STATISTICAL ANALYSIS PLAN

Study Title:	A Randomized, Double-blind, Placebo- and Active -controlled, Multicenter, Phase 3 Study to Assess the Efficacy and Safety of Filgotinib Administered for 52 weeks Alone and in Combination with Methotrexate (MTX) to Subjects with Moderately to Severely Active Rheumatoid Arthritis Who Are Naïve to MTX Therapy
Name of Test Drug:	Filgotinib
Study Number:	GS-US-417-0303
Protocol Version (Date):	Amendment 1: 05 July 2016
Analysis Type:	Week 24 Analysis and Final Analysis
Analysis Plan Version:	Final Draft
Analysis Plan Date:	06 August 2018
Analysis Plan Author(s):	PPD

CONFIDENTIAL AND PROPRIETARY INFORMATION

TABLE OF CONTENTS

LIST OF IN-TEXT TABLES

LIST OF IN-TEXT FIGURES

LIST OF ABBREVIATIONS

ACR	American College of Rheumatology
ACR20/50/70	American College of Rheumatology 20/50/70% improvement
AE	adverse event
AESIs	adverse events of special interest
ALP	alkaline phosphatase
ALT	alanine aminotransferase
ANCOVA	analysis of covariance
anti-CCP Ab	anti-citrullinated peptide antibody
AST	aspartate aminotransferase
ATC	anatomical therapeutic chemical drug class
bDMARD	biological disease modifying anti-rheumatic drug
BLQ	below the limit of quantitation
BMI	body mass index
CBC	complete blood count
CDAI	clinical disease activity index
CI	confidence interval
CPK	creatine phosphokinase
CSR	clinical study report
CTCAE	Common Toxicity Criteria for Adverse Events
DAS28	disease activity score for 28 joint count
DMC	data monitoring committee
ECG	electrocardiogram
eCRF	electronic case report form
EQ-5D	EuroQoL 5 Dimensions
ET	early termination
EULAR	European League Against Rheumatism
FACIT-Fatigue	Functional Assessment of Chronic Illness Therapy-Fatigue
FAS	full analysis set
Gilead	Gilead Sciences
HAQ-DI	health assessment questionnaire-disability index
HLGT	high-level group term
HLT	high-level term
HRQoL	health-related quality of life
HRUQ	healthcare resource utilization questionnaire
hsCRP	high=sensitivity c-reactive protein
IAC	independent adjudication committee
ID	identification
IWRS	interactive web response system

LLOQ	lower limit of quantitation
LLT	lower-level term
LOCF	last observation carried forward
LOQ	limit of quantitation
LS	least squares
LTE	long-term extension
MedDRA	Medical Dictionary for Regulatory Activities
MI	multiple imputation
MMRM	mixed model repeated measures
MNAR	missing not at random
MST	MedDRA search listing
MTX	methotrexate
NRI	non-responder imputation
OC	observed case
PD	pharmacodynamics
PGA	physician's global assessment
PK	pharmacokinetic
PP	per-protocol
PT	preferred term
PTM	placebo to match
Q1, Q3	first quartile, third quartile
QD	quaque die (each day)
RA	rheumatoid arthritis
RF	rheumatoid factor
SAE	serious adverse event
SAP	statistical analysis plan
SD	standard deviation
SDAI	simplified disease activity index
SE	standard error
SF-36	36-item short form survey
SGA	subject's global assessment
SJC	swollen joint count
SMQ	standardized MedDRA query
SOC	system organ class
TEAE	treatment-emergent adverse event
TFLs	tables, figures, and listings
TNFα	tumor necrosis factor-alpha
TJC	tender joint count
TSQM	treatment satisfaction questionnaire for medication
ULN	upper limit of normal

VAS	visual analog scale
VR	ventricular rate
WBC	white blood cell
WHO	World Health Organization
WPAI	Work Productivity and Activity Impairment

PHARMACOKINETIC ABBREVIATIONS

AUC_{last}	area under the concentration versus time curve from time zero to the last quantifiable concentration
AUC_{tau}	area under the concentration versus time curve over the dosing interval
C_{last}	last observed quantifiable concentration of the drug
C_{max}	maximum observed concentration of drug
C_{tau}	observed drug concentration at the end of the dosing interval
CL_{ss}/F	apparent oral clearance after administration of the drug: at steady state: CL_{ss}/F = Dose/AUC_{tau} , where “Dose” is the dose of the drug
$t_{1/2}$	estimate of the terminal elimination half-life of the drug, calculated by dividing the natural log of 2 by the terminal elimination rate constant (λ_z)
T_{last}	time (observed time point) of C_{last}
T_{max}	time (observed time point) of C_{max}
V_z/F	apparent volume of distribution of the drug
λ_z	terminal elimination rate constant, estimated by linear regression of the terminal elimination phase of the concentration of drug versus time curve

1. INTRODUCTION

This statistical analysis plan (SAP) describes the statistical analysis methods and data presentations to be used in tables, figures, and listings (TFLs) in the clinical study report (CSR) for Study GS-US-417-0303. This SAP is based on the study protocol Amendment 1 dated 05 July 2016 and the electronic case report form (eCRF). The SAP will be finalized before database finalization of Week 24 Analysis. Any changes made after the finalization of the SAP will be documented in the CSR.

