Protocol A4091058

A Phase 3 Randomized, Double-Blind, Active-Controlled, Multicenter Study of the Long-Term Safety and Efficacy of Subcutaneous Administration of Tanezumab in Subjects with Osteoarthritis of the Hip or Knee

Statistical Analysis Plan (SAP)

Version: 3.0

Author: PPD (Statistics, GPD)

Date: 01 March 2019

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

1. AMENDMENTS FROM PREVIOUS VERSION(S)

Table 1. Summary of Changes

Version/ Date	Associated Protocol Amendment	Rationale	Specific Changes
1 12 Feb 2014	Original 01 Apr 2015	N/A	N/A
2 28 Mar 2016	1 10 Jul 2015	Regulatory input, clarification of prior version	• Clarifications on various analysis populations or subsets for efficacy, safety, etc; • Analysis windows for visit-based assessments and for pain diary weekly means; • Updates and clarifications on various endpoints for analysis purpose including PK, PD and OA biomarkers; • Specifications on the safety and efficacy analysis and summary periods; • Revising the Tier 3 AE definition from 1% to 5%; • Miscellaneous updates or changes to align with the amended protocol.
3 01 Mar 2019	2 15 May 2016	Blinded data review, Clarification of prior version, Program decisions	• Various formatting and editorial clarifications (throughout); • Updating of definitions and analyses for consistency with protocol and program (throughout); • Clarification of treatment period and safety follow-up period as main periods for safety summaries (throughout); • Specification of gatekeeping testing strategy for co-primary and key secondary endpoints (Section 4.2); • Clarification of analysis set definitions (Section 5);

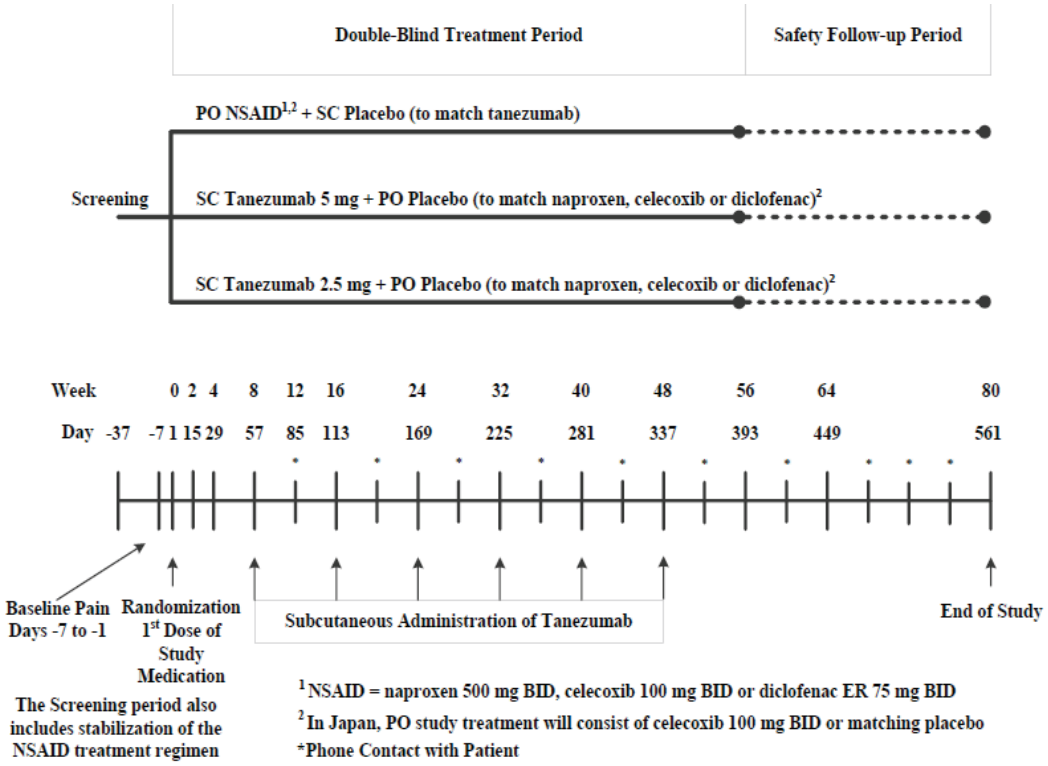
			• Clarification of baseline definition (Section 6); • Updated adverse event terms (Section 6.2); • Specification of seeds for efficacy datasets (Section 7); • Clarification of on-/off-treatment data and windowing (Section 8); • Addition of sensitivity analysis excluding sites with potential compliance issues (Section 8.2.1); • Clarification of joint space width analysis (Section 8.2.3.1).

2. INTRODUCTION

Note: in this document any text taken directly from the protocol is ***italicized***.

2.1. Study Design

The study design is summarized in the diagram below.



Subjects will be randomized to one of three treatment groups (placebo SC + NSAID, tanezumab 2.5 mg SC + oral placebo, or tanezumab 5 mg SC + oral placebo). NSAID dosing will be BID, and SC dosing will be every 8 weeks at Baseline and Weeks 8, 16, 24, 32, 40 and 48 (a maximum of 7 doses of SC tanezumab or placebo).

The end of treatment period is at Week 56, with the safety follow-up period up to Week 80. The primary time point for efficacy is Week 16. The period of interest for most safety results is the treatment period. Selected safety results will also be provided separately for the safety follow-up period, and some results will be provided for the combined overall study period comprising the treatment and safety follow-up periods.

The randomization is stratified by Index Joint (Knee or Hip), the NSAID treatment that the subject was taking prior to the study start (naproxen, celecoxib or diclofenac), and the highest Kellgren-Lawrence grade (Grade 2, 3 or 4) for the subject for any knee or hip joint, which may not necessarily relate to the index joint.

2.2. Study Objectives

Primary Objectives

- *Characterize the long-term risk of joint safety events in subjects with osteoarthritis of the knee or hip who receive tanezumab 2.5 mg or tanezumab 5 mg SC versus NSAID treatment (naproxen 500 mg BID, celecoxib 100 mg BID, or diclofenac ER 75 mg BID) over the course of 56-weeks of treatment using a composite endpoint (includes adjudication outcomes of rapidly progressive osteoarthritis type-1 or type-2, subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture).*

- *Demonstrate superior efficacy of tanezumab 2.5 mg and tanezumab 5 mg SC versus NSAID treatment (naproxen 500 mg BID, celecoxib 100 mg BID, or diclofenac ER 75 mg BID) at Week 16.*

Secondary Objectives

- *Characterize the long-term joint safety risk using a composite endpoint (includes adjudication outcomes of rapidly progressive osteoarthritis (type-2 only), subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture).*

- *Characterize the long-term risk of the following individual adjudication outcomes occurring: rapidly progressive osteoarthritis (type-1 only), rapidly progressive osteoarthritis (type-2 only), rapidly progressive osteoarthritis (type-1 or type-2 combined), subchondral insufficiency fracture, primary osteonecrosis, and pathological fracture.*

- *Characterize the long-term risk of all-cause total joint replacements (subjects who undergo total joint replacement plus subjects who have an adjudicated outcome of rapidly progressive osteoarthritis type-1 or type-2, subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture whether they undergo total joint replacement or not) occurring.*

- *Characterize joint space width changes in subjects with Kellgren-Lawrence Grade 2 or 3 osteoarthritis of the index knee or index hip.*

- *Demonstrate superior efficacy of tanezumab 5 mg and tanezumab 2.5 mg SC versus each separate NSAID treatment group (naproxen 500 mg BID, celecoxib 100 mg BID and diclofenac ER 75 mg BID) at Week 16.*

- *Demonstrate the efficacy of tanezumab 2.5 mg and tanezumab 5 mg SC versus NSAID (combined) treatment at all time points to Week 56.*

- *Evaluate the long-term safety of tanezumab 2.5 mg and tanezumab 5 mg SC.*

- *Explore relationships between adjudicated outcomes of rapidly progressive osteoarthritis (type-1 or type-2), subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture and variables that may be associated with these orthopedic risks.*

- *Characterize changes in physical activity level.*

3. INTERIM ANALYSES, FINAL ANALYSES AND UNBLINDING

There is no interim analysis for efficacy data planned for this study. The final analysis will be performed after the database is released.

Safety data will be subject to regular and ongoing reporting and review throughout the study. The details of these interim analyses will be documented in a separate Statistical Analysis Plan. Review of the safety data will be by the tanezumab external Data Monitoring Committee (E-DMC).

A blinded Adjudication Committee will be convened and asked to review all possible or probable joint-related safety events identified by the Central Reader (rapidly progressive osteoarthritis (type-1 or type-2), subchondral insufficiency fractures, primary osteonecrosis, or pathological fracture), total joint replacement as well as investigator reported adverse events of osteonecrosis, rapidly progressive osteoarthritis, subchondral insufficiency fracture or pathologic fracture. Adverse events related to joint safety that the investigator or sponsor considers medically important may also be reviewed by the Adjudication Committee. A stopping rule relating to a set of adjudicated outcomes has been defined, and is described below.

If the blinded Adjudication Committee identifies adjudicated events of rapidly progressive osteoarthritis type 2, subchondral insufficiency fractures, primary osteonecrosis, or pathological fracture occurring at a combined rate that could trigger the protocol-based stopping criteria, an urgent, ad hoc assessment of the events will be made by the Data Monitoring Committee.

The protocol (or treatment group) stopping rule has three components; the difference in the number of subjects with an adjudicated joint safety event, the exposure-adjusted risk difference (RD) and the exposure adjusted risk ratio (RR) between each tanezumab treatment group and the active comparator group. The exposure-adjusted RD will be calculated as the difference in the ratios of the number of subjects with an adjudicated joint safety event divided by exposure (patient-years) between each tanezumab group and the comparator group. The exposure-adjusted RR will be similarly calculated using the ratio of exposure adjusted event rates (number of subjects with an adjudicated joint safety event divided by exposure) for each tanezumab group relative to the comparator group. The exposure will be calculated as the combined treatment and follow-up periods.

If the protocol-based stopping rule is triggered, the E-DMC will formulate a recommendation whether it is safe to continue dosing in some or all treatment groups or whether the study should be terminated completely. This decision will be made by Pfizer in consultation with the E-DMC.

A separate set of dosing suspension rules for specified Serious Adverse Events and events consistent with Hy's Law are described in Section 9.6.1 of the protocol.

4. APPROPRIATE HYPOTHESES AND DECISION RULES

4.1. Statistical Hypotheses

The treatment comparisons being made in this study are tanezumab 2.5 mg and 5 mg versus NSAID (for combined, and individual NSAIDs). The co-primary efficacy endpoints are changes from baseline to Week 16 in the WOMAC pain subscale and physical function subscale and in the Patient's Global Assessment of Osteoarthritis, respectively. For these treatment comparisons, the null and alternative hypotheses are shown below (note $\mu_{TREATMENT}$ relates to the mean change from Baseline for the specified treatment group). All tests will be 2-sided.

Null Hypotheses	H0: $\mu_{\text{TANEZUMAB 2.5mg}} - \mu_{\text{NSAID}} = 0$
	H0: $\mu_{\text{TANEZUMAB 5mg}} - \mu_{\text{NSAID}} = 0$
Alternative Hypotheses	H1: $\mu_{\text{TANEZUMAB 2.5mg}} - \mu_{\text{NSAID}} \neq 0$
	H1: $\mu_{\text{TANEZUMAB 5mg}} - \mu_{\text{NSAID}} \neq 0$

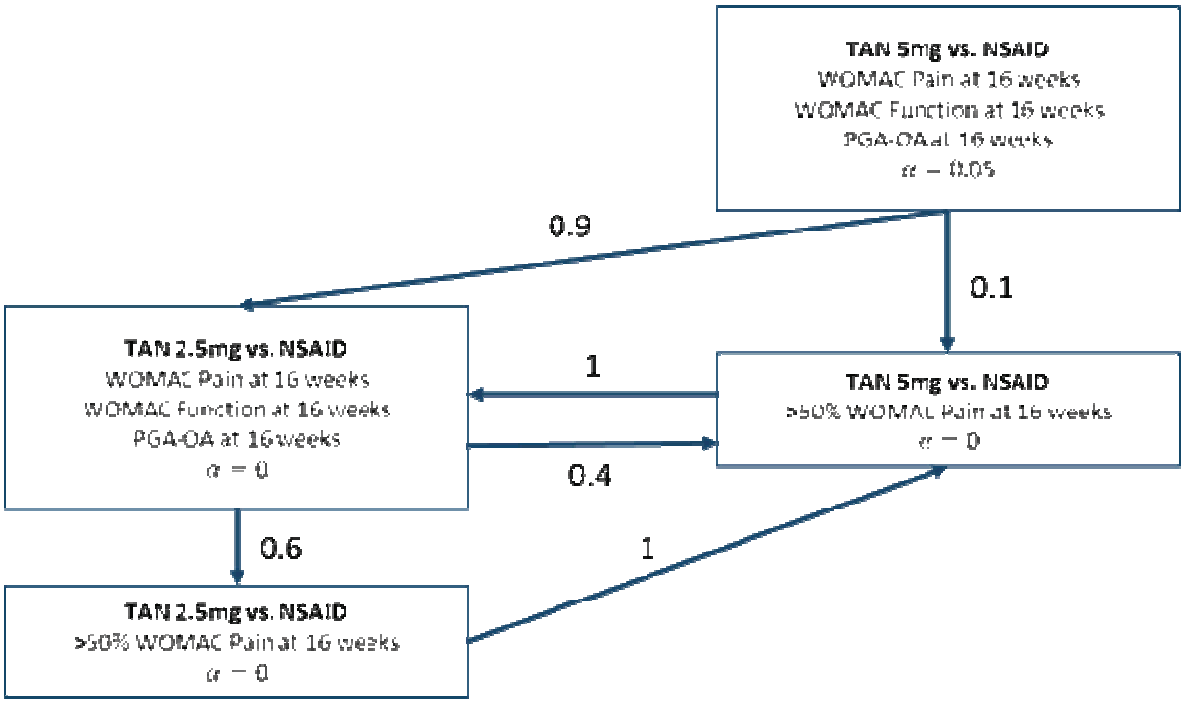
The hypotheses for other types of analyses (eg, for the binary response endpoints) would be similar to those shown above.

4.2. Statistical Decision Rules

The Type I error rate (α-level) used in the assessment of pair-wise treatment comparisons for the primary efficacy endpoints is 5%. The testing strategy of the co-primary and key secondary efficacy endpoints is described below.

