristics

6.2.1 Baseline values for primary efficacy variable components

Baseline values for the 7 components for primary efficacy variable will be summarized by treatment group and overall. The following variables will be summarized:

- TJC,
- SJC,
- PGA-PsA,
- PhGA-PsA,

- PtAAP,
- HAQ-DI,
- hs-CRP (mg/L).

For hs-CRP, SJC, and TJC the following rule will be applied: If Baseline value is missing, then take the most recent value (taken at previous unscheduled visit or at Screening); If value is still missing, the Baseline value will be considered missing. hs-CRP will be summarized using descriptive statistics and in classes (≥6mg/L, <6mg/L).

Efficacy variables not listed above will not be presented in the Baseline characteristics section. They will be presented in by-visit tables in the context of the analysis of secondary and other efficacy variables.

To understand the study population during the different periods of the COVID-19 pandemic, the reporting of Baseline demographics will be repeated by COVID-19 pandemic period (pre-, during, and post- as determined by the Baseline visit date) on the SS.

6.2.2 Other baseline characteristics

The following variables will be summarized by treatment group and overall:

- Prior TNFα exposure (Intolerance to TNFα inhibitor, Inadequate response to 1 TNFα inhibitor, Inadequate response to 2 TNFα inhibitors) – IXRS randomization strata
- Prior TNFα exposure (Intolerance to TNFα inhibitor, Inadequate response to at least 1 TNFα inhibitor, Inadequate response to 2 prior or more TNFα inhibitors) – actual randomization strata
- % of BSA affected by PSO (<3%, ≥3% to ≤10%, >10%)
- PASI score at Baseline (for subjects with PSO involving at least 3% of BSA at Baseline). This variable will be summarized as a continuous variable and in classes (<10, 10 to 20,>20)
- BASDAI (<4, ≥4)
- Rheumatoid factor (positive, negative). A rheumatoid factor value < 30 IU/mL is defined as negative and a value ≥ 30 IU/mL is defined as positive. (Chemistry results / Proteins / Test=RF).
- Anti-cyclic citrullinated peptide (CCP) antibodies (Positive, Negative). Anti-CCP antibodies negative is a value < 20 U/mL and a value ≥ 20 U/mL is defined as positive.
- Nail psoriasis (yes, no)
- Dactylitis ("yes" if LDI score>0, "no" if LDI score=0 or missing)
- Enthesitis based on LEI ("yes" if LEI score>0, "no" if LEI score=0 or missing)
- Enthesitis based on SPARCC ("yes" if SPARCC>0, "no" if SPARCC=0 or missing)
- Prior NSAID (yes, no)
- NSAID at Baseline (yes, no)
- Past cDMARDs therapy (yes, no)

- Prior cDMARDs (0, 1, ≥2) (see Section 4.8)
- cDMARDs at Baseline (yes, no)
- HLA-B27 status (positive, negative)
- Methotrexate at Baseline (yes, no)
- Prior oral corticosteroids (yes, no)
- Oral corticosteroids at Baseline (yes, no)

The above data including extra-articular assessments will also be listed.

Efficacy variables not listed above will not be presented in the Baseline characteristics section. They will be presented in by-visit tables in the context of the analysis of secondary and other efficacy variables.

To understand the study population during the different periods of the COVID-19 pandemic, the reporting of Baseline demographics will be repeated by COVID-19 period (pre-, during, and post- as determined by the Baseline visit date) on the SS.

6.2.3 History of psoriatic arthritis and psoriasis

The following PsA history variables will be summarized overall and by treatment group:

- Time since first diagnosis of PsA (years). Time since first diagnosis will be summarized as a continuous variable and as categorical variables based on following categories: <1, 1 to <2, ≥2 years.
- Time since first diagnosis of PSO (years),

  Time since first diagnosis of PsA/PSO (years) will be calculated as follows:

$$\frac{Date\ of\ informed\ consent - Date\ of\ diagnosis}{365.25}$$

- Age at first diagnosis of PsA (years),

  Age at first diagnosis (years) of PsA will be calculated as:

$$\frac{Date\ of\ diagnosis\ of\ PsA - Date\ of\ birth}{365.25}$$

If the date of diagnosis is partial, it should be imputed to the most recent feasible date (ie, last day of the month if only day is missing, or the last day of the year if day and month are missing). Partial dates should not be imputed later than the subject's screening date. For subjects enrolled at German sites, only the year of birth may be entered into the eCRF for this study. For these subjects age will be calculated after imputing their date of birth to be on January 1st.

- PsA subtype (Polyarticular – symmetric arthritis, Oligoarticular – asymmetric arthritis, Distal interphalangeal joint predominant, Spondylitis predominant, Arthritis mutilans)

The data on history of psoriasis and psoriatic arthritis will be listed for all screened subjects.

6.3 Medical history and concomitant diseases

Medical history conditions are defined as conditions that have resolved prior to study entry.

Ongoing medical conditions are defined as conditions that are ongoing at the time of study entry.

Medical history and on-going medical conditions will be summarized on the SS by MedDRA System Organ Class (SOC) and Preferred Term (PT), by treatment and overall including the number and percentage of subjects with each condition. Psoriatic arthritis history will not be included in this table as it will be presented separately (Section 6.2.3). The denominator of the percentages will be the number of subjects in the population considered. The table summaries will be ordered alphabetically for SOC and in terms of decreasing frequency for PT within SOC in the BKZ treatment group. In the event of ties, PT will be ordered alphabetically.

Medical history and ongoing medical conditions will be listed by treatment and subject including the reported term, PT, and SOC for the RS. The start date (month and year only) and the end date (or ongoing if applicable) will be included in the listing. Partial dates will not be imputed. A glossary of all medical history conditions will be presented including the reported term, PT and SOC. Concomitant medical procedures will be listed on the RS.

6.4 Prior and concomitant medications

Prior medications include any medications that started prior to the start date of the first IMP administration.

Concomitant medications are medications taken at least one day in common with the study medication dosing period.

The IMP dosing period corresponding to the combined Double-Blind Treatment Period will be calculated as follows:

- The study medication dosing period start date is defined as the date of first dose.

The study medication dosing period stop date is defined as follows:

- For subjects who are ongoing at the time of the clinical data cut (not including subjects who are in SFU) use date of last clinical contact
- For subjects who died prior to last visit (not including those who died during SFU), use the minimum of the following:
  - Date of death
  - Date of last dose of any study medication + 28 days
- For subjects who complete the study as planned, the dosing period ends at the later of the following two dates:
  - Date of last administration of IMP + 28 days
  - The last scheduled visit date not including SFU
- For all other subjects, use the maximum of the following:
  - Date of last dose of any study medication + 28 days
  - Date of last visit (not including SFU)

In the case of missing data, the classification of medications as prior or concomitant will be performed as described in Section 4.2.4. Imputations of missing data will be performed before calculation of relative study days.

The number and percentage of subjects taking prior medications and prior Anti-TNFs (excluding past psoriasis medications) will be summarized on the SS by treatment group, overall and by ATC class, presenting Anatomical Main Group (ATC Level 1), Therapeutical/Pharmacological Subgroup (ATC level 3), and preferred term (PT).

The table summary will be ordered alphabetically for the ATC class and in terms of decreasing frequency for PT within ATC class in the BKZ treatment group. In the event of ties, PT will be ordered alphabetically.

A by-subject listing of all prior medications and prior Anti-TNFs and concomitant medications on the SS as well as glossary of all medical medications will be provided.

6.5 Prohibited concomitant medications

Prohibited or restricted medications are defined in the protocol section 7.8.2.

Prohibited medications will be listed on SS.

7 MEASUREMENTS OF TREATMENT COMPLIANCE

Treatment compliance summaries will be performed based on the SS, by randomized treatment and for all subjects.

Due to the method of administration of the treatments, compliance will be examined in terms of completed injections. Treatment compliance will be calculated as:

$$\frac{\text{Total number of actual (completed) injections}}{\text{Total number of expected injections}} \times 100\%$$

The total number of expected injections is derived relative to when the subject finishes treatment.

If a subject completes treatment, 4 injections are expected (at Baseline, Week 4, Week 8 and Week 12).

If a subject discontinues early, then the number of expected injections is based on the time of early discontinuation relative to the dosing visits. This number will be calculated by determining the number of expected dosing visits prior to the date of early discontinuation. If the early discontinuation coincides with an expected dosing visit, this visit will used to determine the number of expected injections.

A summary of percent treatment compliance categorized as <75% and ≥75% will be provided by treatment group, as well as a by-subject listing of treatment compliance.

8 EFFICACY ANALYSES

8.1 Statistical analysis of the primary efficacy variable

8.1.1 Derivation of ACR50 response

The ACR50 response at Week 16 is the primary efficacy variable.

This is a composite endpoint and represents at least 50% improvement from Baseline at Week 16 for each of the following:

- TJC (based on 68 joint count)
- SJC (based on 66 joint count)
- At least 3 of the 5 remaining core set measures:
  - PGA-PsA
  - PhGA-PsA
  - PtAAP
  - HAQ-DI
  - hs-CRP

Prior to deriving ACR50, the percent of change for each component will be calculated as:

$$\text{Percent improvement from Baseline} = 100 \times \frac{\text{Baseline Value} - \text{Week 16 value}}{\text{Baseline Value}}$$

As the ACR response is based on 7 different component scores, it is necessary to consider various data scenarios that could impact the calculation of ACR response. The rules described here are applicable in the context of the calculation of ACR response and may differ from the rules applied for calculating and summarizing the components, individually (some values may need to be imputed for component for component analysis but are not required here to evaluate the ACR response).

The following rules will be applied prior to invoking any imputation analysis at the variable level:

- If a subject has a component value that is equal to 0 at Baseline and the post-Baseline value is greater than or equal to 0, then the percentage of improvement for that component will be treated as 0 for purposes of ACR response calculations.
- If a subject has a component value that is missing at Baseline, then the percentage of improvement for that component will be treated as missing for purpose of ACR response calculations.

Observed data will be used to calculate ACR response where possible. In case of partial missing data where an observed response may be calculated, imputed data will not change the result.

To address possible missing data in ACR components, ACR50 will be derived as below:

- Positive response on ACR50
  - Improvement at Week 16 of at least 50% from baseline on SJC and TJC and at least 3 out of the 5 remaining ACR components (regardless of whether the two remaining components are missing or not)
- Negative response on ACR50

  - No improvement at Week 16 of at least 50% from baseline on SJC (regardless of whether the remaining components are missing or not)
  - No improvement at Week 16 of at least 50% from baseline on TJC (regardless of whether the remaining components are missing or not)
  - No improvement for at least 3 out of the 5 ACR components (PGA-PsA, PhGA-PsA, PtAAP, HAQ-DI, hs-CRP) regardless of whether the remaining components are missing or not

- Missing ACR50 response
  - For all other situations that are not specified above.

For the OC analysis: if a given visit falls after the treatment discontinuation date, the endpoint at this visit and all subsequent visits (whether the data were observed or not) will be derived but reported as missing (Section 8.1.4.6).

For the MI analysis: if a given visit falls after the treatment discontinuation date, the endpoint at this visit and all subsequent visits (whether the data were observed or not) will be imputed. Responses falling after an IE (defined as treatment discontinuation due to AE or lack of efficacy for the modified composite estimand) will be set to nonresponse (Section 8.1.4.3).

8.1.1.1 Derivation of tender joint count (TJC) and swollen joint count (SJC)

Tender and swollen joints are assessed based on the 68 and 66 joints for tenderness and swelling, respectively. The 68/66 joints that are assessed are:

- Upper body (6) - bilateral temporomandibular, sternoclavicular, and acromioclavicular joints.
- Upper extremity (34) - bilateral shoulders, elbows, wrists (includes radiocarpal, carpal and carpometacarpal bones considered as a single unit), metacarpophalangeal (I, II, III, IV, and V), thumb interphalangeals, proximal interphalangeal (II, III, IV, and V), and distal interphalangeals (II, III, IV, and V).
- Lower extremity (28) - bilateral hips, knees, ankles, tarsi (includes subtalar, transverse tarsal, and tarsometatarsal considered as a single unit), metatarsophalangeals (I, II, III, IV, and V), great toe interphalangeals, and proximal interphalangeals (II, III, IV, and V).

The assessment on swelling exclude the hips.

Each of the joints can be graded as follows by the assessor:

- Permanently not assessable
- Temporarily not assessable
- Asymptomatic
- Tender only
- Swollen only
- Tender and swollen

Permanently non-evaluable joints will be considered missing for both tender and swollen joint counts at the visit at which the grading was recorded and all subsequent visits. The asymptomatic joints will be included in the joint count analysis with scores of 0.

For the statistical analysis, swelling and tenderness will be each graded on a 2-point scale as described below in Table 8–1.

Table 8–1: Swelling and Tenderness grading

Present	Tenderness response (68)	Swelling response (66)
No (0)	Not tender	None
Yes (1)	Positive response to questioning (tender), spontaneous response elicited (tender and winced) or withdrawal by subject on examination (tender, winced, and withdrew)	Detectable synovial thickening with or without loss of bony contours, or bulging synovial proliferation with or without cystic characteristics

The joint graded by the assessor will then have its grade mapped as follows:

- The "Tender only" joints will correspond to a tenderness response of 1, and a swelling response of 0.
- The "Swollen only" joints will correspond to a swelling response of 1, and a tenderness response of 0.
- The "Tender and swollen" joints will correspond to a swelling and tenderness response of 1.

Dactylitic digits will be identified as follows and the joints will be counted as follows:

- Fingers 2-5: Swelling Distal Interphalangeal Joint x (side), Swelling Proximal Interphalangeal Joint x (side), Swelling Metacarpophalangeal Joint x (side) – if they are in any of the gradings 'swollen only' or 'tender and swelling' or 'injected' then only 1 finger will be added to the swollen joint count
- Toes 1-5: Swelling Interphalangeal Joint x (side), Swelling Metatarsophalangeal Joint x (side) – if they were in any of gradings 'swollen' or 'tender and swollen' or 'injected' then only 1 toe will be added to the swollen joint count

The tender joint count (TJC) and swollen joint count (SJC) are weighted joint counts. If there are missing observations in the tender or swollen joint assessments (TJ and SJ, respectively), then the remaining observations will be assessed and weighted by the number of the assessed joints (AJ) as shown below:

$$TJC = n \times \frac{\sum_{i=1}^{n} TJ}{\sum_{i=1}^{n} AJ}$$

$$SJC = n \times \frac{\sum_{i=1}^{n} SJ}{\sum_{i=1}^{n} AJ}$$

Where n represents the number of total joints.

However, if a joint is missing at Baseline, then that joint will be set to missing throughout the study. If more than 50% of the planned tender joint assessments (ie, more than 34) or 50% of the

planned swollen joint assessments (ie, more than 33) are missing at any post-Baseline visit, then no imputation will be done and the total TJC or SJC will be set to missing for that visit.

8.1.1.2 Patient and physician reported clinical outcomes on psoriatic arthritis

The physician's global assessment of psoriatic arthritis (PhGA-PsA), the patient's global assessment of psoriatic arthritis (PGA-PsA) and the patient's assessment of arthritis pain (PtAAP) are clinical outcomes based on visual analog scales (VAS) ranged from 0 to 100 (in mm).

- PhGA-PsA: The investigator will assess the overall status of the subject with respect to their PsA signs and symptoms using a numerical rating scale where 0 is "very good, asymptomatic and no limitation of normal activities" and 100 is "very poor, very severe symptoms which are intolerable and inability to carry out all normal activities". This assessment will be based through the following question "Considering all the ways the disease affects your patient, mark a vertical line on the scale for how well his or her condition is today."
- PGA-PsA: Subjects will assess the impact of PsA in answering the following question "Considering all the ways your psoriatic arthritis affects you, please mark a vertical line on the scale below to show how well you are doing today." The subject should be asked to consider all aspects of their disease (including joint and skin components) in their response to this question. Subjects will score PGA-PsA in using a VAS where 0 is "very good, no symptoms" and 100 is "very poor, severe symptoms."
- PtAAP: Subjects will assess their arthritis pain through the PtAAP using a VAS where 0 is "no pain" and 100 is "most severe pain."

8.1.1.3 Derivation of the health assessment questionnaire disability index score (HAQ-DI)

The HAQ-DI contains 20 items that measure the degree of difficulty experienced in the following 8 categories of the daily living activities: dressing and grooming (2 items), arising (2 items), eating (3 items), walking (2 items), hygiene (3 items), reach (2 items), grip (3 items), and common daily activities (3 items).

For each question/item, subjects should indicate the level of difficulty (from 0 to 3) in the past week as follows:

- 0: 'without any difficulty'
- 1: 'with some difficulty'
- 2: 'with much difficulty'
- 3: 'unable to do'

Each category is given a score by taking the maximum value for each question.

For each question related to a specific daily living activity, subjects can specify whether an aid or devices is usually used.

If a category score equals 0 or 1, but a device related to that category is used, or help from another person is provided for that category, then the category score is increased to 2.

If a category score already equals 2, and a device related to that category is used, or help from another person is provided for that category, the score for that category remains 2.

Table 8–2 details how each aid and device is associated with the category scores. Aid or device considered as 'other' than the ones listed will not be considered in the analysis.

Table 8–2: Aid and device associated with HAQ-DI category

Aid or device	Will be associated with category score
[redacted]	Walking
	Dressing and grooming
	Eating
	Arising
	Hygiene
	Hygiene
	Grip
	Hygiene
	Reach
	Hygiene

If all questions within a given category are unanswered, no score will be provided for that category (this rule applies even if aids and devices are non-missing).

The HAQ-DI score will be calculated by dividing the sum of the highest category scores (0 to 24) by the number of categories with at least 1 question answered. If fewer than 6 categories have responses, no HAQ-DI score will be calculated (ie, HAQ-DI score will be considered missing).

The HAQ-DI score ranges from 0 to 3. A lower HAQ-DI score indicates an improvement in function.

8.1.1.4 High-sensitivity C-Reactive Protein (hs-CRP)

In the summary tables presenting hs-CRP values, the observed values and ratio to Baseline values will be displayed.

hs-CRP values which are below the lower limit of quantification (LLOQ) should be set to the midpoint between 0 and the LLOQ prior to the analysis. Listing will show values below the LLOQ. The LLOQ for hs-CRP is 0.10mg/L.

8.1.2 Primary analysis of the primary efficacy variable

The primary efficacy variable will be analyzed for all subjects in the RS.

The primary endpoint is the ACR50 response at Week 16. The primary efficacy analysis will evaluate the composite estimand in the RS. The composite estimand combines the clinically meaningful improvement from Baseline in ACR components as assessed using the ACR50 response at Week 16 and not discontinuing study treatment prior to Week 16.

The following 4 attributes describe the estimand that will be used to define the treatment effect of interest for the primary efficacy analysis:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject-level outcome = ACR50 at Week 16
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to Week 16. A composite strategy will be implemented in which a positive clinical outcome is defined as achieving ACR50 at Week 16 and not discontinuing study treatment through Week 16.
- Population-level summary measure = Conditional odds ratio comparing BKZ to placebo and difference of response rate between BKZ and placebo.

Missing data at Week 16 that are not preceded by an IE and any data after an IE will be imputed as non-responders. This results in a more traditional NRI approach and will be labeled as such throughout the SAP.

The statistical hypothesis for the ACR50 response at Week 16 is that the conditional OR for ACR50 response in the BKZ treatment compared to placebo treatment is equal to 1.

A logistic regression model will be used to assess the treatment effect on ACR50 response at Week 16. The model will include fixed effects for treatment, and prior TNFα inhibitor exposure and region as stratification factors. The suitability of including these variables in the model will be assessed using Pearson and Deviance and the Hosmer-Lemeshow Goodness-of-Fit Tests (Hosmer and Lemeshow, 2000).

P-values below 0.05 would lead to a reconsideration of the model to be used and if the logistic regression model is unable to converge the stratification variables will be dropped (Section 4.1). In that case, all supportive analyses as well as imputation models conducted for the primary endpoint will disregard these exploratory factors from their models.

The country-specific analyses performed on subjects randomized in Japan will not consider the region factor as a covariate for the modelling. Considering that the number of subjects randomized in Japan is low (less of 10% of the RS), statistical models might not converge. If a model (logistic Model or mixed model) is not converging, all related adjusted statistics and p-value will not be presented: "NE" for "Not Evaluable" will be displayed instead.

The SAS® PROC LOGISTIC will be used to run the logistic regression.

The summary table results will present the adjusted responder rates and the associated 95% CI for BKZ and placebo, the adjusted odds ratio and 95% CI for the comparison of BKZ versus placebo, and the p-value that the OR=1 and the difference of response rate between BKZ and placebo and associated 95% CI. The treatment comparison will be made using the 2-sided Wald test at a significance level of α=0.05 (ie, H_1 in Table 7–4).

8.1.3 Subgroup analyses of the primary efficacy variable

Subgroup analyses will be performed on the primary efficacy variable on the RS.

The variables for subgroup analyses are defined in Section 4.8.

For each subgroup analysis (except for the analysis by prior TNFα exposure, by region, and by ADAb status), a logistic regression will be fitted involving the same terms that were retained when running the primary analysis model, plus a term for the subgroup and the subgroup by treatment interaction as detailed below:

- Fixed effect for treatment
- Prior TNFα 1 (1 for inadequate response to at least 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior or more TNFα inhibitors)
- Region
- Term for subgroup
- Terms for subgroup by treatment interaction(s).

For subgroup analyses by Prior TNFα exposure, the terms that will be retained will be:

- Fixed effect for treatment
- Prior TNFα 1 (1 for inadequate response to 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior TNFα inhibitors)
- Region
- Terms for Prior TNFα inhibitor exposure by treatment interaction

For subgroup analyses by region, the terms that will be retained will be:

- Fixed effect for treatment
- Prior TNFα 1 (1 for inadequate response to 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior TNFα inhibitors)
- Region
- Terms for region by treatment interaction

For subgroup analyses by ADAb status, the terms that will be retained will be:

- prior TNFα exposure at Baseline
- Region
- ADAb Status (Positive, Negative)

For all subgroup analyses, the terms for the subgroups will coded as below:

- For Yes/No or positive /negative variables, the category “yes” or “positive” will be coded “1”.
- For continuous variables in 2 classes, the highest category will be coded “1”.
- For continuous variables in 3 classes (or more), 2 (or more) binary variables will be created and coded “1” for the 2 (or more) top categories.

- For gender, “male” will be the category coded “1”.

The same estimand approach as the one used for the primary analysis will be used to handle missing data (NRI).

For each subgroup category and each treatment group (subgroup analysis on the ADAb status for only BKZ group), the mean proportion of responders on imputed datasets, the adjusted responder rates with associated 95% CI will be provided. The adjusted OR (for the comparison BKZ and placebo) and the corresponding 95% CI will also be provided. In addition, the difference in response rates between BKZ and placebo and the corresponding 95% CI will also be presented (except for the subgroup analysis on the ADAb status). The results obtained for each subgroup will be presented in one single table. The ORs and associated 95% CIs for each subgroup category will also be displayed on a single forest-plot.

The observed rates of responses by subgroup categories will be also provided along with the results obtained on imputed data.

In addition, the mean proportion of ACR50 responders at Week 16 separated by the 6 pre-defined ADAb subcategories (as defined in Section 9.2) will be presented graphically for the BKZ-randomized subjects.

8.1.4 Supportive analyses of the primary efficacy variable

Unless otherwise specified, the following supportive analyses planned for the primary efficacy endpoint will be conducted on the RS (except analyses described in Sections 8.1.4.1, 8.1.4.2 and 8.1.4.8). For the below analyses, unless specified:

- the same model as the one specified for the primary analysis will be used
- the same statistics will be provided: the mean proportion of responders (on imputed datasets when applicable) the adjusted responder rates for each treatment group, the adjusted ORs for the comparison BKZ versus placebo and its 95% CI, the p-value for the comparison between BKZ and placebo and the difference of response rates between BKZ and placebo and its corresponding 95% CI

The adjusted OR and 95% CI for the comparison BKZ versus placebo obtained from the primary analysis and from the supportive analyses described below (except for Section 8.1.4.4) will be displayed in a forest-plot.

8.1.4.1 Analysis on the PPS

The primary analysis described in Section 8.1.2 will be repeated based on the PPS to evaluate the effect of IPD on the analysis.

8.1.4.2 Analysis on the FAS

The primary analysis described in Section 8.1.2 will be repeated for all subjects in the FAS to evaluate the consistency between the RS and the more restrictive FAS. This analysis will only be produced if the number of subjects in the RS and FAS differs.

8.1.4.3 Analysis using modified composite estimand where intercurrent events are defined as discontinuation due to AE or Lack of efficacy

The modified composite estimand for this analysis is detailed below

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP.
- Subject-level outcome = ACR50 at Week 16.
- Intercurrent event handling = An intercurrent event is defined as discontinuation due to AE or lack of efficacy. A composite strategy will be implemented in which a positive clinical outcome is defined as achieving ACR50 at Week 16 and not discontinuing study treatment due to an AE or lack of efficacy through Week 16.
- Population-level summary measure = Conditional OR comparing BKZ to placebo.

Any missing data at Week 16 that is not preceded by an IE will be imputed based on a predefined MI model (MI/MCMC - Monotone Regression) (Section 4.2.2.2).

The following categories of subjects will enter the MI/MCMC – Monotone regression process for each ACR component:

- Subjects with all post-Baseline values missing.
- Subjects withdrawn from the study treatment (data collected prior to the treatment withdrawal will be retained) - Those subjects will have values imputed for missing assessments and non missing assessment falling after study treatment discontinuation (falling after or before the study withdrawal).
- Subjects with missing value at Baseline. MI will be applied for each of the components separately before deriving the response.

In the case of partial missing data, ACR will be derived as in Section 8.1.1.

If an IE occurred prior to or at week 16 (or for any other intermittent missing data at Week16), the subjects will be considered as “non-responders”.

In MI, the missing value is replaced by a set of plausible values, where each value is a Bayesian draw from the conditional distribution of the missing data given the observed data.

8.1.4.4 Analysis on individual components of the ACR

The primary comparison (BKZ vs. placebo) will also be repeated for all individual components of the ACR50 response to explore the effect of the signs and symptoms of the individual components on the composite endpoint. Since all ACR components are continuous variables (eg, change from Baseline in TJC), an analysis of covariance (ANCOVA) with treatment, region, and Prior TNFα inhibitor exposure as fixed effects and the Baseline value as covariate will be used for the analysis. Missing data imputation described in Section 4.2.2 will be successively applied for each component.

The following 4 attributes describe the hypothetical estimand that will be used to define the treatment effect of interest for each of the 7 ACR components:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject level outcome = Change from Baseline in the ACR components at Week 16

- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to Week 16. This estimand targets the treatment difference in a scenario where withdrawal from study treatment does not occur, such that outcomes for subjects without an intercurrent event are as observed, and outcomes for subjects with an intercurrent event are treated successively according to the 2 following hypothetical strategies:
  - (a) as they had completed treatment through Week 16. In that strategy, missing data or non missing data after IE (reset as missing) will be imputed using MI as though they had completed the randomized study treatment through Week 16.
  - (b) as they had completed the randomized study treatment through Week 16 but on placebo. In that strategy, missing data or non missing data after IE (reset as missing) will be imputed using Reference-Based MI, in which the MI model is based on data from the Placebo group.
- Population-level summary measure = The difference in the adjusted means between BKZ and placebo.

For the results obtained using the 2 MI methods, the SAS® PROC MIXED will be used to run the ANCOVA model. The inferential statistics obtained from each model will be displayed in a single table.

For all variables, the following statistics will be presented:

- the adjusted least-square means (LSM) and standard error (SE) by treatment group,
- for the comparison between placebo and BKZ: the difference between the LSM, the associated 95% CI for the contrasts, and the corresponding p-value.

An additional table will present the descriptive statistics for the 2 treatment groups for each ACR component variable at Week16. Those statistics will be the mean and SE, median, minimum, and maximum after MI.

8.1.4.5 Analyses using treatment policy strategy imputation for missing data

The treatment policy strategy for addressing the intercurrent events will be considered. In this analysis, all available data at Week 16 will be considered regardless of the occurrence of intercurrent events. This analysis will use the same models specified for the primary analyses where subjects are analyzed according to their randomized treatment, even if they discontinued study treatment for any reason prior to Week 16 and were no longer on the randomized study treatment when the assessment was performed at Week 16.

In the context of this analysis, in addition to the statistics provided in the context of MI, the number of observed ACR50 responses regardless of the existence of IE will be also reported.

The following 4 attributes describe the treatment policy strategy analysis is detailed below:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject-level outcome = ACR50 at Week 16
- Intercurrent event handling = Discontinuation of study medication for any reason will be handled by using the data as observed irrespective of the intercurrent event occurring.

- Population-level summary measure = Conditional odds ratio comparing 'BKZ and other medication as needed' to 'placebo and other medication as needed'.

8.1.4.6 Analysis on Observed cases

An observed case (OC) analysis will additionally be conducted, where only observed data for subjects who are still on the randomized treatment at Week 16 are included. Subjects with missing data at Week 16 or who have prematurely discontinued study treatment at or prior to Week 16 will be treated as missing and missing data will not be replaced.

The following 4 attributes describe the estimand that will be used to define the treatment effect of interest for the OC analysis:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject-level outcome = ACR50 at Week 16 while on randomized treatment
- Intercurrent event handling = A "while on randomized treatment" strategy will be implemented.
- Population-level summary measure = Conditional OR comparing BKZ to placebo.

The same analysis model as used for the primary efficacy analysis will be conducted on the RS.

8.1.4.7 Tipping point analysis

The degree of the departure from the MAR assumption to overturn conclusions from the primary analysis will be investigated in a tipping point analysis. This analysis will be performed on the monotone missing data and only if the primary analysis is significant at α=0.05. Intermittent missing data will be imputed using the MCMC method. In this analysis, it will be assumed that subjects who have missing data and are randomized to BKZ have a lower probability of response compared to subjects who have missing data and are randomized to placebo.

The analysis will be primarily conducted on the binary response endpoint itself in considering the worst-case scenario that subjects who have missing ACR50 response will be set as non-responders if they are randomized to BKZ, and as responders if they are randomized to Placebo. If the result tips, then the tipping point analysis will be reconducted on the ACR50 endpoint after having delta-adjusted each individual component of the ACR50 endpoint. For both analyses, the same model as used for the primary efficacy analysis will be conducted on the RS. If the delta parameter value for which the study conclusion under MAR is reversed is plausible, the conclusion under MAR will then be questionable.

Refer to Section 4.2.2.4 for more details on the methodology.

8.1.4.8 Analyses including COVID-19 impact

To assess the impact of the COVID-19 pandemic on the primary efficacy endpoint, 2 sensitivity analyses will be conducted using the same analysis method as the primary analysis (Section 8.1.2):

- on the COVID-19-free Set

- by COVID-19 period (pre- during -post as determined by the Week 16 date).

8.2 Statistical analysis of the secondary efficacy variables

8.2.1 Derivation of secondary efficacy variables

The primary and supportive analysis methods of secondary efficacy variables for which the derivation method is provided below are described in Section 8.2.2 and Section 8.2.3.

For the other secondary variables which are not part of the testing hierarchy, there is no priority designated between methods described as primary and supportive analysis methods.

8.2.1.1 HAQ-DI: Change from Baseline at Week 16

The derivation for HAQ-DI is described in Section 8.1.1.3 (The analysis of this variable is already covered in Section 8.1.4.4).

Change from Baseline at Week 16 in HAQ-DI is the 2nd endpoint in the sequential testing hierarchy (Table 7–4).

8.2.1.2 Psoriasis Area and Severity Index 90 (PASI90) at Week 16 (in subjects with PSO involving at least 3% of BSA at Baseline)

PASI scoring of psoriatic plaques is based on 3 criteria:

- redness (R),
- thickness (T),
- scaliness (S).

Severity is rated for each index (R, S, T) on a 0-4 scale (0 = none, 1 = slight, 2 = moderate, 3 = marked, and 4 = very marked)).

The body is divided into 4 areas comprising:

- head (h),
- upper extremities (u),
- trunk (t),
- lower extremities (l).

In each of these areas, the fraction of total surface area affected is graded on a 0-6 scale (where 0 = 0% [clear], 1 = >0% to <10%, 2 = 10% to <30%, 3 = 30% to <50%, 4 = 50% to <70%, 5 = 70% to <90%, and 6 = 90% to 100%).

The various body regions are weighted to reflect their respective proportion of BSA. The composite PASI score is then calculated by multiplying the sum of the individual-severity scores for each area by the weighted area-of-involvement score for that respective area, and then summing the 4 resulting quantities as follows:

$$PASI = (0.1 \times (R_h + T_h + S_h) \times A_h) + (0.2 \times (R_u + T_u + S_u) \times A_u) + (0.3 \times (R_t + T_t + S_t) \times A_t)$$

$$+(0.4 \times (R_l + T_l + S_l) \times A_l)$$

where

R_h, R_u, R_t, R_l = redness score of plaques on the head, upper extremities, trunk, and lower extremities, scored 0-4 respectively;

T_h, T_u, T_t, T_l = thickness score of plaques on the head, upper extremities, trunk, and lower extremities, scored 0-4 respectively;

S_h, S_u, S_t, S_l = scaliness score of plaques on the head, upper extremities, trunk, and lower extremities, scored 0-4 respectively;

A_h, A_u, A_t, A_l = numerical value translation of % area of psoriatic involvement score for the head, upper extremities, trunk, and lower extremities, scored from 0-6 respectively.

The highest potential PASI score is 72 for severe disease; the lowest is 0 for no psoriasis lesions. PASI scores are treated as continuous If a subject is missing 1 or 2 severity measurements for a certain region, the average of the remaining non-missing severity measurement(s) within that region will be utilized to substitute for the missing severity measurement(s) in that region. If the area of affected skin and/or all severity measurements for up to 2 regions are missing, then the missing (R+T+S) x A for a region will be substituted by the average of the available (R+T+S) x A. Otherwise, the PASI score will be set to missing.

The percent improvement in PASI scores from Baseline will be then computed as:

$$\text{Percent improvement from Baseline} = 100 \times \frac{\text{Baseline PASI} - \text{Post Baseline PASI}}{\text{Baseline PASI}}$$

If a subject has experienced an improvement, this measure will be positive. If a subject has experienced a worsening in their condition, this measure will be negative.

PASI90 is a binary variable equal to 1 if the percentage improvement from Baseline in the PASI scores is 90% or greater and 0 if the percentage improvement from Baseline is less than 90%.

PASI90 is the 3rd endpoint in the sequential testing hierarchy (Table 7–4).

8.2.1.3 SF-36 PCS: Change from Baseline at Week 16

The SF-36 (Version 2, standard recall) is a 36-items generic health-related quality of life instrument that uses a recall period of 4 weeks. Items are grouped into 8 domains as follows:

- Physical Functioning (10 items)
- Role Physical (4 items)
- Bodily Pain (2 items)
- General Health (5 items)
- Vitality (4 items)
- Social Functioning (2 items)
- Role Emotional (3 items)
- Mental Health (5 items)

[redacted] The concepts represented by these domains contribute to physical, mental, and social aspects of health-related quality of life.

In addition to domain scores, the PCS and MCS scores are calculated from the 8 domains (excluding the Health Transition item). Component scores reflect the impact of each domain on physical and mental health status.

The norm-based T-scores for the 2 SF-36 component summary (PCS and MCS) and the 8 domains are standardized with a mean of 50 and a standard deviation (SD) of 10 in the general US population (Maruish, 2011). An individual respondent's score that falls outside the T-score range of 45 to 55 should be considered outside the average range for the US general population. When considering group-level data, a score below 47 should be considered indicative of impaired functioning within that health domain or dimension. Similar to individual respondent data, group mean scores of 47 or greater should be considered average or above average as compared to the general US population. Higher scores indicate a better health status.