1.1. Study Objectives

The primary objective of this study is as follows:

- To evaluate the effects of filgotinib in combination with MTX versus MTX alone for the treatment of signs and symptoms of rheumatoid arthritis (RA) as measured by the proportion of subjects achieving an American College of Rheumatology 20% improvement response (ACR20) at Week 24

The secondary objectives of this study are as follows:

- To evaluate the effect of filgotinib in combination with MTX versus MTX alone on physical function as measured by change from Baseline in the Health Assessment Questionnaire Disability Index (HAQ-DI) score at Week 24

- To evaluate the effects of filgotinib in combination with MTX versus MTX alone for the treatment of signs and symptoms of RA as measured by the proportion of subjects achieving Disease Activity Score for 28 joint count using CRP (DAS28 [CRP])<2.6 at Week 24

- To evaluate the effects of filgotinib alone versus MTX alone for the treatment of signs and symptoms of RA as measured by the proportion of subjects achieving ACR20 at Week 24

- To evaluate the effect of filgotinib alone versus MTX alone on physical function as measured by change from Baseline in HAQ-DI score at Week 24

- To evaluate the effects of filgotinib alone versus MTX alone for the treatment of signs and symptoms of RA as measured by the proportion of subjects achieving DAS28 (CRP)<2.6 at Week 24

- To evaluate the effects of filgotinib in combination with MTX versus MTX alone on preservation of joint structure as measured by change from Baseline in the mTSS at Week 24 and 52

- To evaluate the effects of filgotinib alone versus MTX alone on preservation of joint structure as measured by change from Baseline in modified Total Sharp Score (mTSS) at Week 24 and 52

- To evaluate the safety and tolerability of filgotinib alone and in combination with MTX
- To evaluate the effects of filgotinib alone and in combination with MTX on work productivity, fatigue, and general quality of life as measured by Short-Form Health Survey, 36 Questions (SF-36), Functional Assessment of Chronic Illness Therapy-Fatigue (FACIT-Fatigue), EuroQol 5 Dimensions (EQ-5D) and Work Productivity and Activity Impairment for RA (WPAI-RA)

The exploratory objectives of this study are as follows:

CCI

- [redacted]
- [redacted]
- [redacted]

1.2. Study Design

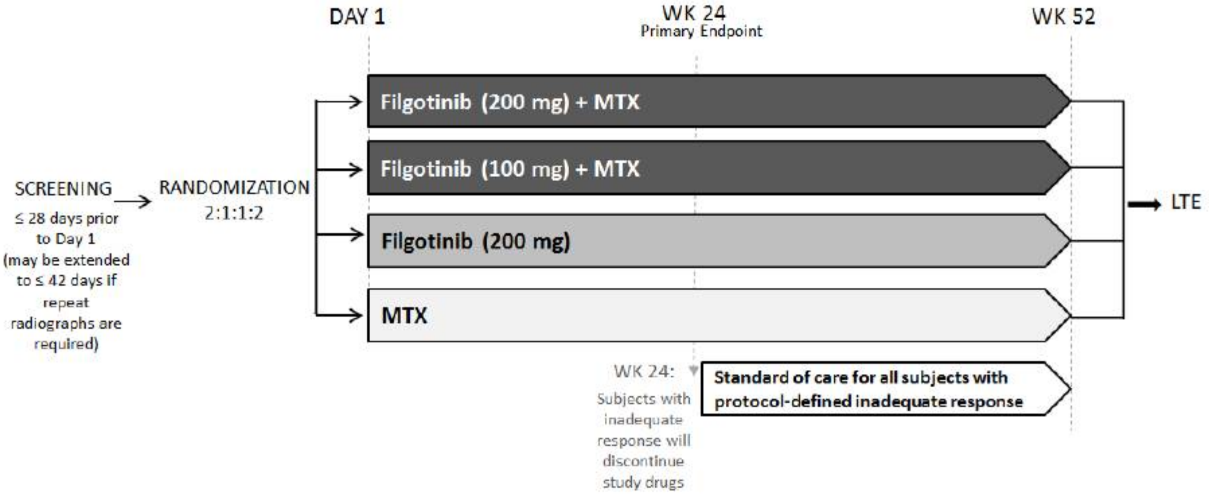
This is a randomized, double-blind, placebo- and active-controlled, add-on, Phase 3 study in adult male and female subjects with moderately to severely active RA who have limited or no prior exposure to MTX therapy. The study is designed to evaluate the efficacy, safety and tolerability of filgotinib as well as its effect on patient-reported outcomes, including work productivity, fatigue, and quality of life. In addition, PK will be assessed.

Approximately 1200 subjects will be randomized in a 2:1:1:2 ratio to filgotinib 200 mg plus MTX, filgotinib 100 mg plus MTX, filgotinib 200 mg alone, or MTX alone, for up to 52 weeks:

- Filgotinib 200 mg + MTX group: filgotinib (200 mg once daily [q.d.]) + PTM filgotinib 100 mg (PTM q.d.) + MTX (once weekly up to 20 mg) (N=400)
- Filgotinib 100 mg + MTX group: filgotinib (100 mg q.d.) + PTM filgotinib 200 mg (PTM q.d.) + MTX (once weekly up to 20 mg) (N=200)
- Filgotinib 200 mg monotherapy group: filgotinib (200 mg q.d.) + PTM filgotinib 100 mg (PTM q.d.) + PTM MTX (PTM once weekly) (N=200)
- MTX monotherapy group: PTM filgotinib 100 mg (PTM q.d.) and PTM filgotinib 200 mg (PTM q.d.) + MTX (once weekly up to 20 mg) (N=400)

Randomization will be stratified by geographic region and presence of rheumatoid factor (RF) or anti-cyclic citrullinated peptide antibody (anti-CCP Ab) at screening.