In addition to the three co-primary endpoints, the percentage of subjects with ≥50% reduction from baseline in WOMAC Pain at Week 16 is identified as a key secondary endpoint. The testing of these co-primary endpoints and key secondary endpoint will follow the graphical approach of gate keeping strategy proposed by Bretz et al (2011), as depicted in the following figure. This will be implemented to control the family-wise type I error rate of 5% (two-sided), and this graphical approach is a closed testing procedure; hence, it strongly controls the family-wise error rate (Alosh et al. 2014).[5]

Figure 1. Graphical Multiple Testing Procedure for Strong Control of Type I Error



The co-primary outcome of tanezumab (TAN) 5 mg versus NSAID will first be tested at a two-sided α=0.05 (ie, WOMAC Pain subscale, WOMAC Physical Function subscale, and PGA at 16 weeks for TAN 5 mg vs NSAID must all be significant at two-sided α=0.05 for the co-primary to be significant). If the null hypothesis is not rejected, no further testing is conducted as the α for that test is considered "spent" and cannot be passed to other endpoints. The testing process continues as long as at least one null hypothesis can be rejected at its assigned α-level. Each time a null hypothesis is rejected, the graph is updated to reflect the reallocation of α assigned to that hypothesis, which is considered "recycled" (Alosh et al. 2014). This iterative process of updating the graph and reallocating α is repeated until all hypotheses have been tested or when no remaining hypotheses can be rejected at their corresponding α level.

In the Figure 1, the three co-primary endpoints are represented by a single graphical node per dose. As described above, a particular tanezumab dose group is declared as superior to placebo on the primary analysis if the corresponding treatment contrast is significant over all three co-primary endpoints. This is equivalent to a sequential gatekeeping approach where each of the three co-primary endpoints are tested in order of WOMAC Pain, WOMAC Physical Function, and PGA. If all three are significant, the alpha is reallocated to the next endpoint(s). For clarity of the graph (and to highlight that all three endpoints must be significant for that dose), the co-primaries are grouped by dose rather than displaying all six tests individually. This testing procedure will maintain the Type I error to 5% or less for each dose's co-primary efficacy endpoint overall, and to less than 5% for each dose's three co-primary efficacy endpoints individually.

For Japan, the procedures outlined above are followed for the two co-primary endpoints, WOMAC Pain and WOMAC Physical Function.

The primary analysis will be that with the multiple imputation approach (see below for details), and thus the overall type I error is controlled for each of the two doses (2.5 mg and 5 mg) since all three (two) co-primary endpoints need to be significant for a single dose. The overall type I error of the study is also controlled given the step-down testing strategy for each of the endpoints. Control of the family wise type I error rate accounting for multiplicity of contrasts will only apply to the three (two) co-primary endpoints (model with the primary imputation analysis) and the key secondary endpoint.

Regardless of the outcome of the primary analyses and the key secondary endpoint, other secondary endpoints will be tested. No adjustment for multiple comparisons will be made for the secondary and safety endpoints. The α-level for each hypothesis test for the secondary and exploratory analyses will be 5%. There is no testing strategy for the primary safety endpoint.

5. ANALYSIS SETS

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

5.1. Full Analysis Set

The intent to treat (ITT) analysis set is the primary analysis set for efficacy and safety analyses. It consists of all randomized subjects who received at least one dose of SC study medication (either tanezumab or placebo SC). This analysis set is used in the presentations of all efficacy data, and all data listings, and is labeled as the 'ITT Analysis Set' or 'ITT Population'.

This analysis set is expected to be the same as the safety analysis set. If a subject is treated without having been randomized, then the analysis sets will be different, and safety analyses will use this different safety analysis set (see Section 5.3). If the sets are the same, safety analyses may still refer to it as the 'Safety Population'.

5.2. Per Protocol Analysis Set

The per-protocol (PP) analysis set is the secondary efficacy analysis set. It is defined as all subjects in the ITT analysis set who are not major protocol deviators (which would potentially affect efficacy). The criteria for major protocol deviators are described below in Section 5.6. The identification of specific subjects included and excluded (and reason for exclusion) for this analysis set will be done prior to unblinding. Protocol deviations for the PP analysis set will be obtained from the collected list of potentially important protocol deviations, and this list will comprise deviations identified from review of programmed listings and study monitoring. This analysis set is used in a specific sensitivity analysis of the co-primary efficacy endpoints, and is labeled as the 'Per Protocol Population'.

5.3. Safety Analysis Set

The safety analysis set is defined as all subjects treated with tanezumab or placebo SC. And it will be labeled as the 'Safety Analysis Set' or 'Safety Population' in the corresponding data analysis and summary presentations.

5.4. Other Analysis Sets

Accelerometry Analysis Set: This is defined as all subjects treated with tanezumab or placebo SC who have any baseline or post-baseline accelerometry data. Only a subset of the enrolled subjects will have accelerometry data collected.

Joint Safety Event Biomarker Analysis Set: This analysis set will be described separately from this analysis plan.

NGF Subgroup Analysis Set: This analysis set includes selected subjects in each treatment group for whom serum NGF results are available.

Synovial Fluid Analysis Set: Subjects who have arthrocentesis performed and for whom synovial fluid data is available.

5.5. Treatment Misallocations

If a subject was:

- Randomized but not treated with SC treatment, then that subject will be excluded from all efficacy and safety analyses.
- Treated but not randomized, then by definition that subject will be excluded from the efficacy analyses, but will be reported under the treatment they actually received for all safety analyses.
- Randomized but received incorrect treatment, then that subject will be reported under their randomized treatment group for all efficacy analyses, but will be reviewed on a case-by-case basis by the study team and a decision on potential changes related to the subject and on how to analyze the data for safety analyses will be made in a timely manner and prior to database unblinding.

5.6. Protocol Deviations

The PP analysis set is the secondary efficacy analysis set. It is defined as all subjects in the ITT analysis set who are not major protocol deviators (which would potentially affect efficacy). The criteria for defining a major protocol deviator are described below in Section 5.6.1 and Section 5.6.2. The identification of specific subjects included and excluded (and reason for exclusion) for this analysis set will be made and documented prior to unblinding. Any other major deviation which is not pre-specified below, but results in a subject being excluded from the PP analysis set, will be specified in the protocol deviations document which is completed prior to unblinding.

The following protocol deviations are defined as 'major' and would exclude a subject from the PP analysis set (see Section 5.2). These deviation criteria can be split into those assessed prior to randomization relating to the protocol inclusion and exclusion criteria, and those assessed post randomization.

5.6.1. Major Deviations Assessed Prior to Randomization

- Inclusion criteria: #3-5.
- Exclusion criteria: #3, 4 (if any of the following conditions are in the index joint: severe chondrocalcinosis, other arthropathies [eg, rheumatoid arthritis], systemic metabolic bone disease [eg, pseudogout, Paget's disease, metastatic calcifications], primary or metastatic tumor lesions, stress or traumatic fracture),10, 16, 17, 18 (if index joint was involved), 19.
- Randomization criteria: #1, 3-6. Note, subjects with missing Baseline data for any of the co-primary endpoints would not meet the randomization/inclusion criteria for Baseline co-primary endpoints and so would be defined as a deviation according to these criteria.

5.6.2. Major Deviations Assessed Post-Randomization

- Mismatch in specification of index joint in the CRF vs. electronic tablet for WOMAC data collection.
- Rescue medication taken within 24 hours prior to the Week 16 visit.
- Prohibited medications that could affect pain and function assessments (protocol section 5.8.1) taken (i) within 48 hours prior to Week 16 visit for non-NSAID medications (or any use if long-acting, eg, Synvisc), or (ii) within 48 hours prior to Week 16 visit or within the wash-out period specified by Appendix 3 of the protocol, for NSAID medications.
- Subjects who were <50% compliant with oral study medication between the baseline and the Week 16 visit.

In addition, unforeseen major protocol deviations may be added to this list. However the final definition of this criteria and the per-protocol population will be made prior to unblinding of this study.

6. ENDPOINTS AND COVARIATES

Baseline is generally defined as the last observation prior to first receipt of study drug, within the baseline window as defined in Appendix 1.1.

For analysis of diary pain intensity scores for the index joint, baseline is defined as the mean average daily Pain NRS score using the last 3 values during the final 7 days of the Initial Pain Assessment Period prior to Randomization/Day 1.

6.1. Efficacy Endpoint(s)

The co-primary efficacy endpoints are listed below.

- *Change from Baseline to Week 16 in the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) Pain subscale.*
- *Change from Baseline to Week 16 in the WOMAC Physical Function subscale.*
- *Change from Baseline to Week 16 in the Patient's Global Assessment of Osteoarthritis* (Note, this is not a primary endpoint for Japan).

The secondary efficacy endpoints are listed below.

- *WOMAC Pain subscale change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, 56 and Week 64.*
- *WOMAC Physical Function subscale change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, 56 and Week 64.*
- *Patient's Global Assessment of Osteoarthritis change from Baseline to Weeks 2, 4, 8, 16 (Japan only), 24, 32, 40, 48, 56 and Week 64.*
- *OMERACT-OARSI responder index at Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*
- *Treatment Response: Reduction in the WOMAC Pain subscale of ≥30%, ≥50%, ≥70% and ≥90% at Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*
- *Cumulative distribution of percent change from Baseline in the WOMAC Pain subscale score to Week 16, 24, and 56 (endpoint for summary only).*
- *Treatment Response: Reduction in the WOMAC Physical Function subscale of ≥30%, ≥50%, ≥70% and ≥90% at Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*
- *Cumulative distribution of percent change from Baseline in the WOMAC Physical Function subscale score to Week 16, 24, and 56 (endpoint for summary only).*
- *Treatment Response: Improvement of ≥2 points in Patient's Global Assessment of Osteoarthritis at Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*
- *Average pain score in the index joint change from Baseline to Weeks 1, 2, 3, 4, 6, 8, 10, 12, 16, 20, 24, 32, 40, 48, 56 and Week 64.*
- *WOMAC Stiffness subscale change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*

- *WOMAC Average change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*

- *WOMAC Pain Subscale Item: Pain When Walking on a Flat Surface, change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*

- *WOMAC Pain Subscale Item: Pain When Going Up or Down Stairs, change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*

- *Work Productivity and Activity Impairment Questionnaire for Osteoarthritis (WPAI:OA) impairment scores change from Baseline to Weeks 16, 24, 56 and 64.*

- *EuroQol 5 Dimension (EQ-5D-5L™) dimensions and overall health utility score at Baseline and Weeks 8, 16, 24, 40, 56 and 64.*

- *Treatment Satisfaction Questionnaire Medicine v.II (TSQM v.II) satisfaction with effectiveness, side effects and convenience, and overall satisfaction at Weeks 16 and 56.*

- *Patient Reported Treatment Impact Assessment-Modified (mPRTI) at Weeks 16 and 56.*

- *Incidence and Time to discontinuation due to Lack of Efficacy.*

- *Usage of rescue medication (incidence and number of days of use) during Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56 and Week 64.*

- *Usage of rescue medication (amount taken) during Weeks 2, 4, 8 and Week 16.*

- *Health Care Resource Utilization at Baseline and Weeks 64, and 80.*

Note, in this document where reference is made to three co-primary efficacy endpoints this relates to all countries except Japan, where there are only two co-primary efficacy endpoints (WOMAC Pain and Physical Function subscale scores).

6.2. Safety Endpoints

The primary joint safety endpoint is listed below:

- *Incidence of a predefined composite endpoint consisting of adjudication outcomes of rapidly progressive osteoarthritis (type-1 or type-2), subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture (primary composite endpoint).*

The secondary bone and joint safety endpoints are listed below:

- *Incidence of a predefined composite endpoint consisting of adjudication outcomes of rapidly progressive osteoarthritis (type-2 only), subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture.*

- *Incidence of individual adjudication outcomes of rapidly progressive osteoarthritis (type-1 only), rapidly progressive osteoarthritis (type-2 only), rapidly progressive osteoarthritis (type-1 or type-2 combined), subchondral insufficiency fracture, primary osteonecrosis, and pathological fracture.*

- *Incidence of all-cause total joint replacements (subjects who undergo total joint replacement plus subjects who have an adjudicated outcome of rapidly progressive osteoarthritis type-1 or type-2, subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture whether they undergo total joint replacement or not).*

Reporting of total joint replacement events including surgery and recovery will be described in a separate Statistical Analysis Plan for Study 1064, and reported in the 1064 study report. Corresponding data from Studies 1056, 1057 and 1058 will be reported under study 1064, as well as patients who enter study 1064 from studies 1059, 1061, and 1063 due to those studies closing out.

The Radiographic Endpoints are listed below:

- *Change from Baseline to Week 56 and Week 80 in Medial or Lateral Minimum Joint Space Width of the index knee (for subjects with Kellgren-Lawrence Grade 2 or 3 medial or lateral osteoarthritis of the index knee).*

- *Change from Baseline to Week 56 and Week 80 in Minimum Joint Space Width of the index hip (for subjects with Kellgren-Lawrence Grade 2 or 3 osteoarthritis of the index hip).*

- *Incidence of subjects with progression of osteoarthritis in the index knee according to Bland and Altman method, at Week 56 and Week 80 (separately) (for subjects with Kellgren-Lawrence Grade 2 or 3 medial or lateral osteoarthritis of the index knee).*

- *Incidence of subjects with progression of osteoarthritis in the index hip according to Bland and Altman method, at Week 56 and Week 80 (separately) (for subjects with Kellgren-Lawrence Grade 2 or 3 osteoarthritis of the index hip).*

The Bland-Altman method for defining progression of Joint Space Width (JSW), is to calculate the t-statistic (based on the number of subjects in the sample for the 0.975 percentile of the t-distribution) multiplied by the standard deviation of the change from Baseline to Week 56 and to Week 80 (two separate values, giving two different definitions of progression for each joint). These values represent the limit of the 95% confidence interval (CI) for a change of 0 over these time points. Note, 95% CI for the observed mean change is

not used because this would be affected by any treatment effect seen in the trial, that this endpoint is aiming to explore, and would be potentially biased. The lower bounds of the 95% CI for change from Baseline to Week 56 and to Week 80 would represent the definitions of progression of Osteoarthritis (OA) for Hip and Knee OA separately. The incidence of subjects with a reduction in JSW will then be compared according to the appropriate definition (Week 56 or Week 80; Hip or Knee). The calculation of the progression of OA definition and analysis of incidence will be performed separately for Weeks 56 and 80 and subjects with OA of the Hip and Knee, for the subset of subjects with Kellgren-Lawrence grade 2 or 3 and with just grade 3 (knee).

The adverse events of Abnormal Peripheral Sensation (APS) are defined in the table below.

Allodynia	Neuralgia
Axonal neuropathy	Neuritis
Burning sensation	Neuropathy peripheral
Decreased Vibratory Sense	Paraesthesia
Demyelinating polyneuropathy	Paraesthesia oral
Dysaesthesia	Peripheral sensorimotor neuropathy
Formication	Peripheral sensory neuropathy
Hyperaesthesia	Polyneuropathy
Hyperpathia	Polyneuropathy chronic
Hypoaesthesia	Sensory disturbance
Hypoaesthesia oral	Sensory loss
Intercostal neuralgia	Thermohypoaesthesia
Sciatica	Carpal tunnel syndrome
	Tarsal tunnel syndrome

Adverse Events of the Sympathetic Nervous System are defined in the table below.