For the calculation of the SF-36 norm-based T-scores for the 8 domains and the PCS and MCS, the scoring software Optum's PRO CoRE (version 1.4) will be used. The software uses updated 2009 U.S. population norms and applies a Full Missing Score Estimation method as follows:

- A health domain score (except the Physical Functioning domain) will be estimated provided that at least one 1 non-missing response is available within that domain.
- For the Physical Functioning domain, item response theory will be used to develop a model for estimates of the missing score.
- Regression methods will then be applied to estimate the PCS and MCS on the basis of the available domains.

Change from Baseline at Week 16 in SF-36 PCS score is the 4th endpoint in the sequential testing hierarchy (Table 7–4).

8.2.1.4 MDA at Week 16

MDA is a state of disease activity deemed a useful target of treatment by both the subject and physician, given current treatment possibilities and limitations. The criteria that need to be met are based on key outcome measures in PsA and are shown below:

- Tender Joint Count ≤1
- Swollen Joint Count ≤1
- PASI ≤1(for subjects with PSO involving at least 3% of BSA at Baseline) or BSA ≤3%
- PtAAP (VAS) ≤15 mm
- PGA-PsA (VAS) ≤20 mm
- HAQ-DI ≤0.5
- Tender enthesial points ≤1(LEI≤1)

A subject has achieved MDA if 5 or more of the 7 above criteria are fulfilled.

The enthesitis instrument of the study (ie,: LEI) will be used to determine whether a subject has ≤1 tender enthesial point.

The following rule will be applied for subjects with BSA <3 at Baseline: Subjects with BSA <3 at Baseline will always meet the criteria PASI ≤1 or BSA ≤3 except in the cases where a post-Baseline BSA score >3 is observed.

MDA at Week 16 is the 5th endpoint in the sequential testing hierarchy (Table 7–4).

8.2.1.5 ACR20 and ACR70 response at Week 16

The ACR20 and ACR70 responses are defined based on respectively 20% and 70% or greater improvement from Baseline in the same measures than the primary endpoint. See Section 8.1.1 for details.

8.2.1.6 Proportion of subjects with an Investigator's Global Assessment (IGA) score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline at Week 4 and Week 16 in the subset of subjects with psoriatic skin lesions at Baseline

A static IGA for psoriasis will be used to assess disease severity in all subjects during the study. Investigator will assess the overall severity of PSO using the following 5-point scale specified in Table 8–3 .

Table 8–3: Five-point IGA

Score	Short description	Definition
0	Clear	No signs of PSO; post-inflammatory hyperpigmentation may be present
1	Almost clear	No thickening; normal to pink coloration; no to minimal focal scaling
2	Mild	Just detectable to mild thickening; pink to light red coloration; predominately fine scaling
3	Moderate	Clearly distinguishable to moderate thickening; dull to bright red, moderate scaling
4	Severe	Severe thickening with hard edges; bright to deep dark red coloration; severe/coarse scaling covering almost all or all lesions

IGA=Investigator's Global Assessment; PSO=psoriasis

The IGA response is defined as clear [0] or almost clear [1] assessment with at least a two-grade category improvement from Baseline, meaning that this parameter will be only evaluable for subjects with psoriatic skin lesions (IGA score ≥2) at Baseline.

Note that per protocol and device setting, only subjects with BSA ≥ 3% at Baseline will have IGA assessed at post-Baseline visits.

For some subjects with BSA ≥ 3% at Baseline, IGA data have not been captured at Week 4 and Week 16 visits when BSA<3% at those visits. To address that, additional supportive analyses will be performed in which these subjects will be considered as IGA responders at the respective visit.

8.2.1.7 Change from Baseline in the Patient's Assessment of Arthritis Pain (PtAAP)

The change from Baseline at Week 16 in PtAAP will be analyzed as described in Section 8.1.4.4.

8.2.1.8 Change from Baseline in PsAID-12 total score at Week 16

The PsAID questionnaires measure the impact of disease on the subject. The impact of PsA is measured by weighting 12 different domains of health to derive a weighted summary score. The domains are pain, fatigue, skin problems, work and/or leisure activities, functional capacity, discomfort, sleep disturbance, coping, anxiety/fear/uncertainty, embarrassment and/or shame, social participation, and depression. The PsAID score is based on the relative importance of each of the domains. Each domain is answered based on a Numeric Rating Scales (NRS) scale ranging from 0 to 10.

The equation for the PsAID-12 scores is as follows:

$$PsAID - 12 = \frac{1}{20}(3 \times Pain + 2 \times Fatigue + 2 \times Skin + 2 \times Work/Leisure + 2 \times Functionnal + 2 \times Discomfort + 2 \times Sleep + Coping + Anxiety + Embarrassment + Social + Depression)$$

The total score ranges from 0 to 10 with higher scores indicating a worse impact of the disease. A score below 4 out of 10 is considered a subject-acceptable status. A change of 3 or more points is considered relevant absolute change.

If one value (from the 12 domains) is missing, the missing value will be imputed in taking the mean value of the remaining (non-missing) domains, and the imputed value will be used to calculate the PsAID with the formula above. If two or more of the domains are missing, the PsAID score will be set to missing.

8.2.1.9 Psoriasis Area Severity Index 90 (PASI90) at Week 4 (in subjects with PSO involving at least 3% of BSA at Baseline)

This variable will be derived as described in Section 8.2.1.2 .

8.2.2 Primary analysis of secondary efficacy variables

The secondary efficacy variables will be analyzed for all subjects in the RS.

As indicated in Table 7–4, there are 4 secondary endpoints included in the testing hierarchy that will be tested in a chronological order until one is failed to be statistically significant.

For the other secondary endpoints (which are not part of the testing hierarchy), the p-values produced by the statistical models will be considered nominal since these endpoints are not controlled for multiplicity.

For the secondary composite and non-composite binary endpoints, the same estimand structure (composite estimand) as the one defined for the primary efficacy analysis of the primary efficacy endpoint (Section 8.1.2) will be used. The NRI approach for handling missing data and the same analysis model will be considered, and the analysis results will be presented similarly.

The statistical hypothesis for the binary variables at Week 16 is that the conditional odds ratio in the BKZ group compared with placebo treatment group is equal to 1. For the secondary continuous endpoints, the analysis will evaluate the hypothetical estimand as defined below:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP.
- Subject level outcome = variable as stated in Section 2.2.2
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to Week 16. A hypothetical strategy for addressing intercurrent events will be implemented. This estimand targets the treatment difference in a scenario where withdrawal from study treatment does not occur, such that outcomes for subjects without an IE are as observed, and outcomes for subjects with an IE are treated as though they had completed treatment through Week 16 but on placebo. A MI strategy will be used to impute any missing data and observed data after IE which will be set to missing prior to running MI. Such data will be imputed using Reference-Based MI, in which the MI model is based on data from the placebo group.
- Population-level summary measure = the difference in the adjusted means between BKZ 160mg Q4W and placebo.

The same analysis model and imputation strategy for handling missing data as in Section 8.1.4.4 will also be considered. The analysis results will be presented similarly as for this analysis on the individual ACR components.

In the case of MDA as the response is based on 7 different component scores, it is necessary to consider various data scenarios that could impact the calculation of response. The rules described here are applicable in the context of the calculation of MDA response and may differ from the rules applied for calculating and summarizing the components individually (some values may need to be imputed for component analysis but are not required here to evaluate MDA response).

The following rules will be applied to complete the derivation of MDA response based on the composite estimand definition:

- If a given visit has been preceded by an IE (treatment discontinuation):
  - The endpoint at all subsequent visits (whether the data were observed or not) will be set to “nonresponse” as the subject has not met the criteria for response.
- If a given visit has not been preceded by an intercurrent event:
  - If a subject satisfies at least 5 of the 7 MDA criteria, the subject is considered as MDA responder
  - If a subject does not satisfy at least 3 MDA criteria, the subject is considered as non-responder
  - In all other cases, the NRI approach will be applied.

An overview table will combine the results of the primary analysis for the primary and secondary efficacy endpoints included in the testing hierarchy.

For the country specific analyses performed on subjects randomized in Japan, the region factor will net be considered as a covariate for the modelling.

8.2.3 Supportive analyses of the secondary efficacy variables

As mentioned in Table 7–2, the supportive analyses for all secondary variables will be performed using the modified composite estimand for binary variables (using MI-MCMC/monotone regression as in Section 8.1.4.3) and on OC.

For PASI90, the MI will be run on the PASI score on subjects involving at least 3% of BSA at Baseline.

For the MDA (and VLDA) MI analysis, subjects with BSA <3 at Baseline will always meet the criteria PASI ≤1 or BSA ≤3 except in the cases where a BSA score >3 is observed. Subjects involving at least 3% of BSA at Baseline will have their BSA values imputed using MI when deriving MDA (VLDA).

Additionally, subgroup analyses will be performed on the PASI90 response at Week 16 (in subjects with PSO involving at least 3% of BSA at Baseline) endpoint. Refer to Section 4.8 for the list of subgroups of interest. The same analysis method than the one used for the subgroup analysis of the primary endpoint will be performed. The results will be presented similarly.

To assess the impact of the COVID-19 pandemic, the analysis of secondary efficacy endpoints with be repeated on the COVID-19-free Set using the primary analysis method.

8.3 Statistical analysis of other efficacy variables

8.3.1 Derivation of other efficacy variables

The analysis of other efficacy variables will be performed in the two randomized treatment groups.

8.3.1.1 Time to ACR20, ACR50, ACR70 response from Baseline

Time to a given response will be defined as the length in days from Baseline until the first date when the response is achieved.

Time from Baseline to censoring will be considered for the following subjects:

- Subjects who discontinue study treatment prior to achieving a response will be censored at the date of study treatment discontinuation.
- Subjects who reach the end of treatment period visit without achieving the given response will be censored at the date of the end of treatment period visit (for the interim analysis, subjects who reached Week 16 without achieving the given response will be censored at that visit).

8.3.1.2 ACR20, ACR50 and ACR70 response

Method for derivation of ACR50 can be found in Section 8.1.1. ACR20 and ACR70 will be derived similarly except that they will be based on a 20% and on 70% improvement from Baseline.

ACR20, ACR50 and ACR70 responses will be analyzed at Week 4, 8, 12, and 16.

8.3.1.3 PASI75, PASI90, and PASI100 response (in subjects with PSO involving at least 3% of BSA at Baseline)

Method for derivation of PASI score can be found in Section 8.2.1.2.

PASI75, PASI90, and PASI100 are binary variables equal to 1 if the percentage improvement from Baseline in the PASI scores is ≥ 75%, ≥90%, 100% and 0 if the percentage improvement from Baseline is < 75%, 90%, 100% respectively.

For PASI75, PASI90, and PASI100, the PASI score and the change from Baseline for subjects with PSO involving at least 3% of BSA at Baseline will be analyzed at Week 4, 8, 12 and 16.

8.3.1.4 Composite endpoints composed of ACR50/PASI90 and ACR50/PASI100 response (in subjects with PSO involving at least 3% of BSA at Baseline)

Composite endpoint composed of ACR50 and PASI90 is a binary variable equal to 1 for subjects responding to both ACR50 (ACR50=1) and PASI90 (PASI90=1).

The composite endpoint will be equal to 0 if ACR50=0 or PASI90=0 regardless of whether any of the two endpoints is missing.

The same rules as above will be applicable for the composite endpoint between ACR50 and PASI100.

The composite endpoints will be missing if both endpoints are missing or if any of the two is missing and the other one equal to 1. Each endpoint (ACR50 and PASI90 or PASI100) will be imputed separately before deriving the composite endpoint. The composite endpoints for subjects with PSO involving at least 3% BSA at Baseline will be analyzed at Week 4, 8, 12 and 16.

8.3.1.5 Proportion of Psoriatic Arthritis Response Criteria (PsARC) responders

The PsARC is based on the TJC (68 joints) and SJC (66 joints), the PGA-PSA and PhGA-PSA.

The PsARC response is defined as improvement from Baseline in at least 2 of the 4 measures (TJC, SJC, PGA-PSA, PhGA-PSA) 1 of which must be TJC or SJC and no worsening from Baseline in any of the 4 measures:

- improvement for TJC and SJC is defined as a reduction of ≥30%. Improvement of PGA-PSA and PhGA-PSA is defined as a reduction of the 100-point VAS of ≥20 points and, worsening is defined as an increase of and,
- worsening for TJC and SJC is defined as an increase of ≥30%. Worsening for PGA-PsA and PhGA-PsA is defined as an increase of the 100-point VAS of ≥20.

PsARC response will be analyzed at Week 4, 8, 12, and 16.

8.3.1.6 Psoriatic Arthritis Disease Activity Score (PASDAS) categories

The PASDAS is a composite score that includes patient and physician global scores of skin and joint disease, SJC, TJC, LEI, tender dactylitis count (sum of tenderness scores), the physical component of the SF-36 Health Survey, and level of hs-CRP. PASDAS is calculated using the following equation:

$$PASDAS = \Big((0.18 \times \sqrt{\text{Physician global VAS}}) + (0.159 \times \sqrt{\text{Patient global VAS}}) - (0.253 \times \sqrt{\text{SF36 PCS}}) + \big(0.101 \times \text{LN}(\text{SJC} + 1)\big) + \big(0.048 \times \text{LN}(\text{TJC} + 1)\big) + \big(0.23 \times \text{LN}(\text{LEI} + 1)\big) + (0.377 \times \text{LN}(\text{Tender dactylitis count} + 1)) + (0.102 \text{ x LN}(\text{hs} - \text{CRP} + 1) + 2)\Big) \times 1.5$$

Note: LN=natural logarithm

If Tender Dactylitis count is missing because LDI=0 at Baseline, Tender Dactylitis count will be replaced by 0 in the formula.

If for any other reason a PASDAS component is missing, PASDAS will be set to missing.

The following categories are used to define the level of disease activity:

- Remission: PASDAS ≤1.9
- Low disease activity: PASDAS >1.9 to <3.2
- Moderate disease activity: PASDAS 3.2 to <5.4
- High disease activity: PASDAS ≥5.4

The PASDAS categories and the change from Baseline in PASDAS score will be analyzed at Week 4, 12, and 16.

8.3.1.7 MDA response

Method for derivation of the MDA response can be found in Section 8.2.1.4.

MDA response will be analyzed at Week 4, 8, 12, and 16.

8.3.1.8 Very Low Disease Activity (VLDA) response

Like MDA, VLDA is a state of disease activity deemed a useful target of treatment by both the subject and physician, given current treatment possibilities and limitations.

A subject is considered as having VLDA if all 7 of the criteria used for MDA and listed in Section 8.2.1.4 are fulfilled. If any of the 7 criteria is not met, then the subject is considered as non-responder.

VLDA response will be analyzed at Week 4, 8, 12 and 16.

8.3.1.9 Proportion of subjects with an IGA score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline in the subset of subjects with psoriatic skin lesions at Baseline

Method for derivation of this variable can be found in Section 8.2.1.6.

The proportion of subjects with an IGA score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline in the subset of subjects with psoriatic skin lesions at Baseline will be analyzed at Week 4, 8, 12 and 16.

For some subjects with BSA ≥ 3% at Baseline, IGA data have not been captured at some post-Baseline visits when BSA<3% at those visits. To address that, an additional supportive analysis will be performed in which these subjects will be considered as IGA responders at the respective visit.

In this context of this analysis, observed data for % of BSA affected by PSO will also be analyzed by visit using descriptive statistics.

8.3.1.10 Disease Activity Index for Psoriatic Arthritis (DAPSA) score categories, state and change from Baseline in DAPSA score

Disease Activity Index for Psoriatic Arthritis is a composite score of patient global and pain VAS, TJC, SJC and hs-CRP that incorporates a pattern of peripheral arthritis that often is encountered in PsA. DAPSA score will be calculated as follows:

$$DAPSA = SJC + TJC + PGA - Arthritis\ (VAS) + PtAAP\ (VAS) + hs - CRP$$

where:

- SJC ranges from 0 to 66
- TJC ranges from 0 to 68
- PGA-Arthritis represents the patient's global assessment of arthritis ranging from 0 to 10; 0=best, 10=worst (using the unrounded VAS that ranges from 0 to 100 mm and dividing this value by 10) (Refer to Section 8.3.1.24).
- PtAAP VAS ranges from 0 to 10 0=best, 10=worst (unrounded, by dividing the original value by 10)
- hs-CRP in mg/L (no upper limit applied).

DAPSA values will be categorized in the following disease activity states:

- Remission (REM) (DAPSA ≤4),
- Low Disease Activity (LDA) (DAPSA from >4 to ≤14),
- Medium Disease Activity (MDA) (DAPSA from >14 to ≤28),
- High Disease Activity (HAD) (DAPSA >28).

For analyses with imputation of missing data, the imputed value for DAPSA will be based on the imputed values for the individual components.

For OC analyses, if any individual component score is missing, the DAPSA score will be set to missing.

DAPSA state, DAPSA scores and changes in scores from Baseline will be analyzed at Week 4, 8, 12, and 16.

8.3.1.11 Change from Baseline in the Disease Activity Score -28 based on C-reactive protein (DAS28[CRP])

The components for DAS28[CRP] include:

- the TJC and SJC based on 28 joints (Section 8.1.1.1):
  - Upper extremity (26) - bilateral shoulders, elbows, wrists (includes radiocarpal, and carpal and carpometacarpal bones considered as a single unit), metacarpophalangeal joints I, II, III, IV, and V, thumb interphalangeals, proximal interphalangeal II, III, IV, and V.

  - Lower extremity (2) - knees.

- the PGA-Arthritis (ranging from 0 to 10 by dividing the original value by 10)
- the hs-CRP (in mg/L)

DAS28[CRP] is calculated as follows:

$$DAS28[CRP] = 0.56 \times \sqrt{TJC} + 0.28 \times \sqrt{SJC} + 0.014 \times PGA - Arthritis + 0.36 \times \ln(hs - CRP + 1) + 0.96$$

For analyses with imputation of missing data, the imputed value for DAS28[CRP] will be based on the imputed values for the individual components.

For OC analyses, if any individual component score is missing, then the DAS28 [CRP] score will be set to missing.

The changes from Baseline in DAS28[CRP] will be analyzed at Week 4, 8, 12, and 16.

8.3.1.12 Change from Baseline in all individual ACR core components

Individual ACR components (as listed in Section 8.1.1) are:

- SJC,
- TJC,
- HAQ-DI,
- PtAAP,
- PhGA-PsA,
- PGA-PsA,
- hs-CRP.

The change from Baseline (ratio from Baseline for hs-CRP) for each of the ACR components will be summarized at Week 4, 8, 12, and 16.

8.3.1.13 Enthesitis-free state based on the LEI and change in LEI from Baseline in the subgroup of subjects with enthesitis at Baseline

The Leeds Enthesitis Index (LEI) is an index composed of 6 items, 3 for the right part and 3 for the left part of the body that measures the severity of enthesitis. The body parts are lateral epicondyles of the humerus, medial femoral condyle, and Achilles tendons. Each item is scored for pain with either 1 (painful) or 0 (no pain). The total score ranges from 0 (no enthesitis) to 6 (severe enthesitis).

Note that LEI is collected for all subjects regardless of their Baseline score.

In case 4 items are missing, the LEI will be set to missing. If at least 3 items are available, the available items will be assessed and scored and then weighted by the number of the assessed items (AI) as shown in the formula below:

$$LEI = n \times \frac{\sum_{i=1}^{n} score}{\sum_{i=1}^{n} AI}$$

Note that *n* refers to the number of total items.

Enthesitis-free state is based on LEI of 0, for subjects with enthesitis at Baseline (LEI>0). Analyses on LEI will be restricted to the subset of subjects with enthesitis at Baseline (defined as LEI score>0).

The percentage of subjects with enthesitis free-state and the change from Baseline in LEI, in the subgroup of subjects with enthesitis at Baseline, will be analyzed at Week 4, 8, 12, and 16.

While not a secondary endpoint, to support the pooled endpoint in PA0010, enthesitis free-state at Week 16 will be analyzed using the same analysis method as the primary analysis.

8.3.1.14 Enthesitis-free state based on the SPARCC index and change in SPARCC index from Baseline in the subgroup of subjects with enthesitis at Baseline

The SPARCC index measures the severity of enthesitis through the assessment of 16 sites, 8 for the right part and 8 for the left part of the body: the greater trochanter (right/left), quadriceps tendon insertion into the patella (right/left), patellar ligament insertion into the patella and tibial tuberosity (right/left), achilles tendon insertion (right/left), plantar fascia insertion (right/left), medial and lateral epicondyles (right/left), and the supraspinatus insertion (right/left). Tenderness on examination is recorded as either present (1) or absent (0) for each of the 16 sites, for an overall score range of 0 (no enthesitis) to 16 (severe enthesitis).

In case 9 or more items are missing, the SPARCC index will be set to missing. If at least 8 items are available, the available items will be assessed and scored and then weighted by the number of the assessed items as shown in the formula below:

$$SPARCC\ index = n \times \frac{\sum_{i=1}^{n} score}{\sum_{i=1}^{n} AI}$$

where *n* refers to the number of total items, and AI the number of assessed items.

Analysis of the change of SPARCC index at Week 16 from Baseline will be restricted to the subset of subjects with enthesitis at Baseline (defined as SPARCC >0).

Enthesitis-free state based on SPARCC index is a status defined as subjects having achieved a SPARCC index score of 0 at Week 16. This endpoint represents the proportion of subjects with enthesitis at Baseline who achieve a SPARCC index score of 0 at Week 16.

The percentage of subjects with enthesitis free-state and the change from Baseline in the SPARCC index (on the subgroup of subjects with enthesitis at Baseline) will be analyzed at Week 4, 8, 12, and 16.

8.3.1.15 Dactylitis-free state based on the LDI and change in LDI from Baseline in the LDI in the subgroup of subjects with dactylitis at Baseline

The LDI measures the percent difference between the circumference of the affected digit and the circumference of the digit on the opposite hand or foot (referred to as the contralateral digit), where circumference is measured in millimeters. A minimum difference of 10% and the assessment of the investigator that the digit was affected by dactylitis will be used to define a dactylitis digit.

The percent difference between circumferences will be multiplied by a tenderness score (0 for the score using the binary tenderness non-tender, 1 for tender). This score using the binary tenderness score is referred to as the LDI basic. The results from each digit with dactylitis will then be summed to produce a final LDI score.

Table 8–4 provides the comparison, if matching digits are thought to be involved.

Table 8–4: Normative values for LDI

	Digit	Men	Women
Hand	Thumb	70	58
	Index	63	54
	Middle	63	54
	Ring	59	50
	Little	52	44
Foot	Great toe	82	72
	Second	52	46
	Middle	50	44
	Fourth	50	44
	Little	52	45

The following rules will be applied for the LDI calculation in case of unclear data.

- Circumferences of 0 will be considered as missing. Circumferences <15 will be assumed to be in cm instead of mm and will be multiplied by 10 before being used in summaries/analyses.
- If both digits of a given pair are recorded as affected, then each digit will be compared to the normative value (shown in the above table). Note that if both comparisons result in a difference ≥10%, then both digits will contribute to the final LDI score.
- If the circumference of the affected digit is smaller than the unaffected digit, then the LDI will be calculated by comparing the smaller digit to the normative value.
- If a digit is recorded as affected and the circumference of the contralateral digit is missing, the normative value will be used for comparison with the affected digit.
- If a digit is selected but recorded as 'not affected' but has circumference and contralateral circumferences collected, a missing tenderness score will be considered as 1.
- If the investigator did not complete the LDI questionnaire at Baseline, it will be assumed that no digits were affected (ie, a LDI score of 0 will be assigned).

Dactylitis-free state is based on a Leeds Dactylitis Index (LDI) of 0, for subjects with dactylitis at Baseline (LDI>0). Analyses will be restricted to the subset of subjects with dactylitis at Baseline (defined as LDI score>0).

Note: A tender dactylitis count will be calculated as the sum of the tenderness scores for all digits and will be used for the PASDAS score calculation (Section 8.3.1.6).

The percentage of subjects with dactylitis free-state and the change from Baseline in LDI (on the subgroup of subjects with dactylitis at Baseline) will be analyzed at Week 4, 8, 12, and 16.

While not a secondary endpoint, to support the pooled endpoint in PA0010, dactylitis free-state at Week 16 will be analyzed using the same analysis method as the primary analysis.

8.3.1.16 Change from Baseline in the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI) in the subgroup of subjects with axial involvement defined by a score of ≥4 at Baseline

The BASDAI measures the disease activity of ankylosing spondylitis. The BASDAI is a validated self-reported instrument which consists of 6 horizontal NRSs, each with 10 units to measure the severity of the following 5 major symptoms over the last week: fatigue, spinal pain, peripheral joint pain and swelling, enthesitis, and morning stiffness (both severity and duration) over the last week.

To give each symptom equal weighting, the average of the 2 scores relating to morning stiffness is taken. The resulting 0 to 50 sum score is divided by 5 to give a final BASDAI score between 0 and 10, with lower scores indicating lower disease activity.

The BASDAI is calculated as follows:

$$BASDAI = \frac{Q1 + Q2 + Q3 + Q4 + \left(\frac{Q5 + Q6}{2}\right)}{5}$$

where Q1 – Q6 are the six questions from the BASDAI questionnaire.

If 1 of the 2 morning stiffness measurements (ie, questions: "How would you describe the overall level of morning stiffness you have had from the time you wake up?" and "How long does your morning stiffness last from the time you wake up?") is missing, the other one will be used for the morning stiffness calculation.

If one major symptom of the BASDAI is missing, the total score of the remaining symptoms will be divided by the number of symptoms assessed. If more than one major symptom is missing, the total BASDAI score will be set to missing.

Analyses will be restricted to the subset of subjects with axial involvement at Baseline (defined as BASDAI score of ≥4).

Change from Baseline in BASDAI will be analyzed for subjects with axial involvement at Baseline (BASDAI≥4) at Week 4, 8, 12, and 16.

8.3.1.17 Change from Baseline in the modified Nail Psoriasis Severity Index (mNAPSI) score at Week 4, 8, 12 and 16 in the subgroup of subjects with psoriatic nail disease at Baseline

Subjects with psoriatic nail disease will have a target nail selected at the Baseline Visit for evaluation using the mNAPSI. The nail selected should be the most affected nail observed at Baseline and should be the only one assessed throughout the study. The target nail will be scored (from 0 to 3) for onycholysis/oil drop dyschromia, nail plate crumbling, and pitting and will be

scored (0 for "no" or 1 for "yes") for leukonychia, nail bed hyperkeratosis, splinter hemorrhages, and red spots in the lunula. The total score (also called the mNAPSI score) will then range from 0 to 13 with higher scores indicative of more severe nail disease.

If any of the 7 response items that contribute to mNAPSI is present, while other items are missing (ie, partial mNAPSI data), then the missing items are assumed to be 0 for the mNAPSI calculation. In some cases, the data may be captured in such a way that only non-zero component scores are present in the database. Again, those components that are not present are simply assumed to be 0 for the mNAPSI calculation

Analyses will be restricted to the subset of subjects with psoriatic nail disease at Baseline (defined as mNAPSI score>0).

Change from Baseline in mNAPSI and mNAPSI resolution (mNAPSI=0 at post-Baseline assessments) will be analyzed for subjects with psoriatic nail disease at Baseline (mNAPSI score>0) at Week 4, 8, 12, and 16.

8.3.1.18 Change from Baseline in PsAID-12 total score

The PsAID questionnaires measure the impact of disease on the subject. The impact of PsA is measured by weighting 12 different domains of health to derive a weighted summary score. The domains are pain, fatigue, skin problems, work and/or leisure activities, functional capacity, discomfort, sleep disturbance, coping, anxiety/fear/uncertainty, embarrassment and/or shame, social participation and depression. The PsAID score is based on the relative importance of each of the domains. Each domain is answered based on a numeric rating scale ranging from 0 to 10. The equation for the PsAID-12 score is as follows:

$$PsAID-12 = \frac{1}{20}(3 \times Pain + 2 \times Fatigue + 2 \times Skin\ Problems + 2 \times Work/Leisure + 2 \times Functional\ Capacity + 2 \times Discomfort + 2 \times Sleep\ Disturbance + Coping + Anxiety/Fear + Embarrassment/Shame + Social\ Participation + Depression)$$

The total score also ranges from 0 to 10 with higher scores indicating a worse impact of the disease. A score below 4 out of 10 is considered a subject-acceptable status. A change of 3 or more points is considered relevant absolute change.
If 1 value (from the 12 domains) is missing, the missing value will be imputed in taking the mean value of the remaining (non-missing) domains, and the imputed value will be used to calculate the PsAID with the formula above. If 2 or more of the domains are missing, the PsAID score will be set to missing

The proportion of subjects achieving a total score of PsAID-12 ≤ 4 - Patient Acceptable Symptom State (PASS), the proportion of subjects achieving a decrease of 3 or more points from Baseline on the total score (for subjects with minimal score of 3 at Baseline), the change from Baseline in the total score and in the individual scores for each PsAID-12 domains, will be analyzed at Week 4, 12, and 16 (the proportion of subjects achieving PASS will also be analyzed at Baseline).

8.3.1.19 Change from Baseline in the Psoriatic Arthritis Quality of Life (PsAQoL) total score

PsAQoL is a disease specific Health-Related Quality of Life questionnaire comprised of 20 items. Each item is scored as 0 for "not true" or 1 for "true", so that the score ranges from 0 to 20 with higher scores indicating worse health-related quality of life.

If 6 or less item responses are missing, the missing responses will be imputed with the mean of available responses to calculate a total score. If more than 6 items responses are missing, the total score will be left missing.

Change from Baseline in PsAQoL total score will be analyzed at Week 4, 12, and 16.

8.3.1.20 Change from Baseline in the SF-36 PCS and MCS scores, as well as the 8 domain scores (Physical Functioning, Role Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role Emotional, and Mental Health)

Scoring instructions of SF-36 norm-based T-scores for the 8 domains and the SF-36 PCS and MCS are described in Section 8.2.1.3.

Change from Baseline in SF-36 individual domain scores, MCS and PCS scores will be analyzed at Week 4, 12, and 16.

8.3.1.21 Change from Baseline in Functional Assessment of Chronic Illness Therapy (FACIT) – Fatigue subscale score

The FACIT-Fatigue is 40-item measure that assesses self-reported fatigue and its impact upon daily activities and function over the past 7 days. The scale consists of 5 subscales (physical well-being, social/family well-being, emotional well-being, functional well-being, and fatigue). The fatigue subscale is the only one used in this study. It is composed of 13 items, all scored from 0 (Not at all) to 4 (Very much). The FACIT-fatigue subscale score ranges from 0 to 52 with 0 being the worst possible score and 52 being the best possible score. To obtain a score from 0 to 52, all negatively worded questions have to be recoded, so that responses range from worst (0) to the best (4) outcome.

Items 1 to 6, and 9 to 13 are negatively framed questions for which the coding needs to be reversed before deriving the score.

The FACIT-Fatigue subscale score is obtained by summing up the responses of all reversed / non reversed items with equal weight. In cases where some answers are missing, a total score is re-scaled from the score of the answered items, so long as more than 50% of the items (ie, at least 7 of 13) were answered.

The formula to derive the FACIT-Fatigue subscale score is then:

$$FACIT - Fatigue\ subscale = 13\ x\ \frac{\sum Score\ of\ items}{number\ of\ item\ answered}$$

The minimum clinically-important difference for FACIT-Fatigue subscale in patients with PsA was determined to be a 4-point positive change.

Change from Baseline in FACIT-Fatigue subscale score as well as the proportion of FACIT-Fatigue subscale responders (subjects with minimum clinically-important difference for FACIT-Fatigue subscale score, defined as an increase of ≥4) in subjects with FACIT-Fatigue subscale score ≤48 at Baseline will be analyzed at Week 4, 12, and 16.

8.3.1.22 Change from Baseline in Work Productivity and Activity Impairment Questionnaire–Specific Health Problem (WPAI-SHP) v2.0 adapted to PsA scores

The WPAI-SHP is a subject-reported questionnaire that assesses subject's employment status, work absenteeism, work impairment while working, overall work, and daily activity impairment attributable to a specific health problem.

Five out of 6 items of the WPAI-SHP are regrouped into the 4 dimensions (absenteeism, presenteeism, work productivity and activity impairment) with scores expressed as percentage, where higher numbers indicate greater impairment and less productivity (ie, worse outcomes) as described in the WPAI-SHP scoring rules.

The scoring rules for the WPAI-SHP are as follows:

[redacted]

Scores:

- Percent work time missed due to problem from the target disease:

$$\frac{Q2\ hours}{Q2\ hours + Q4\ hours} * 100$$

- Percent impairment while working due to problem from the target disease:

$$\frac{Q5\ \text{score}}{10} * 100$$

- Percent overall work impairment due to problem from the target disease:

$$\left[\frac{Q2\ \text{hours}}{Q2\ hours + Q4\ hours} + \left(1 - \frac{Q2\ \text{hours}}{Q2\ hours + Q4\ hours}\right) * \frac{Q5\ \text{score}}{10}\right] * 100$$

- Percent activity impairment due to problem from the target disease:

$$\frac{Q6\ \text{score}}{10} * 100$$

In order to make data consistent and amenable to statistical analysis, the following counting rules will be applied to handle out of range and ambiguous answers of the WPAI-SHP. These rules will be applied prior to conducting any type of statistical analysis of the data. Due to the inter-

relation between certain questions of the WPAI-SHP, the priority order for implementing these specific counting rules is as in the listed order below.

- Employment status:
  - If (Q1=missing) and (Q2>0 or Q3>0 or Q4>0), then Q1=YES.
- Hours missed due to psoriatic arthritis:
  - If Q1 = NO, then Q2 = missing.
  - If (Q2 =0 or missing) and Q1 = missing, then Q2=missing.
- Hours missed due to other reasons:
  - If Q1 = NO, then Q3 = missing.
  - If (Q3 =0 or missing) and Q1 = missing, then Q3=missing.
- Hours actually worked:
  - If Q1 = NO, then Q4 = missing.
  - If (Q4 =0 or missing) and Q1 = missing, then Q4=missing.
- Work productivity:
  - If Q1 = NO, then Q5 = missing.
  - If (Q5 =0 or missing) and Q1 = missing, then Q5=missing.
  - If Q4=0, then Q5 = missing.

In the listings, the original values will be kept and displayed.

WPAI-SHP scores are based on 1-item (presenteeism, activity impairment), 2-items (absenteeism) and multiple items (overall work productivity). A score cannot be calculated if there is a missing response to the corresponding item.

A negative number will indicate a reduction in the score/improvement for subjects.

The change from Baseline score is derived as post Baseline score minus Baseline score. A negative change score indicates a reduction in the score/improvement for the subject.

The change from Baseline in scores will be analyzed at Week 12 and 16.

The analysis of WPAI-SHP scores will be performed on OC only.

8.3.1.23 Responses to the EuroQol-5D-3-Level (EQ-5D-3L) dimensions, change from Baseline in VAS score

The Euro-Quality of Life 5-Dimensions 3 Level (EQ-5D-3L) version health questionnaire provides a descriptive profile and a single index value for health status (using a VAS).

The instrument is composed of a 5-item health status measure (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) and a VAS. Each of the 5 health states is divided into 3 levels (no problem, some or moderate problems, and extreme problems) and is scored as 1, 2, and 3, respectively. These scores are referred as “utility scores”.

The EQ-5D VAS records the respondent's self-rated health status on a vertical 20 cm visual analogue scale, graduated from 0 to 100 (0=worst imaginable health status, 100=best imaginable health status).

The response in each of the 5 items of the EQ-5D-3L and the values and the changes in VAS from Baseline will be presented at Week 4, 12, and 16.

8.3.1.24 Change from Baseline in Patient's Global Assessment of Arthritis (PGA-arthritis) and Physician's Global Assessment of Arthritis (PhGA-Arthritis)

The physician's global assessment of arthritis (PhGA-Arthritis) and the patient's global assessment of arthritis (PGA-Arthritis) are clinical outcomes based on visual analog scales (VAS) ranged from 0 to100 (in mm).