Figure 1-1. Study Design


DAY 1
WK 24
Primary Endpoint
WK 52
SCREENING
≤ 28 days prior to Day 1 (may be extended to ≤ 42 days if repeat radiographs are required)
RANDOMIZATION
2:1:1:2
Filgotinib (200 mg) + MTX
Filgotinib (100 mg) + MTX
Filgotinib (200 mg)
MTX
LTE
WK 24:
Subjects with inadequate response will discontinue study drugs
Standard of care for all subjects with protocol-defined inadequate response

At Week 24, subjects who have not achieved at least 20% improvement from Day 1 in both swollen joint count (SJC) and tender joint count (TJC) will discontinue investigational study drug dosing but will continue with study visits and assessments per protocol. All subjects meeting this criterion who discontinue from investigational therapy are to receive standard of care treatment for their RA as determined by the investigator.

All subjects who attain responder status at Week 24 will continue on their assigned study drug, in a blinded fashion through Week 52. Placebo-responders at Week 24 will continue on placebo in a blinded fashion through Week 52.

All subjects who have received at least one dose of study drug and exit the study early will complete an early termination (ET) visit at the time of study discontinuation, with a follow up visit four weeks after the last dose of study drug (Post Treatment visit Week 4), regardless of dosing duration.

At completion of the 52-week dosing period all subjects (who did not discontinue assigned study drug or have not met criteria for loss of therapeutic response) will be provided the option to enroll into a separate Long Term Extension (LTE) study (GS-US-417-0304).

1.3. Sample Size and Power

Sample size is determined based on the superiority test of filgotinib in combination with MTX compared to MTX alone based on the change from Baseline in mTSS at Week 24. When assuming a difference of 0.62 between the two groups and a common standard deviation of 2.7, 400 subjects in the filgotinib 200 mg in combination with MTX group and 400 in the MTX alone group are required to obtain 90% power at a 2-sided 0.05-level.

The total sample size will be 1200 (400 subjects for filgotinib 200 mg in combination with MTX group, 200 subjects for filgotinib 100 mg in combination with MTX group, 200 subjects for filgotinib 200 mg alone group and 400 subjects for MTX alone group). This sample size will provide over 90% power to detect an increase in ACR20 response rate of 62% to 78% between the MTX alone group and each of the filgotinib groups respectively, using a 2-sided 0.05-level test.

2. TYPE OF PLANNED ANALYSIS

2.1. Data Monitoring Committee Analyses

An external multidisciplinary Data Monitoring Committee (DMC) will review the progress of the study and perform interim reviews of safety data in order to protect subject welfare and preserve study integrity. To ensure the best interests of the participants, the DMC will recommend to the sponsor if the nature, frequency, and severity of adverse effects associated with the study treatment warrant the early termination of the study, the continuation of the study, or the continuation of the study with modifications.

The initial review was conducted after approximately 100 subjects completed at least 12 weeks of dosing in any one of the three RA phase 3 studies (GS-US-417-0301, GS-US-417-0302, and GS-US-417-0303). Regular DMC review of safety data will be scheduled approximately every 6 months following the initial data review meeting including all 3 RA phase 3 studies and a long-term extension study (GS-US-417-0304).

The DMC's role and responsibilities and the scope of analysis to be provided to the DMC are provided in a mutually agreed upon charter, which defines the DMC membership, meeting logistics, and meeting frequency.

2.2. Week 24 Analysis

A planned Week 24 analysis will be conducted after all subjects have either completed their Week 24 visit or prematurely discontinued from the study. A prespecified sponsorteam including members of which are not actively involved in the conduct of the study, will review the Week 24 unblinded safety and efficacy analysis results. A memo and a list will be maintained which document the individuals who have been granted access to the Week 24 unblinded results along with the justification for unblinding in accordance with Gilead SOPs. An interim analysis communication plan will be developed prior to unblinding. The Study Management Team with direct involvement in the conduct of the study will remain blinded to treatment assignments throughout the trial, until all subjects have completed the planned study Week 52 visits and the database has been locked.

2.3. Final Analysis

After all subjects have completed the study, outstanding data queries have been resolved or adjudicated as unresolvable, and the data have been cleaned and finalized, the study blinding will be broken and the final analysis of the data will be performed.

3. GENERAL CONSIDERATIONS FOR DATA ANALYSES

Analysis results will be presented using descriptive statistics. For categorical variables, the number and percentage of subjects in each category will be presented; for continuous variables, the number of subjects (n), mean, standard deviation (SD) or standard error (SE), median, first quartile (Q1), third quartile (Q3), minimum, and maximum will be presented.

All statistical tests will be 2-sided and performed at the 5% significance level unless otherwise specified.

By-subject listings will be presented for all subjects in the All Randomized Analysis Set and sorted by subject identification (ID) number, visit date, and time (if applicable). Data collected on log forms, such as AEs, will be presented in chronological order within the subject. The treatment group to which subjects were randomized will be included in the listings, as well as age, sex at birth, race, and ethnicity.

3.1. Analysis Sets

Analysis sets define the subjects to be included in an analysis. Analysis sets and their definitions are provided in this section. Subjects included in each analysis set will be determined before the study blind is broken for analysis. The analysis set will be identified and included as a subtitle of each TFL.

For each analysis set, the number and percentage of subjects eligible for inclusion, as well as the number and percentage of subjects who were excluded and the reasons for their exclusion, will be summarized by treatment group.

A listing of reasons for exclusion from analysis sets will be provided by subject.

3.1.1. All Randomized Analysis Set

All Randomized Analysis Set includes all subjects who are randomized in the study. This is the primary analysis set for by-subject listings.

3.1.2. Full Analysis Set

The Full Analysis Set (FAS) includes all randomized subjects who received at least 1 dose of study drug. The study drugs in this study are filgotinib 200 mg, filgotinib 100 mg, MTX, and PTMs. This is the primary analysis set for efficacy analyses.

3.1.3. Per-Protocol Analysis Set

The Per-Protocol (PP) Analysis Set includes subjects in the FAS who did not have major protocol deviations occurring prior to or at Week 24 which will affect the efficacy analysis, and were compliant (with on-treatment adherence ≥ 80% in the first 24-week period for all study drugs except MTX) to study treatment.