Abdominal discomfort	Micturition urgency
Anhidrosis	Nausea
Blood pressure orthostatic decreased	Nocturia
Bradycardia	Orthostatic hypotension
Diarrhoea	Presyncope
Dizziness postural	Respiratory distress
Early satiety	Respiratory failure
Ejaculation delayed	Sinus bradycardia
Ejaculation disorder	Syncope
Ejaculation failure	Pollakiuria
Anal incontinence	Urinary hesitation
Heart rate decreased	Urinary incontinence
Hypertonic bladder	Vomiting
Hypohidrosis	

A smaller set of the above Adverse Events will also be summarized. These are defined below.

Anhidrosis	Orthostatic hypotension
Bradycardia	Syncope
Hypohidrosis	

The lists given above may be updated depending on any additional adverse events observed in any tanezumab study. There are a number of summaries based on these groupings of adverse events.

The Neuropathy Impairment Score (NIS) is the sum of scores over all 37 items from both the left and right side. Items 1-24 are scored on a 0-4 scale (0, 1, 2, 3, 3.25, 3.5, 3.75, 4) and items 25-37 are scored on a 0-2 scale (0, 1, 2). The possible range of the NIS is 0-244.

A 3-tier approach will be used to summarize AEs. Under this approach, AEs are classified into 1 of 3 tiers. Different analyses will be performed for different tiers. A description of the three tiers and analyses are given in Section 8.2.3.

All summaries of adverse events will be shown for adverse events that begin or worsen from the first SC dose (treatment-emergent) up to the end of the treatment period. In addition a selection of adverse event tables will be produced for the safety follow-up period and for the whole period up to the end of the study, including the treatment period and safety follow-up period.

The incidence of orthostatic hypotension at each visit, at any treatment period visit (including unscheduled visits), and at any safety follow-up period visit (including unscheduled visits), will be summarized. The definition of orthostatic hypotension is:

- For subjects with Baseline supine systolic Blood Pressure ≤150 mmHg:
  - Reduction in sBP (standing minus supine) ≥20, OR
  - Reduction in dBP (standing minus supine) ≥10
- For subjects with Baseline supine systolic Blood Pressure >150 mmHg:
  - Reduction in sBP (standing minus supine) ≥30, OR
  - Reduction in dBP (standing minus supine) ≥15.

An additional summary will be provided of outcomes of assessments resulting from an incident of orthostatic hypotension or other events of interest, using data from both the CRF database and the consultation database, as appropriate.

The Survey of Autonomic Symptoms (SAS) is a 12 item (11 for females) questionnaire. From this the total number of symptoms (0-12 for males and 0-11 for females) will be calculated. Where a subject has a symptom then the impact of that symptom is then rated from 1 ('not at all') to 5 ('a lot'). The total impact score is calculated using this 1-5 scale, with 0 assigned where the subject does not have the particular symptom. The range for the total impact score is 0-60 for males and 0-55 for females.

Additional safety endpoints include anti-drug antibody (ADA) assessments.

6.3. Other Endpoints

6.3.1. PK Endpoints

- *Plasma tanezumab concentrations.*
- *Synovial fluid tanezumab concentrations (for a subset of subjects).*

6.3.2. PD Endpoints

The following assessments will be made:

- *Serum NGF assessment.*
- *Serum and urine osteoarthritis biomarker concentrations.*
- *Synovial fluid NGF assessment (for a subset of subjects).*

6.3.3. Outcomes Research Endpoints

The Baseline and Weeks 8, 16, 24, 40, 56 and 64 responses in the five dimensions (mobility; self-care; usual activities; pain/discomfort; anxiety/depression) and overall health utility score from the EuroQol 5 Dimensions (EQ-5D-5L), and the EQ-VAS will be summarized. The overall health utility score is calculated using the EuroQol value sets, and is described in Appendix 1.4.

The change from baseline to Weeks 16, 24, 56 and 64 in the impairment scores of the Work Productivity and Activity Impairment Questionnaire for Osteoarthritis (WPAI:OA) will be summarized. These are listed below:

- Percent work time missed due to Osteoarthritis.
- Percent impairment while working due to Osteoarthritis.
- Percent overall work impairment due to Osteoarthritis.
- Percent activity impairment due to Osteoarthritis.

The calculation of these endpoints is described in Appendix 1.4.

The 11 questions of the TSQM are used to calculate the 4 endpoints of Effectiveness, Side Effects, Convenience and Global Satisfaction, each scored on a 0-100 scale with 100 being the best level of satisfaction. The calculation of these 4 parameters are described in Appendix 1.4.

6.3.4. Activity Level Monitoring Endpoints

The Lower Extremity Activity Scale (LEAS) is an 18-point integer scale reflecting the subjects physical function status, ranging from 1 (bed bound) to 18 (competitive athlete).

The accelerometry endpoints of average daily minutes of physical activity, average daily physical activity counts, minutes of moderate to vigorous physical activity (MVPA), minutes of bouted MVPA, and average daily step count will be calculated from the raw data.

The Activity Level Monitoring endpoints are listed below:

- *Lower Extremity Activity Scale: change from Baseline to Weeks 4, 8, 16, 24, 56 and Week 80 (all subjects).*
- *Change from Baseline to Weeks 16 and 56 in average daily minutes of physical activity (a subset of subjects).*
- *Change from Baseline to Weeks 16 and 56 in average daily physical activity counts (a subset of subjects).*
- *Change from Baseline to Weeks 16 and 56 in average daily minutes of moderate to vigorous physical activity (a subset of subjects).*
- *Change from Baseline to Weeks 16 and 56 in average daily minutes of bouted (sustained) moderate to vigorous physical activity (a subset of subjects).*
- *Change from Baseline to Weeks 16 and 56 in average daily step count (a subset of subjects).*

A valid day of monitoring will be defined as 10 or more wear hours in a 24-hour period as verified from accelerometer output. During Screening, a complete monitoring period will be defined as containing at least 1 valid weekend day of data and a minimum of 4 valid days of monitoring. During activity level monitoring between Week 14 and the Week 16 visit and between Week 54 and the Week 56 visit, a complete monitoring period will be defined as containing at least 2 valid weekend days of data and a minimum of 8 valid days of monitoring.

For the purposes of the MVPA endpoints, the three intensity levels of activity are defined as light (100 – <1,500 counts), moderate (1,500 – <6,500 counts), and vigorous (≥6,500 counts). The data will be further summarized as total daily time (minutes) for each intensity level.

A "bout" of moderate to vigorous activity is defined as 10 or more consecutive minutes above the moderate physical activity level threshold, with allowance for interruptions of 1 or 2 minutes below the threshold.

The daily data for each endpoint will be calculated as indicated above. These daily scores will be used to calculate the score for Baseline, and Weeks 16 and 56 relative to the appropriate visit (ie, 7 days within the screening period for Baseline and 14 days prior to the Weeks 16 and 56 visits). All valid data will be used even if it exceeds the intended 7/14 days specified in the protocol.

6.4. Covariates

For all models analyzing the continuous primary and secondary efficacy endpoints (except rescue medication) the corresponding Baseline value will be used as a covariate, in addition to Baseline diary average pain. Study site will be fitted as a random effect in the ANCOVA models. The randomization stratification variables of index joint (hip or knee), NSAID cohort (celecoxib, naproxen and diclofenac), and highest Kellgren-Lawrence grade of any Hip/Knee joints (grade 2, 3 or 4) will be included as fixed effects.

A listing of subjects with mis-matches between the stratification variables entered at randomization and the case report form data (including central lab data for Kellgren-Lawrence grade) will be provided. In analysis models, the strata entered at randomization will be used, but for descriptive summarization of the population and identification of subgroups, the strata as indicated on the case report form data will be used.

For the models analyzing the amount and number of days of rescue medication use the model will include terms for Baseline WOMAC Pain, Baseline diary average pain and stratification factors.

The analysis of the incidence of treatment *discontinuation due to lack of efficacy will* include *model terms for baseline WOMAC Pain subscale score, Baseline Diary Average Pain, index joint, Kellgren-Lawrence grade, NSAID group and treatment group.*

For treatment response endpoints relating to WOMAC Pain and PGA, the Baseline WOMAC Pain and PGA values will be used respectively as covariates in the analysis model, in addition to the stratification parameters of index joint, highest Kellgren-Lawrence grade of any hip or knee joint, NSAID group and Baseline diary average pain. For the OMERACT response value, the Baseline WOMAC Pain subscale score will be used to represent the Baseline score covariate, in addition to the covariates named above (Baseline diary average pain, and the three stratification factors).

Additional analyses of the three co-primary endpoints will examine the treatment interactions with Study site and Country. Note, separate analyses will be performed for the individual NSAID groups.

The Baseline diary average pain is used in the analysis of most endpoints. However if a patient has a missing value for this covariate then to avoid exclusion of the patient for the endpoint then a Baseline value will be imputed as the patient's WOMAC Pain subscale score. This imputed value will not be used in the analysis of the Average Pain from the diary, but as a covariate for other endpoints.

6.5. Subgroup Analyses

Separate tables will be produced for subjects in the ITT analysis set in Japanese sites. These tables will be defined prior to the unblinding of the study.

The accelerometry data is collected in the subset of subjects (the accelerometry analysis set), so analysis of these subjects will occur in this subset.

Analysis of tanezumab versus individual NSAID groups will occur in the individual NSAID cohort subsets.

7. HANDLING OF MISSING VALUES

The three co-primary efficacy endpoints are the changes from Baseline to Week 16 in the WOMAC Pain subscale, the WOMAC Physical Function subscale, and the Patient Global Assessment of Osteoarthritis. For Japan, the two co-primary efficacy endpoints are the change from Baseline to Week 16 in the WOMAC Pain and Physical Function subscales.

The primary analysis of the co-primary endpoints will use multiple imputation for missing data at Week 16 (where the method for imputation will be dependent on the reason for missing data) followed by the ANCOVA analysis with the model described below for the multiple imputed datasets. The imputation strategies are described in the following table. While the table describes the multiple imputation strategy specifically for the Week 16 time point, multiple imputation analysis at other time points will use the same strategy but with the appropriate time point, eg, 'Week 2' substituted for 'Week 16' in the table below. Efficacy data missing from windows after the Week 56 window, eg, Week 64, will not be imputed for any summary or analysis unless otherwise indicated.

Type of Missing Data	Imputation Method
Missing data resulting from discontinuation due to Death, Adverse Events (AEs) or Insufficient Clinical Response (Lack of Efficacy, LoE, including Patient meets protocol specified pain criteria for discontinuation) prior to or during the Week 16 visit reporting window*.	Multiple imputations will be created by sampling from a normal distribution based on the subject's baseline score and the standard deviation (over all treatment groups) of the observed efficacy data at Week 16 over all ITT subjects. This is a multiple imputation version of BOCF single imputation method. [Seeds 1, 3, and 5 below].
Missing data for other reasons, ie, • Subject did not discontinue on or before Week 16 (includes discontinuation for any reason after the end of the Week 16 visit reporting window*) • Subject discontinued for a different reason prior to or during the Week 16 visit reporting window*.	Multiple imputations will be created by sampling from a normal distribution based on the subject's last score and the standard deviation (over all treatment groups) of the observed efficacy data at Week 16 over all ITT subjects. For example if last observation for a subject is at Week 12, then the imputation sample for that subject is created using the subject's Week 12 observation and the standard deviation of the Week 16 observations for all subjects. Note, a subject's last observation may be the Baseline observation. This is a multiple imputation version of LOCF single imputation method. [Seeds 2, 4, and 6 below].

* See Appendix 1.1 for a definition of the reporting windows.

The imputation of baseline-like data for subjects with missing data due to discontinuation due to Death, AE or LoE is intended to impute conservative efficacy values for those subjects who discontinue because of a reason that is considered to be a poor outcome for the subject, and so a poor outcome is imputed. For those subjects with missing data that is likely to not be related to treatment group, the intention is that missing data should be imputed based on a 'missing at random' assumption taking into account the subject's previous available data.

One hundred imputed samples will be used in this analysis. In order to pre-define the analysis (and not to allow the results to change if run again), the following seeds will be used in the creation of the multiple imputed data: WOMAC scores [1] 7001-7100; and [2] 8001-8100; PGA scores: [3] 9001-9100 and [4] 10001-10100; and diary pain scores: [5] 11001-11100 and [6] 12001-12100. Imputed data for the PGA will be rounded to integer scores in the range 1 to 5. Imputed data for the WOMAC subscale and Average scores, and for the diary pain scores <0 and >10 will be truncated to 0 and 10, respectively. Imputed data for the WOMAC items of Pain when Going Up or Down Stairs and Pain when Walking on a Flat Surface will be rounded to integer scores in the range 0 to 10. The ANCOVA analysis described in Section 8.1.1 (with covariates in Section 6.4) will be used for each imputation dataset, and the overall results will be calculated to take account of the variability both within and between imputation datasets using standard methods (Little & Rubin, 2002), which are described in Appendix 3.2.

This analysis will be used for the co-primary efficacy endpoints at Week 16, plus secondary analyses at other time points, and also for a range of secondary efficacy endpoints at all time points up to Week 56. When using the multiple imputation method described above for time points earlier than Week 56, then the reason for missing data is assessed up to the end of the window for that particular time point (see Appendix 1.1).

Three additional methods will explore the sensitivity of the effect of missing data. The first method of Baseline Observation Carried Forward (BOCF) for missing data at the primary time point of Week 16 will impute the subject's Baseline value for the Week 16 time point, and therefore a zero change from baseline. If a subject's baseline data is also missing then that subject's data remain missing for the post-baseline time point. The second method of Last Observation Carried Forward (LOCF) for missing data at the primary time point of Week 16 will impute the subject's last observed data value for the efficacy endpoint. With LOCF, if a subject is missing all post-baseline efficacy data for a given efficacy endpoint, then baseline will be carried forward (if baseline is missing then the subject would have no contributing data to be included in the analysis). In both the BOCF and LOCF imputation analyses, the same main effects ANCOVA model as described below will be used. The third method will use Mixed Model for Repeated Measurements (MMRM) utilizing all observed data up to and including Week 16, including data considered off-treatment (retrieved dropout; see Appendix 1.1 for details on windows; if multiple observations are within a window, only the single observation selected for analysis by the windowing algorithm will be used in the MMRM analysis).

Analyses of the three co-primary measures at secondary time points will use the BOCF and LOCF imputation methods for missing data, and use the same (main effects) ANCOVA model as described for the primary analyses.

The responder endpoints will be analyzed using logistic regression for binary data, using both BOCF and LOCF separately for missing data of the response endpoint at a particular time point. Imputation using BOCF will lead to the subject being assessed as a non-responder. In addition, in order to closely match the primary imputation analysis, a mixed BOCF/LOCF imputation for response endpoints will be used. In this analysis BOCF imputation (ie, a subject would be a non-responder) would be used for missing data due to discontinuation for reasons of lack of efficacy (Insufficient clinical response on the end of treatment Subject Summary Case Report Form, and also including Patient meets protocol specified pain criteria for discontinuation), adverse event or death up to the time point of interest, and LOCF imputation would be used for missing data for any other reason.