- PGA-Arthritis: Subjects will assess the impact of their arthritis pain in answering the following question "Considering all the ways your arthritis affects you, please mark a vertical line on the scale below to show how you are feeling today." The subject should be asked to consider their arthritis symptoms and functional capacity in their response to this question. Subjects will score PGA-Arthritis in using a VAS where 0 is "very good, no symptoms" and 100 is "very poor, severe symptoms."
- PhGA-Arthritis: The investigator will assess how the subject's overall arthritis appears at the time of the visit using a numerical rating scale where 0 is "very good" and 100 is "very poor". This assessment will be based through the following question: "The Patient's arthritis at this time is: "Please mark a vertical line on the scale below to assess the overall status of the subject's arthritis signs and symptoms and the functional capacity of the subject."

Values and changes from Baseline in PhGA-Arthritis and PGA-Arthritis will be analyzed at Week 4, 8, 12, and 16.

8.3.1.25 Proportion of subjects with a decrease of HAQ-DI from Baseline of at least 0.35 in the subgroup of subjects with Baseline HAQ-DI≥0.35

The proportion of HAQ-DI responders (subjects with a decrease of HAQ-DI from Baseline of at least 0.35 for subjects with HAQ-DI≥0.35 at Baseline) will be analyzed at Week 4, 8, 12, and16.

In addition, subgroup analyses will be performed on the HAQ-DI response at Week 16. Refer to Section 4.8 for the list of subgroups of interest. The same analysis method used for the subgroup analysis of the primary endpoint will be performed and the results will be presented similarly.

8.3.2 Analyses of other efficacy endpoints

Unless specified above, all other efficacy variables will be analyzed for all subjects in the RS.

8.3.2.1 Time to ACR20/50/70 response

Each time to ACR20/50/70 response will be estimated and presented using the Kaplan-Meier (KM) product-limit method for each treatment group. 1- KM estimates will be presented in tables.

The KM plots of time to ACR20,50,70 response will be presented by treatment group. The line will start at 0 and will increase over time, representing time to achieving the response.

The median time to response, including the 2-sided 95% CI, will be calculated for each treatment group. Comparisons of BKZ versus placebo will be analyzed using a log-rank test stratified by region and prior TNFα inhibitor exposure (considering that the same stratification factors were retained for the primary efficacy analysis). This comparison will be performed using PROC LIFETEST on subjects from both treatment groups. The hazard ratio as well as the corresponding 95% CI and p-value will be displayed (using PROC PHREG stratified by region and prior TNFα inhibitor exposure at Baseline).

The time to ACR20/50/70 endpoints will be analysed using the OC data.

8.3.2.2 Other endpoints (excluding time to ACR20/50/70 response)

For the binary variables, the following estimand structure (composite estimand) will be defined:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject-level outcome = The given variable and time point being summarized
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to the time point being summarized. A composite strategy will be implemented in which a positive clinical outcome is defined as achieving the given variable at the specified time point and not having an intercurrent event through that time point.
- Population-level summary measure = Unadjusted proportion of responder.

Subjects who have an IE which will be considered associated to a treatment failure and thus considered as non-responders. Any missing data will also be considered as nonresponders (NRI approach). For categorical variables, the worst category will be imputed similarly instead of non-response.

For the continuous variables, the following estimand structure (hypothetical estimand) will be defined:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject-level outcome = The given variable and time point being summarized
- Intercurrent event handling = An IE is defined as discontinuation of study treatment prior to the time point being summarized. A hypothetical strategy will be implemented in which outcomes for subjects without an IE are as observed at the given time point, and outcomes for subjects with an IE are treated as though they had completed the randomized study treatment through the time point being summarized. A MI strategy will be used to impute data following an IE.
- Population-level summary measure = Unadjusted mean.

The other efficacy variables will also be summarized based on OC and the modified composite estimand (for binary variables) and summarized by visit for the 2 treatment groups as detailed below:

<u>For binary endpoints using:</u>

- The number and proportion of responders using the NRI approach
- The mean proportion of responder in the multiple imputed data sets and the 95% CI (modified composite estimand)
- The observed number and proportion of responder on subjects with observed data.

For categorical endpoints (ie, EQ-5D-3L dimensions score, DAPSA state, PASDAS categories) using:

- The number and percentage by endpoint category using the NRI approach (as performed for the primary analysis of the primary endpoint with the exception that the worst category for the imputation = Worst Category imputation)
- The mean percentage and corresponding 95% CI by endpoint category in the multiple imputed data sets (modified composite estimand). Regardless of imputed values, value of the categorical endpoint after IE are by default set to the worst category (ie, High Disease Activity for DAPSA state)
- The observed number and percentage by endpoint category on subjects with observed data.

For continuous endpoints, on absolute values and changes from Baseline (or ratio from Baseline for hs-CRP) using:

- The mean (with geometric mean for hs-CRP and it 95% CI), SE, median, minimum, and maximum in the multiple imputed data sets (hypothetical estimand) (except for WPAI-SHP scores)
- The observed number of observations (n), the mean (with the geometric mean for hs-CRP), SD, Median, CV (geometric CV for hs-CRP), minimum and maximum.

For ACR20, ACR50, ACR70, PASI75, PASI90, and PASI100 responders over time, a line plot by treatment group will be produced.

8.3.2.3 Additional statistical analysis for other efficacy endpoints

Although statistical testing is not planned for other efficacy variables as per protocol, for selected other efficacy variables, it is of interest to perform statistical testing and calculate inferential statistics. The associated p-values will be considered nominal and not controlled for multiplicity.

For continuous variables, the MI-MCMC/monotone regression approach will be applied for the imputation model on the change from Baseline (hypothetical estimand). The analysis model will be based on ANCOVA with fixed effect of treatment, region, and prior TNFα inhibitor exposure and Baseline value as covariates.

For responder variables, the analysis will follow the NRI approach (composite estimand). The analysis model will be based on a logistic model with fixed effect for treatment and prior TNFα inhibitor exposure at Baseline and region as stratification variables.

Below is a list of other secondary efficacy variables for which these nominal p-values (with the timepoints in parentheses) will be calculated.

The results of these inferential tests (LSM differences for continuous variables, OR for binary variables, along with 95% CI and p-value) will be presented in a single table:

For the following endpoints, the p-value provided (based on the change from Baseline for continuous variables) will be based on the comparison of the 2 treatment groups using the same statistical methods for binary variables as used for the primary analysis of ACR50 (Section 8.1.2) and for continuous variables as used for the analysis of ACR components (Section 8.1.4.4):

- ACR20/50/70 (Week 4, 8, 12, 16)
- PASI75/90/100 (Week 4, 8, 12, 16)
- ACR50/PASI90 composite (Week 4, 8, 12, 16)
- ACR50/PASI100 composite (Week 4, 8, 12, 16)
- TJC (Week 4, 8, 12, 16)
- SJC (Week 4, 8, 12, 16)
- MDA (Week 4, 8, 12, 16)
- PtAAP (Week 4, 8, 12, 16)
- PsAQoL (Week 4, 12, 16)
- PsAID-12 (Week 4, 12, 16)
- FACIT-Fatigue subscale (Week 4, 12, 16)
- WPAI-SHP (Week 12, 16) (on observed data)
- DAPSA (Low Disease Activity or better) "response" (Week 4, 8, 12, 16)
- PsARC Response (Week 4, 8, 12, 16)

9 PHARMACOKINETICS AND IMMUNOLOGICAL (IMMUNOGENICITY) ANALYSES

9.1 Pharmacokinetics

Pharmacokinetic variables will be analyzed for all subjects in the PK-PPS.

Only subjects treated with BKZ will be included in the PK analyses. The Baseline observation for the Placebo group will be the Week 16 pre-dose observation. PK variables will be summarized at each scheduled visit under the "BKZ 160mg Q4W" column. BKZ concentrations will also be summarized similarly for the subgroup of Japan.

No imputation will be used for missing samples. However, if plasma concentration measurements are below the level of quantification (BLQ), then for the calculation of the derived statistics, the result will be set to ½ of LLOQ (ie, 1/2*0.250 = 0.125ug/mL). Descriptive statistics including number of values, geometric mean, its 95% CI, geometric CV, mean, SD, median, minimum, and maximum. Geometric mean and its 95% CI, geometric CV, mean and SD will be calculated if applicable will be calculated if at least ⅔ of the values of interest is above the LLOQ and number of values>=3; otherwise, only number of values, median, minimum, and maximum will be presented.

In addition, geometric mean of BKZ plasma concentration time curves (with its 95% CI) will be plotted on linear and semi-logarithmic scales for all subjects dosed with 160mg Q4W.

The table summaries and figures will be primarily repeated by anti-BKZ antibody status (positive, negative, or missing) for the total BKZ 160mg Q4W concentration data. The Missing group will not be displayed if ≥95% of subjects are categorized in the non-missing groups.

The ADAb status (positive, negative, or missing) will be considered in a cumulative manner at each time point:

- a subject will be counted as positive from the first visit at which the subject achieved a positive ADAb sample result to the end of the treatment period (regardless of any missing/inconclusive or negative ADAb sample result).
- If a subject has only negative ADAb samples or only one missing/inconclusive sample with negative ADAb samples up to that timepoint, the subject will be classified as negative. An exception remains for the Baseline Visit where only one sample could be available. If the sample is missing/inconclusive, then the sample will be classified as being negative for the cumulative ADAb status.
- Otherwise, the subject will be classified in the missing ADAb category.

In addition, the table summaries and figures will be repeated by neutralizing anti-bimekizumab antibodies (NAb) status (ADAb negative, NAb positive, ADAb positive / NAb negative, missing, Section 9.2.2). Tables and figures will be generated for concentration data up to weeks 16. For these tables and figures: the NAb status will be determined on a subject-level basis (ie not on a visit level) according to the categories in Section 9.2.2. For BKZ160mg Q4W dose group: the NAb status will be determined based on data from the Baseline visit (pre-dose sample on study day 1) up to Week 16 NAb sample (eg if a subject has no positive NAb samples for IL17AA or IL17F from Baseline up to Week 16, that subject will be assigned a NAb negative status for all visits of the initial treatment period).

The NAb status will be re-derived and updated with SFU data for the final study analysis, followed by an update to the relevant tables and figures.

PK samples collected at scheduled visits (except the SFU visit) and meeting the following requirements will be included in summaries and corresponding figures. Samples not meeting the following requirements will be excluded:

- For PK samples associated with the ET visit: include samples collected >14 days and <42 days after the last/previous dose received
- For PK samples associated with all other visits (Week 4, 8, 12, 16): include samples collected >14 days after the preceding dose, <42 days after the preceding dose, and no later than 1 hour after the current visit dose including unscheduled assessments as described in Section 3.1.
- For the SFU visit, all concentrations obtained at the SFU visit will be included in the summary tables but will not be included in the figures.

When multiple samples meeting the criteria above are associated with the same visit (either because multiple samples were collected or due to remapping or unscheduled visits), only results from one valid sample will be included in summaries and figures using the following rules. All others will be excluded and only listed:

- For PK samples with different dates, the sample that is closest to the target visit date will be included. For samples with different dates that are the same distance to the target visit date, the sample collected prior to IMP dosing will be included.
- For PK samples with the same date, the first sample as provided in the raw data will be included.

The BKZ concentrations will also be listed for all subjects in the SS. All concentrations will be listed as received, prior to substitution of any BLQ values. The listing will include flags for concentrations that were excluded from the summary statistics where the reason for exclusion will be one or more of the following:

- Dosing performed out of window
  - Sample collected out of window relative to current dose: for PK samples associated with visits other than ET, or SFU: includes samples >1 hour after the current visit dose
  - Sample collected out of window relative to previous dose: For PK samples associated with all visits except SFU visit; include samples collected <14 days or >42 days after the previous dose received
  - More than 1 sample obtained at the same visit: includes all samples excluded due to multiple valid samples associated with a visit.

All plasma concentration data will be reported in ug/mL in the tables, figures, and listings.

If more than 10% of the PK concentration results have been excluded from the table summaries, the PK excluded results will be listed in a separate listing.

9.2 Immunological (immunogenicity) analyses

9.2.1 Anti-bimekizumab antibody

ADAb will be assessed using a 3-tiered assay approach: screening, confirmatory and titration assay. Samples confirmed as positive within the confirmatory assay will be further evaluated for the presence of NAb specific to binding of bimekizumab to IL-17AA, IL-17FF or both.

ADAb (including positivity) will be summarized by BKZ treatment at each scheduled visit at which samples are collected.

Screening, confirmatory and titer cut points of the respective assays will be determined by the bioanalytical laboratory based either on commercially available drug-naïve samples or on the pre-dose samples of a given study. The relevant statistical reports will be provided as part of the bioanalytical reports. ADAb samples are analyzed on the SS, but are not analyzed when subjects are on a treatment other than BKZ. Subjects assigned to the placebo arm will have only their Baseline (study day 1) and Week 16 samples analyzed for ADAb.

The Screening cut point will be used to determine the ADAb status in the test sample as "Positive Screen" (PS) or "Negative Screen" (NS). For samples presenting anti-BKZ antibody levels that are PS, further confirmatory assay will be performed, and the result of which will be reported as either "Positive Immunodepletion" (PI) or "Negative Immunodepletion" (NI).

ADAb status for each visit will be derived as follows:

- Samples that are either NS or PS and NI will be defined as ADAb negative.
- Sample values that are either NS or PS and NI and where the BKZ concentration exceeds the validated ADAb assay drug tolerance limit (200 µg/mL) will be defined as inconclusive.
- Sample values that are PS and PI will be defined as ADAb positive (regardless of availability of a titer value)
- Missing if it does not go into one of the above categories.

PI samples will be titrated, and the ADAb titer (reciprocal dilution factor including MRD) reported. The PI samples will also be subject to a neutralizing assay to evaluate whether the anti-BKZ antibody neutralizes the target binding of bimekizumab (IL17A or IL17F or both) in-vitro. The following rule will be implemented for by-visit ADAb summaries where applicable:

- If the ADAb sample is collected within ±14 days (inclusive) relative to the visit date at which the drug was administered (or ±14 days from a scheduled visit at which dosing was not performed), the ADAb result for that sample will be associated with the scheduled visit and summarized accordingly. This will include unscheduled assessments (if a dose was administered at an unscheduled visit). Samples collected outside this window will be excluded from the by-visit ADAb summaries and will be listed only.

The rule above will apply to by-visit summaries only; all other summaries of ADAb status will use all available data (scheduled and unscheduled). This rule will not apply for the ET or SFU visits. Thus, all ADAb data obtained at these visits will be included in the by-visit summaries.

In addition, the anti-BKZ antibody status will be further classified as outlined below:

- **Category 1: Pre ADAb negative – treatment emergent ADAb negative**: Includes subjects who are negative at Baseline and antibody negative at all sampling points post treatment (including SFU), one post-baseline missing/inconclusive sample is allowed for subjects with pre-ADAb negative sample. This group also includes subjects who have missing/inconclusive pre-treatment sample (eg, either missing/inconclusive or insufficient volume) at baseline with all post-Baseline samples as ADAb negative.
- **Category 2: Pre ADAb negative – treatment emergent ADAb positive**: Includes subjects who are negative at Baseline and antibody positive at any sampling point post treatment (including SFU). This group also includes subjects who have a missing pre-treatment sample (either missing or insufficient volume) at Baseline with one or more ADAb positive post-treatment samples.
- **Category 3: Pre ADAb positive – treatment emergent reduced ADAb**: Includes subjects who are positive at Baseline, and antibody negative at all sampling points post treatment (including SFU).
- **Category 4: Pre ADAb positive – treatment unaffected ADAb positive**: Includes subjects who are positive at Baseline and are positive at any sampling point post treatment (including SFU) with titer values of the same magnitude as Baseline (ie, less than a predefined fold difference from the Baseline value). For the purpose of this study, this is set at an increase of less than the validated Minimum Significant Ratio (MSR) of 2.07-fold from Baseline.

- **Category 5: Pre ADAb positive – treatment boosted ADAb positive**: Includes subjects who are positive at Baseline and are positive at any sampling point post treatment (including SFU) with increased titer values compared to Baseline (greater than or equal to a predefined fold difference increase from Baseline value which is defined within the validation of the assay). For the purposes of this study, this is set at an increase greater than or equal to the validated Minimum Significant Ratio (MSR) of 2.07-fold from Baseline.

  Note: For any subject who is positive at Baseline and positive at a post-Baseline time point, but for whom titers are not available to determine treatment unaffected or treatment boosted status, the subject will be considered as treatment boosted (ie, Category 5), assuming no other samples are available.

- **Category 6: Inconclusive**: Includes subjects who have a positive pre-treatment sample and some post-treatment samples are missing, while other post-treatment samples are ADAb negative or missing.

- **Category 7: Total treatment-emergent:** Category 2 and 5 combined: Includes subjects who are pre ADAb negative – treatment emergent ADAb positive (Category 2) and pre ADAb positive – treatment boosted ADAb positive (Category 5).

- **Category 8: Total prevalence of pre-ADAb positivity**: (Categories 3, 4, 5 and 6 combined): Subjects that are tested ADAb positive at Baseline.

- **Category 9: Missing:** Includes subjects who have a negative or a missing/inconclusive pre-treatment sample and some post-treatment samples are missing/inconclusive, while other post-treatment samples are ADAb negative or missing.

All the following analyses will be prepared on the SS:

- Summary tables displaying the number and percentage of subjects with ADAb status (positive, negative, or missing) at the time of each visit by treatment group (160mg Q4W) will be provided. For the placebo arm, the by-visit summary of ADAb positive status will include a summary of subjects with ADAb status (positive, negative, or missing) at Baseline and Week 16. For the overall summary at any visit (only applicable to 160mg Q4W dose arm), two summaries will be presented: one that will exclude data obtained at the SFU visit and will include data obtained at Baseline and one that will include data obtained at both the SFU visit and at Baseline. Subjects who rolled over to the OLE study will not have a SFU visit per study protocol. For these subjects, the 'overall ADAb status including SFU' will then be considered as being identical to the 'overall ADAb status up to Week 16'.

  For the overall ADAb status, a subject will be classified as:

  – Positive if the subject has at least one positive sample at any time in the treatment period (regardless of having missing /inconclusive data)

  – Negative if the subject has all the samples negative or only one missing/inconclusive sample with negative ADAb samples up to the timepoint of interest

  – Missing if the subject has missed more than one sample result for ADAb assessment (or have more than one inconclusive sample) and all other available ADAb samples are negative during the period of interest (If there are ≥95% of subjects included in the non-missing groups, the missing group will not be displayed on the table).

The above summary table displaying the number and percentage of subjects in each of the ADAb status (positive, negative, total of positive and negative, missing) will be repeated for the subgroup of Japan.

- A summary table displaying the number and percentage of subjects in each of the ADAb status (positive, negative, total of positive and negative, missing) by concomitant medications (use of cDMARDs at entry, use of methotrexate at entry, use of oral/systemic corticosteroids at entry) will be provided (overall summaries only).
- A table displaying the number and percentage of subjects with the first occurrence of any ADAb positivity during the study (ie, including Baseline visit) will be summarized. This summary will include the following categories:
  - Any ADAb+: ADA positive sample regardless of category during the treatment period
  - ADAb category 2: Pre ADAb negative – treatment-emergent ADAb positive
  - ADAb category 5: Pre ADAb positive – treatment-boosted ADAb positive (if this category represents at least 10% of the subjects, otherwise results are to be merged with category 2). For this category, the first occurrence of a boosted result is considered.

  The table will include the number and percentage of subjects with first occurrence of any ADAb+ sample, and also subjects who are either treatment-emergent ADAb positive or treatment-boosted ADAb positive for the first time at the specified time point in the study and will include the cumulative number and percentage of subjects with treatment-emergent ADAb positive results at each time point.
- A boxplot of the ADAb titer by time of occurrence of ADAb positivity will be created. The ADAb titer results will be presented on the log-scale. The time represents the visit week for the ADAb positive sample. Subjects who do not have any ADAb positivity will be excluded from the plot.
- A summary table of the number and percentage of subjects in each of the 9 ADAb categories will be tabulated.
- Figure summarizing the time to achieve ADAb positivity and treatment-emergent ADAb positivity on a cumulative basis (2 lines per plot) will be presented. Subjects will be considered to have an event at the time point at which an ADAb+ result or treatment emergent ADAb positive is first achieved. Treatment-emergent ADAb+ in this plot will be based on ADAb category 7 (total treatment-emergent ADAb).

  In the event that ≥10% of subjects are classified as ADAb category 5, the lines will be further split by the following categories (thus the plot will include 3 lines):
  - ADAb category 2: Pre ADAb negative – treatment-emergent ADAb positive
  - ADAb category 5: Pre ADAb positive – treatment-boosted ADAb positive
- Individual plots (1 plot per subject) of BKZ concentrations/ADAb titer and the ACR50 response (based on NRI) will be created. Each time, all 3 endpoints will be plotted on the Y-axis by visit (X-axis) for the full treatment period, including SFU where a subject has not progressed into the OLE. Plots should be labeled and grouped into the 9 ADAb categories and display vertical lines on the X-axis corresponding to the dosing incidences/visits.

- Individual plot (1 plot per subject) of BKZ concentrations/ADAb titer and PASI90 response will also be created similarly as described above.
- Spaghetti plots of ADAb titer (Y-axis) will be reported on the log-scale by visit (X-axis) for all ADAb positive subjects, including Baseline positive subjects. This plot will include the following ADAb categories:
  - ADAb category 2: Pre ADAb negative – treatment-emergent ADAb positive
  - ADAb category 5: Pre ADAb positive – treatment-boosted ADAb positive

  Plots will be presented using a semi-logarithmic scale for the ADAb titers (ADAb negative samples will therefore be excluded from the plot). Furthermore, all ADAb titer values <100 will be represented as 1 in these plots.
- Figures will summarize efficacy response (ACR50, ACR20 and PASI90 responders based on NRI) versus ADAb titer quartiles. The X-axis will display the ADAb titer quartiles at Week 16 (categorized as negative, Q1, Q2, Q3 and Q4) and the y-axis will display percentage of ACR50 (ACR20 or PASI90) responders at Week 16. The plot will therefore display the percentage of ACR50 (ACR20 or PASI90) responders as a function of the number of subjects within each ADAb and titer categories. Subjects with negative ADAb results at Week 16 will be included in the 'negative' category on the X-axis.
- A figure summarizing time course of efficacy response (ACR50 responders based on NRI) for the BKZ 160mg Q4W group by the following ADAb groups (3 lines per plot) will be created:
  - ADAb positive - defined as subjects having at least 2 ADAb positive samples during the study (including baseline and SFU) regardless of other ADAb negative samples and/or missing or inconclusive samples
  - ADAb negative - defined as subjects for whom either (1) all samples (including baseline and SFU) are ADAb negative and there are no missing or inconclusive samples, (2) only 1 sample is ADAb positive and all other samples (including baseline and SFU) are ADAb negative or missing/inconclusive or 3) only one sample is missing/inconclusive and the remaining ADAb samples are negative.
  - Missing - defined as subjects who do not fulfil the criteria for one of the 2 groups listed above.

  If ≥95% of subjects are included in the non-missing groups, the missing group will not be displayed in the figure.

  The data will also be presented in a tabular format. The figure and table will be repeated for ACR20 and PASI90.

  Note: for efficacy by ADAb status summary up to a particular timepoint (Weeks 16), only ADAb data up to the timepoint of interest will be utilized in deriving the overall ADAb status during that period. The percent of subjects in each of the ADAb categories considered for efficacy sub-group analyses will be summarized.
- All individual subject-level ADAb results will be listed including the Screening assay, confirmatory assay, ADAb status, and titers if applicable. Note, that titer results will only be

available, if the confirmatory assay is positive. The listing will also include flags for ADAb measurements that were excluded from the by-visit summaries. The reason for exclusion will be one of the following:

  - Sample collected out of window relative to current dose (or visit),
  - More than one sample obtained at the same visit.

- Finally, if more than 10% of the ADAb results have been excluded from the table summaries, the ADAb excluded results will be listed in the same listing as the one mentioned in Section 9.1 for excluded PK results.

9.2.2 Neutralizing anti-bimekizumab antibodies

NAb will be assessed using IL-17AA- and IL-17FF-specific assay method, respectively. NAb results will be positive or negative to each specific NAb assay or both.

All NAb results derived from samples with drug concentrations > the drug tolerance limits of the NAb assays (100µg/mL) will be labeled 'inconclusive'. All inconclusive results will be regarded as missing.

Subjects will be assigned an overall Neutralizing anti-BKZ antibodies (NAb) classification, inclusive of Baseline and post-Baseline results from NAb assay:

- NAb negative: No NAb positive samples for IL-17AA and IL-17FF at Baseline or post-Baseline. This group will also include subjects who have only 1 missing sample and all other available samples during the period of interest are negative. Subjects who are NAb negative will be classified as follows:
  - ADAb positive / NAb negative: ADAb positive subjects who are 1) NAb negative for all available ADAb positive samples or 2) with only 1 missing NAb sample and all other evaluated ADAb positive samples are NAb negative.
  - ADAb negative: if the subject has all the samples as ADAb negative or only 1 missing/inconclusive sample with all other available samples as negative ADAb. Note that ADAb negative samples are not subject to the neutralizing assay.
- NAb positive: One or more positive samples (IL-17AA positive, IL-17FF positive, or both) at Baseline or post-Baseline (regardless of missing samples). Subjects who are NAb positive will be further classified as follows:
  - Positive for IL-17AA only: 1 or more positive samples for IL-17AA at Baseline or post-Baseline. No positive samples for IL-17FF
  - Positive for IL-17FF only: 1 or more positive samples for IL-17 FF at Baseline or post-Baseline. No positive samples for IL-17AA
  - Positive for both IL-17AA and IL-17FF: 1 or more positive samples for both IL-17AA and IL-17FF at Baseline or post-Baseline

- NAb Missing: more than one relevant NAb samples are missing/inconclusive and other available NAb samples during the period of interest are negative, missing or insufficient sample left for NAb testing.

Note: For PK and efficacy analyses up to a certain timepoint (Week 16) by NAb status, only data up to the timepoint of interest will be utilized in deriving the subject's overall NAb status during that period (eg for the ACR50 by NAb up to Week 16, only NAb data from Baseline up to Week 16 will be used to derive subject's overall NAb status during the 16-week Double-Blind Treatment Period).

A listing will be produced to summarize the NAb status overall in the study. The listing will be sorted by treatment group, subject identifier and visit and will summarize the following information for each subject assessed for NAbs:

- Visit
- Study week
- Laboratory sampling date and time
- Time since previous dose (weeks)
- The corresponding BKZ plasma concentration level at each visit (ug/mL)
- Anti-BKZ anti-body titer at each visit
- IL-17AA NAb status and corresponding IL-17AA signal/negative control result
- IL-17FF NAb status and corresponding IL-17FF signal/negative control result

A summary table will provide the following overall summary statistics by treatment group (based on the total number of subjects in the SS):

- The number and percentage of subjects confirmed as anti-BKZ antibody positive and anti-BKZ antibody negative to Week 16 (excluding SFU), 16 (including SFU).
- The number and percentage of subjects who are NAb positive, NAb negative and missing to Week 16 (excluding SFU), 16 (including SFU).

The NAb summary tables will be repeated with percentages calculated on a denominator based on the number of ADAb positive subjects during the above periods.

In addition, the following analyses will be performed to assess the impact of NAb status on efficacy variables:

- A figure summarizing efficacy response (ACR50 based on NRI) versus time for each treatment group by NAb status (ADAb negative, NAb positive, ADAb positive/ NAb negative, missing) will be created. If ≥95% of subjects are included in the non-missing groups, the missing group will not be displayed in the figure. Separate figures will be presented for data up to Week 16. The data will also be presented in a tabular format. The analyses will be repeated for ACR20 and PASI90. The data will be also presented in a tabular format.

10 SAFETY ANALYSES

All safety summaries will be done on all subjects in the SS.

10.1.1 Extent of exposure and time at risk

The extent of exposure will be evaluated based on the duration of exposure to IMP and the time at risk. Two summary tables will be created to describe the extent of exposure.

The first table consists of a descriptive summary of study medication duration (in days), total study medication duration (in subject-years), and total time at risk (in subject-years).

The sum of all durations of exposure (in subject-years) will be calculated as:

$$\frac{\sum_{Subjects} Duration\ of\ exposure\ (days)}{365.25}$$

The sum of all times at risk (in subject-years) will be calculated as:

$$\frac{\sum_{Subjects} Times\ at\ risk\ (days)}{365.25}$$

A second summary will be created on the cumulative study medication duration for subjects exposed to the following durations of time:

- > 0 week
- ≥ 4 weeks
- ≥ 8 weeks
- ≥ 12 weeks
- ≥16 weeks

The study medication duration (in days) will be calculated as follows:

$$Date\ of\ last\ dose - Date\ of\ first\ dose + 28$$

- If the date of last dose + 28 days extends to a date beyond the final visit date (not including SFU), then this calculation reverts to:

$$Date\ of\ last\ visit\ (not\ including\ SFU) - Date\ of\ first\ dose + 1$$

- For subjects who die, this calculation reverts to:

$$Date\ of\ death - Date\ of\ first\ dose + 1$$

The time at risk (in days) will be calculated as follows:

- For subjects who complete the study as planned and continue into the OLE study (and, therefore do not have a SFU visit):

$$Date\ of\ last\ visit - Date\ of\ first\ dose + 1$$

- For subjects who die prior to the final visit, this calculation reverts to:

$$Date\ of\ death - Date\ of\ first\ dose + 1$$

- For all other subjects, this calculation reverts to the minimum of the following:

$$Date\ of\ last\ dose - Date\ of\ first\ dose + 141$$

$$Date\ of\ last\ clinical\ contact - Date\ of\ first\ dose + 1$$

Note: 140 days refer to 5 times the half-life of BKZ.

Date of last clinical contact for each subject is defined as the maximum of [last visit date including SFU visit, last AE start date (including imputed AE start dates), date of study termination or completion, last date of study drug administration].

This group of subjects could include subjects who discontinue early from the study, subjects who complete study as planned but chose not to continue into the OLE study, or subjects who are ongoing in the SFU Period at the time of the data snapshot.

A by-subject listing of date of first and last dose and the duration of exposure will be performed.

10.1.2 Time at risk in subject-years by COVID-19 pandemic period

The time at risk will be determined by COVID-19 pandemic period as defined in Section 3.11.

10.1.2.1 Time at Risk in the pre-COVID-19 pandemic period

Only subjects starting study drug during the pre-COVID-19 pandemic period will have a time at risk calculated during that period.

- For subjects with date of last dose of study drug on or after 11Mar2020:

  10Mar2020 – date of first dose + 1

- For subjects who have died on or before 10Mar2020:

  Date of death – date of first dose + 1 (If the date of death is later than 140 days following the last dose of BKZ, this rule does not apply).

- For subjects with date of last dose of study drug on or before 10Mar2020 but with the earliest date between (date of last dose + 141) and (date of last clinical contact) on or after 11Mar2020:

  10Mar2020 – date of first dose + 1

- For subjects with date of last dose of study drug on or before 10Mar2020 but with the earliest date between (date of last dose + 140) and (date of last clinical contact) on or before 10Mar2020, minimum between:

  date of last dose + 141 – date of first dose

  date of last clinical contact – data of first dose +1

10.1.2.2 Time at Risk in the COVID-19 pandemic period

Subjects with study drug period overlapping the COVID-19 pandemic periods will have a time at risk calculated during that period.

- For subjects with date of last dose of study drug on or after 11Mar2020, minimum between:

  Minimum (date of last dose + 141, COVID-19 pandemic end date) – maximum of (11Mar2020, date of first dose)

Minimum (date of last clinical contact, COVID-19 pandemic end date) – maximum of (11Mar2020, date of first dose) +1

- For subjects who have died on or after 11Mar2020:

  Minimum (Date of death, COVID-19 pandemic end date) – maximum of (11Mar2020, date of first dose) + 1 (If the date of death is later than 140 days following the last dose of BKZ, this rule does not apply).

- For subjects with date of last dose of study drug on or before 10Mar2020 but with the earliest date between (date of last dose + 141) and (date of last clinical contact) on or after 11Mar2020, minimum between:

  Minimum (date of last dose+ 141, COVID-19 pandemic end date) – 11Mar2020

  Minimum (date of last clinical contact, COVID-19 pandemic end date) – 11Mar2020+1

10.1.2.3 Time at Risk in the post COVID-19 pandemic period

Subjects with study drug period continuing after the COVID-19 pandemic periods will have a time at risk calculated during that period.

- For subjects with date of last dose of study drug on or after the end date of the COVID-19 pandemic period, minimum between:

  date of last dose + 141 – maximum of (COVID-19 pandemic end date, date of first dose)

  date of last clinical contact – maximum of (COVID-19 pandemic end date, date of first dose) +1

- For subjects who have died on or after *COVID-19 pandemic end date*:

  Date of death – maximum of (COVID-19 pandemic end date, date of first dose) + 1 (If the date of death is later than 140 days following the last dose of BKZ, this rule does not apply).

- For subjects with date of last dose of study drug on or before *COVID-19 pandemic end date* but with the earliest date between (date of last dose + 140) and (date of last clinical contact) on or after *COVID-19 pandemic end date*, minimum between:

  date of last dose + 141 – COVID-19 pandemic end date

  date of last clinical contact – COVID-19 pandemic end date+1

10.2 Adverse events

An AE is any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product that does not necessarily have a causal relationship with this treatment. An AE can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether related or not to the IMP.

AEs (including SAEs) are characterized as either non-treatment or treatment emergent according to the following criteria:

- Non-treatment emergent are the events with onset date and time prior to the very first administration of study medication (BKZ or placebo) or after a 140-day period after the final drug administration.
- Treatment-emergent AEs (TEAEs) are those with onset date at or after the very first administration of study medication. The events that emerge within 140 days after the final drug administration, will also be considered as treatment emergent (eg, in the case of premature discontinuation or during the 20-week SFU Period).

The incidence of TEAEs will be summarized descriptively by treatment group, MedDRA primary system organ class (SOC), high-level term (HLT), and preferred term (PT).

AEs will be presented as "number of subjects (percentage of subjects) [number of events]". In this style of output, "[number of events]" will include all cases of an AE including repeat occurrences in individual subjects, while "number of subjects" will count each subject only once.

Adverse event summaries by SOC, HLT and PT will be ordered alphabetically for SOC and HLT within SOC and in terms of decreasing frequency for PT within HLT in the BKZ treatment group and then similarly by decreasing frequency in the placebo group.

AE summaries by PT will be ordered in terms of decreasing frequency for PT within the BKZ treatment group, and in the event of ties, PT will be sorted alphabetically.

TEAEs will be classified according to their date of emergence.

All summaries will be provided by treatment group on the SS.

10.2.1 Standard AE summaries

Tables for incidence of the following categories of TEAES will be provided:

- TEAEs (Overview)
- TEAEs by SOC, HLT, and PT
- TEAEs by decreasing frequency of PT
- TEAEs above reporting threshold of 5% by PT, by SOC and PT
- TEAEs by maximum intensity
- TEAEs by relationship to IMP
- Related TEAEs by SOC, HLT, and PT
- Related TEAEs above reporting threshold of 5% by PT, by SOC and PT
- TEAEs leading to study discontinuation, by SOC, HLT, and PT
- TEAEs leading to study drug discontinuation, by SOC, HLT, and PT
- Treatment-emergent SAEs (TESAEs) by SOC, HLT, and PT
- TESAEs by relationship by SOC, HLT, and PT
- TEAEs leading to death by SOC, HLT, and PT - deaths will be also be tabulated and listed
- Non-Serious TEAEs above reporting threshold of 5% by relationship by PT

Some selected tables of incidence will include the exposure adjusted incidence rate (EAIR) with associated 95% confidence interval and the exposure adjusted event rate (EAER) (see Section 10.2.3). The categories of TEAEs concerned are listed below.