Major protocol deviations include the following

- Violate at least one of the following key eligibility criteria

  1) Have a diagnosis of RA (2010 ACR/EULAR criteria for RA), and are ACR functional class I-III

  2) Have ≥ 6 swollen joints (from swollen joint count based on 66 joints [SJC66]) and ≥ 6 tender joints (from tender joint count based on 68 joints [TJC68]) at both screening and Day 1 (need not be the same joints)

  3) Have serum CRP ≥ 4 mg/L

  4) Have limited or no prior treatment with MTX, ie, no more than 3 doses of MTX ≤ 25 mg each in the subject's lifetime for the treatment of RA, with the last dose at least 28 days prior to Day 1, and are an appropriate candidate for MTX therapy, as per investigator judgment

  5) Did not receive protocol prohibited medications

- Received prohibited concomitant medications

- Background medications were changed in violation of protocol

- Data were questionable because of significant site quality or compliance issues

Qualifications and identification of the specific major protocol deviations that result in exclusion from the PP Analysis Set will be determined while the study remains blinded, prior to database finalization.

The PP Analysis Set is the secondary analysis set for efficacy analyses on primary and key secondary endpoints. Statistical analyses using PP Analysis Set are specified in Section 6.

3.1.4. Safety Analysis Set

The Safety Analysis Set includes all subjects who received at least 1 dose of study drug. This is the primary analysis set for safety analyses.

3.1.5. Pharmacokinetic Analysis Set

The Pharmacokinetic (PK) Analysis Set includes all subjects in the Safety Analysis Set who have at least 1 nonmissing plasma concentration data for filgotinib and/or its metabolite GS-829845. This is the primary analysis set for general PK analyses.

CCI [redacted]

[redacted]

3.2. Subject Grouping

For analyses based on the All Randomized Analysis Setand FAS, subjects will be grouped according to the treatment to which they were randomized. For analyses based on the Safety Analysis Set and PP Analysis Set, subjects will be grouped according to actual treatment received. The actual treatment received will differ from the randomized treatment only when their actual treatment differs from randomized treatment for the entire treatment duration.

For the PK Analysis Set, subjects will be grouped according to the actual treatment they received.

3.3. Strata and Covariates

Subjects will be randomly assigned to treatment groups via the interactive web response system (IWRS) in a 2:1:1:2 ratio to filgotinib 200mg + MTX, filgotinib 100mg + MTX, filgotinib 200mg, and MTX using a stratified randomization schedule. Stratification will be based on the following variables:

- Geographic region (**Group A** includes the following countries: Australia, Belgium, Canada, France, Germany, Ireland, Israel, Italy, Netherlands, New Zealand, Republic of Korea, Singapore, South Africa, Spain, Sweden, Switzerland, United Kingdom, and USA; **Group B** includes the following countries: Bulgaria, Croatia, Czech Republic, Estonia, Georgia, Hungary, India, Latvia, Moldova, Poland, Romania, Russia, Serbia, Slovakia, and Ukraine; **Group C** includes the following countries: Argentina, Brazil, Chile, Colombia, Mexico, Peru, and Puerto Rico; **Group E** includes Japan; **Group D** includes the following contries/regions China (was originally planned but no subjects were screened or enrolled from this region), Hong Kong, Malaysia, Philippines, Taiwan, Thailand, and Vietnam.

- Presence of RF or anti-CCP Ab (Yes or No) (Presence of RF = No if RF < 15 IU/mL; Presence of anti-CCP Ab = No if anti-CCP Quant < 17 U/mL)

If there are discrepancies in stratification factor values between the IWRS and the clinical database, the values recorded in the clinical database will be used for analyses.

For efficacy endpoints, stratification factors and the baseline value of the efficacy variable(s) will be included as covariates in the efficacy analysis model, as specified in Section 6.

3.4. Examination of Subject Subgroups

The primary and key secondary efficacy endpoints will be examined using the following subgroups (but not limited to the ones listed below):

- Age (on the first dosing date of any study drug, < 65 or ≥ 65)
- Sex at birth (male, female)
- Race
- Baseline weight (< 60 kg, ≥ 60 to < 100 kg, ≥ 100 kg)
- Geographic region (A, B, C, or E)
- Presence of RF or anti-CCP Ab (Yes or No)
- Duration of RA diagnosis on the first dosing date of any study drug (< 1 year, ≥ 1 to < 2 years, ≥ 2 years)
- Disease activity at Baseline (DAS28[CRP] ≤ 5.1 or DAS28[CRP] > 5.1)
- Concurrent use of oral systemic corticosteroids on the first dosing date of any study drug (Yes or No)
- High-sensitivity C-reactive protein (hsCRP) at baseline (≥ 4 mg/L, < 4 mg/L)

Lists of oral systemic corticosteriods are provided in Appendix 1.

3.5. Multiple Comparisons

The primary endpoint for the study is the proportion of subjects who achieve ACR20 response at Week 24. The primary analyses will consist of a superiority test of filgotinib 200 mg in combination with MTX compared to MTX alone based on the ACR20 response rate at Week 24.

The following hypothesis testing for secondary analyses will commence after the primary analysis reaches statistical significance, and will be tested according to the hierarchical testing principle at the 2-sided 0.05 level. If a null hypothesis is not rejected, formal sequential testing will be stopped and only nominal significance will be reported for the remaining hypotheses:

1) Superiority of filgotinib 100 mg in combination with MTX compared to MTX alone based on ACR20 response rate at Week 24.
2) Superiority of filgotinib 200 mg in combination with MTX compared to MTX alone based on the change from Baseline in HAQ-DI at Week 24.
3) Superiority of filgotinib 100 mg in combination with MTX compared to MTX alone based on the change from Baseline in HAQ-DI at Week 24.