Note, if Baseline is missing then the subject data for the change from Baseline will be set to missing for all efficacy analyses for that parameter. A subject who has a missing Baseline score will be missing for the response criteria for endpoints where the response is based on one parameter. The OMERACT-OARSI responder index is based on 3 parameters. It is set to missing if two or three out of these three parameters are missing at baseline (per its definition, a response can be still be achieved if only one parameter is missing, regardless of which one it is).

The individual WOMAC subscales are calculated as the mean of the individual items (5 for Pain, 17 for Physical Function and 2 for Stiffness). CCI [redacted] The WOMAC Average score is calculated as the mean of the three WOMAC subscale scores of Pain, Physical Function and Stiffness. CCI [redacted] Missing WOMAC subscale or WOMAC Average scores will be subject to the imputation method of the analysis as described above.

For the analysis of the rescue medication endpoints, missing data is imputed for daily missing scores first, and then the last available weekly score (after daily missing data is imputed) will be used for subsequent missing weekly scores, as described below. For the analysis of the rescue medication endpoints while subjects are still in the study any missing data will be imputed by carrying forward the last recorded daily data up to Week 16 (LOCF daily data). Imputation using the daily data will occur up to the end of the last week when the subject is in the study (see Appendix 1.1 for definitions of the last study day in each week). For example if a subject discontinues on study day 10, then data up to the end of Week 2 will be imputed in this way. The weekly scores for the rescue medication endpoints can then be calculated for each week the subject is in the study. Rescue medication endpoints are summarized and analyzed using LOCF, and so the last weekly score for the rescue medication will be used for LOCF after the subject has discontinued from the study (note, imputation is taken from the last week with non-missing data and not necessarily from the last available study week, eg, if Week 8 is missing then Week 7 data can be used). The baseline observation will not be carried forward in the case where a post-baseline observation is not available for the LOCF imputation. In the example above, the subject who discontinued in Week 2 (Study Day 10) will have their Week 2 value used as the LOCF value for all Weeks 3-16. The BOCF imputation rule will not be used for the subject because rescue medication is collected during the Initial Pain Assessment Period only (days -7 to -1) and subjects should not be taking rescue medication within 24 hours of the Baseline visit (so part of day -1), therefore Baseline rescue medication use is not an accurate reflection of subjects' true Baseline use of rescue medication. Imputation of weekly diary data after Week 56 will not be performed.

The electronic diary data are a mix of daily and weekly average pain assessments for the index hip or knee, although the recall assessment period is the past 24 hours for both daily and weekly assessments. A weekly mean score will be calculated from the available daily pain scores. Any missing daily pain scores will be left as missing in the weekly pain score calculated. If there are no non-missing observations then the weekly score will be missing. The Baseline mean will be calculated using the last 3 actual values from the last 7 days of the Initial Pain Assessment Period (IPAP). The weekly pain scores (either calculated from the daily scores when available or directly from the weekly pain assessments) will then be utilized for the multiple imputation, and the LOCF and BOCF imputations in the standard way. Note, for the weekly pain score, a pain score being carried forward with LOCF might not be a visit week assessment (eg, carry forward Week 3 for missing Week 4 data). For the purposes of the imputation analyses, where there is no post-baseline observation available to

carry forward, then the baseline score carried forward will be the baseline average pain score, being the mean of the last 3 pain scores in the baseline assessment period. If there are less than 3 baseline pain scores then the baseline is calculated over the remaining non-missing values.

Missing values in standard summaries of AEs, lab values, vital signs and ECGs will be handled per Pfizer standard algorithms. For the analysis of NIS the Baseline observation will not be carried forward in the case where a post-baseline observation is not available for the LOCF imputation.

8. STATISTICAL METHODOLOGY AND STATISTICAL ANALYSES

A modified treatment-policy estimands strategy is applied as the main strategy to assess effectiveness of tanezumab. Data collected will be included for efficacy assessment regardless of rescue medication being used or not.

The general study design for efficacy includes a planned treatment period through the Week 56 visit, and a planned 24-week post-treatment safety follow-up period. Efficacy data planned to be collected during this post-treatment safety follow-up period are intended to have efficacy measures contemporaneous to safety observations during this period. They are not intended to assess treatment effects or compare treatment groups. *All endpoints up to Week 64 will be summarized (where available), and endpoints up to Week 56 will be analyzed*.

The method and definition of reporting windows for assigning efficacy data to particular time points is described in Appendix 1.1.

All efficacy assessments are made on the analysis windows defined in Appendix 1. Using these windows we find the analysis window for a patient's last subcutaneous (SC) dose. Any data included in a window up to 8 weeks from this last SC dose window is 'on-treatment', and any in a window more than 8 weeks after the last SC dose window is off treatment. Data in on-treatment analysis windows will be used in summaries and analyses, while data in off-treatment analysis windows will be excluded from all summaries and analyses of treatment period efficacy data, ie, up to Week 56.

For example the table below shows on-treatment and off-treatment windows for the planned collection visits for the WOMAC data during the treatment period:

Last SC Dose Analysis Window	On-treatment Analysis Window Data	Off-treatment Analysis Window Data
Baseline	Weeks 2, 4, 8	Weeks 16, 24, 32, 40, 48, 56
Week 2	Weeks 2, 4, 8	Weeks 16, 24, 32, 40, 48, 56
Week 4	Weeks 2, 4, 8,	Weeks 16, 24, 32, 40, 48, 56
Week 8	Weeks 2, 4, 8, 16	Weeks 24, 32, 40, 48, 56
Week 16	Weeks 2, 4, 8, 16, 24	Weeks 32, 40, 48, 56
Week 24	Weeks 2, 4, 8, 16, 24, 32	Weeks 40, 48, 56
Week 32	Weeks 2, 4, 8, 16, 24, 32, 40	Weeks 48, 56
Week 40	Weeks 2, 4, 8, 16, 24, 32, 40, 48	Week 56
Week 48	Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56	None
Week 56	Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56	None

Efficacy data at Week 64 is planned to be off-treatment so will not be subject to the above handling, ie, all available data in the Week 64 window will be used in summaries.

Efficacy data collected via subject diary (NRS pain scores and rescue medication use) are collected daily or weekly, not at study visits. Diary efficacy data will be considered on-treatment if it is collected up to 12 weeks (84 days) after the last SC dose. Diary efficacy data collected more than 12 weeks (84 days) after the last SC dose will be considered off-treatment and excluded from summaries and analyses of treatment period efficacy data, ie, for presentations up to Week 56.

Diary data after Week 56 is planned to be off-treatment so will not be subject to the above handling, ie, all available data in windows after Week 56 will be used in summaries.

8.1. Statistical Methods

8.1.1. Analyses for Continuous Data

The mixed model ANCOVA, with multiple imputation, will be used with continuous change from Baseline endpoints for landmark (single time point) analyses. The model will include the covariates described in Section 6.4, including study site as a random effect. Estimates of treatment effects and pair wise treatment comparisons will be done using least squares means (LS means) and 95% CIs.

Under the primary analysis using multiple imputation defined in Section 7, the multiple ANCOVA results will be combined using standard methods (Little & Rubin, 2002)[3], which are described in Appendix 3.2.

A sensitivity analysis for the primary endpoints will use a mixed model repeated measures analysis, with covariate terms for Time (study week, treated as a categorical variable), Treatment Group and Time-by-Treatment interaction, as well as the covariates described in Section 6.4. The unstructured covariance will be used in the modeling of the within-subject errors in the analysis. Even though this is a sensitivity analysis for the primary endpoints, estimates for the time points of Weeks 2, 4, and 8, in addition to Week 16 will be shown from this analysis.

Interaction analyses will be performed for the primary endpoints, exploring the effect of Study site and Country. These analyses will fit the covariate terms described in Section 6.4 (except for use of Study site as a covariate in the Country interaction analysis, where Country will be used instead [as a fixed term]), in addition to the interaction term of treatment group by factor.

The interaction of Treatment with Study site will be fitted as a random effect (in addition to Study site itself), with the resulting estimated treatment differences being shown for the largest study sites to illustrate the level of consistency of treatment benefit across the larger study sites. The study sites to be examined in this way will be any site with an average of ten or more subjects per treatment group within the site, which for this study relates to any site with 30 or more subjects in total. This assessment will be made prior to unblinding, therefore a study site in this group may still have fewer than ten subjects in one or more of the treatment groups, however that site will still be included in this summary of efficacy of the

largest study sites. To aid the interpretation of the treatment-site and treatment-country interactions, a summary of the efficacy data for each co-primary endpoint by treatment group will be shown for the sites with ≥30 subjects and also for the countries with ≥30 subjects over all treatment groups.

The time to discontinuation from treatment due to lack of efficacy will be analyzed using the log-rank test, and estimated failure curves will be produced using Kaplan-Meier estimation. The time to selected percentiles will also be shown. These are influenced by the frequency of discontinuation, but are expected to be shown for the 1st, 2nd, 5th, 10th and 25th percentiles, in addition to the minimum and maximum time to discontinuation. Any subject who discontinues for any other reason prior to the planned Week 56 visit will be censored at the time of discontinuation. Subjects who complete the study or who discontinue for any reason after the Week 56 visit (including lack of efficacy) will be censored at the Week 56 visit. This analysis will be performed using 'Insufficient clinical response' alone, and either 'Insufficient clinical response' or 'Patient meets protocol specified pain criteria for discontinuation', as reasons for discontinuation.

The time to joint safety events will be summarized and analyzed in the same way as described above for the time to discontinuation due to lack of efficacy. For these analyses, subjects without events would be censored up to the end of the observation period (defined as end of study completion or discontinuation for subjects who did not have an event). Time to the earliest event would be considered for those who had multiple events.

8.1.2. Analyses for Categorical Data

The number of days and amount (mg of acetaminophen/paracetamol) of rescue medication used per week will be analyzed using a negative binomial model with model terms of treatment group and covariates as described in Section 6.4. In this model the error term is defined with a negative binomial distribution, and 'log' is used as the link function. Output from this analysis will be the estimated number/amount of rescue medication use per week in each treatment group, and (following the exponential back transformation) the ratio of rescue medication use for the treatment comparisons shown in Section 4.1. The 95% CIs will be given for the estimates of both the individual treatment groups and the treatment group ratios.

The change from Baseline in the NIS will be analyzed using a Cochran-Mantel-Haenszel (CMH) test for 'row mean scores differ', using change from Baseline categories as the scores in the analysis, and stratified by the combined levels of the three stratification factors (Section 6.4). Output will show number and percentage of subjects whose NIS score worsened (change>0), improved (change<0) or had no change, in addition to the mean (with standard deviation) and median change, and minimum and maximum change. This analysis will be performed for the two treatment comparisons separately, and shown by visit and worst change (largest change from baseline to any post-baseline visit), and by last change (summary statistics only).

The change from Baseline in the Patient's Global Assessment of Osteoarthritis to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 will also be analyzed using the CMH test and stratified by the combined levels of the stratification factors. Changes by each level of improvement will be summarized, as well as any improvement (change<0), and any worsening (change>0). For this analysis imputation for missing data will use mixed BOCF/LOCF, BOCF, and LOCF.

The change from Baseline in the Lower Extremity Activity Scale to Weeks 4, 8, 16, 24, 56 and 80 and worst post-Baseline score will be analyzed using CMH test and stratified by the combined levels of the three stratification factors (Section 6.4). LOCF will be used for imputation of missing data. The change from Baseline to all weeks (including Week 80) will be summarized showing the number and percentage of subjects with any worsening score (change<0), no change, or improvement (change>0), as well as each category of change, and mean (with standard deviation) and median change.

The CMH test will be stratified by the combinations of the three stratification factors (Section 6.4). For any analysis using the CMH test, if there are too few subjects in any stratification combination group (defined as <15 subjects in any stratification factor) then an unstratified test will be performed.

8.1.3. Analyses for Binary Endpoints

Binary response parameters, and the incidence of rescue medication use and treatment discontinuation due to lack of efficacy will be analyzed using logistic regression for binary data, with covariates described in Section 6.4. Output will show the number and percentage of subjects in each response category, and odds ratios (with 95% CIs) for the treatment comparisons shown in Section 4.1.

Separate analyses of the joint safety endpoints will use exposure time as the denominator (giving results in terms of events per 1000 patient years of exposure) and then number of subjects as the denominator (giving results in terms of percentages). Both these analyses will be performed for risk difference and risk ratio. The analyses using number of subjects as the denominator will analyze risk difference and ratio between both tanezumab groups and NSAID using exact methods. The analyses using exposure as the denominator will analyze risks using Poisson models. The model for risk difference will have a linear link, poisson errors, and model terms for exposure and exposure-by-treatment (where treatment is 1 for tanezumab group and 0 for NSAID). The model for risk ratio will have a log-link, poisson errors, and model term for treatment group with log-exposure as the offset variable.

Results will show event rates by treatment group, and difference/ratio of tanezumab doses versus NSAID, all shown with 95% confidence intervals, and significance tests for the treatment comparisons.

Events will be included in summaries if they occur up to the end of the safety follow-up period or 26 weeks (planned duration of the follow-up period + 2 weeks) after the end of the treatment period, whichever is later.

For the joint safety event analyses, the observation period is defined as the time from first SC dose to study completion or discontinuation for subjects who did not have an event, or time from first SC dose to the earliest event for subjects who did have at least one event.

8.2. Statistical Analyses

A summary of all analyses is given in Section 8.2.5. In all tables the treatment group ordering will be: tanezumab 2.5 mg, tanezumab 5 mg, NSAID.

8.2.1. Primary Analysis (Efficacy)

The primary analysis for the co-primary endpoints will use ANCOVA with covariates of Baseline score, Baseline diary average pain, Index Joint (Knee or Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group, and study site as a random variable. The primary analysis set is the ITT analysis set. The primary analysis will use multiple imputation as described in Section 7, and analysis using the ANCOVA model, with combination of results from imputation analyses using standard methods as described in Section 8.1.1. These three primary endpoint analyses (two for Japan) will be used to assess the primary objective of the study.

Primary Endpoint Sensitivity Analyses

A number of sensitivity analyses will be performed on the primary efficacy endpoints in order to assess the robustness of the conclusions for the primary objective. These relate to the analyses for missing data and the analysis population, the homogeneity of the results across factors that may influence efficacy, and for a secondary analysis of the PGA. The analyses described below will not be subject to the testing strategy described for multiple comparisons of the primary analyses (given in Section 4.2). As such, assessment of all treatment comparisons will be made independent of results over the co-primary endpoints or the two treatment comparisons for each analysis.

The ITT analysis set is used in the analyses numbered 2 to 5 below, and the PP analysis set is used in analysis number 1 below.

1. **Per-Protocol Analysis Set.**

The primary analysis described above will be repeated, but using the Per-Protocol analysis set (as described in Section 5.2 and Section 5.6) in place of the ITT analysis set. This analysis will assess the robustness of the efficacy conclusions to subjects who have more strictly adhered to protocol inclusion and exclusion criteria, and to protocol defined study procedures.

2. **Alternative Missing Data Analyses.**

There are three additional analyses that will assess the robustness of the efficacy conclusions to the choice of multiple imputation as the primary method for accounting for missing data. These analyses are described in detail in Section 7.