- TEAEs by SOC, HLT, PT,
- TEAEs leading to study discontinuation by SOC, HLT, PT,
- TEAEs leading to study drug discontinuation,
- TEAEs by timing of onset relative to ADAb status by SOC, HLT, PT (on subjects treated with BKZ). This will include columns for the following:
  - TEAEs starting before the first ADAb positive result (includes ADAb categories 2 and 5) where TEAEs have occurred before the following events: a) the first positive ADAb result for subjects in category 2 and b) the first post-Baseline boosted ADAb titer result for subjects with titer results and the first post-Baseline positive ADAb result for subjects with positive ADAb at Baseline with no other samples with titer available for subjects in category 5
  - TEAEs starting on the same date or after the first ADAb positive result (includes ADAb Categories 2, 3, 4, 5 and 6) where TEAEs have occurred on or after the following events: a) the first positive ADAb results for subjects in categories 2, 3, 4 and 6, and b) the first post-Baseline boosted ADAb titer result for subjects with titer results and the first post-Baseline positive ADAb result for subjects with positive ADAb at Baseline with no other samples with titer available for subjects in category 5
  - TEAEs for subjects who are ADAb negative at all timepoints (includes ADAb Category 1)
- TEAEs by NAb status (ADAb negative, NAb positive, ADAb positive / NAb negative) on subjects treated with BKZ. TEAEs will be sorted by system organ class, high level term and preferred term.

  Note: for TEAE by onset relative to ADAb positivity status and by NAb status, all available ADAb and NAb data at the time of IA cut-off, respectively, will be utilized to derive the subject-level ADAb/NAb status categories.
- TESAEs by SOC, HLT, PT.

By-subject listings on all AEs and all deaths will be provided on the ES. AEs leading to study discontinuation will also be listed on the SS. AEs occurring at first study drug administration will also be listed. A glossary listing will also be created for all TEAEs.

10.2.2 Exposure duration of AEs

The duration of each AE will be calculated as follows:

$$Duration\ of\ AE\ (days) = End\ date\ of\ AE - Start\ date\ of\ AE + 1$$

For the calculation of duration for AEs with missing start dates, the dates will be imputed as described in Section 4.2.3.

The time since first dose of IMP for each AE will be calculated as follows for all TEAEs:

$$Time\ since\ first\ dose\ (days) = Start\ date\ of\ AE - Date\ of\ first\ dose + 1$$

The time since the most recent dose of IMP for each AE will be calculated as follows for all TEAEs:

$$Time\ since\ the\ most\ recent\ dose\ (days) = Start\ date\ of\ AE - Date\ of\ most\ recent\ dose + 1$$

Time to first dose and time to most recent dose will not be calculated for pretreatment AEs.

For the calculation of time to first BKZ and time to most recent BKZ dose the same formulas as above will be used but the date of first BKZ dose will be used instead of the date of first dose of IMP. The time to first dose and time to first BKZ dose differs only for those subjects who receive placebo at Baseline after the time they switch to BKZ.

10.2.3 Exposure adjusted incidence rate (EAIR) and exposed adjusted event rate (EAER)

The EAIR is defined as the number of subjects (k) with a specific AE adjusted for the exposure and will be scaled to 100 patient-years:

$$EAIR = 100 \times \frac{k}{\sum_{i=1}^{n}\left(T_{Exp(i)}\right)}$$

where n is the total number of subjects and $T_{Exp(i)}$ is the time of exposure for each subject defined as:

- For subjects with the AE: years since first dose of IMP to the first occurrence for the AE of interest at the level of coding evaluated,
- For subjects without the AE: the time at risk in years.

Exact Poisson 95% CIs for incidence rates are calculated using the relationship between the Poisson and the Chi-square distribution (Ulm, 1990; Fay and Feuer, 1997):

$$LCL = \frac{\chi^2_{2k,\alpha/2}}{2}$$

$$UCL = \frac{\chi^2_{2(k+1),1-\alpha/2}}{2}$$

where k is the number of subjects with a specific AE for the incidence rate of interest and is the basis for the number of the degrees of freedom for the chi-square quantile for the upper tail probability $\chi 2$.

The EAER will be the number of AEs (N_{AE})including repeat occurrences in individual subjects divided by the total time at risk scaled to 100 patient-years and calculated using:

$$EAER = 100 \times \frac{N_{AE}}{\sum_{i=1}^{n}\left(T_{Risk(i)}\right)}$$

where N_{AE} is the number of AE occurrences, n is the total number of subjects and $T_{Risk(i)}$ is the time at risk in years for each subject.

No CI will be computed for EAER.

10.2.4 Other safety topics of interest

The other safety topics of interest for this study include:

- Infections (serious, fungal, opportunistic and tuberculosis)
- Malignancies
- Major adverse cardiac events (MACE)
- Neutropenia
- Suicidal Ideation and Behavior (SIB)Inflammatory bowel disease
- Hypersensitivity (including anaphylaxis)
- Hepatic events and drug-induced liver injury (DILI)

The analyses produced for the above TEAE are based on the specifications described in the BKZ safety topic of interest document version from Apr-2021.

The incidence of other safety topics of interest will be summarized by MedDRA system organ class, high level term, and PT. EAIR with associated 95% CI and the EAER will be included in the summary tables.

Separate tables will be created to summarize the incidence of each of the above AEs (unless otherwise specified) per 100 subject years, by SOC, HLT, and PT. A listing will also be created.

10.2.4.1 Infections (serious, opportunistic, fungal and TB)

Serious infections will be identified based on MedDRA classification (SOC "Infections and infestations") using the "Any SAE" table. A separate table does not need to be produced to summarize these events.

- Fungal infections will be summarized in a stand-alone table. The table will include all TEAEs (serious and non-serious) which code into the High-Level Group Term (HLGT) "Fungal infectious disorders".
- Opportunistic infections (including tuberculosis) will be summarized in a stand-alone table and will include all opportunistic infection TEAEs identified as per the UCB-defined search criteria which were adjudicated as opportunistic infections (refer to Section 12.2 for further details on the process for identifying opportunistic infections).

10.2.4.2 Malignancies

Two tables will be created on malignancies:

- One table will be based on the criteria SMQ= "Malignant or unspecified tumors (SMQ)".
- One table will be based on the criteria SMQ=" Malignant tumors (SMQ)".

SMQ search will include all TEAEs which code to a PT included in the Scope=Narrow group within each SMQ.

Note that the events included in the "Malignancies" table will be a subset of the events included in the "Malignancies (including unspecified)" table. While the "Malignant tumors (SMQ)" is most relevant, "Malignant or unspecified tumors (SMQ)" must be reviewed for potential malignancies.

The output tables will include 2 different overall incidence rows:

- The first overall incidence row will summarize "Any Malignancy" and this row will summarize the incidence of all AEs flagged for inclusion in the table (using the appropriate SMQ depending on the table), regardless of the High Level Term (HLT) it codes to.
- The second overall incidence row will summarize "Any Malignancy excluding non melanomic skin cancers HLT" and this row will summarize the incidence of AEs flagged for inclusion in the table (using the appropriate SMQ depending on the table), excluding those which code to an HLT of "skin neoplasms malignant and unspecified (excl melanoma)".

10.2.4.3 Major adverse cardiac events (MACE)

Major adverse cardiac events will be presented in a table. The classification of an event as MACE is determined by external cardiovascular event adjudication committee.

Extended MACE events will be presented in a separate table and listing. All events which are classified by the adjudication committee as any of the following events types will be considered an extended MACE event.

Table 10–1: Type of Extended MACE events

Event Type Code	Event Type
1	Non-Fatal Myocardial Infarction (MI)
2	Non-Fatal Stroke: hemorrhagic
3	Non-Fatal Stroke: ischemic
4	Non-Fatal Stroke: embolic
5	Non-Fatal Stroke: undeterminable
6	Hospitalization or ER for Unstable Angina with urgent revascularization
8	Hospitalization for Heart Failure
10	Coronary Revascularization Procedures (eg, percutaneous coronary intervention, coronary artery bypass grafting)
11	Urgent Revascularization Procedures (ie, due to symptoms of brain ischemia or pending infarction)
18	Death due to Myocardial Infarction (MI)
19	Death due to Stroke
20	Sudden Cardiac Death

Event Type Code	Event Type
21	Other CV Death (eg, heart failure, pulmonary embolism, cardiovascular procedure-related)
22	Cardiovascular: Undetermined Cause of Death (ie, cause of death unknown)

A separate table and listing will present the adjudicated cardiovascular events by type. For each cardiovascular event type (24 in all, for full list refer to Bimekizumab-Event-Adjudication-Committee-Analysis- Plan), the individual PTs which fall within each event type will be summarized.

Additionally, a listing of all events identified for potential review by the cardiovascular event adjudication committee will be produced. This listing will indicate whether each event was escalated to the committee for formal review/adjudication.

10.2.4.4 Neutropenia

The table on neutropenia will be based on the following PTs (regardless of seriousness):

- Autoimmune neutropenia
- Band neutrophil count decreased
- Cyclic neutropenia
- Febrile neutropenia
- Idiopathic neutropenia
- Neutropenia
- Neutropenic infection
- Neutropenic sepsis
- Neutrophil count decreased

10.2.4.5 Suicidal Ideation and Behavior Neuropsychiatric events

An external neuropsychiatric adjudication committee will evaluate potential neuropsychiatric events and determine whether any of those events were associated with suicidal ideation and behavior (SIB). If an event is adjudicated as SIB, further information will be provided. A table and a listing for SIB events as determined by the adjudication committee will be produced.

A separate table will present the adjudicated neuropsychiatric events by type. For each neuropsychiatric event type (6 total), the individual PTs which fall within each event type will be summarized. It will include events adjudicated as SIB and events adjudicated as non-suicidal. Note that the event type Suicidal ideation can be classified as either SIB or non-suicidal.

Additionally, a listing of all events identified for potential review by the neuropsychiatric adjudication committee will be produced. This listing will indicate whether each event was escalated to the committee for formal review/adjudication. A separate listing will also be produced to summarize the adjudicated results of all events escalated to the full committee.

10.2.4.6 Inflammatory bowel disease

An external inflammatory bowel disease (IBD) adjudication committee will evaluate potential IBD events and will classify each one as follows:

- Event Type Code 1: Possible IBD – Crohn's Disease
- Event Type Code 2: Probable IBD – Crohn's Disease
- Event Type Code 3: Definite IBD – Crohn's Disease
- Event Type Code 4: Possible IBD – Ulcerative Colitis
- Event Type Code 5: Probable IBD – Ulcerative Colitis
- Event Type Code 6: Definite IBD – Ulcerative Colitis
- Event Type Code 7: Possible IBD – Unclassified
- Event Type Code 8: Probable IBD – Unclassified
- Event Type Code 9: Definite IBD – Unclassified
- Event Type Code 10: Symptoms not consistent with IBD
- Event Type Code 11: Possible Inflammatory Bowel Disease – Microscopic Colitis
- Event Type Code 12: Probable Inflammatory Bowel Disease – Microscopic Colitis
- Event Type Code 13: Definite Inflammatory Bowel Disease – Microscopic Colitis
- Event Type Code 14: Possible Inflammatory Bowel Disease – no further differentiation possible
- Event Type Code 15: Probable Inflammatory Bowel Disease – no further differentiation possible
- Event Type Code 16: Definite Inflammatory Bowel Disease – no further differentiation possible
- Event Type Code 99: Not enough information to adjudicate

A table for adjudicated IBD events (event type codes 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15 and 16) as determined by the adjudication committee will be produced. It will summarize events determined by the adjudication committee as definite IBD (event type codes 3, 6, 9, 13, and 16), probable IBD (event type codes 2, 5, 8, 12, and 15) and possible IBD (event type codes 1, 4, 7, 11, and 14). Definite and probable IBD will also be aggregated and summarized. This table will be stratified by subjects with or without a previous medical history of IBD. Previous medical history of IBD will be determined using the information recorded on the Extra-Articular Assessment at Screening CRF page ("Does subject have a history of IBD?").

A separate table will present the adjudicated gastrointestinal events by type. For each gastrointestinal event type (17 total), the individual PTs which fall within each event type will be summarized. It will include events determined by the adjudication committee as definite IBD

probable IBD and possible IBD. It will also include events determined as Symptoms not consistent with IBD (event type code 10) and Not enough information to adjudicated (event type code 99).

A listing of all events identified for potential review by the IBD adjudication committee will be produced. This listing will indicate whether each event was escalated to the committee for formal review/adjudication.

A listing will present the adjudicated IBD events by type. For each IBD event type (event type codes 1 through 16 and 99; 17 total), the individual PTs which fall within each event type will be listed.

A third listing will present the individual diagnostic criteria met for each adjudicated IBD event.

10.2.4.7 Hypersensitivity (including Anaphylaxis)

A separate table will be prepared based on the MedDRA anaphylaxis Algorithm (see Section 12.1) for acute anaphylactic events (reported on the same day as when an injection was administered or 1 day after). An AE glossary table will also be produced to summarize the MedDRA coding for these events. The glossary table will include the following fields: reported term, PT, LLT, HLT, and SOC.

A separate table will be prepared to summarize all hypersensitivity events and serious hypersensitivity events, identified using the SMQ "Hypersensitivity (SMQ)". All TEAEs which code to a PT included in the Scope=Narrow search will be included in this table. In addition, a separate table will be prepared to summarize serious hypersensitivity events, identified using the SMQ "Hypersensitivity (SMQ)". All serious TEAEs which code to a PT included in the Scope=Narrow search will be included in this table. An AE glossary table will also be produced to summarize the MedDRA coding for these events. The glossary table will include the following fields: reported term, PT, LLT, HLT, and SOC.

Furthermore, a separate table will be prepared to summarize injection site reactions, identified by looking under the HLTs "Administration site reactions NEC" and "Injection site reactions".

10.2.4.8 Hepatic events and DILI

The table for hepatic events will be created based on the SMQ of "Drug related hepatic disorders - comprehensive search (SMQ)" excluding the 2 sub-SMQs "Liver neoplasms, benign (incl cysts and polyps) (SMQ)" and "Liver neoplasms, malignant and unspecified (SMQ)". For each of the above SMQs, include all TEAEs which code to a PT included in the Scope=Broad and/or Scope=Narrow.

Note that all AEs meeting the above criteria are to be included. It should be limited to events that the investigator determined to be related to study drug.

Also, potential Hy's Law cases will be reported separately in a liver function test table (see Section 10.3)

10.2.5 COVID-19 impact

To assess the impact of COVID-19 pandemic on safety, additional summaries and listings will be presented.

For reporting purposes, AEs will be assigned to either Pre-COVID-19, COVID-19, or post-COVID-19 pandemic period by comparing the AE start date (based on imputed date) to the COVID-19 pandemic period dates (AE allocated to a period if it starts during the period):

The following categories of TEAEs will be summarized by MedDRA SOC, HLT and PT, including EAIR and EAER.

- All TEAEs by time of onset relative to COVID-19 pandemic period (pre – during – post) by region and overall,
- All TEAEs leading to study discontinuation and/or permanent withdrawal of study medication by time of onset relative to COVID-19 pandemic period (pre – during – post) by region and overall,
- All Serious TEAEs by time of onset relative to COVID-19 pandemic period (pre – during – post) by region and overall,
- All COVID-19 related TEAEs by treatment group, by region and overall – COVID-19 related TEAEs will be identified by the two preferred terms "Corona virus infection" and "coronavirus test positive" and will include confirmed or suspected COVID-19 infections.

A listing of COVID-19 related AEs will be presented where these AEs are identified as described above. Additionally, the listing of AEs will include a column for COVID-19 relatedness and the time of onset of each AE relative to the COVID-19 pandemic will be included in all AE listings.

For the purpose of calculating EAIR and EAER by COVID-19 pandemic period, the calculation of exposure time at risk presented in Section 10.1.2 will be used. An individual subject may therefore be counted in the denominator for several COVID-19 pandemic periods dependent on whether the subject is still considered at risk on the COVID-19 pandemic start and stop dates. In this case time at risk will be calculated separately for each period.

10.3 Clinical laboratory evaluations

The list of the laboratory measurements/variables that will be evaluated in the study is specified in Table 10–2 per laboratory category.

Table 10–2: Laboratory measurements

Hematology	Biochemistry	Urinalysis
Basophils	Calcium	pH
Eosinophils	Chloride	Albumin (protein)
Lymphocytes	Magnesium	Glucose
Monocytes	Potassium	Blood
Neutrophils	Sodium	Leukocyte esterase
Hematocrit	Glucose	Nitrite
Hemoglobin	BUN	Urine dipstick for pregnancy testing[b]
MCH	Creatinine	Urine drug screen[c]
MCHC	hs-CRP	
MCV	AST	

Hematology	Biochemistry	Urinalysis
Platelet count	ALT	
RBC count	GGT	
WBC count	ALP	
	Total bilirubin	
	Triglycerides[a]	
	Cholesterol[a]	
	HDL cholesterol[a]	
	LDL cholesterol[a]	
	LDH	
	Serum pregnancy testing[b]	

ALP=alkaline phosphatase; ALT=alanine aminotransferase; AST=aspartate aminotransferase; BUN=blood urea nitrogen; CRP=C-reactive protein; ET=early termination; GGT=gamma-glutamyltransferase; HDL=high density lipoprotein, IMP=investigational medicinal product; LDH=lactate dehydrogenase; MCH=mean corpuscular hemoglobin; MCHC=mean corpuscular hemoglobin concentration; MCV=mean corpuscular volume; RBC=red blood cell; SFU=Safety Follow-Up; WBC=white blood cell.

[a] Biochemistry will include triglycerides, cholesterol, HDL cholesterol, and LDL cholesterol at Baseline, Weeks 12 and 24, and at the ET Visit.

[b] Pregnancy testing will consist of serum testing at the Screening Visit for all women of childbearing potential. The pregnancy test will be urine at all other visits.

[c] Urine drug screen will be performed at Screening Visit.

Biochemistry and hematology values will be flagged based on their reference range:

- Values that are below the lower limit of the reference range will be flagged as ‘L’ (low)
- Values that are above the upper limit of the reference range will be flagged as ‘H’ (high)
- All other values will be referenced as normal.

Markedly abnormal values for blood chemistry and hematology will be defined as laboratory values graded 3 or 4 according to the Rheumatology Common Toxicity Criteria (RCTC). Definitions of the markedly abnormal values are given in Table 10–3 and Table 10–4 and are based on the RCTC units. All units in the tables below will be converted to the standard units based on Clinical Data Interchange Standards Consortium (CDISC) standards.

Table 10–3: Definitions of markedly abnormal biochemistry values

Parameter name	Unit	Criteria	Abnormal designation
Creatinine	mmol/L	>3.0 x ULN or >3.0 x Baseline	AH
Glucose	mmol/L	<2.2 >13.9	AL AH
Calcium	mmol/L	>3.1 <1.75	AH AL

Parameter name	Unit	Criteria	Abnormal designation
Magnesium	mmol/L	>1.23 <0.4	AH AL
Potassium	mmol/L	>6.0 <3.0	AH AL
Sodium	mmol/L	>155 <130	AH AL
Triglyceride	mmol/L	>5.7	AH
Cholesterol	mmol/L	>10.34	AH

AH=abnormal high; AL=abnormal low; dL = deciliter; L = liter;
mg = milligram; mmol = millimoles; ULN = upper limit of normal.

Table 10–4: Definitions of markedly abnormal hematology values

Parameter name	Unit	Criteria	Abnormal Designation
Hemoglobin	g/L	<80 >40 above ULN or >40 above Baseline if Baseline is above ULN	AL AH
Lymphocytes Absolute	10^9/L	<0.5 >20.0	AL AH
Neutrophils Absolute	10^9/L	<1.0	AL
Platelets	10^9/L	<50	AL
WBC/ Leukocytes	10^9/L	<2.0 >100	AL AH

Definitions of the markedly abnormal values for liver enzyme elevation are given in Table 10–5.

Table 10–5: Definitions of markedly abnormal liver enzyme elevation values

Parameter name	Unit	Criteria	Abnormal Designation
Alkaline Phosphatase		>5.0 x ULN	AH
ALT	U/L	>5.0 x ULN	AH
AST	U/L	>5.0 x ULN	AH
Total Bilirubin	umol/L	>3.0 x ULN	AH
GGT	U/L	>5.0 x ULN	AH

ALT= alanine aminotransferase; AST = aspartate aminotransferase; GGT=gamma-glutamyltransferase.

All laboratory summaries will be presented in standard international (SI) units and will be based on observed case values. In the case where laboratory values are below the lower limit of quantification, then these will be set to the midpoint between 0 and the lower limit of quantification for the purpose of summarizing the data.

In the case where laboratory data are expressed as above a fixed value (ie,: ">xx.xx"), then the numeric portion of the value will be used (ie, : xx.xx) to present and calculate statistics.

To define the minimum/maximum post-Baseline values, all post-Baseline assessments including those at unscheduled visits (but excluding any that occur more than 140 days after the last administration of study medication) will be used.

Different summary tables for hematology, biochemistry variables (except hs-CRP which will be analyzed in the efficacy section) and urinalysis (pH, Albumin (protein), glucose, blood, leukocyte esterase and nitrite) will be provided in the SS. For shift tables, results for the treatment period will be presented together for the overall study period.

As indicated in Section 3.1, for tables where data are summarized by visit only values occurring at scheduled visits will be included.

For each laboratory parameter, the below results will be presented for the overall study period, by treatment group:

- Observed values at Baseline and observed values and changes from Baseline at each scheduled post-Baseline visit.
- Number and percentage of subjects by CTCAE grade (version 4.03) (when applicable) based on the minimum/maximum post-Baseline value (for blood chemistry and hematology).
- Shift in CTCAE grade from Baseline to minimum/maximum post-Baseline (for blood chemistry and hematology). Subjects who meet the decreased potassium criterion of 3.0-<LLN, which is specified as the decreased potassium lab criterion for both CTCAE Grade 1 and Grade 2, will be counted as Grade 2.
- Number and percentage of subjects meeting the below criteria at any time during the study:
  - AST: >2xULN, >3xULN, >5xULN, >10xULN, >20xULN

  - ALT: >2xULN, >3xULN, >5xULN, >10xULN, >20xULN
  - AST or ALT: >2xULN, >3xULN, >5xULN, >10xULN, >20xULN
  - Bilirubin: >1xULN, >1.5xULN
  - ALP: >2xULN
- Number and percentage of subjects with potential Hy's Law cases based on the two following definitions:
  - [AST ≥3xULN or ALT ≥3xULN] and Total Bilirubin ≥2xULN
  - [AST ≥3xULN or ALT ≥3xULN] and Total Bilirubin ≥2xULN in the absence of ALP ≥2xULN

In order to meet either of the above criteria, a subject must experience the elevation in bilirubin and ALT or AST (and the absence of ALP elevation, if applicable) at the same visit.

The below results will be presented on the SS. All values observed at any time while on treatment (ie, occurring at scheduled and unscheduled visits) will be included in the below tables. Baseline values and values observed more than 140 days after the last administration of study medication are not considered:

- Number and percentage of subjects with treatment emergent markedly abnormal laboratory data overall (for blood chemistry and hematology). All values observed at any time while on treatment (ie, occurring at scheduled and unscheduled visits) will be included in this table summarizing markedly abnormal values. For this summary, Baseline values and values observed more than 140 days after the last administration of study medication are not considered.
- Number and percentage of subjects with markedly abnormal liver enzyme elevation tests (data beyond the CTCAE Grade 3 thresholds as outlined in Table 10–5)

All laboratory data will be listed by treatment, subject and visit including changes from Baseline for numeric variables, flags for measurements outside the normal ranges, the relative study day, a flag for whether the test was not done and a flag for whether the subject was fasting.

The markedly abnormal laboratory data will also be listed separately.

Listings will include Hepatitis B and C, human immunodeficiency virus (HIV), genomic proteomic/metabolomics, and genetic/epigenetic tests.

In addition, the laboratory results classified as Grade 3 or Grade 4 will be listed separately.

Urinalysis laboratory results and hepatic events will be listed separately.

10.4 Vital signs, physical findings, and other observations related to safety

10.4.1 Vital signs, weight and physical findings

The following results will be tabulated for all vital sign variables (systolic blood pressure (mmHg), diastolic blood pressure (mmHg), and pulse rate (beats/min), by treatment group and visit:

- A summary of the absolute and change from Baseline value

- A summary of the number and percentage of subjects experiencing at least one markedly abnormal value for a vital sign variable as defined in Table 10–6.

Table 10–6: Definitions of markedly abnormal blood pressure values

Parameter (unit)	Markedly abnormal low	Markedly abnormal high
Systolic blood pressure (mmHg)	<90 and a decrease from Baseline of ≥20	>180 and an increase from Baseline of ≥20
Diastolic blood pressure (mmHg)	<50 and a decrease from Baseline of ≥15	>105 and an increase from Baseline of ≥15

For the Baseline visit where several BP assessments are taken, the individual observed data assessments as well as the mean of these assessments will be summarized and listed (pre-dose, 30 min post-dose and 1-hour post-dose).

Treatment-emergent markedly abnormal values will be analyzed on the SS.

Vital signs measurements will also be listed by visit and timing relative to dosing including changes from Baseline. The listing will also include details of abnormal value based on flags as defined in Table 10–6 ("L" for markedly abnormal low and "H" for markedly abnormal high).

Similarly, physical examination findings together with the details of abnormalities (when applicable) will be listed by treatment group, subject and visit.

Body weight will also be listed.

10.4.2 Electrocardiograms

Since it is only planned to assess electrocardiogram (ECG) at the Screening Visit, the Baseline value will be derived based on the Screening assessment.

A summary of the absolute and change from Baseline values at Week 16 and at SFU in each ECG variable will also be created by treatment group. The following ECG variables will be assessed:

- Heart Rate (bpm)
- PR interval (ms)
- RR interval (ms)
- QRS duration (ms)
- QT interval (ms)
- QTcF interval (ms)
- QTcB interval (ms)

The date and time of the ECG will be recorded in the eCRF together with the Investigator interpretation. A summary of the number and percentage of subjects with normal, abnormal not clinically significant, and abnormal clinically significant ECG results will be presented for Baseline, Week 16 and SFU timepoints.

QTc outliers are defined as QTcF (or QTcB) values following dosing that are greater than 450 ms or are increases from Baseline greater than 30 ms. QTcF and QtcB outliers will be summarized in a dedicated table using the following categories:

- Values >450 ms, >480 ms, >500 ms
- Increase from Baseline of >30 ms, >60 ms, >90 ms
- Values >450 ms and increase of >30 ms. Values >500 ms and increase from Baseline >60 ms.

Number and percentage of subjects who meet the ECG outlier criteria at any assessment post first dose will be summarized.

All ECG tables will be provided on the SS by randomized treatment.

All 12-lead ECG data where QTcF and QTcB outliers are highlighted will also be listed.

10.5 Other safety variables

10.5.1 Electronic Columbia Suicide Severity Rating Scale (eC-SSRS)

The eC-SSRS is an assessment tool that evaluates suicidal ideation and behavior.

The eC-SSRS contains 9 categories with binary responses (yes/no):

- Category 1 - Wish to be dead
- Category 2 - Non-specific active suicidal thoughts
- Category 3 - Active suicidal ideation with any methods (not plan), without intent to act
- Category 4 - Active suicidal ideation with some intent to act, without specific plan
- Category 5 - Active suicidal ideation with specific plan and intent
- Category 6 - Preparatory acts or behavior
- Category 7 - Aborted attempt
- Category 8 - Interrupted attempt
- Category 9 - Actual attempt

Following composite endpoints based on the above categories are defined as:

- Suicidal ideation: A “yes” answer at any time during treatment to any one of the five suicidal ideation questions (Categories 1-5).
- Suicidal behavior: A “yes” answer at any time during treatment to any one of the four suicidal behavior questions (Categories 6-9).
- Suicidal ideation or behavior: A “yes” answer at any time during treatment to any one of the nine suicidal ideation and behavior questions (Categories 1-9).

Self-injurious behavior without suicidal intent is defined as event in the category non-suicidal self-injurious behavior (in category 9).

The visit identification for each assessment of eC-SSRS will not be part of the data collected and the visit date in eCRF will then be used to assign a visit identification to the eC-SSRS data.

For each subject, eC-SSRS assessment dates will be compared to eCRF visit dates.

In the Study Data Tabulation Model (SDTM):

If an eC-SSRS assessment date is matching an eCRF scheduled visit date, then the eC-SSRS visit identification will be the one corresponding to eCRF visit date. Otherwise, the eC-SSRS visit identification will be considered as "unscheduled".

In the analysis data sets (ADaM):

If an eC-SSRS assessment date is matching an eCRF scheduled visit date within a window of +/- 3 days, then the eC-SSRS visit identification will be the one corresponding to the eCRF visit date. Otherwise, the eC-SSRS visit identification will stay as "unscheduled".

The incidence of subjects with suicidal ideation, suicidal behavior, suicidal behavior or ideation, and self-injurious behavior without suicidal intent will be summarized by treatment and by visit for the SS. A by-subject listing of the eC-SSRS questionnaire data will be provided.

10.5.2 Assessment of tuberculosis

The laboratory test results (negative, positive or indeterminate) from the TB assessment performed by interferon gamma release assay will be summarized by treatment group for Baseline (Screening) and Week 12 timepoints, as well as be listed.

The results from the 'Evaluation of signs and symptoms of tuberculosis' questionnaire data will be also be listed.

10.5.3 Change from Baseline in the Patient Health Questionnaire-9 (PHQ-9)

The PHQ-9 is a multipurpose instrument for Screening, diagnosing, monitoring, and measuring the severity of subject's depression. The PHQ-9 score is based on nine questions assessing the depression over the last two weeks. Each criterion is scored from 0 (not at all) to 3 (nearly every day). The total score for the nine questions is added up and provide the PHQ-9 score and can range from 0 to 27 with higher scores indicating worse state. If one of the nine criteria is missing, the PHQ-9 score will be set to missing.

The following depression states are defined based on the PHQ-9 score:

- A score of 0-4 is considered to be none or minimal depression severity
- A score of 5-9 is considered to be mild depression severity
- A score of 10 to 14 is considered to indicate moderate depression severity
- A score of 15 to 19 is considered to indicate moderately severe major depression
- A score ≥20 is considered to be severe major depression

The percentage of subjects with PHQ-9 scores below 5, between 5 and 9, between 10 and 14, between 15 and 19, and 20 or greater will be summarized by visit and treatment group, as well as the change from Baseline over time.

Only observed data will be presented for PHQ-9 results.

11 REFERENCES

Fleiss JL, Tytun A, Ury SHK. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics. 1980;36:343-6.

Moser, BK; Stevens, GR; Watts, CL. The Two-Sample T Test Versus Satterthwaite's Approximate F Test. Communication in Statistics- Theory and Methods 18(11):3963-3975 ·

Mallinckrodt CH. Preventing and Treating Missing Data in Longitudinal Clinical Trials: A Practical Guide. New York, NY: Cambridge University Press; 2013.

O'Kelly, Michael, and Bohdana Ratitch. Clinical trials with missing data: A guide for practitioners. John Wiley & Sons, 2014.

Rubin, Donald B. "The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm." Journal of the American Statistical Association 82, no. 398 (1987): 543-546.

Eberl G, Studnicka-Benke A, Hitzelhammer H, et al. Development of a disease activity index for the assessment of reactive arthritis (DAREA). Rheumatology (Oxford) 2000;39:148–55.

Ulm K. A simple method to calculate the confidence interval of a standardized mortality ratio. American Journal of Epidemiology 1990;131(2):373-375.

Hosmer, D. W., Jr., and Lemeshow, S. (2000). Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons.

Maruish M, Kosinki M, Bjorner J, Gandek B, Turner-Bowker D, Ware J. User's mamual for the SF36v2 Health Survey.

12 APPENDICES

12.1 Appendix1: MedDRA algorithmic approach to anaphylaxis

The SMQ Anaphylactic reaction consists of 3 parts:

- A narrow search containing PTs that represent core anaphylactic reaction terms (Cat A)

- SMQ Anaphylactic reaction (SMQ)
  - PT Anaphylactic reaction
  - PT Anaphylactic shock
  - PT Anaphylactic transfusion reaction
  - PT Anaphylactoid reaction
  - PT Anaphylactoid shock
  - PT Circulatory collapse
  - PT Dialysis membrane reaction
  - PT Kounis syndrome
  - PT Shock
  - PT Shock symptom
  - PT Type I hypersensitivity

- A **broad search** that contains additional terms that are added to those included in the narrow search. These additional terms are signs and symptoms possibly indicative of anaphylactic reaction and categorized in B, C, or D

Cat B

- PT B Acute respiratory failure
- PT B Asthma
- PT B Bronchial oedema
- PT B Bronchospasm
- PT B Cardio-respiratory distress
- PT B Chest discomfort
- PT B Choking
- PT B Choking sensation
- PT B Circumoral oedema
- PT B Cough
- PT B Cyanosis
- PT B Dyspnoea
- PT B Hyperventilation
- PT B Irregular breathing
- PT B Laryngeal dyspnoea
- PT B Laryngeal oedema
- PT B Laryngospasm
- PT B Laryngotracheal oedema
- PT B Mouth swelling
- PT B Nasal obstruction
- PT B Oedema mouth
- PT B Oropharyngeal spasm
- PT B Oropharyngeal swelling
- PT B Respiratory arrest
- PT B Respiratory distress
- PT B Respiratory dyskinesia
- PT B Respiratory failure
- PT B Reversible airways obstruction
- PT B Sensation of foreign body
- PT B Sneezing
- PT B Stridor
- PT B Swollen tongue
- PT B Tachypnoea
- PT B Throat tightness
- PT B Tongue oedema
- PT B Tracheal obstruction
- PT B Tracheal oedema
- PT B Upper airway obstruction
- PT B Wheezing

Cat C

- Allergic oedema
- Angioedema
- Erythema
- Eye oedema
- Eye pruritus
- Eye swelling
- Eyelid oedema
- Face oedema
- Flushing
- Generalised erythema
- Injection site urticaria
- Lip oedema
- Lip swelling
- Nodular rash
- Ocular hyperaemia
- Oedema
- Periorbital oedema
- Pruritus
- Pruritus allergic
- Pruritus generalised
- Rash
- Rash erythematous
- Rash generalised
- Rash pruritic
- Skin swelling
- Swelling
- Swelling face
- Urticaria
- Urticaria papular

Cat D

- Blood pressure decreased
- Blood pressure diastolic decreased
- Blood pressure systolic decreased
- Cardiac arrest
- Cardio-respiratory arrest
- Cardiovascular insufficiency
- Diastolic hypotension
- Hypotension

- An **algorithmic approach** which combines a number of anaphylactic reaction symptoms in order to increase specificity. A case must include either **(on the same day):**
  - A narrow term or a term from Category A;
  - A term from Category B - (Upper Airway/Respiratory) AND a term from Category C - (Angioedema/Urticaria/Pruritus/Flush);
  - A term from Category D - (Cardiovascular/Hypotension) AND [a term from Category B - (Upper Airway/Respiratory) OR a term from Category C - (Angioedema/Urticaria/ Pruritus/Flush)]
- Hypersensitivity events will be identified using the “Hypersensitivity (SMQ)”. All TEAEs which code to a PT included in the Scope=Narrow search will be included.

12.2 Appendix2: UCB-defined search criteria for identifying Opportunistic infections

Opportunistic infections are identified in two steps:

Step 1: Refer to column B of the UCB Opportunistic infections document (Opportunistic infections MedDRA v 19.xlsx) which identifies the Preferred Terms (PTs) to be classified as opportunistic infections using either a single 'x' or a double 'xx'.

TEAEs which code to a PT flagged with a single 'x' need to also be serious to be considered an opportunistic infection.

All TEAEs which code to a PT flagged with a double 'xx' are considered to be an opportunistic infection, regardless of seriousness.

Step 2: Refer to column C of the spreadsheet which identifies the PTs that need to be evaluated on a case-by-case basis by the study physician to determine whether or not it is an opportunistic infection. If Column C has a single 'x', then the corresponding preferred term should be flagged for case-by-case review by the study physician.