4) Superiority of filgotinib 200 mg in combination with MTX compared to MTX alone based on the change from Baseline in mTSS at Week 24.

5) Superiority of filgotinib 100 mg in combination with MTX compared to MTX alone based on the change from Baseline in mTSS at Week 24.

6) Superiority of filgotinib 200 mg alone compared to MTX alone based on ACR20 response rate at Week 24.

7) Superiority of filgotinib 200 mg alone compared to MTX alone based on the change from Baseline in HAQ-DI to Week 24.

8) Superiority of filgotinib 200 mg in combination with MTX compared to MTX alone based on the proportion of subjects with DAS28(CRP) < 2.6 at Week 24.

9) Superiority of filgotinib 100 mg in combination with MTX compared to MTX alone based on the proportionof subjects with DAS28(CRP) < 2.6 at Week 24.

10) Superiority of filgotinib 200 mg alone compared to MTX alone based on the proportion of subjects with DAS28(CRP) < 2.6 at Week 24.

11) Superiority of filgotinib 200 mg alone compared to MTX alone based on the change from Baseline in mTSS at Week 24.

12) Superiority of filgotinib 200 mg in combination with MTX compared to MTX alone based on the change from Baseline in SF-36 physical component score (PCS) at Week 24.

13) Superiority of filgotinib 100 mg in combination with MTX compared to MTX alone based on the change from Baseline in SF-36 PCS at Week 24.

14) Superiority of filgotinib 200 mg alone compared to MTX alone based on the change from Baseline in SF-36 PCS at Week 24.

15) Superiority of filgotinib 200 mg in combination with MTX compared to MTX alone based on the change from Baseline in FACIT-Fatigue at Week 24.

16) Superiority of filgotinib 100 mg in combination with MTX compared to MTX alone based on the change from Baseline in FACIT-Fatigue at Week 24.

17) Superiority of filgotinib 200 mg alone compared to MTX alone based on the change from Baseline in FACIT-Fatigue at Week 24.

3.6. Missing Data and Outliers

3.6.1. Missing Data

In general, missing data will not be imputed unless methods for handling missing data are specified. Exceptions are presented in this document.

For missing last dosing date of study drug, imputation rules are described in Section 4.2.1. For partial date of initial RA diagnosis, imputation rules are described in Section 5.2. The handling of missing or incomplete dates for AE onset is described in Section 7.1.5.2, and for prior and concomitant medications in Section 7.4. Imputation rules adopted in the efficacy analyses are specified in Section 6.

3.6.2. Outliers

Outliers will be identified during the data management and data analysis process, but no sensitivity analyses will be conducted. All data will be included in the data analyses.

3.7. Data Handling Conventions and Transformations

In general, age (in years) on the date of the first dose of study drug will be used for analyses and presentation in listings. If an enrolled subject was not dosed with any study drug, the randomization date will be used instead of the first dosing date of study drug. For screen failures, the date the last informed consent was signed will be used for age calculation. If only birth year is collected on the CRF, "01 July" will be used for the unknown birth day and month for the purpose of age calculation. If only birth year and month are collected, "01" will be used for the unknown birth day.

Non-PK Data that are continuous in nature but are less than the lower limit of quantitation (LOQ) or above the upper LOQ will be imputed as follows:

- A value that is 1 unit less than the LOQ will be used to calculate descriptive statistics if the datum is reported in the form of "< x" (where x is considered the LOQ). For example, if the values are reported as < 50 and < 5.0, values of 49 and 4.9, respectively, will be used to calculate summary statistics. An exception to this rule is any value reported as < 1 or < 0.1, etc. For values reported as < 1 or < 0.1, a value of 0.9 or 0.09, respectively, will be used for calculate summary statistics.

- A value that is 1 unit above the LOQ will be used to calculate descriptive statistics if the datum is reported in the form of "> x" (where x is considered the LOQ). Values with decimal points will follow the same logic as above.

- The LOQ will be used to calculate descriptive statistics if the datum is reported in the form of "≤ x" or "≥ x" (where x is considered the LOQ).

Natural logarithm transformation will be used for plasma/blood concentrations and analysis of PK parameters. Plasma concentration values that are below the limit of quantitation (BLQ) will be presented as "BLQ" in the concentration data listing. Values that are BLQ will be treated as 0 at predose time points, and one-half the value of the LOQ at postbaseline time points.

The following conventions will be used for the presentation of summary and order statistics:

- If at least 1 subject has a concentration value of BLQ for the time point, the minimum value will be displayed as "BLQ."
- If more than 25% of the subjects have a concentration data value of BLQ for a given time point, the minimum and Q1 values will be displayed as "BLQ."
- If more than 50% of the subjects have a concentration data value of BLQ for a given time point, the minimum, Q1, and median values will be displayed as "BLQ."
- If more than 75% of the subjects have a concentration data value of BLQ for a given time point, the minimum, Q1, median, and Q3 values will be displayed as "BLQ."
- If all subjects have concentration data values of BLQ for a given time point, all order statistics (minimum, Q1, median, Q3, and maximum) will be displayed as "BLQ."

PK parameters that are BLQ will be imputed as one-half LOQ before log transformation or statistical model fitting.

3.8. Analysis Visit Windows

3.8.1. Definition of Study Day

The first dose date of individual study drug will be calculated separately for each study drug (ie, filgotinib 200 mg, filgotinib 100 mg, MTX, and PTMs) in a treatment group. Study Day 1 is defined as the first dose date of any study drug, which is the minimum of the first dose dates of individual study drugs in a treatment group.