In the first and second analyses, the primary ANCOVA analysis model described in Section 8.1.1 will be repeated, but using BOCF and LOCF respectively for missing data (note these are single imputation analyses).

In the third analysis the mixed model Repeated Measures (MMRM) analysis will be performed using observed data up to Week 16 (ie, retrieved dropout), as described in Section 8.1.1 (see Appendix 1.1 for details on windows).

A summary of the missing data pattern will be shown for the WOMAC Pain and Physical Function subscales, and the PGA for Baseline and Weeks 2, 4, 8 and 16. This summary will show the incidence of subjects with each pattern of observed and missing data over these visits and endpoints. This summary will be shown overall, and split by treatment group.

3. Sensitivity Analysis Excluding Subjects from Sites with Potential GCP Compliance Issues.

During the conduct of the study, it was identified that there may be GCP compliance issues at Sites 1714 and 1730. A sensitivity analysis will be performed excluding the patients from both sites, using the same approach as the primary analysis for the primary endpoint. Additional sites may be excluded from this sensitivity analysis prior to unblinding if other issues deemed important are identified.

4. Interaction Analyses.

There will be 2 analyses for each of the co-primary endpoints to explore the interaction of treatment with Study Site and Country, as described in Section 8.1.1. Estimates for Study sites with 30 or more subjects, and for Countries with 30 or more subjects will be shown. Estimates will be shown within each level of each factor, ie,:

- Study Site: Individual sites with ≥30 subjects in the ITT set;
- Country: Individual countries with ≥30 subjects in the ITT set.

5. CMH test for PGA.

The CMH test stratified by the combined levels of the three stratification factors (Section 6.4) will be performed for the PGA, with additional summaries for the change from Baseline to Week 16 as described in Section 8.1.2. This analysis will provide a sensitivity analysis for the primary analysis of the PGA. The missing data imputation used for this analysis will be mixed BOCF/LOCF.

8.2.2. Secondary Analyses (Efficacy)

The following secondary endpoint analyses support the primary endpoints in the assessment of efficacy. All analyses in this section use the ITT analysis set only. Unless otherwise stated, efficacy data will be summarized up to Week 64, and analyzed up to Week 56.

1. Other time points for the primary efficacy measures.

The ANCOVA model described above for the co-primary endpoints, using covariates of Baseline score, Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment, with Study Site as a random effect, will be used in the analysis of WOMAC Pain, WOMAC Physical Function and PGA for the change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56. This analysis will be produced using multiple imputation (as described in Section 7 and Section 8.1.1), and BOCF and LOCF for missing data.

The MMRM analysis described above in Section 8.1.1 will also analyze results for the secondary time points of Weeks 2, 4 and 8, for the co-primary efficacy endpoints.

The CMH test for the PGA with corresponding summary will be performed for the change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56, as described above for the primary PGA endpoint.

2. Secondary endpoints analyzed using ANCOVA.

The ANCOVA model described above for the primary endpoint using covariates of Baseline score, Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment, with Study Site as a random variable, will be used in the analysis of WOMAC Stiffness subscale, WOMAC Average score, WOMAC Pain item "Pain when walking on a flat surface", WOMAC Pain item "Pain when going up or down stairs" for the change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56. This analysis will be produced using multiple imputation for missing data.

The ANCOVA analysis (with covariates of Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment, with Study Site as a random variable) will be produced for the Average Pain in the Index Joint from the diary, for the change from Baseline to Days 1, 2, 3, 4, 5, 6, and 7, and to Weeks 1, 2, 3, 4, 6, 8, 10, 12, 16, 20, 24, 32, 40, 48, and 56, using multiple imputation for missing data.

3. Response and Incidence endpoints.

The response endpoints of OMERACT, improvement in PGA ≥2 and WOMAC Pain and Physical Function ≥30, 50, 70 and 90% improvements are analyzed using logistic regression with covariates of Baseline score (WOMAC Pain for OMERACT), Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment Group, for response at Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56. These analyses use mixed BOCF/LOCF, and BOCF and LOCF for missing data. The use of BOCF for missing data implies subjects with missing data are included in the analysis as non-responders. Similarly the use of LOCF in the case where subjects have no post-Baseline data (and Baseline would be carried forward) again implies those subjects are included in the analysis as non-responders.

Incidence of rescue medication use will be analyzed using the logistic regression as described above up to Week 56, but only using LOCF imputation for missing data.

The incidence of treatment discontinuation due to lack of efficacy will be analyzed using logistic regression with model terms of Baseline WOMAC Pain, Baseline diary average pain, the three stratification parameters and treatment group, and using incidence up to the end of treatment period (the Week 56 visit or early termination). Discontinuation in the post-treatment safety follow-up period will not be included in this endpoint for analysis, but will be summarized as part of the safety tables. Lack of efficacy is indicated as 'Insufficient Clinical Response' on the Subject Summary Case Report Form. An additional analysis will be performed also including the discontinuation reason of 'Patient meets protocol specified pain criteria for discontinuation'.

The cumulative WOMAC Pain and Physical Function response at Weeks 16, 24 and 56 using response definitions from a reduction of >0% to =100% (in steps of 10%) will be summarized, using mixed BOCF/LOCF, and also LOCF and BOCF imputation. Imputation with BOCF for subjects with missing data at that time point will lead to the subjects being assessed as non-responders for the response endpoint.

4. Time to Event.

The time to treatment discontinuation due to lack of efficacy will use the log-rank test. Survival curve estimates (time to 1^{st}, 2^{nd}, 5^{th}, 10^{th} and 25^{th} percentiles, and minimum and maximum values) and a plot of the time to discontinuation (failure) will be shown using the Kaplan-Meier estimates. Only treatment discontinuation up to the end of treatment period (Week 56 visit or early discontinuation) will be used in this analysis. Discontinuation due to lack of efficacy after the end of treatment visit will be included in the standard safety tables. Imputation of time to event for discontinued subjects (discontinuing for reasons other than lack of efficacy) prior to the Week 56 visit uses censoring at the time of discontinuation. Imputation of time to event for completed subjects or discontinued subjects (for any reason) after the Week 56 visit uses censoring at the Week 56 visit time point. Lack of efficacy is indicated as 'Insufficient Clinical Response' on the Subject Summary Case Report Form. An additional analysis will be performed also including the discontinuation reason of 'Patient meets protocol specified pain criteria for discontinuation'.

5. Number of Days and Amount of Rescue Medication Use.

The rescue medication data will be converted to Weekly scores for the week prior to the timepoint of interest. Calculation of the endpoints for both the IPAP and the concomitant medication log data collection is described in Appendix 1.4.

The number of days and amount of rescue medication endpoints will be analyzed using the negative binomial model, with model terms for Baseline WOMAC Pain score, Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment Group as described in Section 8.1.2.

The number of days of rescue medication use per week endpoint will be analyzed for the Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56. The amount of rescue medication use per week will be analyzed for the Weeks 2, 4, 8 and 16. Missing data will be imputed using LOCF only. For this analysis, Baseline data will not be carried forward in the case of a post-Baseline observation not being available for use in LOCF.

6. EuroQol 5 Dimensions 5 Level (EQ-5D-5L).

The Baseline and Weeks 8, 16, 24, 40, 56 and 64 responses in the five dimensions (mobility; self-care; usual activities; pain/discomfort; anxiety/depression) and overall health utility score from the EuroQol 5 Dimensions 5 Level (EQ-5D-5L) will be summarized by treatment group. This summary will use observed data only (no imputation for missing data). The calculation of the overall health utility score is described in Appendix 1.4.

An additional question, called the EQ-VAS asks the subject to rate their health today using a VAS scale from 0 (the worst health you can imagine) to 100 (the best health you can imagine). This will be summarized along with the health utility score. A table showing number and percentage of subjects will summarize the response for each dimension (item) of the EQ-5D at each time point. These summary tables will be shown by treatment group. In addition, for each treatment and each time point assessed, descriptive statistics (mean, standard deviation, median, number of subjects) will be shown for the health utility score, and the EQ-VAS measure of health today.

7. Work Productivity and Activity Impairment Questionnaire for Osteoarthritis (WPAI:OA).

The change from Baseline to Weeks 16, 24, 56 and 64 in the 4 impairment scores of the Work Productivity and Activity Impairment Questionnaire for Osteoarthritis (WPAI:OA) will be summarized by treatment group.

This summary will use observed data only (no imputation for missing data). The calculation of these endpoints is described in Appendix 1.4.

The summary will show number and percentage of subjects with a decrease, no change, and an increase in score for the change from Baseline to each time point, as well as descriptive statistics (mean, standard deviation, median, number of subjects) of the Baseline and change at Weeks 16, 24, 56, and 64.

Change from baseline at Weeks 16, 24, and 56 in the 4 parameters will be analyzed using the ANCOVA model described above for the primary endpoint using covariates of the corresponding Baseline score, Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment, with Study Site as a random variable.

8. Health Care Resource Utilization.

The Health Care Resource Utilization data at Baseline, Weeks 64 and 80 will be summarized.

9. Treatment Satisfaction Questionnaire Medicine v.II (TSQM v.II).

The Weeks 16 and 56 responses in the 4 TSQM parameters of satisfaction with effectiveness, side effects and convenience, and overall satisfaction will be summarized. The 11 questions of the TSQM are used to calculate the 4 endpoints of Effectiveness, Side Effects, Convenience and Global Satisfaction, each scored on a 0-100 scale with 100 being the best level of satisfaction.

The four parameters of the TSQM will be analyzed using the ANCOVA model described above for the primary endpoint using covariates of the Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment, with Study Site as a random effect at both Weeks 16 and 56. Summary tables showing number and percentage of subjects by value and treatment group will be shown for the TSQM items 1-11, and the four satisfaction parameters.

This summary and analysis will use observed data only (no imputation for missing data). The calculation of these endpoints is described in Appendix 1.4.

10. Patient Reported Treatment Impact Assessment-Modified (mPRTI) at Weeks 16 and 56.

The mPRTI is collected at Weeks 16 and 56. The two endpoints derived from this questionnaire are described below:

- Patient willingness to use drug again. This comes from the question "In the future, would you be willing to use the same drug that you have received in this study for your osteoarthritis pain?" This is rated on a 5 point Likert scale from 1 ('Yes, I would definitely want to use the same drug again') to 5 ('No, I definitely would not want to use the same drug again').

- Patient preference of drug versus prior treatment. This comes from the question "Overall, do you prefer the drug that you received in this study to the treatment you received before this clinical trial?" This is rated on a 5 point Likert scale from 1 ('Yes, I definitely prefer the drug I am receiving now') to 5 ('No, I definitely prefer my previous treatment').

The two parameters of the mPRTI will be analyzed using the CMH test (stratified by the combinations of the three stratification factors) at both Weeks 16 and 56. If there are too few subjects in any stratification combination group (defined as <15 in any strata level) then the analysis will be modified to an unstratified test. Summary tables showing number and percentage of subjects by value and treatment group will be shown for all mPRTI questions.

This summary and analysis will use observed data only (no imputation for missing data).

11. Lower Extremity Activity Scale.

The LEAS is an 18-point scale ranging from 1 ('I am confined to bed all day') to 18 ('daily vigorous physical activity'), and is collected at Baseline and Weeks 4, 8, 16, 24, 56 and 80.

The number and percentage of subjects whose activity score is improved (increased), worsened (decreased) or have no change, and each level of change will be shown by treatment group for each week, using LOCF for any missing data. The mean (with standard deviation) and median change will also be summarized. The change from baseline to each post-Baseline visit (using LOCF for missing data), and for the worst change from Baseline (over all post-Baseline visits) will be summarized, and analyzed using CMH test (stratified by the combinations of the three stratification factors) (note, Week 80 data will be summarized, but not analyzed). If there are too few subjects in any stratification combination group (defined as <15 in any strata level) then the analysis will be modified to an unstratified test.

12. Accelerometry Endpoints.

The five accelerometry endpoints will be collected prior to Baseline (at least 4 days, with 1 being a weekend day) and prior to Weeks 16 and 56 (at least 8 days, with 2 being weekend days). The change from Baseline to Week 16 and 56 will be calculated for these parameters. Summaries and analyses of these endpoints will use observed data for change from Baseline to Weeks 16 and 56, and imputation using LOCF for Week 56.

These parameters will be analyzed using a negative binomial model using covariates of the corresponding Baseline accelerometry score, Baseline diary average pain, Index Joint, Highest KL grade (2, 3 or 4), NSAID cohort and Treatment.

8.2.3. Safety Analyses

8.2.3.1. Safety Endpoints (including Primary Joint Safety Endpoint)

1. Joint Safety Endpoints.

The primary joint safety endpoint is the incidence of subjects with any of the adjudication outcomes of rapidly progressive osteoarthritis (type-1 or type-2), subchondral insufficiency fracture, primary osteonecrosis, or pathological fracture.

The primary endpoint *will be shown by number of subjects treated and subject years of exposure (treatment plus follow-up periods), for individual treatment groups and differences between tanezumab treatment groups and the NSAID treatment group. The risk ratio and risk difference* (using number of subjects as the denominator and then exposure as the denominator) *with 95% confidence intervals will be calculated for the comparisons of each tanezumab treatment group versus the NSAID treatment group, as well as significance tests for each treatment comparison. The time to each event will be summarized, and (where there are sufficient numbers of subjects) Kaplan-Meier estimates of the time to event will be produced, together with an analysis of each tanezumab treatment group versus the NSAID treatment group using the log-rank test.*

The primary analysis of the primary joint safety endpoint is the analysis of the exposure-adjusted risk difference. The analyses of the exposure-adjusted risk ratio, and risk difference and ratio based on percentage of subjects with the event will be secondary analyses.

The summary and analysis of the secondary joint safety endpoints will follow the same method described above for the primary joint safety endpoint.

2. Neuropathy Impairment Score.

The change from Baseline in the NIS for Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56, 64 and 80 will be analyzed using a CMH test (stratified by the levels of the stratification factors) with change categories. Missing data will be imputed using LOCF only. For this analysis, Baseline data will not be carried forward in the case of a post-Baseline observation not being available for use in LOCF. An additional analysis will use the change from Baseline to the largest (worst) post-Baseline value and to the last value (summary, not analysis).

3. Radiographic Endpoints.

The change from Baseline to Weeks 56 and 80 in the Minimum Joint Space Width (JSW) for subjects with Kellgren-Lawrence grades of 2 or 3 in the index joint will be analyzed for subjects with measurements in the knee and hip separately. The analysis will also be performed for subjects with a Kellgren-Lawrence grade of 3 in the index knee. For subjects with an index joint of a knee, where both medial and lateral measurements are collected, if the baseline medial JSW is ≤ the baseline lateral JSW, the medial view will be followed through the study for these analyses. If the baseline lateral JSW < the baseline medial JSW, the lateral view will be followed through the study for these analyses. The percentage of subjects with narrowing over certain intervals will be shown, in addition to summary statistics of the mean (with standard deviation) and median change from Baseline.