13 AMENDMENT(S) TO THE STATISTICAL ANALYSIS PLAN (SAP)

13.1 Amendment 1

Rationale for the amendment

The primary purpose of this amendment was to implement changes in response to protocol amendments and discussions and feedback provided at meetings between UCB and Parexel technical teams for procedural clarifications and due to COVID-19 global pandemic impact. The main changes are changing the statistical hierarchy, rules for handling missing data and guidelines on the implementation of multiple imputation, modifying the secondary variables, and incorporate latest guidelines from the BKZ AE of special monitoring convention document.

Modifications and changes

Global changes

- Study participants was changed to subjects
- Re-classification of Change from Baseline in Patient Health Questionnaire-9 (PHQ-9) variable from other efficacy variable to other safety variable
- New rules for handling of missing data for the primary and key secondary efficacy variables
- Use of latest guidelines from the BKZ AE of special monitoring convention document
- Clarification on use and multiple imputation approaches in supportive analyses
  - Observed cases
  - NRI
  - Reference Based imputation
- Imputation of partial start dates
- FACIT-F has been corrected to FACIT-Fatigue subscale, as it is the fatigue subscale that is being assessed in this study
- Deleted Asia/Australia to represent only Asia now which includes Australia

Specific changes

Change #1

Section 1 Introduction

The SAP is based on the final Protocol (26 NOV 2018)

Has been changed to:

The SAP is based on the Protocol Amendment 2 (1 APR 2021).

Change #2

Section 2.2.2.1 Secondary efficacy variables

The following has been changed:

Was section 2.2.1.2

- PASI90 response at Week 4 and Week 16 in the subgroup of subjects with PSO involving at least 3% BSA at Baseline
- Change from Baseline in Health Assessment Questionnaire—Disability Index (HAQ-DI) at Week 16
- Change from Baseline in the Short-Form 36-item Health Survey (SF-36) Physical Component Summary (PCS) at Week 16

Etc

Changed ordering to:

Section 2.2.2.1

- Change from Baseline in Health Assessment Questionnaire—Disability Index (HAQ-DI) at Week 16
- PASI90 response at Week 4 and Week 16 in the subgroup of subjects with PSO involving at least 3% BSA at Baseline
- Change from Baseline in the Short-Form 36-item Health Survey (SF-36) Physical Component Summary (PCS) **score** at Week 16
- Change from Baseline in Psoriatic Arthritis Impact of Disease-12 (PsAID-12) **total score** at Week 16

Change #3

Section 2.2.2.2 Secondary Safety variables

The section was moved from section 2.2.5.1 to 2.2.2.2 and the following was updated:

Treatment-emergent **a**dverse events leading to withdrawal from IMP

Change #4

Section 2.2.3.1 Other efficacy variables

The section was moved from section 2.2.1.3 to 2.2.3.1 and the following was added:

- Composite endpoint composed of ACR50 and PASI100 response in subjects with PSO involving at least 3% BSA at Baseline
- Proportion of Psoriatic Arthritis Response Criteria (PsARC) responders
- Psoriatic Arthritis Disease Activity Score (PASDAS) categories
- Change from Baseline in the PASDAS
- Disease Activity Index for Psoriatic Arthritis (DAPSA) score **categories**
- Proportion of subjects with a decrease of HAQ-DI from Baseline of at least 0.35 in the subgroup of subjects with Baseline HAQ-DI≥0.35
- Proportion of FACIT-Fatigue subscale responders (subjects with a minimum clinically important difference for FACIT-Fatigue subscale score, defined as an increase of ≥4) in subjects with FACIT-Fatigue subscale score ≤48 at Baseline

Change #5

Section 2.2.3.1 Other efficacy variables

- Disease Activity Index for Psoriatic Arthritis (DAPSA) score
- Change from Baseline PsAID-12
- Change from Baseline in the Psoriatic Arthritis Quality of Life (PsAQoL)
- Change from Baseline in Functional Assessment of Chronic Illness Therapy—Fatigue (FACIT-F)

Changed to:

- Disease Activity Index for Psoriatic Arthritis (DAPSA) score **categories**
- Change from Baseline in PsAID-12 **total score**
- Change from Baseline in the individual domain scores of PsAID-12
- Proportion of subjects achieving PsAID-12 total score ≤4
- Proportion of PsAID-12 responders (decrease ~~Change~~ from Baseline in PsAID-12 total score ≥3) ~~(PsAID-12 responders)~~ in subjects with PsAID-12 total score >3 at Baseline
- Change from Baseline in the Psoriatic Arthritis Quality of Life (PsAQoL) **total score**
- Change from Baseline in Functional Assessment of Chronic Illness Therapy (FACIT)—Fatigue ~~(FACIT-F)~~ **subscale score**

Change #6

Section 2.2.5.2 Other safety variables

Has been moved to

Section 2.2.3.2 Other safety variables

Change #7

Moved PHQ-9 variable

Section 2.2.1.3 Other efficacy variables

Moved to section 2.2.3.2 Other safety variables

Change #8

Moved and updated the following:

Section 2.2.2 Pharmacokinetic variable

Section 2.2.3 Pharmacogenomic variables

Section 2.2.4 Immunological variables

Moved to:

Section 2.2.3.3 Pharmacokinetic variable

Section 2.2.3.4 Pharmacogenomic variables

Where local regulations permit, additional blood samples will be collected at specific time points from study participants consenting to participate in a pharmacogenomic substudy. These samples will be stored at -80°C for up to 20 years to allow for potential, exploratory analyses of genomic, genetic, epigenetic, proteins, and metabolite biomarkers to evaluate the relationship with response to treatment with bimekizumab, psoriatic arthritis disease biology, and inflammatory and immune response processes.

The pharmacogenomic variables are the blood or blood derivative (eg, serum) concentrations of cytokines and chemokines of relevance to IL-17A/F signaling pathway and psoriatic arthritis biology and will be assessed at Baseline and at Week 16.

These variables will be described in a separate statistical analysis plan. The nature and format of these substudy analyses will be determined at a later date.

Updated to the following:

Additional blood samples will be collected from subjects who consent to participate in the sub study at specific time points and stored at -80°C for up to 20 years.
Genomic, genetic, epigenetic, proteins, and metabolite biomarkers may be measured to evaluate the relationship with response to treatment with BKZ, PsA disease biology, and inflammatory and immune response processes. The nature and format of these tentative sub study analyses will be determined when the results of the main study are made available.

The candidate exploratory variables are the blood or blood derivative (eg, serum) concentrations of cytokines and chemokines of relevance to IL-17A/F signaling pathway and PsA biology. Additional variables may include but will not be limited to serum complement concentrations

Section 2.2.3.5 Other Immunological variables

Added the following:

- The neutralizing anti-drug antibody (NAb) status

These 3 variables will be assessed at all visits except Screening

Change #9

Moved section 3.2.2 to section 2.3.2 Study Periods

The following study periods are defined for the classification by study period:

Screening Period (≥14 days to ≤35 days)

Starts at time of the informed consent date (Screening Visit -Visit 1) and ends the day before the first dose administration of study drug (Baseline Visit – Visit 2) (ie, generally the day before Visit 2 date). This period should last between 14 to 35 days and will involve obtaining laboratory data and verifying that the doses of nonsteroidal anti-inflammatory drugs (NSAIDs) or permitted DMARDs, if used to treat PsA, are stable. The Screening Period will also enable washout of any medications not permitted for use during the study.

Double-Blind Treatment Period (16 weeks)

Starts on the day of first dose administration of study drug (Visit 2) and ends as described in Section 3.2.1.2. This period should last 16 weeks maximum. During the Double-Blind Treatment Period, subjects will be randomized 2:1 (stratified by region and prior TNFα inhibitor exposure [inadequate response to 1 or 2 prior TNFα inhibitors, or intolerance to TNFα inhibitors]) to receive 1 of 2 blinded treatment regimens:

- Bimekizumab 160mg sc Q4W,
- Placebo

Visit windows of ±3 days are allowed from the first dose at all visits through Week 16. The visit window is relative to the Day 1 (Baseline) Visit.
BKZ and placebo will be administered sc by unblinded study personnel at the clinical site.
The time between IMP doses should be ≥21 days and ≤35 days during the Double-Blind Treatment Period.
Subjects discontinuing IMP during the Double-Blind Treatment Period will return for all scheduled visits through Week 16 and the SFU Visit (20 weeks after the final dose of IMP) as applicable. Subjects withdrawing from the study will have an Early Termination (ET) Visit and a SFU Visit 20 weeks after the final dose of IMP, as applicable.

The end of the 16-week Double-Blind Treatment Period will be either the date of Week 16 visit (Visit 6) for subjects completing the treatment period, or the date of the early termination (ET) visit for subjects who discontinued during the Treatment Period.
If a subject does not have a Visit 6 or ET visit, then either the date of the last scheduled or unscheduled visit during the Treatment Period or the date of last known dose of study drug during the Treatment Period, whichever is later, will define the end date of the Treatment Period.

Safety Follow-Up

The SFU Period will be defined for all subjects who complete the study and do not enter the OLE study, or for subjects who discontinue early including those withdrawn from IMP. This period should last 20 weeks.

- For subjects who complete the study and do not enter the OLE study, the SFU Period starts the day after Visit 6 date and ends on the day of the SFU Visit.
- For subjects who discontinue early, the SFU Period starts the day after the ET date and ends on the day of the SFU Visit

A subject will be considered to have completed a study period if they complete the last scheduled study visit for that period. Note that subject participants who previously discontinued IMP and

are continuing for all scheduled visits through Week 16 will also be considered as completed the Double-Blind Treatment Period.

Change #10

Section 2.4 Determination of sample size

Was moved to section 2.6 Determination of sample size and following text was added:

Approximately 390 subjects will be randomly assigned in a 2:1 ratio to the following treatment groups: BKZ 160mg sc Q4W (260 subjects) and placebo (130 subjects).
The primary efficacy analysis is based on the primary comparison of BKZ versus placebo for ACR50 response at Week 16.
All sample size and power calculations were done at a significance level of 0.05.
All sample size and power calculations were performed using the software nQuery Advisor® 7.0.

Change #11

Section 3.1 General presentation of summaries and analyses

The following text was updated:

- Percentages will be summarized based on all subjects in the analysis set and a "Missing" category (corresponding to subjects with missing data for the variable being summarized) will be included as the last row in the list of categories being summarized **(this approach will be used when reporting demographics and Baseline characteristics, some NAb tables and some shift tables from Baseline for laboratory data).**

For hs-CRP variable, the summary statistics should contain **arithmetic mean**, geometric mean, geometric coefficient of variation (CV), median, first and third quartile (Q1 and Q3), minimum and maximum **(for the value and ratio to Baseline)**.

Decimal places for descriptive statistics will always apply the following rules:

- "n" will be an integer.
- Mean, SD, **SE**, and median (Q1 and Q3 where applicable) will use 1 additional decimal place compared to the original data **(for a derived score, number of decimals of the original data considered is the one obtained where deriving the score in the absence of missing data)**.
- Coefficient of variance (CV [%]) will be presented with 1 decimal place.
- Minimum and maximum will have the same number of decimal places as the original raw value.

If the number of decimal places reported in the raw data is varied then either the maximum raw number of reported decimal places or 3 will be used, whichever is the lowest, as a guide for the descriptive statistics.

Table 3.1: List of decimals for continuous efficacy endpoints

Continuous efficacy endpoint	Decimal places used for Minimum and maximum	Decimal places used for Mean, SD (or SE) and median

BASDAI	1	2
DAPSA	1	2
DAS28[CRP]	2	3
EQ-5D-3L dimension scores	0	1
EQ-5D-3L (VAS)	0	1
FACIT-Fatigue subscale score	1	2
HAQ-DI	3	4
hs-CRP	2	3
IGA	1	2
LDI	1	2
LEI	0	1
mNAPSI	1	2
PASDAS	1	2
PASI	0	1
PGA-Arthritis /PhGA-Arthritis	0	1
PGA-PsA / PhGA-PsA	1	2
PsAID-12	1	2
PsAQoL	1	2
PsARC	0	1
PtAAP	2	3
SF-36	1	2
SPARCC	1	2
TJC / SJC	1	2
WPAI-SHP	1	2

BASDAI=Bath Ankylosing Spondylitis Disease Activity Index; DAPSA=disease activity index for psoriatic arthritis; DAS28[CRP]=disease activity score-28 based on C-reactive protein; EQ-5D-3L=Euro-Quality of Life 5-Dimensions 3 Level version; FACIT=Functional Assessment of Chronic Illness Therapy; HAQ-DI=Health Assessment Questionnaire – Disability Index; hs-CRP=High sensitivity C-reactive protein; IGA=Investigator Global assessment; LDI=Leeds Dactylitis Index; LEI=Leeds Enthesitis Index; mNAPSI=Modified Nail Psoriasis Severity Index; PASDAS= Psoriatic Arthritis Disease Activity Score; PASI=Psoriasis Area and Severity Index; PGA-Arthritis=Patient's Global Assessment of Arthritis; PGA-PsA=Patient's Global Assessment of Psoriatic Arthritis; PhGA-Arthritis=Physician's Global Assessment of Arthritis; PhGA-PsA=Physician's Global Assessment of Psoriatic Arthritis; PHQ-9=Patient Health Questionnaire-9; PSAID-12=Psoriatic Arthritis impact of disease-12; PsAQoL=Psoriatic Arthritis Quality of Life; PsARC= Psoriatic Arthritis Response Criteria; PtAAP=Patient's assessment of Arthritis Pain; SF-36=Short-Form 36-item Health Survey; SJC= Swollen Joint Count; SPARCC= Spondyloarthritis Research Consortium of Canada ; TJC=Tender Joint Count; WPAI-SHP=Work Productivity and Activity Impairment questionnaire – Specific Health Problem

Unless stated otherwise, statistical tests of efficacy variables will be performed 2-sided and p-values will be rounded to 3 decimal places. P-values less than 0.001 will be presented as "<0.001" and p-values greater than 0.999 will be presented as ">0.999." Statistical comparison will be performed at the 0.05 level of significance. Least square means (LSM), odds-ratios (OR) and corresponding confidence intervals (CI) will be presented using 3 decimalsUCB uses SAS in a 64-bit Windows environment, and it is well-documented that in this environment the maximum accuracy of any numeric value is 15 significant digits. However, SAS by default does not limit

the accuracy of numeric values to 15 significant digits which, in certain instances, may result in inaccurate representation of the data and cause errors when used in subsequent calculations, particularly when comparing a value to a chosen threshold. This, in turn, could potentially result in a change in classification of a subject from a responder to a nonresponder (and vice versa) if these values occur on a threshold used in the evaluation of response (or a critical laboratory value for example).

Therefore, in order to avoid issues caused by inaccurate floating point representation of numeric values, temporary variables are created (i.e., for absolute values, change and percentage change from Baseline) during programming which are rounded to 12 decimal places prior to comparison to a specific threshold in the derivation of a response parameter. This does not imply inherent rounding on the analysis variables for absolute value, change or percentage of change which are retained unrounded in the final analysis dataset. Thus, rounding is applied exclusively during the derivation of new response parameters or critical value variables, and the rounded values are created on a temporary basis only.

The SAS® outputs supportive of any inferential statistics that are part of the hierarchical testing procedure (ie, all inferential statistics associated with the endpoints in Table 7–4), will be provided as a separate PDF document in addition to TFLs. These outputs will be included in the 'Documentation of Statistical Methods' section of the clinical study report.

The order of treatment groups to be presented in tables from left to right will be:

- Placebo
- Bimekizumab 160mg Q4W

Selected tables may also include columns for all subjects (regardless of study treatment).

The abbreviation for bimekizumab is BKZ and will be used in tables and listings headers. In the TLFs subjects **randomized to** bimekizumab 160mg will be labeled as "BKZ 160mg Q4W".

Change #12

Section 3.2.3 Mapping of assessments performed at Early Termination Visit

The following text was added:

If there is an existing scheduled site visit in the window, then the assessments at the ET visit will be mapped to the next scheduled site visit.

The anti-bimekizumab antibody (ADAb) assessments and PK are exceptions to this rule:

- ADAb levels, PK from an ET visit will be mapped to the next visit where antibody levels / PK are measured.

Change #13

Section 3. 3 Definition of Baseline values

The following text was added:

By default, for randomized subjects for which no start date of treatment is available, the Baseline value will be considered as the last available value prior to the randomization date

Baseline values **for component scores** should be computed using components from the same visit where the relevant measurements were recorded prior to dosing. For example, if the

Screening Visit has all of the components, **and** the Baseline visit is missing one or more components, the Baseline value **for the component score** should be calculated using the Screening **Visit** values.

Chest x-ray performed within 3 months before the Screening Visit is considered as Baseline measurement.

Change #14

Added the following:

Section 3.4 Study Treatment Discontinuation and Intercurrent Event

The concept of intercurrent event is one of the estimand attributes.

Unless otherwise noted, treatment discontinuation due to any reason will be considered as intercurrent event (IE). The IE date will be defined similarly as the treatment discontinuation date below. Since there is no treatment discontinuation date collected, the treatment discontinuation date is defined as:

- Treatment end date + 31 days.

Change #15

Updated the following:

Section 3.4 Protocol deviations

A specific category of PD, as a consequence of the Coronavirus Disease 2019 (COVID-19) pandemic, called COVID-19 related PD will be assessed based especially on the information collected on a dedicated eCRF page (and other available sources).

Change #16

Modified the following:

Section 3.6 Analysis sets

Deleted the following text:

The primary efficacy variable will be analyzed for all study participants in the Randomized Set (RS), and the primary efficacy analysis will be repeated on the Full Analysis Set (FAS) and the PPS. All other efficacy variables will be based on the RS.

Demographics tables will be performed on the RS as well as on the Safety Set (SS), if SS is different from RS. Safety variables will be summarized on the SS. PK variables will be analyzed for all study participants in the PK-PPS.

Updated Section 3.6.2 Randomized Set

The Randomized Set (RS) consists **of all enrolled subjects who have been** randomized**.**

Demographic tables, primary, secondary and other efficacy variables will be presented on the RS.

Updated Section 3.6.3 Safety Set

The Safety Set (SS) consists of **all subjects** who received at least one dose of the IMP.

Demographic tables, study treatment compliance and exposure and safety variables will be presented on the SS. The anti-BKZ antibody will also be analyzed on the SS for subjects receiving BKZ.

Updated Section 3.6.4 Full-Analysis Set

Supportive analysis of the primary efficacy variable will be performed on the FAS.

Updated Section 3.6.5 Per-Protocol Set

Exclusion from the FAS will be considered as important protocol deviations that also result in exclusion from the PPS.

Supportive analysis of the primary efficacy variable will be performed on the PPS.

Updated Section 3.6.6 Pharmacokinetic Per-Protocol Set

PK variables will be analyzed for all subjects in the PK-PPS.

Added Section 3.6.7 COVID-19-free Set

The COVID-19 free set will consist of all subjects in the RS who had no COVID-19 impact up to the primary efficacy endpoint. This will be defined as subjects:

- not having a COVID-19 related IPD
- not having an impact based on the COVID-19 eCRF.
- not having an AE related COVID-19 (Section 10.2.5)
- not discontinuing due to COVID-19

The disposition data, the primary efficacy endpoint and the secondary efficacy endpoints included in the testing hierarchy will be analyzed on the COVID-19-free Set.

Change #17

Updated the following:

Section 3.10 Changes to protocol-defined analyses

The following changes from the protocol will be considered:

- The subgroup analysis on BASDAI will be performed on the categories: <4 vs. ≥4 rather than ≤4 vs. >4
- Proportion of subjects with a decrease of HAQ-DI from Baseline of at least 0.35 (HAQ-DI responders) in those subjects with HAQ-DI ≥0.35 instead of >0.35
- Proportion of PsAID-12 responders (decrease from Baseline in PsAID-12 total score ≥3) in subjects with PsAID-12 total score ≥3 at Baseline instead of >3 at Baseline.
- An additional supportive analysis for primary endpoint based on the analysis of the individual components of ACR will be performed using the Reference-Based imputation method.
- The main analyses of the secondary continuous variables included in the testing hierarchy will be performed using the Reference-Based imputation method.
- Additional subgroup analyses will be performed for HAQ-DI responders at Week 16.

- Additional variable added for subgroups: Concomitantly receiving MTX vs. cDMARDs at Baseline (concomitant MTX, no concomitant MTX and cDMARDs at Baseline, no concomitant MTX and no cDMARDs at Baseline

Change #18

Added the following section:

Section 3.11 Changes related to COVID-19

The impact of the COVID-19 pandemic on study procedures/conduct as well as the efficacy and safety endpoints will be investigated, and additional analysis outputs will be provided as appropriate. These additional analyses were not planned as part of the original protocol as the pandemic was not ongoing at the time of protocol finalization. These additional analyses will include analyses by period of the COVID-19 pandemic (pre/during/post) as defined below:

- Pre COVID-19 pandemic period: Period prior to COVID-19 pandemic start date defined as 11-Mar-2020.
- COVID-19 pandemic period: Period from 11-Mar-2020 though the COVID-19 pandemic end date which is currently not defined at the time of approval of the SAP.
- Post COVID-19 pandemic period: Period after the declaration of the end of the pandemic

The additional analyses are described in the following sections:

- Subject disposition (Section 5.1)
- Details of impacted visits and effects on collection and reporting of efficacy data (Section 5.2)
- Protocol deviations (Section 5.3)
- Efficacy analyses related to the hierarchy endpoints (Section 8.1.4.8 and Section 8.2.3)
- Adverse events (Section 10.2.5)

Change #19

Modified the following section:

Section 4.1 Adjustments for covariates

The primary efficacy analysis that will investigate the treatment effect will be adjusted for:

- Prior TNFα inhibitor exposure (inadequate response to 1 prior TNFα inhibitor, inadequate response to 2 prior TNFα inhibitors, intolerance to TNFα inhibitors), and
- Region (Asia/Australia, Eastern Europe, North America, Western Europe)

However, prior TNFα inhibitor exposure may be dropped from the model in order to achieve model convergence (ie, in case the likelihood maximization algorithm failed to converge). If convergence is still not obtained, then region may also be dropped from the model.

Prior TNFα inhibitor exposure and region are the stratification factors of the randomization. In the event that a study participant is stratified in the incorrect stratum (i.e. the stratum recorded in the Interactive voice or web Response System (IXRS) differs from the actual stratum the study participant belongs to), the actual stratum will be used for the analysis.

The secondary analyses will be adjusted on the same categorical factors as the primary analysis. For continuous variables, Baseline value (of the variable of interest) will also be included as covariates where appropriate.

Modified to the following:

The primary efficacy analysis will investigate the treatment effect, adjusting on the two randomization stratification variables:

- Region
  - North America, Western Europe, Eastern Europe, Asia

  or

  - North America, Eastern Europe, Western Europe + Asia if the percentage of randomized subjects is less than 10% in either of the Asia or Western Europe regions
- Prior TNFα inhibitor exposure (inadequate response to 1 prior TNFα inhibitor, inadequate response to 2 prior TNFα inhibitors, intolerance to TNFα inhibitors).

When adjusting on these 2 variables, a statistical model might not converge (ie, in case the likelihood maximization algorithm failed to converge). In that case, the statistical model will be run after dropping stratification variables in their order of appearance, successively region (one first and the other after) and prior TNFα inhibitor exposure.

In the event that a subject is stratified in the incorrect stratum (ie, the stratum recorded in the Interactive voice or web Response System (IXRS) differs from the actual stratum the subject belongs to), the actual stratum will be used for the analysis.

The secondary analyses will be adjusted on the same categorical factors as the primary analysis. For continuous variables, Baseline value (of the variable of interest) will also be included as covariate where appropriate.

Change #20

Section 4.2.1 Handling missing data for efficacy analyses

Added the following table as follows:

Section 4.2.1 Strategy for handling missing data for efficacy analyses

Different approaches will be used to handle missing data including how the intercurrent events will be considered (Section 3.4).

In this section, 3 terms will be defined for binary endpoints:

- Non-composite binary endpoint: Binary endpoints derived based on 1 continuous measurement (eg, PASI90).
- Composite binary endpoint: Binary endpoints derived based on several continuous measurements (eg, ACR50).
- Composite continuous endpoint: Continuous endpoint derived based on several continuous measurements (eg, PASDAS).

The table below summarizes all continuous endpoints (excluding composite continuous endpoints) along with their observed ranges and indicate whether they are represented by an

integer value. These endpoints will be analyzed and / or used in the derivation of composite or non-composite endpoints:

Table 4.1: List of non-composite continuous endpoints

Efficacy endpoint	Minimum	Maximum	Integer Value
BASDAI	0	10	
EQ-5D-3L dimension scores	1	3	Yes
EQ-5D-3L (VAS)	0	100	Yes
FACIT-Fatigue subscale score	0	52	
HAQ-DI	0	3	
hs-CRP	LLOQ/2 where LLOQ=0.10mg/L	No maximum	
IGA	0	4	Yes
LDI	0	No maximum	
LEI	0	6	
mNAPSI	0	13	Yes
PASI	0	72	
PGA-Arthritis /PhGA-Arthritis	0	100	Yes
PGA-PsA / PhGA-PsA	0	100	Yes
PsAID-12	0	10	
PsAQoL	0	20	
PtAAP	0	100	Yes
SF-36[a] Physical Functioning (PF) Role Physical (RP) Bodily Pain (BP) General Health (GH) Vitaly (VT) Social Functioning (SF) Role Emotional (RE) Mental Health (MH) Physical component summary (PCS) Mental component summary (MCS)	 19.26 21.23 21.68 18.95 22.89 17.23 14.39 11.63 5.02 -3.33	 57.54 57.16 62.00 66.50 70.42 57.34 56.17 63.95 79.78 80.09	
SPARCC	0	16	
TJC / SJC	0	68/66	
WPAI-SHP	0	100	

BASDAI=Bath Ankylosing Spondylitis Disease Activity Index; EQ-5D-3L=Euro-Quality of Life 5-Dimensions 3 Level version; FACIT=Functional Assessment of Chronic Illness Therapy; HAQ-DI=Health Assessment Questionnaire – Disability Index; hs-CRP=High sensitivity C-reactive protein; IGA=Investigator Global assessment; LDI=Leeds Dactylitis Index; LEI=Leeds Enthesitis Index; LLOQ=lower limit of quantification; mNAPSI=Modified Nail Psoriasis Severity Index; PASI=Psoriasis Area and Severity Index; PGA-Arthritis=Patient's global assessment of Arthritis; PGA-PsA=Patient's Global assessment of Psoriatic Arthritis; PhGA-Arthritis=Physician's Global Assessment of Arthritis; PhGA-PsA=Physician's Global Assessment of Psoriatic Arthritis; PHQ-9=Patient Health Questionnaire-9; PSAID-12=Psoriatic Arthritis Impact of Disease-12; PsAQoL=Psoriatic Arthritis Quality of Life; PtAAP=Patient's Assessment of Arthritis Pain; SF-36=Short-Form 36-item Health Survey; SJC= Swollen Joint Count; SPARCC= Spondyloarthritis Research Consortium of Canada ; TJC=Tender Joint Count; WPAI-SHP=Work Productivity and Activity Impairment questionnaire – Specific Health Problem
[a] minimum and maximum of the norm-based scores – Normal range for SF-36 scores is 0-100

Change #21

Section 4.2.1 Handling of missing data for efficacy analyses

Main analysis method

"Main analysis" refers in this section to the primary method of analysis used for a given endpoint, regardless of whether it be a primary, secondary, or other efficacy endpoint.Binary endpoints

For binary efficacy endpoints, the main analysis method is as follows:

- If study participants have missing data that is not preceded by an intercurrent event, then:
  - For non-composite binary endpoints, the standard MI (MI) methods (See Section 4.2.1.1) will be implemented on the raw score (eg, PASI for the PASI90 endpoint) before deriving the binary endpoint based on the imputed score.
  - For composite binary endpoints, where the composite binary endpoint cannot be constructed (i.e. either TJC/SJC are missing or >2 out of 5 PRO measures are missing), the MI will be implemented on each component variable that has missing data (e.g. TJC, SJC, PhGA-PsA, PGA-PsA, HAQ-DI, PtAAP, hs-CRP for the ACR50 endpoint), before deriving the binary endpoint based on the imputed components.
- If study participants have an intercurrent event, then the endpoint at all subsequent visits (whether or not the data were observed) will be set to 'nonresponse' as the study participant has not met the criteria for response based on the composite estimand definition (See Section8.1.2).

Note that in the case of a binary endpoint that is based on the combination of a non-composite binary endpoint with a composite binary endpoint (as with the ACR50 and PASI90 combined endpoint), each binary endpoint is derived independently, as described above. Then, the combined binary endpoint is derived based on the imputed data from each of the component binary endpoints.

Continuous endpoints

For continuous efficacy endpoints, the main analysis method is as follows:

- Missing data will be imputed based on the standard MCMC/Monotone Regression MI approach described in Section4.2.1.1, regardless whether missing data is preceded by an intercurrent event.

Supportive analysis methods

For the primary endpoint, which is a composite binary endpoint, several supportive analyses assuming different missing data mechanisms will be conducted:

- The MCMC/Logistic Regression approach described in Section 4.2.1.2.
- The reference-based MI approach described in Section 4.2.1.3.
- The tipping point approach implemented within the MI framework, but only if the primary endpoint analysis result is statistically significant at alpha=0.05. The method is specified in Section 4.2.1.4 below.
- The treatment policy strategy described in Section 4.2.1.5.
- The observed case (OC) analysis that will only include the observed data for study participants who did not discontinue study treatment. Data collected after study treatment discontinuation for any reason and all other missing data will be excluded from this analysis.
- Non-responder imputation (NRI) analysis will consider that study participants with missing data at the time point of interest did not respond to the treatment (ie, non-responder). This also include time points following study medication discontinuation for study participants who have stopped treatment but remain in the study. In the case of ACR50, nonresponse is imputed if any of the following are true (otherwise, ACR50 is evaluated based on the available data):
  - All ACR component variables are missing, or
  - TJC or SJC components are missing, or
  - TJC or SJC data are present, but more than 2 of the other 5 components are missing

For all secondary endpoints, the OC analysis will be done as a supportive analysis. For binary secondary endpoints, NRI will in addition be performed as a supportive analysis.

The OC analysis will be performed as a supportive analysis for other efficacy variables whose value at Week 16 are part of the hierarchical procedure.

See Section 8.1.4, 8.2.3 and 8.3.2 for further details on these analyses.

Table 7–1 summarizes which missing data handling approaches will be used for each type of efficacy endpoint.

Table 4.1: Missing data handling approaches for efficacy endpoints

Variable Priority	Variable Type	Missing data handling approach						
		MI (MCMC/ Monotone Regression)	MI (MCMC/ Logistic Regression)	NRI	MI (MCMC/ Reference-based)	OC	Treatment Policy	Tipping Point Analysis
Primary	Composite Binary	P[a]	S	S	S[a]	S	S[a]	S[a]
Secondary	Binary	P[a]		S		S		
Secondary	Continuous	P				S		
Other	Binary	P[ab]				S[c]		

Variable Priority	Variable Type	Missing data handling approach						
		MI (MCMC/ Monotone Regression)	MI (MCMC/ Logistic Regression)	NRI	MI (MCMC/ Reference-based)	OC	Treatment Policy	Tipping Point Analysis
Other	Continuous	P[b]				S[c]		

[a] Imputation method is applied on continuous data, and binary variable is derived from the continuous variable based on complete data set.
[b] MI will be used to provide point estimates at each timepoint only, i.e. mean ACR50 proportion
[c] Required only for by-visit summaries of variables whose value at Week 16 is part of the hierarchical testing procedure.
P=Primary method, S=Supportive method.

Was modified to:

Section 4.2.1.1 Primary analysis of binary endpoints that are part of the testing hierarchy (primary and secondary endpoints)

All binary endpoints (composite & non-composite) are based on continuous component variables.

The primary analysis of binary efficacy endpoints that are part of the testing hierarchy is considered under a framework estimand in which missing data (due to study treatment discontinuation) is considered indicative of failed treatment and imputed to nonresponse. Further data missing while on treatment, or data observed while not on treatment, will also be imputed to nonresponse.

This composite estimand approach to handling missing data which is the primary analysis method for binary efficacy endpoints is similar to the non-responder imputation (NRI) method.

Section 4.2.1.2 Primary analysis of continuous endpoints that are part of the testing hierarchy (secondary endpoints)

For continuous efficacy endpoints that are part of the testing hierarchy, the main analysis method is based on the hypothetical estimand approach as follows:

- If subjects have missing data regardless whether the missing data is preceded by an IE, then missing data will be imputed based on the MI-MCMC/Reference-Based imputation method (Section 4.2.2.3).
- If subjects have non-missing data after IE, then such data will be set to missing prior to running MI.

Section 4.2.1.3 Supportive analyses of primary efficacy endpoints

For the primary endpoint, which is a composite binary endpoint, several supportive analyses assuming different missing data mechanisms will be conducted:

- The modified composite estimand approach: the IE is changed from all treatment discontinuation to discontinuation of treatment due to AE or lack of efficacy. The same as the primary method will be used with the difference that only data after discontinuation due to AE or lack of efficacy is set to nonresponse. All other imputed data will be used.

- For the analysis of individual components of ACR, under the hypothetical estimand approach using the Reference-Based imputation method (Section 4.2.2.3).
- The tipping point approach implemented within the MI framework, but only if the primary endpoint analysis result is statistically significant at alpha=0.05. (Section 4.2.2.4).
- The treatment policy strategy (Section 4.2.2.5).
- The observed case (OC) analysis that will only include the observed data. For subjects who had treatment discontinuation, only observed data prior to treatment discontinuation date will be analyzed. For visits with missing data without treatment discontinuation, the data will remain missing. For observed data after treatment discontinuation, the corresponding visit data and subsequent visit data will be treated as missing.

Section 4.2.1.4 Supportive analyses of endpoints that are part of the testing hierarchy (secondary endpoints)

The following supportive analysis methods will be performed for the secondary and the other efficacy variables that are part of the testing hierarchy:

For binary endpoints:

- The modified composite estimand approach: the IE is changed from all treatment discontinuation to discontinuation of treatment due to AE or lack of efficacy. For composite or non-composite binary endpoints, the standard MI approach) will be implemented (similarly as in Section 4.2.1.3) on the raw continuous score(s) (eg, PASI for the PASI90 endpoint) before deriving the binary endpoint based on the imputed score(s).
- The OC analysis.

For continuous endpoints:

- The hypothetical estimand approach:
  - If subjects have missing data regardless whether the missing data is preceded by an IE, then missing data will be imputed based on the MI-MCMC/Monotone regression method (Section 4.2.2.2).
  - If subjects have non-missing data after IE, then such data will be set to missing prior to running MI.
- The OC analysis.

Section 4.2.1.5 Analyses of secondary endpoints that not part of the testing hierarchy and other efficacy endpoints

The other efficacy endpoints including the secondary efficacy endpoints that are not part of the testing hierarchy will be analyzed as the secondary efficacy endpoints (with the exclusion of the MI-MCMC/referenced based imputation method for continuous variables) but with no designated priority, with the exclusion of the MI-MCMC/reference-based imputation method for continuous variables.

For a combined binary endpoint based on the combination of a non-composite and a composite binary endpoint (as with the ACR50 and PASI90 combined endpoint), each binary endpoint will be derived independently, as described above, before deriving the combined endpoint.

Section 4.2.1.6 Summary table of missing data handling approaches for efficacy analyses

Table 4.2 summarizes which missing data handling approaches will be used for each type of efficacy endpoint.