The last dose date of individual study drug will be calculated separately for each study drug in a treatment group. The last dose date for an individual study drug will be the end date on study drug administration CRF for the record where the "study drug was permanently withdrawn" flag is "Yes". The last dose date of any study drug will be defined as the maximum of the last dose dates of individual study drugs in a treatment group.

Study Day will be calculated from the Study Day 1 and derived as follows:

- For postdose study days: Assessment Date – Study Day 1 + 1
- For days prior to the first dose: Assessment Date – Study Day 1

3.8.2. Analysis Visit Windows

Subject visits may not occur on protocol-specified days. Therefore, for the purpose of analysis, observations will be assigned to analysis windows.

In general, the baseline value will be the last nonmissing value on or prior to the first dose date of study drug, except for the mTSS baseline value. The mTSS baseline value will be derived based on the measurements collected on or prior to the first dose date of study drug + 14 days given the potential delay in scheduling an x-ray assessment..

For efficacy endpoints, vital signs, electrocardiogram (ECG), weight, lipid, and safety laboratory data, the analysis visit windows will be applied to data collected during the 'on-treatment' period. The 'on-treatment' period is defined as the last dose date of any study drug plus 7 days.

The analysis windows for on-treatment efficacy endpoints including joint count assessment, HAQ-DI (including subject's pain assessment), subject's global assessment (SGA), physician's global assessment (PGA), serum CRP, CCI [redacted] and vital signs, weight, and safety laboratory data (lipid data excluded) are provided in Table 3-1.

Table 3-1. Analysis Visit Windows for On-treatment Joint Count Assessment, HAQ-DI, SGA, PGA, CRP, CCI [redacted] and Vital Signs, Weight, and Safety Laboratory Data (Lipid Data Excluded)

Analysis Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 2	15	2	22
Week 4	29	23	43
Week 8	57	44	71
Week 12	85	72	99
Week 16	113	100	127
Week 20	141	128	155
Week 24	169	156	190
Week 30	211	191	232
Week 36	253	233	281
Week 44	309	282	337
Week 52	365	338	≥ 365

The analysis windows for on-treatment ECG data are provided in Table 3-2.

Table 3-2. Analysis Visit Windows for On-Treatment ECG Data

Analysis Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 12	85	2	127
Week 24	169	128	211
Week 36	253	212	309
Week 52	365	310	≥ 365

The analysis windows for on-treatment lipid data are provided in Table 3-3.

Table 3-3. Analysis Visit Windows for On-Treatment Lipid Data

Analysis Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 12	85	2	127
Week 24	169	128	267
Week 52	365	268	≥ 365

The analysis windows for on-treatment mTSS data are provided in Table 3-4.

Table 3-4. Analysis Visit Windows for On-treatment mTSS Data

Analysis Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	14
Week 24	169	15	267
Week 52	365	268	≥ 365

The analysis windows for on-treatment SF-36, FACIT-Fatigue, EQ-5D and WPAI-RA are provided in Table 3-5. The analysis windows for on-treatment Healthcare Resource Utilization Questionnaire (HRUQ) and Treatment Satisfaction Questionnaire for Medication (TSQM) are provided in Table 3-6.

Table 3-5. Analysis Visit Windows for On-treatment SF-36, FACIT-Fatigue, EQ-5D, and WPAI-RA

Analysis Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 4	29	2	57
Week 12	85	58	127
Week 24	169	128	211
Week 36	253	212	309
Week 52	365	310	≥ 365

Table 3-6. Analysis Visit Windows for On-Treatment HRUQ and TSQM

Analysis Visit	Nominal Study Day	Lower Limit	Upper Limit
Baseline	1	(none)	1
Week 12	85	2	127
Week 24	169	128	211
Week 36	253	212	309
Week 52	365	310	≥ 365

Vital signs, weight, ECGs, lipids, and safety laboratory data collected in the posttreatment followup period will also be summarized. The analysis window for posttreatment followup period is defined as from (the last dose date of any study drug + 8 days) to (the last dose date of any study drug + 30 days). If multiple valid, nonmissing measurements exist in the posttreatment followup visit window, then the latest record will be selected for analysis. If the chronological order cannot be determined (eg, more than 1 record on the same day with time missing), for any given subject, the value with the worst severity will be selected for categorical variable, and the average will be taken for continuous variable. Data obtained after last dose date plus 30 days will be excluded from the summaries, but will be included in the listings.

For efficacy endpoints, the same analysis visit windows will be applied to all available data collected during standard of care.

3.8.3. Selection of Non-Efficacy Data in the Event of Multiple Records in an Analysis Visit Window

Depending on the statistical analysis method, single values may be required for each analysis window. For example, change from baseline by visit usually requires a single value, whereas a time-to-event analysis would not require 1 value per analysis window.

If multiple valid, nonmissing, continuous measurements exist in an analysis window, records will be chosen based on the following rules if a single value is needed:

- In general, the baseline value will be the last nonmissing value on or prior to the first dosing date of study drug, unless specified differently. If multiple measurements occur on the same day, the last nonmissing value prior to the time of first dosing of study drug will be considered as the baseline value. If these multiple measurements occur at the same time or the time is not available, the average of these measurements (for continuous data) will be considered the baseline value.

- For postbaseline visits:

  - The record closest to the nominal day for that visit will be selected.

  - If there are 2 records that are equidistant from the nominal day, the later record will be selected.

  - If there is more than 1 record on the selected day, the average will be taken, unless otherwise specified.

If multiple valid, nonmissing, categorical measurements exist in an analysis window, records will be chosen based on the following rules if a single value is needed:

- For baseline, the last available record on or prior to the date of the first dose of study drug will be selected. If there are multiple records with the same time or no time recorded on the same day, the value with the lowest severity will be selected (eg, normal will be selected over abnormal for safety ECG findings).