Significant progression of osteoarthritis will be defined using the Bland-Altman method, as proposed by OARSI-OMERACT. Progression will be defined as 1.96 times the within-subject standard deviation of the change in JSW. The incidence of subjects with JSW narrowing greater than or equal to these values will be shown (with Kellgren-Lawrence grades of 2 or 3 in the index joint, and separately with Kellgren-Lawrence grade of 3 in the index knee), and incidence analyzed using logistic regression for binary data, taking into account Baseline JSW as a covariate.

The Radiographic endpoint summaries and analyses will be performed separately for subjects with osteoarthritis of the hip and knee. These analyses will use the Week 56 End of Treatment/Early Termination data and then Week 80 End of Study/Early Termination Visit 3 regardless of the study day of assessment and/or where subjects have discontinued early from the study. In the event of missing data, baseline data will not be carried forward for Radiographic data. Additional analyses will be performed for assessments at Week 56 and Week 80 (equivalent to observed data analysis for subjects who reach these time points), subject to a window of ±4 weeks and ±6 weeks for the Week 56 and Week 80 analyses respectively.

8.2.3.2. Other Safety Assessments

Pfizer standard safety data presentations will be made for demography data, discontinuation data, adverse event data, laboratory test data, vital signs data and ECG data.

For the 3-tier adverse event reporting, tier 1 adverse events are defined in the tanezumab Safety Review Plan, and this definition of tier-1 adverse events for the report of study 1058 tables will be finalized prior to the unblinding of this study.

Tier 2 AEs are those with a frequency of ≥3% in any treatment group.

Tier 3 AEs are those not in tier 1 or tier 2, and will be summarized using standard Pfizer data standards tables, where all Adverse Events will be included (ie, tier 3 AEs will not be shown separately).

Adverse events within tier 1 and 2 will be summarized using Risk Differences between each tanezumab group and NSAID, together with 95% confidence intervals, using exact methods. Significance tests will be performed for tanezumab versus NSAID comparisons using exact methods for the tier 1 adverse events. There will be no multiplicity adjustment for these significance tests. These tables will be produced for the comparisons of tanezumab 2.5 mg versus NSAID and tanezumab 5 mg versus NSAID.

The following footnote will be used in the Tier 1 AE tables: "P-values and confidence intervals are not adjusted for multiplicity and should be used for screening purpose only. 95% CIs are provided to help gauge the precision of the estimates for Risk Difference. Risk Difference is computed as 'Tanezumab 5 mg versus NSAID' and 'Tanezumab 2.5 mg versus NSAID'. Exact methods are used for 95% confidence intervals and significance tests."

Similarly the following footnote will be used in the Tier 2 AE tables: "Confidence intervals are not adjusted for multiplicity and should be used for estimation purposes only. 95% CIs are provided to help gauge the precision of the estimates for Risk Difference. Risk Difference is computed as 'Tanezumab 5 mg versus NSAID' and 'Tanezumab 2.5 mg versus NSAID'. Exact methods are used for 95% confidence intervals."

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

All summaries of adverse events will be shown for adverse events that begin or worsen after the first dose of SC study drug (treatment-emergent) up to the end of the treatment period. In addition a selection of adverse event tables will be produced for the safety follow-up period and for the whole period up to the end of the study, including the treatment period and safety follow-up period.

Adverse events (AEs), concomitant medications, laboratory safety tests, physical and neurological examinations (NIS), vital signs, electrocardiogram (ECG), and the ADA test will be collected for each subject during the study according to the schedule of assessments.

The following non-standard safety tables will also be included

1. A summary of baseline characteristics, including Kellgren-Lawrence grade of the index joint (for subjects with Hip and Knee OA separately and then overall), highest Kellgren-Lawrence grade for each subject, Summary of WOMAC subscales at Baseline and Screening (for Pain subscale only), diabetes status (from medical history and/or pre-treatment HbA1c≥6.5%), and the PGA at Baseline. This summary will also include a summary of the number of subjects who are ≥75 years old.

2. Summary of number of subjects treated by country and treatment group, and by NSAID cohort.

3. Summary of failed drug treatments for protocol qualification, with reasons for discontinuation.

4. Incidence and severity of AEs leading to discontinuation.

5. Summary of AEs, Incidence of AEs, Incidence of AEs leading to discontinuation and summary of Serious AEs will be shown for the whole study period (including the safety follow-up period).

6. Summary of AEs and Serious Adverse Events by 1000 patient years of exposure.

7. Summary of evidence of neurological examination abnormalities by visit and final assessment, and incidence of neurological findings over consecutive visits. Further details of this summary are given below.

8. Summary of final outcome of neurological consultation.

9. Summary of the Incidence of sympathetic neuropathy based on investigator assessment and, if performed, expert consultant assessment.

10. 'Incidence and severity' tables of treatment-emergent AEs of Abnormal Peripheral Sensation (APS) and Sympathetic Nervous Function, as defined above. Other adverse events may be added to these groupings if they are observed in this study or other studies in the tanezumab program.

11. Summary table and listing of inclusion and exclusion criteria that are not met by subjects who were screened (but not randomized).

12. Summary of discontinuation by treatment group and reason, and study week of discontinuation for the treatment period (Weeks 1-2, 3-4, 5-8, 9-12, 13-16, 17-24, 25-32, 33-40, 41-48, 49-56, >56) and for the safety follow-up period (Weeks 1-8, 9-16, 17-24, >24).

13. A summary of the maximum increase from baseline in the sitting systolic and diastolic blood pressure. The categories used are: (systolic BP) only decreases or no change, >0 to 10, >10-20, >20-30, >30, and (diastolic BP) only decreases or no change, >0 to 10, >10-20, >20.

14. A summary of the maximum decrease from baseline in the sitting systolic and diastolic blood pressure. The categories used are: (systolic BP) <-30, -30 to <-20, -20 to <-10, -10 to <0, only increases or no change, and (diastolic BP) <-20, -20 to <-10, -10 to <0, only increases or no change.

15. A summary of the change from baseline to last observation in the sitting systolic and diastolic blood pressure. The categories used for these summaries are: (systolic BP) ≤-40, >-40 to -30, >-30 to -20, >-20 to -10, >-10 to 0, >0 to <10, 10-<20, 20-<30, 30-<40, ≥40, and (diastolic BP) ≤-30, >-30 to -20, >-20 to -10, >-10 to 0, >0 to <10, 10-<20, 20-<30, ≥30.

16. A summary of incidence of subjects with confirmed orthostatic hypotension, for each visit and any post-baseline incidence of orthostatic hypotension.

17. A summary of discontinuation up to End of Treatment period, and up to End of Study period.

18. Incidence of musculoskeletal physical examination findings at screening.

19. Summary of Dual Energy X-Ray Absorptiometry (DEXA) at screening, showing summary measures of bone mineral density, and t-scores of the spine and hip.

20. Summary of the Survey of Autonomic Symptoms (SAS) number of symptoms reported and total symptom impact score, at each visit, and for the change from Baseline score.

21. Summary of concomitant medications for Osteoarthritis for non-NSAID and NSAID medications (shown separately).

22. Summary of number of days of non-study NSAID use per dosing interval (eg, Day 1 to Week 8, Week 8 to Week 16, etc.) and for the first 8-week interval in the safety follow-up period. This will show the number and percentage of subjects in an interval who exceeded the limit of 10 days of NSAID use. If an interval exists, the visits will be used to define the interval, otherwise calendar time will be used. A summary of average number of days of NSAID use will be displayed by interval. Also, a summary of the overall number of days of NSAID use from Day 1 to Week 64 will be shown, as well as the number and percentage of subjects who exceeded the limit of 80 days of NSAID use during this interval.

23. Summary of oral study medication compliance. This is calculated for the interval from Baseline to Week 16 and the entire post-Baseline period up to Week 56 (or end of treatment visit). Compliance is calculated as number of tablets dispensed minus the number returned divided by the number of days in the interval, all divided by 2 (for the daily regimen frequency), multiplied by 100, to get a percentage compliance for each subject for each time period.

Neurological-related safety data

The "conclusion from the neurological examination" data will be summarized for each time point and the last subject assessment. In addition the persistence of any neurological examination finding will be summarized, showing the incidence of subjects with new or worsened neurological examination abnormalities (both clinically significant only and also for any finding) for 2, 3, 4, and ≥5 consecutive visits.

Immunogenicity

The following assessments of ADA data will be made:

- A listing of individual serum ADA results sorted by treatment group, subject ID and planned visit. The listing will also include the actual test date/times.
- The proportion of subjects who test positive (ie, develop anti-tanezumab antibodies) and negative will be summarized by treatment group and planned visit. The summary will also include the proportion of subjects who test positive and negative overall in the study.
- Subjects who develop anti-tanezumab antibodies after treatment will be evaluated for the presence of anti-tanezumab neutralizing antibodies, and individual results will be listed.
- Individual subjects with positive ADA results will be evaluated for potential ADA impact on the individual's PK, efficacy and safety profile.

8.2.4. Other Analyses

8.2.4.1. Pharmacokinetics

The following reporting of PK data will be done:

- A listing of all plasma tanezumab concentrations sorted by subject, active treatment group and nominal time post dose. The listing of concentrations will also include the actual times post dose.
- A descriptive summary of the plasma tanezumab concentrations based on nominal time post dose for each treatment group.
- Boxplots of plasma tanezumab concentrations at the nominal times for the tanezumab treatment groups.

- A listing of available synovial fluid tanezumab concentrations at the time of the arthrocentesis only sorted by subject, treatment group and actual collection date and time.

8.2.4.2. Pharmacodynamics (NGF)

Serum samples from a subset of patients will be run in the bioanalytical assays for assessment of NGF and the measurements will be summarized in the following tables and figures.

- A listing of individual serum NGF concentrations sorted by subject, treatment group and time post dose.
- Descriptive statistics of serum NGF concentrations based on time post dose for each treatment group.
- Boxplots of serum NGF over time post dose for each treatment group.

If enough synovial fluid volume is left after determining the tanezumab concentration in the sample that is collected in subjects who have arthrocentesis performed, synovial fluid NGF concentrations will be determined as well at the time of the arthrocentesis only, and the measurements will be summarized as follows:

- A listing of available synovial fluid NGF concentrations at the time of the arthrocentesis only sorted by subject, treatment group and actual collection date and time.

8.2.4.3. Osteoarthritis Biomarkers

The biomarker data analysis related to joint safety events will be performed separately from the A4091058 study CSR.

8.2.5. Summary of Efficacy Analyses

Note: BL=Baseline

Endpoint	Analysis Set	Statistical Method	Model/Covariates	Missing Data	Objective
Change from Baseline to Week 16 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Primary Analysis
Change from Baseline to Week 16 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Primary Analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Primary Analysis
Change from Baseline to Week 16 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Study Site, Treatment Group, Study Site x Treatment Group (Study site and interaction as random effects)	Multiple Imputation	Additional (Interaction) Analysis
Change from Baseline to Week 16 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Country, Treatment Group, Country x Treatment Group.	Multiple Imputation	Additional (Interaction) Analysis
Change from Baseline to Week 16 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4),	Multiple Imputation	Additional (Interaction) Analysis

			NSAID cohort, Study Site, Treatment Group, Study Site x Treatment Group (Study site and interaction as random effects)		
Change from Baseline to Week 16 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Country, Treatment Group, Country x Treatment Group.	Multiple Imputation	Additional (Interaction) Analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Study Site, Treatment Group, Study Site x Treatment Group (Study site and interaction as random effects)	Multiple Imputation	Additional (Interaction) Analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Country, Treatment Group, Country x Treatment Group	Multiple Imputation	Additional (Interaction) Analysis
Change from Baseline to Week 16 in WOMAC Pain subscale	PP	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Sensitivity Analysis (Per protocol)
Change from Baseline to Week 16 in WOMAC Physical Function subscale	PP	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Sensitivity Analysis (Per protocol)
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis	PP	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Multiple Imputation	Sensitivity Analysis (Per protocol)

			(Study site as a random effect)		
Change from Baseline to Week 16 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	BOCF	Sensitivity Analysis
Change from Baseline to Week 16 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	BOCF	Sensitivity Analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	BOCF	Sensitivity Analysis
Change from Baseline to Week 16 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	LOCF	Sensitivity Analysis
Change from Baseline to Week 16 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	LOCF	Sensitivity Analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	LOCF	Sensitivity Analysis
Change from Baseline to Weeks 2, 4, 8 and 16 in WOMAC Pain subscale	ITT	MMRM	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Time, Treatment Group, Time x Treatment Group (Study site as a random effect)	Observed Data	Sensitivity Analysis for Week 16 (Secondary Endpoint Analysis for other time points)
Change from Baseline to Weeks 2, 4,	ITT	MMRM	BL Score, BL Diary Average	Observed Data	Sensitivity Analysis

8 and 16 in WOMAC Physical Function subscale			Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Time, Treatment Group, Time x Treatment Group (Study site as a random effect)		for Week 16 (Secondary Endpoint Analysis for other time points)
Change from Baseline to Weeks 2, 4, 8 and 16 in Patient Global Assessment of Osteoarthritis	ITT	MMRM	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Time, Treatment Group, Time x Treatment Group (Study site as a random effect)	Observed Data	Sensitivity Analysis for Week 16 (Secondary Endpoint Analysis for other time points)
Change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in Patient Global Assessment of Osteoarthritis	ITT	CMH test	Treatment Group [1]	Mixed BOCF/LOCF	Sensitivity Analysis for PGA
Change from Baseline to Week 16 in WOMAC Pain subscale, shown by site (sites with n≥30)	ITT	None (summary)	NA	Multiple Imputation	Supportive summary for interaction analysis
Change from Baseline to Week 16 in WOMAC Physical Function subscale, shown by site (sites with n≥30)	ITT	None (summary)	NA	Multiple Imputation	Supportive summary for interaction analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis, shown by site (sites with n≥30)	ITT	None (summary)	NA	Multiple Imputation	Supportive summary for interaction analysis
Change from Baseline to Week 16 in WOMAC Pain subscale, shown by country (countries with n≥30)	ITT	None (summary)	NA	Multiple Imputation	Supportive summary for interaction analysis
Change from Baseline to Week 16 in WOMAC Physical Function subscale, shown by country (countries with n≥30)	ITT	None (summary)	NA	Multiple Imputation	Supportive summary for interaction analysis
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis, shown by country (countries with n≥30)	ITT	None (summary)	NA	Multiple Imputation	Supportive summary for interaction analysis
Missing data pattern for WOMAC Pain subscale for Baseline and Weeks 2, 4, 8 and 16	ITT	None (summary)	NA	Observed Data	Supportive summary for missing data

Missing data pattern for WOMAC Physical Function subscale for Baseline and Weeks 2, 4, 8 and 16	ITT	None (summary)	NA	Observed Data	Supportive summary for missing data
Missing data pattern for Patient Global Assessment of Osteoarthritis for Baseline and Weeks 2, 4, 8 and 16	ITT	None (summary)	NA	Observed Data	Supportive summary for missing data
Change from Baseline to Week 16 in WOMAC Pain subscale by NSAID cohort	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), Treatment Group (Study site as a random effect)	Multiple Imputation	Primary endpoint analysis for individual NSAID comparisons
Change from Baseline to Week 16 in WOMAC Physical Function subscale by NSAID cohort	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), Treatment Group (Study site as a random effect)	Multiple Imputation	Primary endpoint analysis for individual NSAID comparisons
Change from Baseline to Week 16 in Patient Global Assessment of Osteoarthritis by NSAID cohort	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), Treatment Group (Study site as a random effect)	Multiple Imputation	Primary endpoint analysis for individual NSAID comparisons
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	BOCF	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in WOMAC Pain subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	LOCF	Secondary Endpoint Analysis

Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	BOCF	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in WOMAC Physical Function subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	LOCF	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	BOCF	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 24, 32, 40, 48, and 56 in Patient Global Assessment of Osteoarthritis	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	LOCF	Secondary Endpoint Analysis
The OMERACT-OARSI response at Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56	ITT	Logistic Regression	BL Score (WOMAC Pain), BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Mixed BOCF/LOCF	Secondary Endpoint Analysis

The OMERACT-OARSI response at Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56	ITT	Logistic Regression	BL Score (WOMAC Pain), BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	BOCF	Secondary Endpoint Analysis
The OMERACT-OARSI response at Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56	ITT	Logistic Regression	BL Score (WOMAC Pain), BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
Percentage of subjects with reduction of ≥30/50/70/90% from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Pain subscale	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Mixed BOCF/LOCF	Secondary Endpoint Analysis
Percentage of subjects with reduction of ≥30/50/70/90% from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Pain subscale	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	BOCF	Secondary Endpoint Analysis
Percentage of subjects with reduction of ≥30/50/70/90% from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Pain subscale	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
Percentage of subjects with reduction of ≥30/50/70/90% from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Physical Functioning subscale	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Mixed BOCF/LOCF	Secondary Endpoint Analysis
Percentage of subjects with reduction of ≥30/50/70/90% from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Physical Functioning subscale	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	BOCF	Secondary Endpoint Analysis
Percentage of subjects with reduction of ≥30/50/70/90% from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Physical Functioning subscale	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis

Percentage of subjects with an improvement of ≥2 points from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the Patient Global Assessment of Osteoarthritis	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Mixed BOCF/LOCF	Secondary Endpoint Analysis
Percentage of subjects with an improvement of ≥2 points from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the Patient Global Assessment of Osteoarthritis	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	BOCF	Secondary Endpoint Analysis
Percentage of subjects with an improvement of ≥2 points from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the Patient Global Assessment of Osteoarthritis	ITT	Logistic Regression	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
Reduction of >0%, ≥10%, to ≥90% (in steps of 10%) and =100% from Baseline to Week 16, 24 and 56 in the WOMAC Pain subscale	ITT	None (summary and plot)	NA	Mixed BOCF/LOCF	Secondary Endpoint Analysis
Reduction of >0%, ≥10%, to ≥90% (in steps of 10%) and =100% from Baseline to Week 16, 24 and 56 in the WOMAC Pain subscale	ITT	None (summary)	NA	BOCF	Secondary Endpoint Analysis
Reduction of >0%, ≥10%, to ≥90% (in steps of 10%) and =100% from Baseline to Week 16, 24 and 56 in the WOMAC Pain subscale	ITT	None (summary)	NA	LOCF	Secondary Endpoint Analysis
Reduction of >0%, ≥10%, to ≥90% (in steps of 10%) and =100% from Baseline to Week 16, 24 and 56 in the WOMAC Physical Function subscale	ITT	None (summary and plot)	NA	Mixed BOCF/LOCF	Secondary Endpoint Analysis
Reduction of >0%, ≥10%, to ≥90% (in steps of 10%) and =100% from Baseline to Week 16, 24 and 56 in the WOMAC Physical Function subscale	ITT	None (summary)	NA	BOCF	Secondary Endpoint Analysis

Reduction of >0%, ≥10%, to ≥90% (in steps of 10%) and =100% from Baseline to Week 16, 24 and 56 in the WOMAC Physical Function subscale	ITT	None (summary)	NA	LOCF	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Stiffness subscale	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Average Score	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Item: Pain When Walking on a Flat Surface	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56 in the WOMAC Item: Pain When Going Up or Down Stairs	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Change from Baseline to Days 1, 2, 3, 4, 5, 6, 7, and to Weeks 1, 2, 3, 4, 6, 8, 10, 12, 16, 20, 24, 32, 40, 48, and 56 in the weekly average pain score in the index joint	ITT	ANCOVA	BL Score, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Multiple Imputation	Secondary Endpoint Analysis
Time to treatment discontinuation due to lack of efficacy (up to Week 56/End of Treatment)	ITT	Log-Rank (with KM estimates)	Treatment Group	Observed	Secondary Endpoint Analysis
Incidence of treatment discontinuation due to lack of efficacy (up to Week 56/End of Treatment)	ITT	Logistic Regression	BL WOMAC Pain, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Observed	Secondary Endpoint Analysis

Incidence of rescue medication use during Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56	ITT	Logistic Regression	BL WOMAC Pain, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
Number of days of rescue medication use during Weeks 2, 4, 8, 16, 24, 32, 40, 48, and 56	ITT	Negative Binomial	BL WOMAC Pain, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
Amount (mg) of rescue medication taken during Weeks 2, 4, 8 and 16	ITT	Negative Binomial	BL WOMAC Pain, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
EQ-5D-5L dimensions (Mobility; Self-care; Usual activity; Pain/Discomfort; Anxiety/Depression) and Overall Health Utility at Baseline and Weeks 16, 24 and 56	ITT	Summary	NA	Observed	Secondary Endpoint Analysis
EQ-VAS at Baseline and Weeks 16, 24 and 56	ITT	Summary	NA	Observed	Secondary Endpoint Analysis
mPRTI endpoints (willingness to re-use; patient preference) at Weeks 16 and 56	ITT	CMH test	Treatment group [1]	Observed	Secondary Endpoint Analysis
TSQM endpoints (satisfaction with effectiveness, side effects and convenience, and overall satisfaction) at Weeks 16 and 56	ITT	ANCOVA	BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Observed	Secondary Endpoint Analysis
WPAI endpoints (% work time missed; % impairment while working; % overall work impairment; % activity impairment) at Weeks 16, 24 and 56	ITT	ANCOVA	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group (Study site as a random effect)	Observed	Secondary Endpoint Analysis
Healthcare Resource Utilization at baseline, weeks 56 and 64	ITT	Descriptive Summary	Treatment group	Observed	Secondary Endpoint Analysis

Change from Baseline in LEAS to Weeks 4, 8, 16, 24, and 56, and worst post-baseline score.	ITT	CMH test	Treatment group [1]	LOCF (weeks 4, 8, 16, 24, and 56) and worst score	Secondary Endpoint Analysis
Change from Baseline in Accelerometry endpoints (daily minutes of physical activity, MVPA and bouted MVPA; daily step count; daily physical activity counts) to Weeks 16 and 56	Accelerometry analysis set	negative binomial	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	Observed	Secondary Endpoint Analysis
Change from Baseline in Accelerometry endpoints (daily minutes of physical activity, MVPA and bouted MVPA; daily step count; daily physical activity counts) to Week 56	Accelerometry analysis set	negative binomial	BL Score, BL Diary Average Pain, Index Joint (Knee/Hip), Highest KL grade (2, 3 or 4), NSAID cohort, Treatment Group	LOCF	Secondary Endpoint Analysis
Incidence of primary composite adjudication joint safety endpoint	ITT	Poisson model for risk difference (exposure-adjusted)	Exposure, exposure*treatment	Observed	Primary Joint Safety Analyses
Incidence of primary composite adjudication joint safety endpoint	ITT	Poisson model for risk ratio (exposure-adjusted)	Treatment group, offset by log-exposure	Observed	Secondary Joint Safety Analyses
Incidence of primary composite adjudication joint safety endpoint	ITT	Exact analyses for risk difference and ratio (percentage)	Treatment group	Observed	Secondary Joint Safety Analyses
Incidence of secondary composite endpoint, individual adjudication joint safety endpoints (RPOA-1, RPOA-2, RPOA-1/2, SIF, Primary ON, Pathological fracture), TJR, and TJR and any adjudication outcome	ITT	Poisson model for risk difference (exposure-adjusted)	Exposure, exposure*treatment	Observed	Secondary Joint Safety Analyses
Incidence of secondary composite endpoint, individual adjudication joint safety endpoints (RPOA-1, RPOA-2, RPOA-1/2, SIF, Primary ON, Pathological fracture), TJR, and TJR and any adjudication outcome	ITT	Poisson model for risk ratio (exposure-adjusted)	Treatment group, offset by log-exposure	Observed	Secondary Joint Safety Analyses

Incidence of secondary composite endpoint, individual adjudication joint safety endpoints (RPOA-1, RPOA-2, RPOA-1/2, SIF, Primary ON, Pathological fracture), TJR, and TJR and any adjudication outcome	ITT	Exact analyses for risk difference and ratio (percentage)	Treatment group	Observed	Secondary Joint Safety Analyses
Time to joint safety event (Primary Composite, Secondary Composite, Individual Adjudication Events, TJR, TJR and any adjudication outcome)	ITT	Log-Rank (with KM estimates)	Treatment Group	Observed	Secondary Joint Safety Analyses
Incidence of Bland-Altman defined progression of OA for subjects with OA of the Hip (KL2,3) at Weeks 56/ET and 80/ET, and Weeks 56 and 80 (based on window)	ITT	Logistic Regression	BL JSW, Treatment Group	Observed	Radiographic safety endpoint
Incidence of Bland-Altman defined progression of OA for subjects with OA of the Knee (KL2,3) at Weeks 56/ET and 80/ET, and Weeks 56 and 80 (based on window)	ITT	Logistic Regression	BL JSW, Treatment Group	Observed	Radiographic safety endpoint
Change from Baseline in JSW for subjects with OA of the Hip (KL 2, 3) at Weeks 56/ET and 80/ET, and Weeks 56 and 80 (based on window)	ITT	ANCOVA	BL JSW, Treatment Group (Study site as a random effect)	Observed	Radiographic safety endpoint
Change from Baseline in JSW for subjects with OA of the Knee (KL 2, 3) at Weeks 56/ET and 80/ET, and Weeks 56 and 80 (based on window)	ITT	ANCOVA	BL JSW, Treatment Group (Study site as a random effect)	Observed	Radiographic safety endpoint
Change from baseline to Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56, 64 and 80 in the NIS score, and Change from Baseline to Worst post-Baseline NIS score	ITT	CMH test	Treatment Group [1]	LOCF (Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56, 64, 80), Worst post-baseline score	Safety Analysis
Survey of Autonomic Symptoms (SAS), number of symptoms and total impact score (by visit and change from Baseline)	ITT	Summary	NA	Observed	Safety Analysis

[1] CMH test will be stratified by the levels of the combined stratification parameters (18 levels). If there are <15 subjects in any combined stratification level then the CMH test will be unstratified.

9. REFERENCES

1. Atkinson, MJ et al (2005). Hierarchical construct validity of the treatment satisfaction questionnaire for medication (TSQM Version II) among outsubject pharmacy consumers. *Value in Health*. **8(Supp 1)**, S9-24.

2. EuroQol Group. EuroQol: a new facility for the measurement of health related quality of life. *Health Policy* 1990; 16:199-208.

3. Little RJ & Rubin DB (2002). *Statistical Analysis with Missing Data*. New Jersey: Wiley.

4. Tudor-Locke C, et al. A Catalog of Rules, Variables, Definitions Applied to Accelerometer Data in National Health and Nutrition Examination Survey, 2003-2006. *Prev Chronic Dis* 2012,9:110332. DOI: http://dx.doi.org/10.5888/pcd9.110332).

5. Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. *Statist. Med.* (2014):33:693–713.

6. Bretz F, Posch M, Glimm E, Klingmueller F, Maurer W, Rohmeyer K. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests *Biometrical Journal*. 53(2011) 6, 894–913.

10. APPENDICES

Appendix 1. DATA DERIVATION DETAILS

Appendix 1.1. Definition and Use of Visit Windows in Reporting

Study visits are planned at Screening, Baseline and then at post-baseline Weeks 2, 4, 8, 16, 24, 32, 40, 48, 56, 64 and 80. If a subject discontinues from the trial then there will be an Early Termination Follow-Up period and for those who refuse, ideally, an Early termination visit. To account for this visit and any early or late scheduled visits (compared to the target study days) we define 'windows' to be able to allocate each efficacy observation to a single specific study visit. For the assessments made at each planned study visit (eg, WOMAC subscales, Patient Global Assessment of Osteoarthritis etc.) these visit windows are shown below. When multiple observations occur in a visit window, the observation closest to the protocol specified target day will be used, noting that the latter will be used in the case of a tie.

Visit	Target Study Day	Window
Screening [1]	Variable (up to 37 days prior to baseline visit)	[No lower limit, Day -8]
Baseline	1 (defined as initial day of study drug administration)	[-7,1]
Week 2	15	[2,22]
Week 4	29	[23,43]
Week 8	57	[44,85]
Week 16	113	[86,141]
Week 24	169	[142,197]
Week 32	225	[198, 253]
Week 40	281	[254, 309]
Week 48	337	[310, 365]
Week 56	393	[366, 421]
Week 64	449	[422, 477]

[1] Only efficacy data collected at screening is WOMAC Pain subscale.

One additional window will be created relative to the date of last SC dose for summaries of efficacy data collected beyond the planned treatment period. This window will include data from 16 +/- 4 weeks past the date of the last SC dose. The target day is 113 days after the last SC dose, with a window of [85, 141] days after the last SC dose. If multiple observations occur in this visit window, the observation closest to the specified target day will be used, noting that the latter will be used in the case of a tie.

For the assessments not made at each planned study visit, broader visit windows are shown below. When multiple observations occur in a visit window, the observation closest to the protocol specified target day will be used, noting that the latter will be used in the case of a tie.

LEAS

Visit	Target Study Day	Window
Baseline	1 (defined as initial day of study drug administration)	[-7,1]
Week 4	29	[2,43]
Week 8	57	[44,85]
Week 16	113	[86,141]
Week 24	169	[142,281]
Week 56	393	[282, 477]
Week 80/End of Study	561	[478, no upper limit]

EQ-5D-5L

Visit	Target Study Day	Window
Baseline	1 (defined as initial day of study drug administration)	[-7, 1]
Week 8	57	[2, 85]
Week 16	113	[86, 141]
Week 24	169	[142, 225]
Week 40	281	[226, 337]
Week 56	393	[338, 421]
Week 64	449	[422, no upper limit]

WPAI: OA

Visit	Target Study Day	Window
Baseline	1 (defined as initial day of study drug administration)	[-7,1]
Week 16	113	[2,141]
Week 24	169	[142,281]
Week 56	393	[282, 421]
Week 64	449	[422, no upper limit]

Actigraphy, TSQM, mPRTI

Visit	Target Study Day	Window
Baseline	1 (defined as initial day of study drug administration)	[no lower limit, 1]
Week 16	113	[2,253]
Week 56	393	[254, no upper limit]

HCRU

Visit	Target Study Day	Window
Baseline	1 (defined as initial day of study drug administration)	[no lower limit, 1]
Week 64	449	[2,505]
Week 80	561	[506, no upper limit]

For the average pain in the index joint, the data are collected daily via electronic diary up to the end of Week 16, and thereafter weekly up to Week 80. Data up to Week 64 will be reported as part of the efficacy assessment (summary up to Week 64; analysis up to Week 56).