Table 4.2: Missing data handling approaches for efficacy endpoints

Efficacy Endpoint Priority	Variable Type	Missing data handling approach						
		Composite estimand (NRI)	Modified Composite estimand (MI)	Hypothetical estimand (MI)	Hypothetical estimand Reference-Based (MI)	OC	Tipping Point analysis (MI)	Treatment Policy (MI)
Primary endpoints	Composite Binary	P	S[a]			S	S[a]	S[a]
Secondary endpoints included in the testing hierarchy	Binary	P	S[a]			S		
	Continuous			S	P	S		
Other Secondary endpoints (not included in the testing hierarchy	Binary	X	X[a]			X		
	Continuous			X		X		
Other endpoints	Binary	X	X[a,b]			X		
	Continuous			X[b]	X[d]	X		
	Categorical	X[c]	X[c]			X		

P=Primary method, S=Supportive method, X=Method to be used (no priority designated) MI=multiple imputation, NRI=non-responder imputation, OC=observed cases.
Note: Composite estimand (NRI) refers to the approach in which data preceded by the intercurrent event of study treatment discontinuation are imputed as nonresponse, and other missing data are also imputed as nonresponse.
Note: Modified Composite Estimand (MI) refers to the approach in which data preceded by the intercurrent event of study treatment discontinuation due to AE or lack of efficacy are imputed as nonresponse (or worst category for categorical variable), and other missing data are imputed via a MI model.
Note: Hypothetical Estimand (MI) refers to the approach where outcomes for subjects without an intercurrent event of study treatment discontinuation are as observed, and outcomes for subjects with the intercurrent event are imputed via a MI model.
Note: Referenced Based imputation refers to the approach in which it is assumedthat the statistical behavior of the bimekizumab and placebo-treated subjects after discontinuing study medication becomes that of the placebo-treated subjects that remain in the study.[a] Imputation method is applied on continuous data, and binary variable is derived from the continuous variable based on complete imputed data set.
[b] MI will be used to provide point estimates at each timepoint only.
[c] Missing data to be imputed to the worst category.
[d] For ACR components analysis at Week 16.

Change #22

Added section 4.2.2 Methods for handling missing data for efficacy analyses

The sections below describe the method to be used for missing efficacy data handling.

Note: For a specific endpoint, if analyzed subjects in a one or several treatment groups have no missing data, then analysis of observed cases will be performed in the concerned treatment group(s).

Change #23

Added section 4.2.2.1 Non-responder imputation

For binary endpoints, the nonresponder imputation (NRI) analysis, also described above as the primary composite estimand, will consider the following subjects as "non-responders" for the timepoint of interest:

- Subjects with missing data at the timepoint of interest. In the case of ACR50, nonresponse is imputed if ACR50 cannot be derived (i.e. derived as "missing") based on available ACR components. (Section 8.1.1).
- Subjects remaining in the study at the timepoint of interest but discontinuing study treatment before the timepoint of interest (data after IE).
- Subjects with missing Baseline value (for composite or non-composite binary variables based on change(s) from Baseline of continuous endpoint(s)).

Change #24

Section 4.2.1.1 MI-MCMC/Monotone Regression

This section describes the algorithms to be implemented for the MI – MCMC/Monotone Regression procedures for non-composite binary endpoints, composite binary endpoints, and continuous endpoints. These descriptions focus on the multiple imputation procedure itself and do not specifically account for dealing with intercurrent events, though it is assumed that those will be addressed as previously described.

Non-composite binary endpoints

For non-composite binary endpoints, the MI method will be applied as follows based on raw values:

- Step 1: Create a dataset, one for each treatment group, of study participants with observed values and those needing estimation by multiple imputation. For the imputation step, missing values will be separated into non-monotone missing values (ie, intermittent missing values between completed assessments) and monotone missing values (ie, where all study participant data is missing after a given time point).
  - **For the intermittent missing values**, the missing value in each dataset will be filled in using the Markov-Chain Monte Carlo (MCMC) method with multiple chains, monotone missing data imputing pattern, and non-informative priors for all parameters. Unless specified differently, the first 200 iterations will not be used (the 'burn-in' option). A total of 100 sets of imputations will be performed. The seed used for these imputations will be 2017 and all other multiple imputation procedures described in this SAP will use this same seed as well.

The resulting 100 imputed data sets will have a monotone missing pattern and will be imputed using a method for monotone missingness:

  - **For monotone missing data,** monotone regression will be used to impute missing data. A separate regression model is estimated for each variable with missing values (i.e., measurement at each time point). Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable. Since the data set has a monotone missing data pattern, the process is repeated sequentially for variables with missing values. The procedure will be based on the 100 imputed datasets generated

from the MCMC procedure and will be performed by Imputation. This will be based on 100 sets of imputations.

The SAS® PROC MI procedure will be used for the imputation.

In both cases, the imputation model will include prior TNFα inhibitor exposure, region, the value at Baseline and at each post-Baseline visit (prior to week of interest). The stratification factors (prior TNFα inhibitor exposure and region) may be dropped from the imputation model to facilitate model convergence if required (and this holds true for all imputation models). The post-Baseline values will need to be specified in chronological order in the imputation model so that the SAS® PROC MI imputes variables from left to right (i.e. the Week 4 value will be first imputed using regression based on Baseline value, and then the Week 5 value will be imputed using regression based on Baseline and Week 4 values, etc). In case a Baseline raw value is missing, the study participant will be excluded. The resulting datasets for each treatment arm will be combined into one complete dataset based on each of the 100 imputations (containing 100 times the number of study participants analyzed).

The imputation model based on the MCMC method will only allow multivariate normal variables. Therefore, prior TNFα inhibitor exposure and region will be re-coded as indicator variables:

- For prior TNFα inhibitor exposure, this will be 0 for intolerance to TNF, and 1 for inadequate response to 1 or 2 prior TNFs.
- For region, which has 4 levels, one indicator variable will be defined as 0 for regions other than North America and 1 for North America. Three more indicator variables will be defined similarly replacing North America with Eastern Europe, Western Europe, and Asia/Australia respectively.

- Step 2: Standard rounding rules will be applied to the imputed values. If the MI procedure yields a value outside of the pre-defined range for the given variable, the value will be updated to be within the predefined range of values for the endpoint of interest. For example, the imputed value for PASI will be updated to 0 or 72 in the case of an imputed value less than 0 or greater than 72, respectively.
- Step 3: For each of the 100 complete imputed data set obtained from Step 2, the binary responder variable will be derived. In practice, the value at the week of interest (eg, Week 16) in the imputed data sets will be used to categorize the study participant as a responder or not.
- Step 4: For each of the datasets obtained in Step 3, the responder rate will be analyzed using a logistic regression model with factors of treatment group, region, and prior TNFα inhibitor exposure as fixed effects. For the Other efficacy binary endpoints, at each timepoint the MI main method will calculate the mean of the calculated responder rates from the imputed datasets.
- Step 5: The results obtained from the 100 logistic regression analyses in step 4 (ie, the adjusted responder rates for each treatment group and the corresponding 95% CI, the odds ratio of the treatment comparison and corresponding 95% CI) will be combined for overall inference using Rubin's rules, which account for the uncertainty associated with the imputed values (Rubin, 1987). This will be done using SAS® PROC MIANALYZE. The method for

determining the degrees of freedom when combining results will be the freedom default Rubin's method.

Composite binary endpoints

For composite binary endpoints, the MCMC/monotone regression method described above in Step 1 and 2 for handling intermittent and monotone missing data will be performed separately for each individual component variable. Then, based on the multiply imputed data sets obtained for each component, the binary response will be derived as follows:

- The 100 data sets obtained for each component will be merged by imputation number and study participant number.
- For each of the 100 complete imputed data sets, the binary endpoint will be derived for each study participant/visit based on the component values.

Procedures described for Step 4 and Step 5 will then follow.

For the primary endpoint, the 'burn-in' option in PROC MI may be set higher than the default value. The convergence of the MCMC will be assessed graphically, and the results will be provided in a SAS® output.

Calculation of adjusted responder rates, odds ratio and CIs for binary endpoints.

Estimates of the adjusted responder rates for each treatment group and the associated standard errors are obtained from the logistic regression in Step 4 on the logit scale and as such are assumed to follow a normal distribution. These estimates will be combined using Rubin's rules and the combined estimates and associated SEs will be used to construct 95% CIs on the logit scale. The combined estimates and 95% CIs on the logit scale will be back-transformed using the inverse logit link function to obtain the adjusted responder rates (%) and associated 95% CIs.

Some key points to consider relative to the calculation of the odds ratios and corresponding confidence intervals .As the estimates of the odds ratios from the logistic regression model in step 3 follow a lognormal distribution, a log transformation is needed to normalize these 100 odd ratio estimates since the procedures for combining results from multiple imputed datasets are based on the assumption that the statistics estimated from each imputed dataset are normally distributed. Therefore, the log of the odds ratio estimates from the logistic regression model are used when combining into a single inference (step 3). Additionally, the standard errors (SE) for the odds ratios are transformed as follows:

$$SE = \frac{\log(UCL) - \log(LCL)}{2Z_{\alpha/2}}$$

where UCL and LCL are the upper and lower confidence limit, respectively, for the CI of the odds ratio from the logistic regression model, and $Z_{\alpha/2}$ is the relevant critical value from the standard normal distribution (1.96 for a 95% CI). The estimates of the log odds ratio for bimekizumab relative to placebo and the corresponding upper and lower confidence limits will be provided. The odds ratio will be then estimated by exponentiating the estimate of the log odds ratio. The odds ratio and the confidence limits of the odds ratio will be estimated as follows:

$$OR = \exp(Log\ odds\ ratio\ estimate\ )$$

$$LCL = OR * \exp(-SE * Z_{\alpha/2}\ )$$

$$UCL = OR * \exp(SE * Z_{\alpha/2})$$

where OR is the back-transformed estimate of the odds ratio just described, SE is the standard error of the log odds ratio and $Z_{\alpha/2}$ is the relevant critical value from the standard normal distribution (1.96 for a 95% CI). These calculations will be done such that odds ratios and corresponding CIs are calculated for the odds ratio of bimekizumab versus placebo.

Continuous efficacy endpoints

For secondary continuous efficacy endpoints, the same methods as the ones described above for the non-composite binary endpoints will apply based on the values at the timepoint of interest, with the following differences:

- In Step 3, for each of the 100 complete imputed dataset obtained, the change from Baseline will be calculated
- In Step 4, the analysis will be done on the change from Baseline, and the model will be an ANCOVA model with treatment group, region, and prior TNFα inhibitor exposure as fixed effects and the Baseline value as covariate.
- In Step 5, the results obtained from the 100 ANCOVA analyses in Step 4 (i.e. the Least Square Means for the treatment difference and the 95% CI for the contrasts) will be combined with Rubin's rules.

For other continuous efficacy endpoints, the following steps will apply:

- Step 1 and Step 2 as described for secondary continuous efficacy endpoints will be implemented.
- The 100 imputed datasets will be combined and simple means and standard errors will be calculated using Rubin's rules (via SAS® PROC MIANALYZE). For calculation of other descriptive statistics such as the median, min and max, Rubin's rules do not apply. Multiple imputation estimates will be computed by calculating arithmetic means of the estimates from the multiple repetitions of the imputation algorithm.

Was modified to:

Section 4.2.2.2 MI-MCMC/Monotone Regression

In instances where MI is used, the missing value is replaced by a set of plausible values, where each value is a Bayesian draw from the conditional distribution of the missing data given the observed data. Intermittent missing data will be imputed using the MCMC method, followed by regression for monotone missing data (called "MI-MCMC/Monotone regression" method in this SAP). The multiple imputation procedures planned for efficacy analyses are based on an assumption of data missing at random (MAR).

This section describes the algorithms to be implemented for the MI – MCMC/Monotone Regression procedures for non-composite binary endpoints, composite binary endpoints, and continuous endpoints. These descriptions focus on the multiple imputation procedure itself and do not specifically account for dealing with intercurrent events which is addressed in Section 4.2.1 .

Added section 4.2.2.2.1 Non-composite binary endpoints

For non-composite binary endpoints, the MI method will be applied as follows based on raw values:

- Step 1: Imputation of missing data using MI:

Create datasets, one for each treatment group, of subjects with observed values and missing values (needing estimation by MI). For the imputation step, missing values will be separated into 2 categories: intermittent missing values (ie, missing values for a given subject that has available data before and after the missing timepoint, including missing value at Baseline) and monotone missing values (ie, where all subject data is missing after a given time point). Datasets should be designed in a horizontal structure meaning that each subject should be present in a single observation, with a set of values, one for each scheduled visit where the endpoint is supposed to be collected according to the protocol (this excludes the unscheduled visits).

Datasets should also be sorted by subject number before proceeding with the MI process.

a. **For the intermittent missing values**, the missing value in each dataset will be filled in using the Markov-Chain Monte Carlo (MCMC) method with multiple chains, monotone missing data imputing pattern, and non-informative priors for all parameters. Unless specified differently, the first 200 iterations will not be used (the 'burn-in' option). A total of 100 sets of imputations will be performed. The seed used for these imputations will be 2017 and all other MI procedures described in this SAP will use this same seed as well. The procedure will be performed for each treatment group separately.

The resulting 100 imputed data sets will have a monotone missing pattern and will be imputed using a method for monotone missingness:

b. **For monotone missing data,** one monotone regression will be used to impute missing data. A separate regression model is estimated for each variable with missing values (ie, measurement at each time point). Based on the resulting model, a new regression model is then drawn and is used to impute the missing values for the variable. Since the data set has a monotone missing data pattern, the process is repeated sequentially for variables with missing values. The procedure will be based on the 100 imputed datasets generated from the MCMC procedure and will be performed for each treatment group separately byimputation.

The SAS® PROC MI procedure will be used for the imputation.

In both cases, the imputation model will include the randomization strata as stratification variables (region and prior TNFα inhibitor exposure), the value at Baseline and at each post-Baseline visit (up to the week of interest). The imputation model based on the MCMC method will only allow joint multivariate normal variables. For the MCMC method (when imputing intermittent missing values), randomization strata will be re-coded as indicator variables, and will be always specified in the following order if they are retained in the model:

- North America region (1 for North America, 0 otherwise)
- Western Europe region (1 for Western Europe, 0 otherwise)
- Eastern Europe region (1 for Eastern Europe, 0 otherwise)
- Prior TNFα 1 (1 for inadequate response to 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior or more TNFα inhibitors)

No variable needs to be created for the 3rd category of prior TNFα inhibitors (intolerance to TNFα inhibitors) or the 4th region (Asia) which is covered when all above variables are "0".

The randomization strata (region and prior TNFα inhibitor exposure) may be dropped from the imputation model to facilitate model convergence if required (this will be the case for all imputation models) as described below:

- If the MI fails to converge when adjusting on the stratification variables specified above, or if the percentage of randomized subjects is less than 10% in either of the Asia or Western Europe regions, then the Western Europe and Asia region will be combined and the binary variables for region that will be left in the MI will then be for North America and Eastern Europe regions.
- If after doing that, the MI still fails to converge then all region factors will be removed from the MI.
- If the MI still does not converge after dropping the regions, the remaining stratification variables (Prior TNFα 2,1) will also be removed.
- If a variable is dropped in order to allow convergence for one model in the study, that variable does not have to be dropped from other models in the study of the model converges without dropping the variable. That is, model convergence should be evaluated for each efficacy table independently. Furthermore, this means that a table based on a single timepoint may have different variables in the imputation model than a separate by-visit table for the same variable.

The post-Baseline values will need to be specified in chronological order after the Baseline value in the imputation model so that the SAS® PROC MI imputes variables from left to right (ie, the Week 4 value will be first imputed using regression based on Baseline value, and then the Week 8 value will be imputed using regression based on Baseline and Week 4 values, etc.). The resulting datasets for each treatment arm will be combined into 1 complete dataset containing 100 times the number of subjects analyzed.

Note: When imputing missing hs-CRP values, missing values due to values below LLOQ (0.10 mg/L) will be replaced by the midpoint value between 0 and LLOQ prior to running MI.

- Step 2: Standard rounding rules will be applied to the imputed values after each SAS® PROC MI. If an imputed value falls outside of the range for the given variable (as listed in Table 4.1), the value will be updated to be within the predefined range. For example, the imputed value for PASI will be updated to 0 or 72 in the case of an imputed value less than 0 or greater than 72, respectively. For endpoints that can take only integer values (eg,: IGA as listed in Table 4.1), the imputed values will be rounded to the closest integer after each SAS® PROC MI.

  Note: For hs-CRP, the lower limit used the midpoint value between 0 and LLOQ. For some parameters, there is not necessarily an upper limit to the imputed value (ie, LDI)

- Step 3: On the dataset obtained from Step 2, the binary responder variable will be derived. In practice, the value at the week of interest (eg, Week 16) in the imputed data sets will be used to categorize the subject as a responder or not. If subjects have an IE, then the endpoint at all subsequent visits will be set to "nonresponse" and the following will be performed:
  - When date of visit is available: by comparing IE date vs. date of visit

  - When the date of visit is missing in case of fully imputed data at a visit: by comparing the next visit number after the visit where the last study medication occurs vs. the visit number where data are imputed.

- Step 4: At each timepoint the (unadjusted) proportion of responders will be calculated by treatment group from the imputed datasets using SAS® PROC FREQ.

Step 5: (For primary and secondary efficacy endpoints only): For each value of the imputation number from 1 to 100, the adjusted proportion of responders will be analyzed using a logistic regression model with a fixed effect for treatment. The suitability of including region and prior TNFα inhibitor exposure as fixed effects will be assessed using goodness-of-fit tests (Deviance and Pearson's and Hosmer-Lemeshow) and added, if appropriate, if it allows model convergence. Covariates kept in the modelling should also be the same as the ones previously used for the MI. Comparisons will be made using 2-sided Wald test level of significance of 5%. The results obtained from the 100 logistic regression analyses (ie, the adjusted proportion of responders for each treatment group, the OR and the difference of proportions for the BKZ-placebo comparison and corresponding 95% CI) will be combined for overall inference using Rubin's rules, which account for the uncertainty associated with the imputed values (Rubin, 1987). This will be done using SAS® PROC MIANALYZE.

Added section 4.2.2.2.2 Composite binary endpoints

For composite binary endpoints, the MI method will be applied as follows:

- Step1: The MCMC/monotone regression approach described above in Step 1 for non-composite binary endpoints will be performed separately for each individual component variable.
- Step 2: The same rounding rules described in Step 2 above for non-composite binary endpoints will be performed for each component.
- Step 3: Based on the multiply imputed data sets obtained for each component, the binary response will be derived as follows:
  - The dataset obtained for each component will be merged by imputation number and subject number
  - On the dataset obtained, the binary endpoint will be derived for each subject/visit based on the component values.
  - If subjects have an intercurrent event, then the endpoint at all subsequent visits (from the day after the intercurrent event date, whether the data were observed or not) will be set to "nonresponse".
- Steps 4 and 5: Same as Steps 4 and 5 above for non-composite binary endpoints (4.2.2.2.1).

For the primary endpoint, the 'burn-in' option in SAS® PROC MI may be set higher than the default value. The goodness-of-fit tests to assess convergence of the MCMC will be provided in a SAS® output.

Added section 4.2.2.2.3 Calculation of adjusted responder rates, odds ratio and CIs for binary endpoints

Estimates of the adjusted responder rates for each treatment group and the associated standard errors are obtained from the logistic regression in Step 5 on the logit scale and as such are

assumed to follow a normal distribution. These estimates will be combined using Rubin's rules and the combined estimates and associated SEs will be used to construct 95% CIs on the logit scale. The combined estimates and 95% CIs on the logit scale will be back transformed using the inverse logit link function to obtain the adjusted responder rates (%) and associated 95% CIs.

Some key points to consider relative to the calculation of the odds ratios and corresponding confidence intervals. As the estimates of the odds ratios from the logistic regression model in step 5 follow a lognormal distribution, a log transformation is needed to normalize these 100 odd ratio estimates since the procedures for combining results from multiple imputed datasets are based on the assumption that the statistics estimated from each imputed dataset are normally distributed. Therefore, the log of the odds ratio estimates from the logistic regression model are used when combining into a single inference (step 3). Additionally, the standard errors (SE) for the odds ratios are transformed as follows:

$$SE = \frac{\log(UCL) - \log(LCL)}{2Z_{\alpha/2}}$$

where UCL and LCL are the upper and lower confidence limit, respectively, for the CI of the odds ratio from the logistic regression model, and $Z_{\alpha/2}$ is the relevant critical value from the standard normal distribution (1.96 for a 95% CI). The estimates of the log odds ratio for BKZ relative to placebo and the corresponding upper and lower confidence limits will be provided. The odds ratio will be then estimated by exponentiating the estimate of the log odds ratio. The odds ratio and the confidence limits of the odds ratio will be estimated as follows:

$$OR = \exp(Log\ odds\ ratio\ estimate\ )$$

$$LCL = OR * \exp(-SE * Z_{\alpha/2}\ )$$

$$UCL = OR * \exp(SE * Z_{\alpha/2}\ )$$

where OR is the back-transformed estimate of the odds ratio just described, SE is the standard error of the log odds ratio and $Z_{\alpha/2}$ is the relevant critical value from the standard normal distribution (1.96 for a 95% CI). These calculations will be done such that odds ratios and corresponding CIs are calculated for the odds ratio of BKZ versus placebo.

In addition to calculating the odds ratio, associated CIs, and p-values for the comparisons of BKZ vs. placebo, the estimated proportion of responders (ie, estimated responder rate) and the difference in the proportion of responders between BKZ and placebo will be estimated, along with 2-sided 95% CIs. The creation of the estimates of this difference will be completed using the process detailed below:

1. Use the logistic regression model to calculate:

   Least squares mean (LSM) estimates of the log odds of BKZ (G_B) and placebo (G_P), as well as their corresponding standard errors (S_B and S_P, respectively).

   Standard error of the LSM estimate of the log odds ratio (S_R)

2. Compute estimates for predicted proportions using the following transformations:

   $P_B = \exp(G_B)/(1 + \exp(G_B))$

   $P_P = \exp(G_P)/(1 + \exp(G_P))$

   The difference in proportions is then given by:

$D = P_B - P_P$

3. Estimate the standard error of D by:

$S_D = \text{sqrt}[P_B^2(1-P_B)^2S_B^2 + P_P^2(1-P_P)^2S_P^2 + P_B(1-P_B)P_P(1-P_P)S_R^2 - P_B(1-P_B)P_P(1-P_P)(S_B^2 + S_P^2])$

The MI/MCMC monotone regression method, as previously outlined, will be used to account for missing values. The calculation steps described above will be based on the results provided from the logistic regression model of the multiple imputed datasets. An estimate for the difference in proportions, D, and corresponding standard error, SD, will be computed for each of these datasets. The results from these analyses will be combined into a single estimate of the difference in predicted proportion of response and a 2-sided 95% CI interval using SAS PROC MIANALYZE.

Added section 4.2.2.2.4 Continuous efficacy endpoints

For secondary and other continuous efficacy endpoints, the same methods as the ones described above for the non-composite binary endpoints will apply based on the values at the timepoint(s) of interest, with the following differences:

- Step 1: The same imputation procedure as the described in step 1 for the non-composite binary endpoints will be used.
- Step 2: The same rounding rules described in step 2 above for non-composite binary endpoints will be performed.
- Step 3: On the dataset obtained with the imputation number from 1 to 100; change from Baseline will be derived. Simple means and standard errors will be calculated using Rubin's rules (via SAS® PROC MIANALYZE) for each timepoint(s) of interest. For calculation of other descriptive statistics such as the median, min and max, Rubin's rules do not apply. MI estimates will be computed by simply averaging the estimates from the multiple repetitions of the imputation algorithm. For CRP data (that will be presented using the geometric mean and corresponding 95% CI, arithmetic mean, median, Q1, Q3, minimum and maximum). The change from Baseline will be expressed as the ratio to Baseline (value at the visit divided by value at Baseline) in the by-visit summaries. The following approach will be applied:
  - Following the MI procedure, the ratio to Baseline will be calculated for any of the imputed values
  - The natural logarithm of the absolute values and of the ratios to Baseline will be calculated
  - The logged values will be summarized by treatment, visit and imputation
  - The datasets will be combined using PROC MIANALYZE in order to get the mean from the absolute values and ratios to Baseline across imputations
  - The estimates of the mean will be back transformed to obtain the geometric mean on the original scale
  - For the median, Q1, Q3, minimum and maximum the procedure outlined above for the other endpoints will be followed

- Step 4: (for primary and secondary efficacy variables): For each value of the imputation number from 1 to 100, the change from Baseline will be analyzed using an ANCOVA model with treatment group, region, and prior TNFα inhibitor exposure as fixed effects and the Baseline value as covariate. The results obtained from the 100 ANCOVA analyses (ie, the LSM for the treatment difference and the 95% CI for the contrasts) will be combined using Rubin's rules (via SAS® PROC MIANALYZE).

For composite continuous endpoints, the MI will be applied on each component the same way as for composite binary endpoint. The datasets obtained for each component will be merged by imputation number and subject number. On the dataset obtained, the endpoint will be derived for each subject/visit based on the component values. The endpoint obtained will be analyzed as above.

Added section 4.2.2.2.5 Ordinal efficacy endpoints

For ordinal efficacy endpoints, the MI method will be applied as for the continuous efficacy endpoints with the following exceptions:

- After the imputation of the intermittent and monotone missing data , the rounding rules will be performed to align the data with the possible responses for the respective endpoint (eg, If the variable can take the value 1, 2 or 3, the imputed values will be rounded to the closest possible value).
- If the ordinal efficacy endpoint is used to derive a binary endpoint, the same analysis strategy as described above for the binary endpoints will be used.
- If the ordinal efficacy endpoint is not used to derive a binary endpoint, the unadjusted proportions of subject by categories will be presented.
- If subjects have an IE, then the ordinal endpoint at all subsequent visits (from the day after the IE date, whether the data were observed or not) will be set to the value used for the worst category.

Change #25

Deleted section 4.2.1.2 MI-MCMC/Logistic Regression

Change #26

Updated section 4.2.1.3 MI-MCMC/Reference-Based Regression

Updated to the following: Section 4.2.2.3 MI-MCMC/Reference-Based Regression

For the primary analysis of the secondary continuous efficacy variable included in the testing hierarchy and the individual components of the ACR response, the Reference-Based multiple imputation method will be used. In this imputation method, the missing data will be imputed based on data from the placebo group only (Mallinckrodt, 2013).

Reference-Based MI assumes that the statistical behavior of the BKZ and placebo-treated subjects after discontinuing study medication becomes that of the placebo-treated subjects. All time points after study treatment discontinuation for BKZ and placebo groups will be considered missing. MIs are used to replace missing outcomes for BKZ- and placebo-treated subjects who discontinued using multiple draws from the posterior predictive distribution estimated from the placebo arm.

Note: SAS 9.4 may be used for this analysis where equivalent steps to those outlined below will be used to perform the imputation:

In the case of continuous endpoints, the procedure will be implemented as follows on the raw values:

4. Data will be processed sequentially, one timepoint (visit) at a time, by repeatedly calling SAS® PROC MI to impute missing outcome data at visits t=1,…, Week 16 (Week 16 being the time point of interest) using data from the placebo-treated subjects arm only.

c. *Initialization*. Set t=1 (Baseline visit)

d. *Iteration*. Set t=t+1. Create a dataset combining records from BKZ- and placebo-treated subjects with columns for covariates (prior TNFα exposure and region) and outcomes at visits 1 to t. Outcomes for all BKZ-treated subjects are set to missing at visit t and set to observed or previously imputed values at visits 1 to t-1. Outcomes for placebo-treated subjects are set to observed at visit t or observed or previously imputed values at visits 1 to t-1. The outcomes should be sorted in chronological order in the model.

e. *Imputation*. Impute missing values for visit t using previous outcomes for visits 1 to t-1, prior use of TNFα inhibitor, and region. Note that only placebo data will be used to estimate the imputation model since no outcome is available for BKZ- treated subjects at visit t. As a consequence, the input dataset should include all subjects from placebo but only subjects from the BKZ arm that have values at timepoint t missing.

f. Repeat steps 2a-2c for all timepoints, 100 times with different seed values (seeds ranging from 201 to 300) to create 100 imputed complete datasets. Subjects whose missing values were imputed in the last PROC MI call will be included in the input dataset for the next PROC MI call. Standard rounding rules will be applied to the imputed values. If the MI procedure yields a value outside of the pre-defined range for the given variable, the value will be updated to be within the predefined range of values for the endpoint of interest. For example, the imputed value for TJC will be updated to 0 or 68 in the case of an imputed value less than 0 or greater than 68, respectively.

5. The analysis will be done on the change from Baseline, and the model will be an ANCOVA model with treatment group, region, and prior TNFα inhibitor exposure classification as fixed effects and the Baseline value as covariate.

   For generation of summary statistics, the 100 imputed datasets will be combined, and simple means and standard errors will be calculated using Rubin's rules (via SAS® PROC MIANALYZE). For calculation of other descriptive statistics (median, Q1, Q3, minimum and maximum), Rubin's rules do not apply. Multiple imputation estimates will be computed by calculating arithmetic means of the estimates from the multiple repetitions of the imputation algorithm. Thus, for median, Q1, Q3, minimum and maximum the following approach will apply:

   - The data will be summarized by treatment, visit and imputation and the summary statistics will be computed.
   - Results will be summarized by taking the mean value of each summary statistic at each visit across all imputations.

- The number of decimal places will remain the same as the original for display purposes (ie, if the mean was presented to 1 decimal place, the mean of the means will also be presented to 1 decimal place).

The results obtained from the 100 ANCOVA analyses in Step 4 (i.e. the Least Square Means for the treatment difference and the 95% CI for the contrasts) will be combined with Rubin's rules.

Change #27

Section 4.2.1.4 Tipping Point analysis

Tipping point analyses will also be implemented as a supportive analysis for the primary efficacy endpoint.

The objective of the tipping point analyses is to identify assumptions about the missing data under which the conclusions from the main analysis change, ie, under which there is no longer evidence of a treatment effect. These tipping point analyses will only be performed if the primary efficacy analysis results in a statistically significant treatment effect ($p<0.05$).

As a first step, the worst-case scenario will be evaluated for the primary endpoint (ACR50 at Week 16). Specifically, all bimekizumab-randomized study participants with a missing ACR50 at Week 16 will be imputed as non-responders, while all placebo-randomized study participants with a missing ACR50 at Week 16 will be imputed as responders. While there is little justification for such an approach, it makes the most putative assumption possible against a bimekizumab treatment effect. After applying this imputation approach, a logistic regression model consistent with the one described for the primary analysis will be applied. If the p-value for the odds ratio of bimekizumab versus placebo is less than 0.05, then no further tipping point analyses are needed.

If this analysis based on the worst-case scenario results in a p-value greater than 0.05, then additional tipping point analyses will be performed. Various assumptions will be made about average outcomes among the subsets of study participants who prematurely discontinued study treatment and hence have a monotone missing data pattern (O'Kelly, 2014). In practice, it implies different delta adjustments will be made to the assumed responses on the monotone missing data in each treatment group independently with various degrees of plausibility with the goal to find for each treatment group the "tipping point" that will significantly reverse the primary result which yielded a p-value less than 0.05. These delta adjustments will be done on the components of ACR50 and will be implemented for each component as follows:

1. The same MCMC method described in Section 4.2.1.1 will be implemented for non-monotone (intermittent) missing pattern values, using the same imputation model. This will be done only once for each study participant in order to provide a dataset with monotone missing data.
2. Based on the dataset obtained in Step 1, a monotone regression model will be applied (using the same imputation model as in Step 1) while adjusting the imputed values by various delta adjustments. This will be based on 100 sets of imputations.

   These delta adjustments will be made independently in each treatment group, and the delta parameters should be defined within the pre-defined range of values of the endpoint of interest.

Several scenarios will be considered to define these shift parameters. Once defined, the same shift parameter value will be applied on the imputed endpoint value for all visits.

Scenario 1 will be based on the assumption that study participants randomized to bimekizumab and who have missing data have a lower probability of response compared to study participants randomized to placebo with missing data.

- For endpoints for which high scores are associated with a more favorable outcome, it will mean that:

- A negative shift is applied to the imputed value for study participants randomized to bimekizumab in order to decrease the imputed value.
- A positive shift is applied to the imputed value for study participants randomized to placebo in order to increase the imputed value.

- For endpoints for which high scores are associated with a less favorable outcome, it will mean that:

- A positive shift is applied to the imputed value for study participants randomized to bimekizumab in order to decrease the imputed value.
- A negative shift is applied to the imputed value for study participants randomized to placebo in order to increase the imputed value.

For each continuous endpoint, a set of possible values will be first pre-defined for the shift parameter (example: 0, 1, 2, 3)

Following the delta adjustments for the individual components, the composite binary endpoint (ACR50) will then be derived based on the multiply delta-adjusted imputed data sets obtained for each component.

3. Each of the 100 imputed datasets will then be analyzed using a logistic regression model with factors of treatment group, region, and prior TNFα inhibitor exposure as fixed effects.
4. The results obtained from the 100 logistic regression analyses in Step 3 will be combined for overall inference using Rubin's rules, and the results obtained for each shift parameter will be presented in a single table.
5. Step 2 to Step 4 will be repeated so that, at each iteration, missing values are adjusted with a larger delta than at the previous iteration. Depending on the results obtained, shift parameters with more granularity (eg, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9) may also be investigated. The process will go on until the p-value for the odds ratio between bimekizumab and placebo is no longer statistically significant (i.e. ≥ 0.05). The odds ratio, 95% CI, and p-values obtained for each value of delta will be combined in one single table.

Was modified to:

Section 4.2.2.4 Tipping Point analysis

Tipping point analyses will also be implemented as a supportive analysis for the primary efficacy endpoint.

The objective of the tipping point analyses is to identify assumptions about the missing data under which the conclusions from the main analysis change, ie, under which there is no longer evidence of a treatment effect. These tipping point analyses will only be performed if the primary efficacy analysis results in a statistically significant treatment effect ($p<0.05$).

As a first step, the worst-case scenario will be evaluated. Specifically, all BKZ-randomized subjects with a missing ACR50 at Week 16 (or non-missing ACR50 after study treatment discontinuation) will be imputed as non-responders, while all placebo-randomized subjects with a missing ACR50 at Week 16 will be imputed as responders. While there is little justification for such an approach, it makes the most putative assumption possible against a BKZ treatment effect. After applying this imputation approach, a logistic regression model consistent with the one described for the primary analysis will be applied. If the p-value for the odds ratio of BKZ versus placebo is less than 0.05, then no further tipping point analyses are needed.

If this analysis based on the worst-case scenario results in a p-value greater than 0.05, then additional tipping point analyses will be performed. In practice, it implies different delta adjustments will be made to the assumed responses on the missing data (where missing values include observations after the intercurrent event date and any other missing values) in each treatment group independently with various degrees of plausibility with the goal to find for each treatment group the "tipping point" that will significantly reverse the primary result which yielded a p-value less than 0.05. These delta adjustments will be done on the components of ACR50 and will be implemented for each component as follows:

- Step 1: The same MCMC method described in Section 4.2.2.2 will be implemented for non-monotone (intermittent) missing pattern values, using the same imputation model. This will be based on 100 imputations.
- Step 2: Based on the dataset obtained in Step 1, a monotone regression model will be applied (using the same imputation model as in Step 1) and as described in Section 4.2.2.2. This will be based on 1 imputation.
- Step 3: Imputed values will be shifted using a delta adjustment independently in each treatment group.

Once defined, the same delta adjustment value will be applied on the imputed values for the Week 16 visit only. The selected scenario will assume that subjects randomized to BKZ and who have missing data have a lower probability of response compared to subjects randomized to placebo with missing data.

- For ACR components for which high scores are associated with a less favorable outcome, it will mean that:
  - A positive shift is applied to the imputed value for subjects randomized to BKZ in order to increase the imputed value.
  - A negative shift is applied to the imputed value for subjects randomized to placebo in order to decrease the imputed value.

A set of possible values will be first pre-defined for the delta parameter as follows:

- For all ACR components, the value of the initial delta parameter will be equivalent to a specific percentage of the possible range of the component (eg, 5%). For hs-CRP, the range

will be based on the log-transformed values at Week 16. The delta parameters for each endpoint are listed in the Table 7–3.

Table 4.3: Tipping Point analysis: Delta parameters for each ACR component

ACR component[1]	Range	Delta
SJC	0-66	3.3
TJC	0-68	3.4
HAQ-DI	0-3	0.15
PtAAP, PhGA-PsA, PGA-PsA	0-100	5
hs-CRP[2]	Observed range of the $\log_e$ transformed values for overall subjects at Week 16	5% of the observed range

[1]The shifted imputed value should not exceed the range for the ACR component.
[2]The delta adjustment will be applied on the $\log_e$ transformed imputed values which will then be exponentiated prior to deriving endpoint.