- For postbaseline visits:

  - The record closest to the nominal day for that visit will be selected.

  - If there are 2 records that are equidistant from the nominal day, the later record will be selected.

  - If there is more than 1 record on the selected day, the value with the worst severity will be selected (eg, abnormal will be selected over normal for safety ECG findings), unless otherwise specified.

The rules for selecting efficacy data in the event of multiple records in an analysis visit window are specified in Section 6.1.

4. SUBJECT DISPOSITION

4.1. Subject Enrollment and Disposition

A summary of subject enrollment will be provided by treatment group for each country within each geographic region and investigator within a country. The summary will present the number and percentage of subjects enrolled. For each column, the denominator for the percentage calculation will be the total number of subjects analyzed for that column.

A similar enrollment table will be provided by stratification factor stratum. The denominator for the percentage of subjects in the stratum will be the total number of enrolled subjects. If there are discrepancies in the value used for stratification assignment between the IWRS and the clinical database, the value collected in the clinical database will be used for the summary. A listing of subjects with discrepancies in the value used for stratification assignment between the IWRS and the clinical database at the time of data finalization will be provided.

The randomization schedule used for the study will be provided in a listing as an appendix to the CSR.

A summary of subject disposition will be provided by treatment group. This summary will present the number of subjects screened, the number of subjects not randomized, the number of subjects who met all eligibility criteria but were not randomized with reasons subjects were not randomized, the number of subjects randomized, and the number of subjects in each of the categories listed below:

- Safety Analysis Set
- Full Analysis Set
- Per-Protocol Analysis Set
- PK Analysis Set
- CCI [redacted]
- Continuing study drug (Week 24 Analysis only)
- Completed study drug
- Did not complete study drug with reasons for premature discontinuation of study drug
- Continuing study (Week 24 Analysis only)
- Completed study
- Did not complete study with reasons for premature discontinuation from the study

For the status of study drug and study completion, and reasons for premature discontinuation, the number and percentage of subjects in each category will be provided. The denominator for the percentage calculation will be the total number of subjects in the Safety Analysis Set corresponding to that column.

The following by-subject listings will be provided by subject ID number in ascending order to support the above summary tables:

- Reasons for premature study drug or study discontinuation
- Reasons for screen failure (will be provided by screening ID number in ascending order)
- Lot number and kit ID of assigned study drugs

4.2. Extent of Study Drug Exposure and Adherence

Extent of exposure to study drug will be examined by assessing the total duration of exposure to study drug and the level of adherence to the study drug specified in the protocol.

4.2.1. Duration of Exposure to Study Drug

Total duration of exposure to any study drug will be defined as last dosing date of any study drug minus first dosing date of any study drug plus 1, regardless of any temporary interruptions in study drug administration, and will be expressed in weeks using up to 1 decimal place (eg, 4.5 weeks).

Total duration of exposure to individual study drug will be defined as last dosing date of individual study drug minus first dosing date of individual study drug plus 1, regardless of any temporary interruptions in study drug administration, and will be expressed in weeks using up to 1 decimal place.

For subjects with a partial last dosing date (ie, moth and year of last dose are known), the latest of the dispensing dates of study drug bottles, study drug start dates and end dates, and the imputed last dosing date (day imputed as 15) will be used as the final imputed last dosing date. If the subject died and the death date is complete (ie, not partial date) and before the imputed last dosing date, the complete death date will be used as the imputed last dosing date.

If only year is recorded (ie, month and day of last dose are missing), the latest of the dispensing month of study drug bottles, study drug start month, and study drug end month will be used to impute the unknown last dose month. If the subject died and the death date has month and year available and before the imputed last dose month, then the month of death will be used instead. With the month imputed, the aforementioned method will be used to impute the last dose date.

If subjects are continuing on study drug at Week 24 analysis, the earliest of the date of death or data cutoff date for Week 24 analysis will be used to impute the last dosing date for the calculation of duration of exposure to study drug.

The total duration of exposure to any study drug and total duration of exposure to each individual study drug will be summarized using descriptive statistics and using the number (ie, cumulative counts) and percentage of subjects exposed through the following time periods: Baseline (Day 1), Week 2 (Day 15), Week 4 (Day 29), Week 8 (Day 57), Week 12 (Day 85), Week 16 (Day 113), Week 20 (Day 141), Week 24 (Day 169), Week 30 (Day 211), Week 36 (Day 253), Week 44 (Day 309), and Week 52 (Day 365).

Summaries will be provided by treatment group for the Safety Analysis Set. No formal statistical testing is planned.

4.2.2. Adherence to Study Drug

Adherence will be calculated separately for filgotinib 200 mg/PTM (tablets), filgotinib 100 mg/PTM (tablets).

The total number of tablets administered will be summarized using descriptive statistics.

The presumed total number of tablets administered to a subject will be determined by the data collected on the drug accountability CRF using the following formula:

Total Number of Tablets or Capsules Administered

$$= \left(\sum \text{No. of Tablets Dispensed}\right) - \left(\sum \text{No. of Tablets Returned}\right)$$

If a bottle is dispensed and the bottle is returned empty, then the number of tablets returned will be entered as zero. If a bottle is dispensed but not returned (missing), the number of tablets administered will be counted as zero by assuming that a subject did not take study drug as prescribed during the period for which the bottle was dispensed.

4.2.2.1. On-Treatment Adherence

The level of on-treatment adherence to the study drug will be determined by the total amount of study drug administered relative to the total amount of study drug expected to be administered during a subject’s actual on-treatment period based on the study drug regimen.

The level of on-treatment adherence will be expressed as a percentage using the following formula:

$$\text{On-Treatment Adherence (\%)} = \left(\frac{\text{Total Amount of Study Drug Administered}}{\text{Study Drug Expected to be Administered on Treatment}}\right) \text{x } 100$$

Note: If calculated adherence is greater than 100%, the result will be set to 100%.