The Baseline score is the mean of the last 3 non-missing pain scores over study days -7 to -1. If fewer than 3 are available between study days -7 and -1, the baseline will be the mean of the available scores.

The table below describes the visit days for each week (Weeks 1-16). All available on-treatment diary data in each of the weekly intervals will be used to calculate the mean daily pain score for that study week.

Study Week	Days	Study Week	Days
1	1-7	9	57-63
2	8-14	10	64-70
3	15-21	11	71-77
4	22-28	12	78-84
5	29-35	13	85-91
6	36-42	14	92-98
7	43-49	15	99-105
8	50-56	16	106-112

However, if a subject receives the Week 16 injection dose prior to Day 113, the Week 16 score will be calculated using the mean of the available scores from the 7 calendar days immediately prior to the Week 16 injection dose date. Any scores used in this calculation of Week 16 will not also be used in an earlier week calculation, eg, if the Week 16 dose occurs on Day 109, the available scores from Days 102-108 will be used to calculate the average score for Week 16, and the available scores from Days 99-101 will be used to calculate the average score for Week 15.

After the Week 16 visit, pain scores are captured only once a week in the diary. These are grouped in 4-week intervals using visit windows as shown below. If a subject comes in late for a Week 16 visit (or weekly diary is not activated at the visit), and so has daily diary data collected past Day 112, these data will be averaged with any data obtained weekly for any given interval. All available on- or off-treatment data will be used for these windows after the planned treatment period, ie, after Week 56.

Summary Week	Includes Weeks	Days
20	17 - 20	113-140
24	21 - 24	141-168
28	25 - 28	169-196
32	29 - 32	197-224
36	33 - 36	225-252
40	37 – 40	253-280
44	41 – 44	281-308
48	45 – 48	309-336
52	49 – 52	337-364
56	53 – 56	365-392
60	57 – 60	393-420
64	61 – 64	421-448

One additional window will be created relative to the date of last SC dose for summaries of diary pain scores collected beyond the planned treatment period. This window will be identified as 16 Weeks Post Last Dose, and will include the average of all data from 13 to 16

weeks (85 to 112 days)_past the date of the last SC dose. All available on- or off-treatment data will be used for this window after the planned treatment period.

Appendix 1.2. Definition of Protocol Deviations that Relate to Statistical Analyses/Populations

Not applicable.

Appendix 1.3. Definition of Analysis Populations/Sets

Not applicable.

Appendix 1.4. Further Definition of Endpoints

Health State Utility of the EQ-5D-5L

The EQ-5D-5L contains five questions that measure the following dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each of the five dimensions has five levels: (1) no problems; (2), slight problems; (3) moderate problems; (4) severe problems; and (5) extreme problems.

The health utility scores are defined for every possible set of outcome combinations of the five dimensions for the following countries:

- Denmark, France, Germany, Japan, the Netherlands, Spain, Thailand, UK, US and Zimbabwe

This study recruited treated subjects from the following countries.

- Australia, Bulgaria, Brazil, Colombia, Japan, New Zealand, Peru, Philippines, Russia, South Korea, Taiwan, Ukraine, US, Croatia, Lithuania, Serbia and Slovakia

Some of these may not actually recruit or treat subjects, and other countries may be added. As there is a mismatch between countries where subjects are being recruited and the currently available EQ-5D-5L health utility scoring, we will assign subjects to the following scoring countries based on the following assignments.

EQ-5D-5L Scoring Country	Study Recruitment Country
Denmark	-
France	-
Germany	Russia, Ukraine, Croatia, Lithuania, Serbia, Slovakia, Bulgaria
Japan	Japan, South Korea, Taiwan
The Netherlands	-
Spain	Colombia, Peru
Thailand	-
UK	Australia, New Zealand
US	Brazil, US, Philippines

Zimbabwe	-

If more EQ-5D-5L utility scores become available or other countries are added, then this assignment may be modified.

The health utility for a subject with no problems in all 5 items is 1 for all countries (except for Zimbabwe where it is 0.9), and is reduced where a subject reports greater levels of problems across the five dimensions. The minimum score across the countries is -0.654.

Appendix 2. WPAI:OA Endpoints

The tables below summarizes the 6 questions of the WPAI:OA questionnaire, and the four endpoints of the effect of impairment on activity and impairment.

Question	Question Wording	Scoring
1	Are you currently employed? [if No skip to question 6]	Yes, No
2	During the past seven days, how many hours did you miss from work due to problems associated with your OA of the knee or hip	number of hours (free text)
3	During the past seven days, how many hours did you miss from work because of any other reason, such as vacation, holidays, time off to participate in this study?	number of hours (free text)
4	During the past seven days, how many hours did you actually work (if '0' skip to Question 6)	number of hours (free text)
5	During the past seven days, how much did your OA of the knee or hip affect productivity while you were working?	0 to 10 scale with 0 being 'No effect on my work' and 10 being 'Completely prevented me from working'
6	During the past seven days, how much did your OA of the knee or hip affect your ability to do your regular daily activities, other than work at a job?	0 to 10 scale with 0 being 'No effect on my daily activities' and 10 being 'Completely prevented me from doing my daily activities'

WPAI endpoint	Calculation
Percent activity impairment due to Osteoarthritis	Q6*10
Percent impairment while working due to osteoarthritis	Q5*10
Percent overall work impairment due to osteoarthritis	$\left\{\frac{Q2}{Q2+Q4}+\left[1-\left(\frac{Q2}{Q2+Q4}\right)\right]\left(\frac{Q5}{10}\right)\right\}\times 100$
Percent work time missed due to Osteoarthritis	$\frac{Q2}{Q2+Q4} * 100$

TSQM vII

The 11 questions of the TSQM and the scoring are shown below:

Item	Question wording	Likert Scoring
1	How satisfied or dissatisfied are you with the ability of the medication to prevent or treat the condition?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)
2	How satisfied or dissatisfied are you with the way the medication relieves symptoms?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)
3	As a result of taking this medication, do you experience any side effects at all?	0 (No), 1 (Yes)
4	How dissatisfied are you by side effects that interfere with your physical health and ability to function (eg, strength, energy levels)?	1 (Extremely dissatisfied) to 5 (Not at all dissatisfied)
5	How dissatisfied are you by side effects that interfere with your mental function (eg, ability to think clearly, stay awake)?	1 (Extremely dissatisfied) to 5 (Not at all dissatisfied)
6	How dissatisfied are you by side effects that interfere with your mood or emotions and ability to function (eg, anxiety/fear, sadness, irritation/anger)?	1 (Extremely dissatisfied) to 5 (Not at all dissatisfied)
7	How satisfied or dissatisfied are you with how easy the medication is to use?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)
8	How satisfied or dissatisfied are you with how easy it is to plan when you will use the medication each time?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)
9	How satisfied or dissatisfied are you by how often you are expected to use/take the medication?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)
10	How satisfied are you that the good things about this medication outweigh the bad things?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)
11	Taking all things into account, how satisfied or dissatisfied are you with this medication?	1 (Extremely dissatisfied) to 7 (Extremely satisfied)

The scoring of the 4 satisfaction parameters are shown in the table below.

TSQM Parameter	Scoring
Effectiveness	[(Item 1+Item 2) – 2]/12 *100
Side Effects	[(Item 4 + Item 5 + Item 6) – 3]/12 *100 If one item is missing then: [(Sum of two completed items)-2]/8 *100
Convenience	[(Item 7 + Item 8 + Item 9) – 3]/18 *100 If one item is missing then: [(Sum of two completed items)-2]/12 *100
Global Satisfaction	[(Item 10+Item 11) – 2]/12 *100

The four parameters have a scale of 0-100, with 100 being the best (most satisfied) score.

Healthcare Resource Utilization (example using 3 month recall—8 week recall is also used in study)

Question	Response	Scoring
During the last 3 months, what services did you receive directly related to your osteoarthritis? • Primary Care Physician • Neurologist • Rhematologist • Physician Assistant or Nurse Practioner • Pain Specialist • Orthopedist • Physical Therapist • Chiropractor • Alternative Medicine or Therapy • Podiatrist • Nutritionist/Dietician • Radiologist • Home healthcare services • Other	Number of Visits	Response not selected = 0 Number of visits = 1-999
During the past 3 months, have you visited the emergency room due to your osteoarthritis?	Yes, No	No = 0 Yes = 1
How many times?	Number of visits	0-999
During the past 3 months, have you been hospitalized due to your osteoarthritis?	Yes, No	No = 0 Yes = 1
How many nights in total did you stay in hospital due to your osteoarthritis in the last 3 months?	Number of Nights	0-999 (max should be 92)
Did you use these aids or devices to help you in doing things because of your osteoarthritis in the last 3 months? • Walking Aid • Wheelchair • Devices or utensils to help you dress, eat or bathe • Other	Did not use any aids or devices Never, rarely, sometimes, often, always	Did not use any aids or devices = 0 Device not selected = 0 Never = 1 Rarely = 2 Sometimes = 3 Often = 4 Always = 5
Did you quit your job because of your osteoarthritis?	Yes, No	No = 0 Yes = 1 Not applicable = 2
How long ago did you quit your job because of your osteoarthritis?	Years and Months	0-99 Years and 0-99 Months (should be max of 11 months)

Rescue Medication Endpoints

Rescue medication data is collected daily using an electronic system up to Week 16, and weekly after Week 16 and up to Week 80. Daily and weekly collected data will be assigned to a specific study week for summary and reporting. The assignment of daily and weekly data to weeks will use the same principle as described above in Appendix 1.1 for the daily and weekly index joint pain data.

The incidence of rescue medication use will look for any incidence in the week of interest (collected through daily or weekly diary data). The number of days of RM use (using daily and weekly data) and the total amount taken (using daily data up to Week 16 only) over the week will be calculated for the assigned week algorithm described above.

Imputation is described in Section 7 above. Imputation occurs for daily data up to Week 16 where the subject is in the trial and up to the end of that particular week.

An example of imputation and calculating the three endpoints using the daily diary data is shown below.

Example of calculating rescue medication data from Daily Diary Data (Subject does not discontinue)

In this example, a subject has a Week 2 visit on study day 14 (slightly earlier than the nominal day 15). Study days 8-14 would represent Week 2 data.

Using the Week 2 interval described above for a subject, ie, study days [8-14], we have the following rescue medication example data.

The amount taken and number of days of rescue medication use is adjusted for the duration of the Weekly interval.

Study Day (Week)	Number of Doses of RM taken [1]	Number of Doses of RM taken [1] with LOCF imputation
8 (Week 2)	2	2
9 (Week 2)	Missing	2 [2]
10 (Week 2)	0	0
11 (Week 2)	1	1
12 (Week 2)	Missing	1 [2]
13 (Week 2)	2	2
14 (Week 2)	0	0

[1] 500mg tablets of acetaminophen; [2] Using LOCF imputation for missing data

For this subject the following data will be calculated for Week 2:

- Incidence of rescue medication taken in Week 2: Yes. Rescue medication taken on days 8, 9 (imputed), 11, 12 (imputed), 13.

- Number of days of rescue medication use in Week 2: 5. For days 8-14 we have rescue medication taken on days 8, 9 (imputed), 11, 12 (imputed), and 13. The number of days taken for the 7 day period is 5/7*7=5.

- Amount (mg) of rescue medication use in Week 2: For days 8-14 we have the number of doses taken of 2, 2 (imputed), 0, 1, 1 (imputed), 2, and 0. The number of doses taken for the 7 day period is 8/7*7=8, making the amount of acetaminophen dosage of 4000mg.

Example of calculating rescue medication data from Daily Diary Data (Subject discontinues)

In this example, a subject discontinues on study day 62, a few days after a Week 8 visit (which was on study day 60). The Week 5-8 data is calculated as described above (eg, Week 8 using days [50, 56]). The subject has rescue medication data as shown below.

Study Day (Week)	Number of Doses of RM taken [1]	Number of Doses of RM taken [1] with LOCF imputation
57 (Week 9)	1	1
58 (Week 9)	1	1
59 (Week 9)	Missing	1 [2]
60 (Week 9)	Missing	1 [2]
61 (Week 9)	Missing	1 [2]
62 (Week 9)	Missing	1 [2]
63 (Week 9)	Missing	1 [2]

[1] 500mg tablets of acetaminophen; [2] Using LOCF imputation for missing data

Week 9 is calculated as days 57 to 63. The data up to the end of the last week the subject was in the trial is imputed using LOCF as shown above. Therefore the Week 9 scores are then used to impute the Weekly data for summary and analysis for Weeks 10 to 56 (up to Week 16 for the amount of rescue medication use).

As above the incidence of rescue medication for Week 9 would be 'Yes'. The number of days of rescue medication use would be 7, and the average dose would be 7/7*7*500=3500mg for this week.

Appendix 3. STATISTICAL METHODOLOGY DETAILS

Appendix 3.1. Further Details of Interim Analyses

Details of the ongoing review of safety data (including joint safety events) are given in a separate statistical analysis plan for the Data Monitoring Committee.

Appendix 3.2. Further Details of the Statistical Methods

A description of the combination of the ANCOVA results from each of the multiple imputed datasets is given below, and taken from Little & Rubin (2002),[3] page 86-7.

In this analysis we have defined the number of imputations (D) to be 100.

The treatment estimates for individual treatment groups and treatment contrasts are defined as θ_i for $i = 1 \ldots D$. The combined estimate is $\bar{\theta}_D = \frac{1}{D}\sum_{i=1}^{D}\theta_i$. The variability of the combined estimate contains components of both Within- (W) and Between- (B) imputation dataset variability. These are shown below:

$$\bar{W}_D = \frac{1}{D}\sum_{i=1}^{D} W_i \text{ and } B_D = \frac{1}{D-1}\sum_{i=1}^{D}\left(\hat{\theta}_i - \bar{\theta}_D\right)^2$$

where W_i is the variance for the parameter θ_i.

The total variance for $\bar{\theta}_D$ is shown below:

$$T_D = \bar{W}_D + \frac{D+1}{D} B_D.$$

The test statistic $\frac{\left(\theta - \bar{\theta}_D\right)}{\sqrt{T_D}}$ has a t-distribution with v^* degrees of freedom, which is defined below:

$$v^* = \left(\frac{1}{v} + \frac{1}{\hat{v}_{obs}}\right),$$

using

$$v = (D-1)\left(1 + \frac{1}{D+1}\frac{\overline{W}_D}{B_D}\right)^2$$

$$\hat{v}_{obs} = (1-\hat{\gamma}_D)\left(\frac{v_{com}+1}{v_{com+3}}\right)v_{com}$$

$$\hat{\gamma}_D = \left(1+\frac{1}{D}\right)\frac{B_D}{T_D}.$$

This distribution can be used to construct the test statistics and 95% confidence intervals for θ.