- Step 4: Standard rounding rules will be applied to the imputed values. If the SAS® PROC MI yields a value outside of the range for the given component, the value will be updated after the imputation has been performed to be within the predefined range.
- Step 5: Repeat steps 1 to 4 for each ACR component. The composite binary endpoint (ACR50) will then be derived based on the multiple shifted imputed data sets obtained for each component.
- Step 6:Additionally, as the primary endpoint is derived using NRI, subjects randomized to BKZ with missing data (where missing values include observations after the IE date and any other missing values) should be set to non-response, after applying the delta adjustment outlined in Step 3 above. This ensures subjects randomized to BKZ do not have a higher probability of response in the tipping point analyses compared to the primary analysis (ie, a subject randomized to BKZ who is non-responder in the primary analysis cannot become a responder in the tipping point analyses).
- Step 7: In the data obtained, for each value of the imputation number, ACR50 will be analyzed using a logistic regression model with factors of treatment group, region, and prior TNFα inhibitor exposure at Baseline as fixed effects.
- Step 8: The results obtained from the 100 logistic regression analyses in Step 6 will be combined for overall inference using Rubin's rules, and the results obtained for each shift parameter will be presented in a single table.
- Step 9: Step 3 to Step 8 will be repeated so that, at each iteration, missing values are adjusted with a larger shift than at the previous iteration. Depending on the results obtained, delta parameters with more granularity (eg, 2 times, 3 times the initial delta) will be investigated. The process will go on until the p-value for the odds ratio between BKZ and placebo is no longer statistically significant (ie, ≥ 0.05). The odds ratio, 95% CI, and p-values obtained for each value of delta will be combined in one single table.

Change #28

Section 4.2.2 Handling of missing data for AE

Imputation of Partial Start Dates

- If only the month and year are specified and the month and year of first dose is not the same as the month and year of AE onset, then use the 1st of the month
- If only the month and year are specified and the month and year of first dose is the same as the month and year of AE onset, then use the date/time of first dose
- If only the year is specified, and the year of first dose is not the same as the year of onset, then use January 1 of the year of AE onset
- If only the year is specified, and the year of first dose is the same as the year of AE onset, then use the date/time of first dose
- If the AE onset date is completely unknown, then use the date of first dose

Was updated to section 4.2.3 Handling of missing data for AE

Imputation of Partial AE Start Dates

- If only the month and year are specified:
  - if the month and year of first dose of IMP is not the same as the month and year of the start date, then use the 1st of the month,
  - if the month and year of first dose of IMP is the same as the month and year of the start date, then use the date of first dose of IMP,
- If only the year is specified:
  - if the year of first dose is not the same as the year of the start date, then use January 1 of the year of the start date.
  - if the year of first dose is the same as the year of the start date, then use the date of first dose.
- If the date is completely unknown:
  - use the date of first dose of IMP if the stop date is unknown or not prior to the onset date.

For seriousness, no imputation rule will be applied for the interim analysis, but the worst-case approach will be applied for the final analysis.

The following was added:

If the imputed stop date is prior to the imputed start date:

- For missing start day and start month:
  - if the year of start date is the same as the year of first dose and the imputed stop date is on or after the date of first dose then set the start date to the date of first dose
  - otherwise set to the 1st January of the year of the start date
- For missing start day only
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is on or after the date of first dose, then set the start date to the date of first dose

- if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is before the date of first dose, then set the start date to the 1st of that month

In the event of ambiguity or incomplete data which makes it impossible to determine whether an AE is treatment emergent, the AE will be considered as treatment emergent.

For missing seriousness, no imputation rule will be applied for both the interim and the final analysis.

Change #29

Section 4.2.3 Handling of missing data for prior and concomitant medications

Imputation of Partial Start Dates

- If only the month and year are specified and the month and year of first dose is not the same as the month and year of the start date, then use the 1st of the month.
- If only the month and year are specified and the month and year of first dose is the same as the month and year of the start, then use the date of first dose.
- If only the year is specified, and the year of first dose is not the same as the year of the start date, then use January 1 of the year of the start date.
- If only the year is specified, and the year of first dose is the same as the year of the start date, then use the date of first dose.
- If the start date is completely unknown, and the stop date is unknown or not prior to the date of first dose, then use the date of first dose.

Was modified to the following:

Section 4.2.4 Handling of missing data for prior and concomitant medications

Imputation of Partial Start Dates

- If only the month and year are specified:
  - if the month and year of first dose of IMP is not the same as the month and year of the start date, then use the 1st of the month.
  - if the month and year of first dose of IMP is the same as the month and year of the start, then use the date of first dose of IMP.
- If only the year is specified:
  - if the year of first dose of IMP is not the same as the year of the start date, then use the 1st of January of the year of the start date.
  - if the year of first dose of IMP is the same as the year of the start date, then use the date of first dose of IMP.
- If the start date is completely unknown:
  - if the stop date is unknown or not prior to the date of first dose of IMP, then use the date of first dose of IMP.

- if the stop date is prior to the date of first dose of IMP, then set the start date to the 1st of January of the year of the end date.

Added the following:

If the imputed stop date is prior to the imputed start date:

- For missing start day and start month:
  - if the year of start date is the same as the year of first dose and the imputed stop date is on or after the date of first dose, then set the start date to the date of first dose
  - otherwise set to the 1st January of the year of the start date
- For missing start day only
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is on or after the date of first dose, then set the start date to the date of first dose
  - if the month and year of the start date is the same as the month and year of first dose and the imputed stop date is before the date of first dose, then set the start date to the 1st of that month

In the event of ambiguity or incomplete data which makes it impossible to determine whether a medication was concomitant, the medication will be considered as concomitant.

Change #30

Section 4.4 Multicenter studies

Centers will be grouped in the geographic regions North America, Western Europe, Central/Eastern Europe, and Asia/Australia.

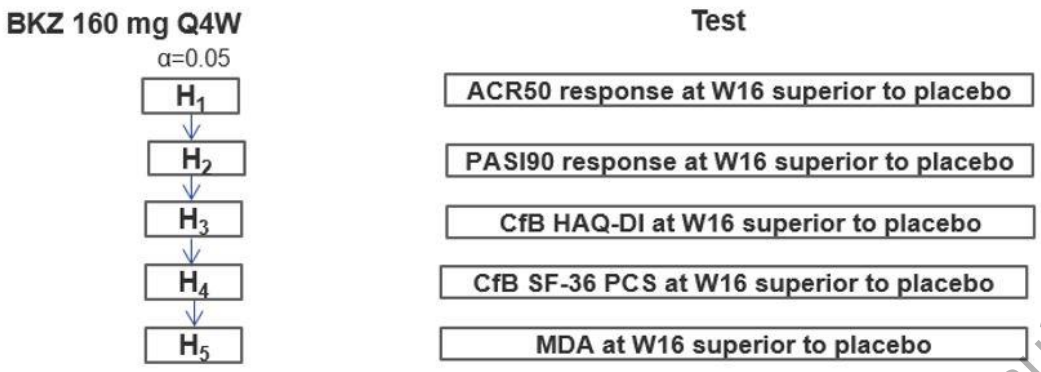
Was modified to:

However, the effect of center (using a pooling of centers by region) on results will be evaluated as mentioned in Section 3.8.

Change #31

Section 4.5 Multiple comparisons/multiplicity

Figure 10.1: Sequential testing procedure of primary and key secondary efficacy endpoints



ACR50=American College of Rheumatology 50% response criteria; BKZ=bimekizumab; CfB=change from Baseline; H=hypothesis; HAQ-DI=Health Assessment Questionnaire-Disability Index; MDA=Minimal Disease Activity; PASI90=Psoriasis Area and Severity Index 90; PCS=Physical Component Summary; Q4W=every 4 weeks; SF-36= Short-Form 36-item Health Survey; W=Week

Was modified to the following:

Table 4.4: Sequential testing procedure of primary/ secondary efficacy endpoints (fixed sequence testing procedure) (All efficacy endpoints at Week 16)

BKZ 160mg Q4W	
H_1	ACR50 Response superior to Placebo
H_2	Change from Baseline HAQ-DI superior to Placebo
H_3	PASI90 Response superior to Placebo in subjects with PSO BSA ≥3% at Baseline
H_4	Change from Baseline SF-36 PCS superior to Placebo
H_5	MDA superior to Placebo

ACR50=American College of Rheumatology 50% response crtiteria; BKZ=bimekizumab; H=hypothesis; HAQ-DI=Health Assessment Questionnaire - Disability Index; MDA=Minimal Disease Activity; PASI90=Psoriasis Area and Severity Index 90; PCS=Physical Component Summary; Q4W=every 4 weeks; SF-36=Short-Form 36-item Health Survey; W=week.

Change #32

Section 4.8 Examination of subgroups

Table 4.2: Categories of variable for subgroup analyses

Subgroup	Categories
Age (years)	<45, ≥45
Gender	Male, Female
Race	Black, White, Other

Subgroup	Categories
Region	Asia/Australia, Eastern Europe, North America, Western Europe
Disease duration (years)	<1, ≥1
Body weight (kg)	≤100, >100
hs-CRP level	< 6 mg/L, ≥ 6mg/L
Prior TNFα exposure	Intolerance to TNFα inhibitor, Inadequate response to 1 TNFα inhibitor, Inadequate response to 2 TNFα inhibitors
Prior cDMARDs[a]	0, 1, ≥2
Concomitant cDMARDs[a]	Yes, No
BSA affected by PSO (%)	<3, ≥3 to ≤10, >10
BASDAI	≤4, >4
ADAb status[b]	Positive, Negative
HLA-B27 status[c]	Positive, Negative

[a] cDMARD= conventional disease-modifying antirheumatic drugs
[b] Refer to Section 9.2 for further details on the definition categories.
[c] HLA-B27= human leukocyte antigen B27

Was modified to the following:

Subgroup analyses will be performed on the variables below. They are all assessed at Baseline, except concomitant cDMARDs, concomitant Methotrexate (MTX) and ADAb status which will be assessed during the 16-week period. Subgroup analyses will be performed on the ACR50 response, the PASI90 response and the HAQ-DI response (subjects with a decrease of HAQ-DI from Baseline of at least 0.35) at Week 16. The ADAb status will also be used for subgroup analysis for the PK endpoints.

The variables for subgroup analyses will be

- Age (<45 years of age, ≥45 years of age),
- Gender (male, female),
- Disease duration (<1 year, ≥1 year),
- Region (eg, North America, Western Europe, Eastern Europe, Asia),
- Race (White, Black and Other),
- Body weight at Baseline (≤100kg, >100kg),
- Prior TNFα exposure (intolerance to TNFα inhibitor, inadequate response to at least 1 TNFα inhibitor, inadequate response to 2 prior or more TNFα inhibitors),
- hs-CRP at Baseline (<6mg/L, ≥6mg/L),
- Prior conventional disease-modifying antirheumatic drugs (cDMARDs) (0, 1, ≥2) (taken prior to Baseline). Such medications will be classified using an adjudicated spreadsheet.
- Concomitantly receiving cDMARDs versus no concomitant cDMARDs,
- Concomitantly receiving MTX versus no concomitant MTX,
- Concomitantly receiving MTX vs. cDMARDs at Baseline (concomitant MTX, no concomitant MTX and cDMARDs at Baseline, no concomitant MTX and no cDMARDs at Baseline),
- PSO affected BSA at Baseline (<3%, ≥3% to 10%, >10%),
- BASDAI Baseline (<4, ≥4),
- ADAb status (positive, negative) (See 9.2) for the BKZ 160mg Q4W group only,
- Human leukocyte antigen B27 (HLA-B27) (positive, negative).

Note: In the context of submission for Japanese health authorities, a subset of tables, listings and figures will be provided for the subgroup of subjects randomized in Japan.

Change #32

Section 5.1 Subject disposition

Included COVID-19-free set and modified to the following:

The disposition of subjects into treatments groups and **analysis sets** will also be summarized on the RS.

The number and percentage of randomized subjects who completed the study, **completed the study not on randomized treatment**, **who discontinued IMP, overall and by reason, and who discontinued the study**, with the primary reason for study discontinuation (as collected on the Study Termination CRF page), will be tabulated overall and by treatment group. The numbers and percentage of **randomized** subjects entering the SFU Period will additionally be presented.

A subject will be said to have completed the study if she/he had completed the last scheduled study visit, not including SFU visits. (The SFU visit is typically 140 days post last dose of IMP, which is roughly equivalent to 5 half-lives of BKZ).For subjects who previously discontinued IMP, the primary reason for discontinuation as collected on the Study Medication termination CRF page will be summarized in a separate table.

Finally, the number of randomized subjects who discontinued the study due to AEs by type of AEs **(serious fatal, non-fatal, other)** will be summarized.

Study disposition and termination details will be listed for all subject screened. Additional listings will be created on study discontinuation, subject analysis sets and visit dates **(both actual visit and visit remapped for analysis)**. A listing will also be provided for subjects who did not meet the study eligibility criteria.

To assess subject disposition (entry and periods in the study) during the COVID-19 pandemic, subject disposition will be assessed by period of the COVID-19 pandemic (pre – during – post) (Section 3.6.7), by comparing the dates of visits (or events) to the dates of the COVID-19 pandemic period.

Change #33
Added the following section 5.2 Impact of COVID-19

A listing of visits affected by COVID-19 will be presented based on the Enrolled Set including the visit, date of visit, relationship to COVID-19, impact category, and a narrative (short description) of the event. These data will be summarized for non-randomized subjects and by treatment group and overall, for enrolled subjects.

A summary of study visits by COVID-19 pandemic period (pre – during – post) will be presented for subjects enrolled prior, during and after the pandemic.

In addition, in order to assess the potential impact of COVID-19 on the collection and reporting of efficacy data, a separate summary on the RS will be presented to display missing data as well as data collected via an alternative modality (e.g.: phone, video call) for efficacy endpoints included in the hierarchy (Table 7–4). For the purpose of these displays, missing data will be presented only for visits affected by COVID-19, as reported on the dedicated eCRF page. Missing data at other visits and for other reasons will not be included.

For visits conducted remotely, it was not possible to assess some endpoints (e.g., TJC, SJC, LDI, LEI, PASI, etc.) and therefore these assessments will be missing for those visits. In addition, for any missed visit or visit conducted remotely, the CRP assessment will also be missing. Such assessments will be considered missing as a result of the COVID-19 pandemic. For these visits, it will therefore not be possible to assess composite endpoints like ACR response. For this summary table, reported on the RS, only visits at which efficacy assessments are scheduled will be included.

Change #34
Added COVID-19 IPD to section 5.3 Protocol deviations

Change #35
Added the following to section 6.1 Demographics

To understand the study population during the different periods of the COVID-19 pandemic, the reporting of Baseline demographics will be repeated by COVID-19 period (pre - during - post) as determined by the Baseline visit date) on the SS.

Change #36
Section 6.2.1 Baseline efficacy variables

The different ACR components at Baseline (TJC, SJC, PGA-PsA, PhGA-PsA, PtAAP, HAQ-DI and hs-CRP (mg/dL)) will be summarized descriptively overall and by treatment group

In addition, hs-CRP will be presented in category (< 6 mg/L, ≥ 6mg/L).

Was modified to the following:

Baseline values for the 7 components for primary efficacy variable will be summarized by treatment group and overall. The following variables will be summarized:

- TJC,
- SJC,
- PGA-PsA,
- PhGA-PsA,
- PtAAP,
- HAQ-DI,
- hs-CRP (mg/L).

For hs-CRP, SJC, and TJC the following rule will be applied: If Baseline value is missing, then take the most recent value (taken at previous unscheduled visit or at Screening); If value is still missing, the Baseline value will be considered missing. hs-CRP will be summarized using descriptive statistics and in classes (≥6mg/L, <6mg/L).

To understand the study population during the different periods of the COVID-19 pandemic, the reporting of Baseline demographics will be repeated by COVID-19 period (pre - during – post as determined by the Baseline visit date) on the SS.

Change #37
Section 6.2.2 Other baseline characteristics

The following categorical variables will be summarized:

- BASDAI (≤4,>4)
- CASPAR criteria fulfilled (score ≥3),
- BSA affected by PSO (<3%, ≥3% to ≤10%, >10%),
- Rheumatoid factor (positive, negative),
- Anti-cyclic citrullinated peptide (CCP) antibodies (Positive, Negative)
- Nail psoriasis (yes, no),
- Dactylitis (yes, no),

- Enthesitis (yes, no),
- Past nonsteroidal anti-inflammatory drug (NSAID) therapy (yes, no)
- Current NSAID therapy (yes, no)
- Prior cDMARD (0, 1, ≥2)
- Current DMARD (yes, no)
- HLA-B27 status (positive, negative)

Was modified to the following:

The following variables will be summarized by treatment group and overall:

- Prior TNFα exposure (Intolerance to TNFα inhibitor, Inadequate response to 1 TNFα inhibitor, Inadequate response to 2 TNFα inhibitors) – IXRS randomization strata
- Prior TNFα exposure (Intolerance to TNFα inhibitor, Inadequate response to at least 1 TNFα inhibitor, Inadequate response to 2 prior or more TNFα inhibitors) – actual randomization strata
- % of BSA affected by PSO (<3%, ≥3% to ≤10%, >10%)
- PASI score at Baseline (for subjects with PSO involving at least 3% of BSA at Baseline). This variable will be summarized as a continuous variable and in classes (<10, 10 to 20,>20)
- BASDAI (<4, ≥4)
- Rheumatoid factor (positive, negative). A rheumatoid factor value < 30 IU/mL is defined as negative and a value ≥ 30 IU/mL is defined as positive. (Chemistry results / Proteins / Test=RF).
- Anti-cyclic citrullinated peptide (CCP) antibodies (Positive, Negative). Anti-CCP antibodies negative is a value < 20 U/mL and a value ≥ 20 U/mL is defined as positive.
- Nail psoriasis (yes, no)
- Dactylitis ("yes" if LDI score>0, "no" if LDI score=0 or missing)
- Enthesitis based on LEI ("yes" if LEI score>0, "no" if LEI score=0 or missing)
- Enthesitis based on SPARCC ("yes" if SPARCC>0, "no" if SPARCC=0 or missing)
- Prior NSAID (yes, no)
- NSAID at Baseline (yes, no)
- Prior cDMARDs (0, 1, ≥2) (see Section 4.8)
- cDMARDs at Baseline (yes, no)
- HLA-B27 status (positive, negative)
- Methotrexate at Baseline (yes, no)
- Prior oral corticosteroids (yes, no)

- Oral corticosteroids at Baseline (yes, no)

The above data including intra-auricular assessments will also be listed.

Efficacy variables not listed above will not be presented in the Baseline characteristics section. They will be presented in by-visit tables in the context of the analysis of secondary and other efficacy variables.

To understand the study population during the different periods of the COVID-19 pandemic, the reporting of Baseline demographics will be repeated by COVID-19 period (pre - during – post as determined by the Baseline visit date) on the SS.

Change #38
Section 6.2.3 History of psoriatic arthritis and psoriasis

For both PsA and PSO, time since first diagnosis (years) will be calculated as follows:

$$\frac{Date\ of\ informed\ consent - Date\ of\ diagnosis + 1}{365.25}$$

Was updated to the following:

Time since first diagnosis of PsA/PSO (years) will be calculated as follows:

$$\frac{Date\ of\ informed\ consent - Date\ of\ diagnosis}{365.25}$$

If the date of diagnosis is partial, it should be imputed to the most recent feasible date (ie, last day of the month if only day is missing, or the last day of the year if day and month are missing). Partial dates should not be imputed later than the subject's screening date. For subjects enrolled at German sites, only the year of birth may be entered into the eCRF for this study. For these subjects age will be calculated after imputing their date of birth to be on January 1st.

The data on history of psoriasis and psoriatic arthritis will be listed for all screened subjects.

Change #39
Section 6.4 Prior and concomitant medications

The following was added and updated:

Prior medications include any medications that started prior to the date of the first IMP administration.

The IMP dosing period corresponding to the combined Double-Blind Treatment Period will be calculated as follows:

The study medication dosing period start date is defined as the date of first dose.

The study medication dosing period stop date is defined as follows:

- For subjects who are ongoing at the time of the clinical data cut (not including subjects who are in SFU) use date of last clinical contact
- For subjects who died prior to last visit (not including those who died during SFU), use the minimum of the following:
  - Date of death

- Date of last dose of any study medication + 28 days

- For subjects who complete the study as planned, the dosing period ends at the later of the following two dates:
  - Date of last administration of IMP + 28 days
  - The last scheduled visit date not including SFU
- For all other subjects, use the maximum of the following:
  - Date of last dose of any study medication + 28 days

Date of last visit (not including SFU)

The number and percentage of subjects taking prior medications **and prior Anti-TNFs (excluding past psoriasis medications)** will be summarized on the RS by treatment group, overall and by ATC class, presenting Anatomical Main Group (ATC Level 1), Therapeutical/Pharmacological Subgroup (ATC level 3), and preferred term (PT).

The table summary will be ordered alphabetically for the ATC class and in terms of decreasing frequency for PT within ATC class in the BKZ treatment group. In the event of ties, PT will be ordered alphabetically.

A by-subject listing of all prior medications **and prior Anti-TNFs, and concomitant medications (on the SS) as well as glossary of all medical medications will be provided**.

Change #40
Section 6.5 Prohibited concomitant medications

Refer to the study protocol for the list of prohibited or restricted medications.

The number and percentage of study participants who used prohibited concomitant medication will be summarized by ATC class, presenting ATC Level 1, ATC level 3, and PT similarly as done for the summary on concomitant medications.

The prohibited medications will be listed on RS.

Was modified to the following:

Prohibited or restricted medications are defined in the protocol (section 7.8.2).

Prohibited medications will be listed on SS.

Change #41
Section 7 Treatment compliance

The following was added:

Treatment compliance summaries will be performed based on the SS, **by randomized treatment and for all subjects.**

If a subject discontinues early, then the number of expected injections is based on the time of early discontinuation relative to the dosing visits. **This number will be calculated by determining the number of expected dosing visits prior to the date of early discontinuation. If the early discontinuation coincides with an expected dosing visit, this visit will used to determine the number of expected injections.**

Change #42
Section 8.1.1 Derivation of ACR50 response

The following was added:

As the ACR response is based on 7 different component scores, it is necessary to consider various data scenarios that could impact the calculation of ACR response. The rules described here are applicable in the context of the calculation of ACR response and may differ from the rules applied for calculating and summarizing the components, individually (some values may need to be imputed for component for component analysis but are not required here to evaluate the ACR response).

The following rules will be applied prior to invoking any imputation analysis at the variable level:

- If a subject has a component value that is equal to 0 or is missing at Baseline and the post-Baseline value is greater than or equal to 0, then the percentage of improvement for that component will be treated as 0 for purposes of ACR response calculations.
- If a subject has a component value that is missing at Baseline, then the percentage of improvement for that component will be treated as missing for purposes of ACR response calculation.

Observed data will be used to calculate ACR response where possible. In case of partial missing data where an observed response may be calculated, imputed data will not change the result.

To address possible missing data in ACR components, ACR50 will be derived as below:

- Positive response on ACR50
    - Improvement at Week 16 of at least 50% from baseline on SJC and TJC and at least 3 out of the 5 remaining ACR components (regardless of whether the two remaining components are missing or not)
- Negative response on ACR50
    - No improvement at Week 16 of at least 50% from baseline on SJC (regardless of whether the remaining components are missing or not)
    - No improvement at Week 16 of at least 50% from baseline on TJC (regardless of whether the remaining components are missing or not)
    - No improvement for at least 3 out of the 5 ACR components (PGA-PsA, PhGA-PsA, PtAAP, HAQ-DI, hs-CRP) regardless of whether the remaining components are missing or not
- Missing ACR50 response
    - For all other situations that are not specified above.

For the OC analysis: if a given visit falls after the treatment discontinuation date, the endpoint at all subsequent visits (whether the data were observed or not) will be derived but reported as missing in corresponding table under "OC" (Section 8.1.4.6).

For the MI analysis: if a given visit falls after the treatment discontinuation date, the endpoint at this visit and all subsequent visits (whether the data were observed or not) will be imputed. Responses falling after an IE (defined as treatment discontinuation due to AE or lack of efficacy) will be set to nonresponse (Section 8.1.4.3).

Change #43
Section 8.1.1.1 Derivation of tender joint count (TJC) and swollen joint count (SJC)

Table 8.1 Swelling and tenderness grading

Grade	Tenderness response (68)	Swelling response (66)
0	Not tender	None
1	Tenderness present	Detectable synovial thickening

Was modified to the following:

Table 8.1 Swelling and tenderness grading

Present	Tenderness response (68)	Swelling response (66)
No (0)	Not tender	None
Yes (1)	Positive response to questioning (tender), spontaneous response elicited (tender and winced) or withdrawal by subject on examination (tender, winced, and withdrew)	Detectable synovial thickening with or without loss of bony contours, or bulging synovial proliferation with or without cystic characteristics

Change #44
Section 8.1.1.4 High-sensitivity C-Reactive Protein (hs-CRP)

Hs-CRP values which are below the lower limit of quantification (LLOQ) should be set to the midpoint between 0 and the LLOQ. Listing will show values below the LLOQ. The limit of quantification for hs-CRP is 0.16 mg/dL.

Was updated to the following:

hs-CRP values which are below the lower limit of quantification (LLOQ) should be set to the midpoint between 0 and the LLOQ prior to the analysis. Listing will show values below the LLOQ. The LLOQ for hs-CRP is 0.10 mg/L.

Change #45
Section 8.1.2 Primary analysis of the primary efficacy variable

The following 4 attributes describe the estimand that will be used to define the treatment effect of interest for the primary efficacy analysis:

- Population = Study participants screened according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Study participant-level outcome = ACR50 at Week 16
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment due to an AE or lack of efficacy prior to Week 16. A composite strategy will be implemented in which a positive clinical outcome is defined as achieving ACR50 at Week 16 and not discontinuing study treatment due to an AE or lack of efficacy through Week 16.

- Population-level summary measure = Conditional odds ratio comparing bimekizumab to placebo.

Intercurrent event will then be acknowledged as an unfavorable outcome for the composite estimand in considering study participants with intercurrent as 'non-responders' to the study treatment.

Any missing data at Week 16 that is not preceded by an intercurrent event (ie discontinuation of study medication due to an AE or lack of efficacy) will be imputed by standard MI techniques (see Section 4.2.1.1 for details). In MI, the missing value is replaced by a set of plausible values, where each value is a Bayesian draw from the conditional distribution of the missing data given the observed data.

As outlined in Section 4.5, the statistical null hypothesis for the ACR50 response at Week 16 is that there is no difference in the proportion of study participants with ACR50 response between the bimekizumab treatment and placebo (i.e. the conditional odds ratio for ACR50 response in the bimekizumab treatment compared with placebo is equal to 1). The alternative hypothesis is that there is a difference between bimekizumab treatment and placebo.

The summary table results will present the adjusted responder rates and the associated 95% CI for bimekizumab and placebo, the adjusted odds ratio and 95% CI for the comparison of bimekizumab versus placebo, and the p-value that the OR=1. The treatment comparison will be made using the 2-sided Wald test at a significance level of α=0.05.

Was updated to the following:

The following 4 attributes describe the estimand that will be used to define the treatment effect of interest for the primary efficacy analysis:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject-level outcome = ACR50 at Week 16
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to Week 16. A composite strategy will be implemented in which a positive clinical outcome is defined as achieving ACR50 at Week 16 and not discontinuing study treatment through Week 16.
- Population-level summary measure = Conditional odds ratio comparing BKZ to placebo and difference of response rate between BKZ and placebo.

Missing data at Week 16 that are not preceded by an IE and any data after an IE will be imputed as non-responders. This results in a more traditional NRI approach and will be labeled as such throughout the SAP.

The statistical hypothesis for the ACR50 response at Week 16 is that the conditional OR for ACR50 response in the BKZ treatment compared to placebo treatment is equal to 1.

A logistic regression model will be used to assess the treatment effect on ACR50 response at Week 16. The model will include fixed effects for treatment, and prior TNFα inhibitor exposure and region as stratification factors. The suitability of including these variables in the model will be assessed using goodness-of-fit tests (Deviance and Pearson and Hosmer-Lemeshow).

P-values below 0.05 would lead to a reconsideration of the model to be used and if the logistic regression model is unable to converge the stratification variables will be dropped (Section 4.1). In that case, all supportive analyses as well as imputation models conducted for the primary endpoint will disregard these exploratory factors from their models.

The country-specific analyses performed on subjects randomized in Japan will not consider the region factor as a covariate for the modelling.

The SAS® PROC LOGISTIC will be used to run the logistic regression.

The summary table results will present the adjusted responder rates and the associated 95% CI for BKZ and placebo, the adjusted odds ratio and 95% CI for the comparison of BKZ versus placebo, and the p-value that the OR=1 and the difference of response rate between BKZ and placebo and associated 95% CI. The treatment comparison will be made using the 2-sided Wald test at a significance level of α=0.05 (ie, H_1 Table 7–4).

Change #46
Section 8.1.3 Secondary analyses of the primary efficacy variable

Unless otherwise notified, for each subgroup analysis, a logistic regression will be fitted involving all model terms from the primary analysis model and additional terms for subgroup and subgroup by treatment interaction. For the analysis by stratification variables (ie, prior TNFα exposure and region), the analysis model will contain terms from the primary analysis model and the subgroup by treatment interaction. The same imputation method as the one used for the primary analysis will be used to handle missing data. In case of model convergence issue, missing data will be imputed using NRI instead of MI. For each subgroup category and each treatment group, the mean proportion of responders, the adjusted responder rates with associated 95% CI will be provided. The adjusted OR (for the comparison bimekizumab and placebo) and 95% CI will also be provided. The results obtained for each subgroup will be presented in one single table. The ORs and associated 95% CIs for each subgroup category will also be displayed on a single forest-plot.

Was modified to the following:

Section 8.1.3 Subgroup analyses of the primary efficacy variable

For each subgroup analysis (except for the analysis by prior TNFα exposure , by region, and by ADAb status), a logistic regression will be fitted involving the same terms that were retained when running the primary analysis model, plus a term for the subgroup and the subgroup by treatment interaction as detailed below:

- Fixed effect for treatment
- Prior TNFα 1 (1 for inadequate response to at least 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior or more TNFα inhibitors)
- Region
- Term for subgroup
- Terms for subgroup by treatment interaction(s).

For subgroup analyses by Prior TNFα exposure, the terms that will be retained will be:

- Fixed effect for treatment

- Prior TNFα 1 (1 for inadequate response to 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior TNFα inhibitors)
- Region
- Terms for Prior TNFα inhibitor exposure by treatment interaction

For subgroup analyses by region, the terms that will be retained will be:

- Fixed effect for treatment
- Prior TNFα 1 (1 for inadequate response to 1 prior TNFα inhibitor)
- Prior TNFα 2 (1 for inadequate response to 2 prior TNFα inhibitors)
- Region
- Terms for region by treatment interaction

For subgroup analyses by ADAb status, the terms that will be retained will be:

- prior TNFα exposure at Baseline
- Region
- ADAb Status (Positive, Negative)

For all subgroup analyses, the terms for the subgroups will coded as below:

- For Yes/No or positive /negative variables, the category "yes" or "positive" will be coded "1".
- For continuous variables in 2 classes, the highest category will be coded "1".
- For continuous variables in 3 classes (or more), 2 (or more) binary variables will be created and coded "1" for the 2 (or more) top categories.
- For gender, "male" will be the category coded "1".

The same estimand approach as the one used for the primary analysis will be used to handle missing data (NRI).

For each subgroup category and each treatment group (subgroup analysis on the ADAb status for only BKZ group), the mean proportion of responders on imputed datasets, the adjusted responder rates with associated 95% CI will be provided. The adjusted OR (for the comparison BKZ and placebo) and the corresponding 95% CI will also be provided. In addition, the difference in response rates between BKZ and placebo and the corresponding 95% CI will also be presented (except for the subgroup analysis on the ADAb status). The results obtained for each subgroup will be presented in one single table. The ORs and associated 95% CIs for each subgroup category will also be displayed on a single forest-plot.

The observed rates of responses by subgroup categories will be also provided along with the results obtained on imputed data.

Change #47
Section 8.1.4 Supportive analyses of the primary efficacy variable

- and the same statistic results as the ones produced for the primary efficacy analysis will be calculated

Was updated to:

- the same statistics will be provided: the mean proportion of responders (on imputed datasets when applicable) the adjusted responder rates for each treatment group, the adjusted ORs for the comparison BKZ versus placebo and its 95% CI, the p-value for the comparison between BKZ and placebo and the difference of response rates between BKZ and placebo and its corresponding 95% CI.

Change #48
Added Section 8.1.4.3 Analysis using modified composite estimand where intercurrent events are defined as discontinuation due to AE or Lack of efficacy

The modified composite estimand for this analysis is detailed below

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP.
- Subject-level outcome = ACR50 at Week 16.
- Intercurrent event handling = An intercurrent event is defined as discontinuation due to AE or lack of efficacy. A composite strategy will be implemented in which a positive clinical outcome is defined as achieving ACR50 at Week 16 and not discontinuing study treatment due to an AE or lack of efficacy through Week 16. Subjects discontinuing study treatment due to AE or lack of efficacy will be imputed based on completer subjects.
- Population-level summary measure = Conditional OR comparing BKZ to placebo

Any missing data at Week 16 that is not preceded by an intercurrent event will be imputed based on a predefined MI model (MI/MCMC - Monotone Regression) (Section 4.2.2.2).

The following categories of subjects will enter the MI/MCMC – Monotone regression process for each ACR component:

- Subjects with all post-baseline values missing.
- Subjects withdrawn from the study treatment (data collected prior to the treatment withdrawal will be retained) - Those subjects will have values imputed for missing assessments and non missing assessment falling after study treatment discontinuation (falling after or before the study withdrawal).
- Subjects with IE - Data collected after the intercurrent event will be set to missing.
- Subjects with missing value at Baseline. MI will be applied for each of the components separately before deriving the response.

In the case of partial missing data, ACR will be derived as in Section 8.1.1.

If an IE occurred prior to or at week 16 (or for any other intermittent missing data at Week16), the subjects will be considered as “nonresponders”.

In MI, the missing value is replaced by a set of plausible values, where each value is a Bayesian draw from the conditional distribution of the missing data given the observed data.

Change #49
Section 8.1.4.3 Analysis on individual components of the ACR

The change from Baseline in each individual component variable of the ACR50 response (TJC, SJC, PGA-PsA and PhGA-PsA) or ratio to Baseline for hs-CRP, will be analyzed at Week 16 using an analysis of covariance (ANCOVA) model. The analyses for the component variables PtAAP and HAQ-DI will be described in Section 8.2.2 where analyses for secondary endpoints are described. For all five variables, the model will include treatment, prior TNFα inhibitor exposure and region as fixed effects and the respective Baseline value as covariate. For hs-CRP, the model will be run on ln-transformed data and the results will be presented back-transformed towards their original scale. These analyses will be done on the RS and will explore the effect of the signs and symptoms of the individual components on the composite endpoint.

The following 4 attributes describe the estimand that will be used to define the treatment effect of interest for this analysis:

- Population = Study participants screened according to the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Study participant level outcome = The change from Baseline in the ACR component variable at Week 16
- Intercurrent event handling = A hypothetical strategy will be implemented in which the estimand targets the treatment difference in a scenario where intercurrent does not occur, such that outcomes for study participants without an intercurrent event are as observed, and outcomes for study participants with an intercurrent event are treated as though they had completed the randomized study treatment through Week 16.
- Population-level summary measure = the difference in the adjusted means between bimekizumab and placebo.

As detailed in Section 4.2.1 any missing data at Week 16 will be imputed by the standard MI techniques described in Section 4.2.1.1.

The SAS® PROC MIXED will be used to run the ANCOVA model.

For each ANCOVA model specified in this analysis (except the one for hs-CRP), the following statistics will be presented:

  - For placebo and bimekizumab: the adjusted least-square means (LSM) and standard error (SE)
  - For the comparison between placebo and bimekizumab: the difference between the LSM, the associated 95% CI for the contrasts, and the corresponding p-value.

For the ANCOVA analysis on the hs-CRP endpoint, the following statistics will be presented:

  - For placebo and bimekizumab: the adjusted geometric mean and geometric CV (%)
  - For the comparison between placebo and bimekizumab: the geometric LSM ratio, the associated 95% confidence intervals for the contrasts, and the corresponding p-value.

An additional table will present the descriptive statistics for the two treatment groups for each ACR component variable at Week16. Those statistics will be the mean and SD (geometric mean and CV (%) for hs-CRP), median, minimum and maximum after MI imputation.

The following supportive analyses described in Section 8.1.4.4 to 8.1.4.8 are planned in order to check the robustness of the primary analysis results to missing data handling.

Was modified to the following:

Section 8.1.4.4 Analysis on individual components of the ACR

The primary comparison (BKZ vs. placebo) will also be repeated for all individual components of the ACR50 response to explore the effect of the signs and symptoms of the individual components on the composite endpoint. Since all ACR components are continuous variables (eg, change from Baseline in TJC), an analysis of covariance (ANCOVA) with treatment, region, and Prior TNFα inhibitor exposure as fixed effects and the Baseline value as covariate will be used for the analysis. Missing data imputation described in Section 4.2.2 will be successively applied for each component.

The following 4 attributes describe the hypothetical estimand that will be used to define the treatment effect of interest for each of the 7 ACR components:

- Population = Subjects who meet the protocol-specified inclusion/exclusion criteria and randomized to IMP
- Subject level outcome = Change from Baseline in the ACR components at Week 16
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to Week 16. This estimand targets the treatment difference in a scenario where withdrawal from study treatment does not occur, such that outcomes for subjects without an intercurrent event are as observed, and outcomes for subjects with an intercurrent event are treated successively according to the two following hypothetical strategies:
  - (a) as they had completed treatment through Week 16. In that strategy, missing data or non missing data after IE (reset as missing) will be imputed using MI as though they had completed the randomized study treatment through Week 16.
  - (b) as they had completed the randomized study treatment through Week 16 but on Placebo. In that strategy, missing data or non missing data after IE (reset as missing) will be imputed using Reference-Based MI, in which the MI model is based on data from the Placebo group.
- Population-level summary measure = The difference in the adjusted means between BKZ and placebo.

For the results obtained using the two MI methods, the SAS® PROC MIXED will be used to run the ANCOVA model. The inferential statistics obtained from each model will be displayed in a single table.

For all variables, the following statistics will be presented:

- the adjusted least-square means (LSM) and standard error (SE) by treatment group,
- for the comparison between placebo and BKZ: the difference between the LSM, the associated 95% CI for the contrasts, and the corresponding p-value.

An additional table will present the descriptive statistics for the 2 treatment groups for each ACR component variable at Week16. Those statistics will be the mean and SE, median, minimum and maximum after MI.

Change #50
Section 8.1.4.5 Analyses using treatment policy strategy imputation for missing data

Added the following:

In the context of this analysis, in addition to the statistics provided in the context of MI, the number of observed ACR50 responses regardless of the existence of intercurrent events will be also reported.

The following 4 attributes describe the treatment policy strategy analysis is detailed below:

Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP.

Change #51

Section 8.1.4.6 Analyses on Observed Cases

Modified the following:

The following 4 attributes describe the estimand that will be used to define the treatment effect of interest for the OC analysis:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP

Change #52

Added the following:
Section 8.1.4.8 Analyses including COVID-19 impact

To assess the impact of the COVID-19 pandemic on the primary efficacy endpoint, 2 sensitivity analyses will be conducted using the same analysis method as the primary analysis (Section 8.1.2):

- on the COVID-19-free Set
- by COVID-19 period (pre- during -post as determined by the Week 16 date).

Change #53

Added the following:
Section 8.2.1. Derivation of secondary efficacy variables

The primary and supportive analysis methods of secondary efficacy variables for which the derivation method is provided below are described in Section 8.2.2 and Section 8.2.3.

For the other secondary variables which are not part of the testing hierarchy, there is no priority designated between methods described as primary and supportive analysis methods.

Change #54

Added the following:
Section 8.2.1.1 HAQ-DI: Change from Baseline at Week 16

The derivation for HAQ-DI is described in Section 8.1.1.3 (The analysis of this variable is already covered in Section 8.1.4.4).

Change from Baseline at Week 16 in HAQ-DI is the 2nd endpoint in the sequential testing hierarchy (Table 7–4).

Change #55

Section 8.2.1.1. Psoriasis Area and Severity Index (PASI), PASI90 has been changed to: Section 8.2.1.2 Psoriasis Area and Severity Index 90 (PASI90) at Week 16 (in subjects with PSO involving at least 3% of BSA at Baseline)

Change #56

Section 8.2.1.2. Short Form-36 (SF-36) Physical Component Summary (PCS)

The SF-36 (Version 2, standard recall) is a 36-items generic health-related quality of life instrument that uses a recall period of 4 weeks. Items are grouped into 8 domains as follows:

- Physical Functioning (10 items)
- Role Physical (4 items)
- Bodily Pain (2 items)
- General Health (5 items)
- Vitality (4 items)
- Social Functioning (2 items)
- Role Emotional (3 items)
- Mental Health (5 items)

One additional item (Question 2) asks respondents about perceived stability or change in health (Health Transition) over the past year. The concepts represented by these domains contribute to physical, mental, and social aspects of health-related quality of life.

In addition to domain scores, the PCS and MCS scores are calculated from the 8 domains (excluding the Health Transition item). Component scores reflect the impact of each domain on physical and mental health status.

The norm-based T-scores for the 2 SF-36 component summary (PCS and MCS) and the 8 domains are standardized with a mean of 50 and a standard deviation (SD) of 10 in the general US population (Maruish, 2011). An individual respondent's score that falls outside the T-score range of 45 to 55 should be considered outside the average range for the US general population. When considering group-level data, a score below 47 should be considered indicative of impaired functioning within that health domain or dimension. Similar to individual respondent data, group mean scores of 47 or greater should be considered average or above average as compared to the general US population. Higher scores indicate a better health status.

For the calculation of the SF-36 norm-based T-scores for the 8 domains and the PCS and MCS, the scoring software Optum's PRO CoRE (version 1.4) will be used. The software uses updated 2009 U.S. population norms and applies a Full Missing Score Estimation method as follows:

- A health domain score (except the Physical Functioning domain) will be estimated provided that at least one 1 non-missing response is available within that domain.

- For the Physical Functioning domain, item response theory will be used to develop a model for estimates of the missing score.
- Regression methods will then be applied to estimate the PCS and MCS on the basis of the available domains.

Change from Baseline at Week 16 in SF-36 PCS score is the 4th endpoint in the sequential testing hierarchy (Table 7–4).

Change #57

Section 8.2.1.3. MDA at Week 16 has been updated to Section 8.2.1.4
And the following has been deleted:
MDA response will be equal to 0 if the condition about achieving at least 5 of the 7 above criteria is not met.
The following was added:
A subject has achieved MDA if 5 or more of the 7 above criteria are fulfilled.
The following rule will be applied for subjects with BSA <3 at Baseline: Subjects with BSA <3 at Baseline will always meet the criteria PASI ≤1 or BSA ≤3 **except in the cases where a BSA score ≥3 is observed.**

Change #58

Section 8.2.1.5. Investigator's Global Assessment (IGA)
Has been updated to:
Section 8.2.1.6 Proportion of subjects with an Investigator's Global Assessment (IGA) score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline at Week 4 and Week 16 in the subset of subjects with psoriatic skin lesions at Baseline

The following has been added:
For some subjects with BSA ≥ 3% at Baseline, IGA data have not been captured at Week 4 and Week 16 visits when BSA<3% at those visits. To address that, an additional supportive analysis will be performed in which these subjects will be considered as IGA responders at the respective visit.

Change #59

Section 8.2.1.5. Investigator's Global Assessment (IGA) has been updated to section 8.2.1.6 Proportion of subjects with an Investigator's Global Assessment (IGA) score of 0 (clear) or 1 (almost clear) AND at least a 2-grade reduction from Baseline at Week 4 and Week 16 in the subset of subjects with psoriatic skin lesions at Baseline

The following has been added:
For some subjects with BSA ≥ 3% at Baseline, IGA data have not been captured at Week 4 and Week 16 visits when BSA<3% at those visits. To address that, an additional supportive analysis will be performed in which these subjects will be considered as IGA responders at the respective visit.

Change #60

Added the following Section 8.2.1.7. Change from Baseline in the Patient's Assessment of Arthritis Pain (PtAAP)

Change #61

Section 8.2.1.6. Psoriatic Arthritis Impact of Disease, PsAID-12 has been updated to section 8.2.1.8 Change from Baseline in PsAID-12 total score at Week 16

Change #62
Added the following sections:
Section 8.2.1.9 Psoraiais Area Severity Index (PASI90) at Week 4 (in subjects with PSO involving at least 3% of BSA at Baseline)

Change #63
Updated the following section:
Section 8.2.2 Primary analysis of secondary efficacy variables

The secondary efficacy variables will be analyzed for all subjects in the RS.

As indicated in Table 7–4, there are 4 secondary endpoints included in the testing hierarchy that will be tested in a chronological order until one is failed to be statistically significant.

For the other secondary endpoints (which are not part of the testing hierarchy), the p-values produced by the statistical models will be considered nominal since these endpoints are not controlled for multiplicity.

For the secondary composite and non-composite binary endpoints, the same estimand structure (composite estimand) as the one defined for the primary efficacy analysis of the primary efficacy endpoint (Section 8.1.2) will be used. The NRI approach for handling missing data and the same analysis model will be considered, and the analysis results will be presented similarly.

The statistical hypothesis for the binary variables at Week 16 is that the conditional odds ratio in the BKZ group compared with placebo treatment group is equal to 1. For the secondary continuous endpoints, the analysis will evaluate the hypothetical estimand as defined below:

- Population = Subjects enrolled according to the protocol-specified inclusion/exclusion criteria and randomized to IMP.
- Subject level outcome = variable as stated in Section 2.2.2
- Intercurrent event handling = An intercurrent event is defined as discontinuation of study treatment prior to Week 16. A hypothetical strategy for addressing intercurrent events will be implemented. This estimand targets the treatment difference in a scenario where withdrawal from study treatment does not occur, such that outcomes for subjects without an IE are as observed, and outcomes for subjects with an IE are treated as though they had completed treatment through Week 16 but on Placebo. A MI strategy will be used to impute any missing data and observed data after IE which will be set to missing prior to running MI. Such data will be imputed using Reference-Based MI, in which the MI model is based on data from the placebo group.
- Population-level summary measure = the difference in the adjusted means between BKZ 160mg Q4W and placebo.

The same analysis model and imputation strategy for handling missing data as in Section 8.1.4.4 will also be considered. The analysis results will be presented similarly as for this analysis on the individual components.

In the case of MDA as the response is based on 7 different component scores, it is necessary to consider various data scenarios that could impact the calculation of response. The rules described here are applicable in the context of the calculation of MDA response and may differ from the rules applied for calculating and summarizing the components individually (some values may need to be imputed for component analysis but are not required here to evaluate MDA response).

The following rules will be applied to complete the derivation of MDA response based on the composite estimand definition:

- If a given visit has been preceded by an IE (treatment discontinuation):
  - The endpoint at all subsequent visits (whether the data were observed or not) will be set to "nonresponse" as the subject has not met the criteria for response.
- If a given visit has not been preceded by an intercurrent event:
  - If a subject satisfies at least 5 of the 7 MDA criteria, the subject is considered as MDA responder
  - If a subject does not satisfy at least 3 MDA criteria, the subject is considered as non-MDA responder
  - In all other cases, the NRI approach will be applied.

An overview table will combine the results of the primary analysis for the primary and secondary efficacy endpoints included in the testing hierarchy.

For the country specific analyses performed on subjects randomized in Japan, the region factor will net be considered as a covariate for the modelling.

Change #64
Updated the following section:
Section 8.2.3 Supportive analyses of the secondary efficacy variables

As mentioned in Table 7–2, the supportive analyses for all secondary variables will be performed using the modified composite estimand for binary variables (using MI-MCMC/monotone regression as in Section 8.1.4.3) and on OC.

For PASI90, the MI will be run on the PASI score on subjects involving at least 3% of BSA at Baseline.

Additionally, subgroup analyses will be performed on the PASI90 response at Week 16 (in subjects with PSO involving at least 3% of BSA at Baseline) endpoint. Refer to Section 4.8 for the list of subgroups of interest. The same analysis method than the one used for the subgroup analysis of the primary endpoint will be performed. The results will be presented similarly.

To assess the impact of the COVID-19 pandemic, the analysis of secondary efficacy endpoints with be repeated on the COVID-19-free Set using the primary analysis method.

Change #65
Updated the following sections:
Section 8.3.1.1 Time to ACR20/50/70 response from Baseline

Time to a given response will be defined as the length in days from Baseline until the first date when the response is achieved.

Time from Baseline to censoring will be considered for the following subjects:

- Subjects who discontinue study treatment prior to achieving a response will be censored at the date of study treatment discontinuation.
- Subjects who reach the end of treatment period visit without achieving the given response will be censored at the date of the end of treatment period visit (for the interim analysis, subjects who reached Week 16 without achieving the given response will be censored at that visit).

Change #66
Added section 8.3.1.2 ACR20, ACR50 and ACR70 response

Change #67
Added sections 8.3.1.3 PASI90 response and 8.3.1.4 PASI75 and PASI100 response

Change #68
Updated section 8.3.1.2 to section 8.3.1.5 to include composite endpoint of ACR50/PASI100

Change #69
Added section 8.3.16 Proportion of Psoriatic Arthritis Response Criteria (PsARC) responders

Change #70
Added section 8.3.1.7 Psoriatic Arthritis Disease Activity Score (PASDAS) categories

Change #71
Section 8.3.2.3 Additional statistical analysis of other efficacy endpoints
Many details have been added to this section. See SAP section in main text.

Change #72
Sections 9 has been updated. See SAP main text

Change #73
Added section 10.1.2 Time at risk in subject-years by COVID-19 pandemic period

Change #74
Updated section 10.2.1 Standard AE summaries

TEAEs by timing of onset relative to ADAb status by SOC, HLT, PT (on subjects treated with BKZ). This will include columns for the following:

- TEAEs starting before the first ADAb positive result (includes ADAb categories 2, 4 and 5) where TEAEs have occurred before the following events: a) the first positive ADAb result for subjects in category 2 and b) the first post-baseline boosted ADAb titer result for subjects with titer results and the first post-Baseline positive ADAb result for subjects with positive ADAb at Baseline with no other samples with titer available for subjects in category 5
- TEAEs starting on the same date or after the first ADAb positive result, including Baseline (includes ADAb Categories 2, 3, 4, 5 and 6) where TEAEs have occurred on or after the following events: a) the first positive ADAb results for subjects in categories 2,

3, 4 and 6, and b) the first post-baseline boosted ADAb titer result for subjects with titer results and the first post-Baseline positive ADAb result for subjects with positive ADAb at Baseline with no other samples with titer available for subjects in category 5

- TEAEs for subjects who are ADAb negative at all timepoints (includes ADAb Category 1)

TEAEs by NAb status (ADAb negative, NAb positive, ADAb positive / NAb, negative) on subjects treated with BKZ. TEAEs will be sorted by system organ class, high level term and preferred term.
Note: for TEAE by onset relative to ADAb positivity status and by NAb status, all available ADAb and NAb data at the time of IA cut-off, respectively, will be utilized to derive the subject-level ADAb/NAb status categories.

Change #75
Section 10.2.4 Other safety topics of interest
The following has been updated
The analyses produced for the above TEAE are based on the specifications described in the BKZ safety topic of interest document version from 8-Apr-2019.
Has been updated to:
The analyses produced for the above TEAE are based on the specifications described in the BKZ safety topic of interest document version from Apr-2021

Subsections were created for all safety topics of interest separately and details have been updated in the sub sections.

Change #76
Added section 10.2.5 COVID-19 impact
See SAP for main text

Change #77
Updated Tables 10.3, 10.4, and 10.5

Change #78
Updated section 11 References

13.2 Amendment 2

Rationale for the amendment

The primary purpose of this amendment was to address discrepancies of the SAP amendment 1 and additional updates as requested by UCB.

Modifications and changes

Change #1

Added the following text to Section 3.5: Protocol deviations

In addition to IPDs resulting in the exclusion from the PPS, subjects who reduce the dose or dosing frequency of certain medications due to intolerance, AE, side-effects or receive new prohibited medication for AE will be removed from the PPS. While this is not an IPD, as it is allowed per protocol for safety reasons, these subjects will be removed from the PPS as this non-PD could have an affect the primary efficacy outcome in the same way as flagged IPDs resulting in exclusion from the PPS.

Change #2

Section 3.6.5: Per Protocol Set

The Per-Protocol Set (PPS) will consist of all subjects in the RS who had no IPD affecting the primary efficacy. The IPDs will be predefined and subjects with IPDs will be evaluated during ongoing data cleaning and data evaluation meetings prior to unblinding of the data (Section 3.5). Exclusion from the FAS will be considered as IPDs that also result in exclusion from the PPS.

Supportive analysis of the primary efficacy variable will be performed on the PPS.

Has been changed to:

The Per-Protocol Set (PPS) will consist of all subjects in the RS who had no IPD **or non-PD related to prohibited medications** affecting the primary efficacy. The ~~**IPDs**~~ **deviations** will be predefined and subjects with ~~**IPDs**~~ **deviations** will be evaluated during ongoing data cleaning and data evaluation meetings prior to unblinding of the data (Section 3.5). Exclusion from the FAS will be considered as **an** IPD~~**s**~~ that also result**s** in exclusion from the PPS.

Supportive analysis of the primary efficacy variable will be performed on the PPS.

Change #3

Section 4.2.1: Strategy for handling missing data for efficacy analyses

Added line to Table 4-1

Efficacy endpoint	Minimum	Maximum	Integer value
Tender dactylitis count	**0**	**20**	**Yes**

Change #4

Section 4.2.2.3: MI-MCMC / Reference-based imputation (steps of the procedure)

In the case of continuous endpoints, the procedure will be implemented as follows on the raw values:

1. Data will be processed sequentially, one timepoint (visit) at a time, by repeatedly calling SAS® PROC MI to impute missing outcome data at visits t=1,…, Week 16 (Week 16 being the time point of interest) using data from the placebo-treated subjects only.

a. Initialization. Set t=1 (Baseline visit).

b. Iteration. Set t=t+1. Create a dataset combining records from BKZ and placebo subjects with columns for covariates (bone erosion and region) and outcomes at visits 1 to t. Outcomes for all BKZ randomized subjects are set to missing at visit t and set to observed or previously imputed values at visits 1 to t-1. Outcomes for placebo-treated subjects are set to observed at visit t or observed or previously imputed values at visits 1 to t-1. The outcomes should be sorted in chronological order in the model.

c. Imputation. Impute missing values for visit t using previous outcomes for visits 1 to t-1, bone erosion at Baseline, and region. Note that only placebo data will be used to estimate the imputation model since no outcome is available for BKZ-treated subjects at visit t. As a consequence, the input dataset should include all subjects from Placebo but only subjects from the BKZ arm that have values at timepoint t missing.

d. Repeat steps 2a-2d for all timepoints, (…)

Has been changed to:

In the case of continuous endpoints, the procedure will be implemented as follows on the raw values:

1. Data will be processed sequentially, one timepoint (visit) at a time, by repeatedly calling SAS® PROC MI to impute missing outcome data at visits t=1 **(Baseline)**,…, Week 16 (Week 16 being the time point of interest) using data from the placebo-treated subjects only.

a. Initialization. Set t=1 (Baseline visit). **Create a dataset combining all records from all subjects with columns for covariates (bone erosion and region) and outcome at Baseline. Impute missing values at Baseline using bone erosion at Baseline, and region.**

b. Iteration. Set t=t+1. Create a dataset combining records from BKZ **subjects with missing data at visit t** and **all** placebo subjects with columns for covariates (bone erosion and region) and outcomes at visits 1 to t. **In this dataset**, outcomes for ~~**all**~~ BKZ randomized subjects are ~~**set to**~~ missing at visit t and ~~**set to**~~ observed or previously imputed ~~**values**~~ at visits 1 to t-1. Outcomes for placebo-treated subjects are ~~**set to**~~ observed **or missing** at visit t ~~**or**~~ **and** observed or previously imputed ~~**values**~~ at visits 1 to t-1. The outcomes should be sorted in chronological order in the model.

c. Imputation. Impute missing values for visit t using previous outcomes for visits 1 to t-1, bone erosion at Baseline, and region. Note that only placebo data will be used to estimate the imputation model since no outcome is available for BKZ-treated subjects at visit t. As a consequence, the input dataset should include all subjects from Placebo but only subjects from the BKZ arm that have values at timepoint t missing.

d. Repeat steps **1b-1c** for all timepoints **after Baseline up to Week 16**, (…)

Change #5

Section 4.8 Examination of subgroups (5th bullet from the end)

- Concomitantly receiving MTX vs. cDMARDs at Baseline (concomitant MTX, no concomitant MTX and cDMARDs at Baseline, no concomitant MTX and no cDMARDs at Baseline)

Has been changed to:

- Concomitantly receiving MTX **at Baseline** vs. **other** cDMARDs at Baseline (~~**concomitant**~~ MTX **at Baseline**, no ~~**concomitant**~~ MTX **at Baseline** and cDMARDs at Baseline, no ~~**concomitant**~~ MTX **at Baseline** and no cDMARDs at Baseline)

Change #6

Added new text to Section 5.1 Subject Disposition

The numbers and percentages of these subjects who either complete the SFU visit or not will additionally be presented. The numbers and percentages of randomized subjects entering the Open Label Extension (OLE) study and those not entering the OLE study will also be presented.

Change #7

Section 6.2.2 Other baseline characteristics

Added**:**

Past cDMARDs Therapy (Yes, No)

Change #8

Following text was added to Section 5.3 Protocol deviations.

number and percentage of subjects excluded from the PPS due to reason other than PD.

A by-subject listing of subjects excluded from the PPS for reasons other than IPD will be provided.

Change #9

Section 7 Measure of treatment compliance

A summary of percent treatment compliance categorized as ≤75% and >75% will be provided by treatment group, as well as a by-subject listing of treatment compliance.

Has been changed to:

A summary of percent treatment compliance categorized as <75% and >=75% will be provided by treatment group, as well as a by-subject listing of treatment compliance.

Change #10

Added following text in Section 8.1.2 Primary analysis of the primary efficacy variable

Considering that the number of subjects randomized in Japan is low (less than 10% of the RS), statistical models might not converge. If a model (logistic model or mixed model) is not converging, all related adjusted statistics and p-value will not be presented: "NE" for "Not Evaluable" will be displayed instead.

Change #11

Section 8.2.1.4 MDA at Week 16 (last but one paragraph)

The following rule will be applied for subjects with BSA <3 at Baseline: Subjects with BSA <3 at Baseline will always meet the criteria PASI ≤1 or BSA ≤3 except in the cases where a BSA score ≥3 is observed.

Has been changed to:

The following rule will be applied for subjects with BSA <3 at Baseline: Subjects with BSA <3 at Baseline will always meet the criteria PASI ≤1 or BSA ≤3 except in the cases where a post-Baseline BSA score >3 is observed

Change #12

Following text has been added in Section 8.2.3 Supportive analyses of the secondary efficacy variables

For the MDA (and VLDA) MI analysis, subjects with BSA <3 at Baseline will always meet the criteria PASI ≤1 or BSA ≤3 except in the cases where a BSA score >3 is observed. Subjects involving at least 3% of BSA at Baseline will have their BSA values imputed using MI when deriving MDA (VLDA).

Change #13

Section 8.3.1.6: Psoriatic Arthritis Disease Activity Score (PASDAS) categories and change from Baseline in PASDAS

The following categories are used to define the level of disease activity:

• Low disease activity: PASDAS 1.9 to <3.2

The PASDAS categories and the change from Baseline in PASDAS score will be analyzed at Week 4, 8, 12, 16.

Has been changed to:

If Tender Dactylitis count is missing because LDI=0 at Baseline, Tender Dactylitis count will be replaced by 0 in the formula.

If for any other reason a ~~any of the~~ PASDAS component is missing, PASDAS will be set to missing.

The following categories are used to define the level of disease activity:

• Low disease activity: PASDAS >1.9 to <3.2

The PASDAS categories and the change from Baseline in PASDAS score will be analyzed at Week 4, 12, 16.

Change #14

Section 8.3.2.2 Other endpoints (excluding time to ACR20/50/70 response) (paragraph about analysis of categorical endpoints)

For categorical endpoints (ie, EQ-5D-3L dimensions score, DAPSA state) using:

- The number and percentage by endpoint category using the NRI approach (as performed for the primary analysis of the primary endpoint with the exception that the NRI is replaced by the worst category imputation)
- The mean percentage by endpoint category in the multiple imputed datasets (modified composite estimand). Regardless of imputed values, the value of the categorical endpoint after IE are by default set to the worst category (ie: High Disease Activity for DAPSA state)

Has been changed to:

For categorical endpoints (ie, EQ-5D-3L dimensions score, DAPSA state, PASDAS categories) using:

- The number and percentage by endpoint category using the NRI approach (as performed for the primary analysis of the primary endpoint with the exception that the NRI is replaced by the worst category imputation)
- The mean percentage and corresponding 95% CI by endpoint category in the multiple imputed datasets (modified composite estimand). Regardless of imputed values, the value of the categorical endpoint after IE are by default set to the worst category (ie: High Disease Activity for DAPSA state)

Change #15

Section 9.1: Pharmacokinetics (paragraph 3)

No imputation will be used for missing samples. However, if plasma concentration measurements are below the limit of quantification (BLQ), then for the calculation of the derived statistics, the result will be set to ½ of LLOQ (ie, 1/2*0.250 = 0.125 ug/mL). Descriptive statistics including number of values, geometric mean, its 95% CI, geometric coefficient of variation, mean, SD, median, minimum, and maximum. Geometric mean and its 95% CI, geometric CV, mean and SD will be calculated if at least ⅔ of the values of interest is above the LLOQ; otherwise, only number of values, median, minimum, and maximum will be presented.

Has been changed to:

No imputation will be used for missing samples. However, if plasma concentration measurements are below the limit of quantification (BLQ), then for the calculation of the derived statistics, the result will be set to ½ of LLOQ (ie, 1/2*0.250 = 0.125 ug/mL). Descriptive statistics including number of values, geometric mean, its 95% CI, geometric coefficient of variation, mean, SD, median, minimum, and maximum. Geometric mean and its 95% CI, geometric CV, mean and SD will be calculated if at least ⅔ of the values of interest is above the LLOQ **and number of values>=3**; otherwise, only number of values, median, minimum, and maximum will be presented.

Change #16

Section 9.1: Pharmacokinetics (Third bullet)

- The table summary and figures will be primarily repeated by anti-BKZ antibody status (positive, negative, missing) by treatment group (3 lines per graph, one graph per treatment group). The missing group will not be displayed if ≥95% of subjects are categorized in the non-missing groups.

  The ADAb status (positive, negative, or missing) will be considered in a cumulative manner at each time point:

  – a subject will be counted as positive from the first visit at which the subject achieved a positive ADAb sample result to the end of the treatment period (regardless of any missing/inconclusive or negative ADAb sample result).

  – If a subject has only negative ADAb samples or only one missing/inconclusive sample with negative ADAb samples up to that timepoint, the subject will be classified as negative.

Has been changed to:

- The table summary and figures will be primarily repeated by anti-BKZ antibody status (positive, negative, missing) by treatment group (3 lines per graph, one graph per treatment group). The missing group will not be displayed if ≥95% of subjects are categorized in the non-missing groups.

  The ADAb status (positive, negative, or missing) will be considered in a cumulative manner at each time point:

  – a subject will be counted as positive from the first visit at which the subject achieved a positive ADAb sample result to the end of the treatment period (regardless of any missing/inconclusive or negative ADAb sample result).

  – If a subject has only negative ADAb samples or only one missing/inconclusive sample with negative ADAb samples up to that timepoint, the subject will be classified as negative. **An exception remains for the Baseline Visit where only one sample would be available. If the sample is missing/inconclusive, then the sample will be classified as being negative for the cumulative ADAb status.**

Change #17

Section 9.1: Pharmacokinetics (last sentence)

All plasma concentration data will be reported in ug/mL in the tables, figures, and listings.

Has been changed to:

All plasma concentration data will be reported in ug/mL in the tables, figures, and listings.

If more than 10% of the PK concentration results have been excluded from the table summaries, the PK excluded results will be listed in a separate listing.

Change #18

Section 9.2.1: Anti-bimekizumab antibody (line before the description of the ADAb categories)

Summaries of cumulative ADAb status and time to treatment-emergent positivity will use all available data.

Has been changed to:

All other summaries of ADAb status will use all available data **(scheduled and unscheduled)**.

Change #19

Section 9.2.1: Anti-bimekizumab antibody (ADAb categories)

- **Category 1: Pre ADAb negative – treatment-emergent ADAb negative**: Includes subjects who are negative at Baseline and antibody negative at all sampling points post treatment (including SFU). This group also includes subjects who have missing/inconclusive pre-treatment sample (eg either missing/inconclusive or insufficient volume) at baseline with all post-baseline samples as ADAb negative.

(…)

- **Category 6: Inconclusive**: Includes subjects who have a positive pre-treatment sample and some post-treatment samples are missing, while other post-treatment samples are ADAb negative or missing.

(…)

- **Category 9: Missing:** Includes subjects who have a negative or a missing/inconclusive pre-treatment sample and some post-treatment samples are missing/inconclusive, while other post-treatment samples are ADAb negative.

Has been changed to:

- **Category 1: Pre ADAb negative – treatment-emergent ADAb negative**: Includes subjects who are negative at Baseline and antibody negative at all sampling points post treatment (including SFU), **one post-baseline missing/inconclusive sample is allowed for subjects with pre-ADAb negative sample**. This group also includes subjects who have missing/inconclusive pre-treatment sample (eg either missing/inconclusive or insufficient volume) at baseline with all post-baseline samples as ADAb negative.

(…)

- **Category 6: Inconclusive**: Includes subjects who have a positive pre-treatment sample and some post-treatment samples are missing, while other post-treatment samples are ADAb negative **or missing**.

(…)

- **Category 9: Missing:** Includes subjects who have a negative or a missing/inconclusive pre-treatment sample and some post-treatment samples are missing/inconclusive, while other post-treatment samples are ADAb negative **or missing**.

Change #20

Added text in Section 9.2.1: Anti-bimekizumab antibody (First Bullet)

Has been changed to:

Subjects who rolled over to the OLE study will not have a SFU visit per study protocol. For these subjects, the 'overall ADAb status including SFU' will then be considered as being identical to the 'overall ADAb status up to Week 16'.

Change #21

Section 9.2.1: Anti-bimekizumab antibody (Second, third and fourth bullet)

The above summary table displaying the number and percentage of subjects in each of the ADAb status (positive, negative, missing) will be repeated by region.

- A summary table displaying the number and percentage of subjects in each of the ADAb status (positive, negative, missing) by concomitant medications (use of cDMARDs at entry, use of methotrexate at entry, use of oral/systemic corticosteroids at entry) will be provided (overall summaries only).

A table displaying the number and percentage of subjects with the first occurrence of ADAb positivity during the study (ie, including Baseline visit) will be summarized.

Has been changed to:

The above summary table displaying the number and percentage of subjects in each of the ADAb status (positive, negative, **total of positive and negative**, missing) will be repeated by region.

- A summary table displaying the number and percentage of subjects in each of the ADAb status (positive, negative, **total of positive and negative**, missing) by concomitant medications (use of cDMARDs at entry, use of methotrexate at entry, use of oral/systemic corticosteroids at entry) will be provided (overall summaries only).
  - A table displaying the number and percentage of subjects with the first occurrence of **any** ADAb positivity during the study (ie, including Baseline visit) will be summarized.

Change #22

Added text in Section 9.2.1: Anti-bimekizumab antibody

Furthermore, all ADAb titer values<100 will be represented as 1 in these plots.

Change #23

Added following text in Section 9.2.1: Anti-bimekizumab antibody (last paragraph)

Finally, if more than 10% of the ADAb results have been excluded from the table summaries, the ADAb excluded results will be listed in the same listing as the one mentioned in Section 9.1 for excluded PK results.

Change #24

Section 9.2.2: Neutralizing anti-bimekizumab antibodies (first bullet)

- ADAb negative: if the subject has all the samples as ADAb negative or only one missing/inconclusive sample with all other available samples as negative ADAb.

Has been changed to:

- ADAb negative: if the subject has all the samples as ADAb negative or only one missing/inconclusive sample with all other available samples as negative ADAb. **Note that ADAb negative samples are not subject to the neutralizing assay.**

Change #25

Section 9.2.2: Neutralizing anti-bimekizumab antibodies (Third bullet)

- NAb Missing:
  - >1 relevant NAb samples are missing and other available NAb samples during the period of interest are negative, eg, missing or insufficient sample left for NAb testing.

Has been changed to:

- NAb Missing:
  - >1 relevant NAb samples are missing**/inconclusive** and other available NAb samples during the period of interest are negative, eg, missing or insufficient sample left for NAb testing.

Change #26

Section 10.2.1: Standard AE summaries (18th bullet)

- TEAEs by timing of onset relative to ADAb status by SOC, HLT, PT (on subjects treated with BKZ). This will include columns for the following:
  - TEAEs starting before the first ADAb positive result (includes ADAb categories 2, 4 and 5) where TEAEs have occurred before the following events: a) the first positive ADAb result for subjects in category 2 and b) the first post-baseline boosted ADAb titer result for subjects with titer results and the first post-Baseline positive ADAb result for subjects with positive ADAb at Baseline with no other samples with titer available for subjects in category 5.

Has been changed to:

- TEAEs by timing of onset relative to ADAb status by SOC, HLT, PT (on subjects treated with BKZ). This will include columns for the following:
  - TEAEs starting before the first ADAb positive result (includes ADAb categories 2,and 5) where TEAEs have occurred before the following events: a) the first positive ADAb result for subjects in category 2 and b) the first post-baseline boosted ADAb titer result for subjects with titer results and the first post-Baseline positive ADAb result for subjects with positive ADAb at Baseline with no other samples with titer available for subjects in category 5.

Change #27

Section 10.3: Clinical laboratory evaluations (2nd bullet after Table 10-5)

- Number and percentage of subjects by CTCAE grade (version 4.03) (when applicable) based on minimum/maximum post Baseline value (for blood chemistry and hematology).

Has been changed to:

Number and percentage of subjects by CTCAE grade (version 4.03) (when applicable) based on minimum/maximum post Baseline value (for blood chemistry and hematology). **Subjects who meet the decreased potassium criterion of 3.0-<LLN, which is specified as the decreased potassium lab criterion for both CTCAE Grade 1 and Grade 2, will be counted as Grade 2**

14 STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

This document has been reviewed and approved per the Review and Approval of Clinical Documents Standard Operating Procedures. Signatures indicate that the final version of the Statistical Analysis Plan (SAP) or amended SAP is released for execution.

Approval Signatures

Name: PA0011 SAP AMENDMENT 2.0

Version: 1.0

Document Number: CLIN-000180308

Title: PA0011 Statistical Analysis Plan Amendment 2.0

Approved Date: 09 Nov 2021

Document Approvals	
Approval Verdict: Approved	Name: [redacted] Capacity: Document Author Date of Signature: 09-Nov-2021 02:19:54 GMT+0000
Approval Verdict: Approved	Name: [redacted] Capacity: Clinical Date of Signature: 09-Nov-2021 07:40:46 GMT+0000
Approval Verdict: Approved	Name: [redacted] Capacity: Clinical Date of Signature: 09-Nov-2021 14:18:11 GMT+0000