Study drug expected to be administered for filgotinib 200 mg/PTM (tablets) = 1× total duration of exposure to any study drug (days).

Study drug expected to be administered for filgotinib 100 mg/PTM (tablets) = 1× total duration of exposure to any study drug (days).

On-treatment adherence will be calculated up to Week 24 and up to Week 52.

Descriptive statistics for the level of on-treatment adherence with the number and percentage of subjects belonging to adherence categories (eg, < 80%, ≥ 80% to < 90%, ≥ 90%) will be provided by treatment group for the Safety Analysis Set.

Categorical displays will be presented for the number of subjects who are at least 80% adherent to their study drug regimen (ie, adherence is ≥ 80% for each study drug).

No formal statistical testing is planned.

A by-subject listing of study drug administration and drug accountability will be provided separately by subject ID number (in ascending order) and visit (in chronological order).

4.3. Protocol Deviations

Subjects who did not meet the eligibility criteria for study entry, but enrolled in the study will be summarized regardless of whether they were exempted by the sponsor or not. The summary will present the number and percentage of subjects who did not meet at least 1 eligibility criterion and the number of subjects who did not meet specific criteria by treatment group based on the All Randomized Analysis Set. A by-subject listing will be provided for those subjects who did not meet at least 1 eligibility (inclusion or exclusion) criterion. The listing will present the eligibility criterion (or criteria if more than 1 deviation) that subjects did not meet and related comments, if collected.

Protocol deviations occurring after subjects entered the study are documented during routine monitoring. The number and percentage of subjects with important protocol deviations by deviation reason (eg, nonadherence to study drug, violation of select inclusion/exclusion criteria) will be summarized by treatment group for the All Randomized Analysis Set. A by-subject listing will be provided for those subjects with important protocol deviation.

5. BASELINE CHARACTERISTICS

5.1. Demographics and Other Baseline Characteristics

Subject demographic variables will be summarized by treatment group and overall using descriptive statistics for continuous variables, and using number and percentage of subjects for categorical variables. The summary of demographic data will be provided for the Safety Analysis Set for the following:

- Age (on the first dose date of any study drug)
- Age group (< 65 years, ≥ 65 years)
- Sex at birth (male, female)
- Race
- Ethnicity (Hispanic or Latino, not Hispanic or Latino)
- Geographic region and country
- Weight
- Height
- Body mass index (BMI; in kg/m^2)
- Smoking status

A by-subject demographic listing, including the informed consent date, will be provided by subject ID number in ascending order.

5.2. Baseline Disease Characteristics

Baseline disease characteristics include:

- Duration of RA from diagnosis (years)

  Calculated as ((first dose date) – (date of initial diagnosis) + 1 day) / 365.25. If the date of initial diagnosis is incomplete, then the following rules will be applied:

  — missing day: use the first of the month.

  — missing month: use January.
- Presence of RF (Yes/No)
- Presence of anti-CCP Ab (Yes/No)

- Presence of RF and anti-CCP Ab (Yes/No)
- Concurrent oral corticosteroid use on the first dosing date (Yes/No): n(%)
  - Oral corticosteroids dose, mg/day, expressed as prednisone-equivalent dose, mean (SD)
- Concurrent anti-malarials use on the first dosing date (Yes/No): n(%)
- DMARD and investigational bDMARD naïve
- Swollen joint count based on 66 joints (SJC66)
- Tender joint count based on 68 joints (TJC68)
- Swollen joint count based on 28 joints (SJC28)
- Tender joint count based on 28 joints (TJC28)
- HAQ-DI total score
- DAS28(CRP)
- SF-36 PCS score
- SF-36 MCS score
- FACIT-Fatigue
- Subject’s global assessment of disease activity (SGA, by visual analog scale [VAS] in mm)
- Physician’s global assessment (PGA, by VAS in mm)
- Subject’s pain assessment (by VAS in mm)
- Simplified Disease Activity Index (SDAI)
- Clinical Disease Activity Index (CDAI)
- hsCRP (mg/L)
- mTSS
  - Erosion score
  - Joint space narrowing score

These baseline characteristics will be summarized by treatment group and overall using descriptive statistics for continuous variables and using number and percentage of subjects for categorical variables. The summary of baseline characteristics will be provided for the Safety Analysis Set.

A by-subject listing of other baseline characteristics will be provided by subject ID number in ascending order.

5.3. Medical History

Medical history collected at screening will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) 21.0.

Medical history will be summarized by system organ class (SOC), preferred term (PT), treatment group, and overall. Subjects who report 2 or more medical history items that are coded to the same SOC and/or PT will be counted only once by the unique coded term in the summary.

In addition, numbers and percentages of subjects who have any first degree relatives that had experienced myocardial infarction or stroke before the age of 50 years, experienced myocardial infarction or stroke before the age of 50 years, or have diabetes (type I or II) will be summarized.

The summary will be provided for the Safety Analysis Set. No formal statistical testing is planned.

A by-subject listing of medical history will be provided by subject ID number.

6. EFFICACY ANALYSES

6.1. General Considerations

The primary analysis set for efficacy analyses will be the FAS, defined in Section 3.1.2.

Efficacy analysis will generally be conducted on the following 2 datasets:

1) On-treatment data as specified in Section 3.8.2 with exclusion of the observations described as the following. This is the primary analysis dataset for efficacy analysis.

   For subjects who continued study drug BUT did not achieve at least 20% improvement from Baseline in both SJC66 and TJC68 at

   - Nominal Week 24 visit, then efficacy data collected after nominal Week 24 visit will be excluded from the primary analysis, OR

   - Nominal Week 30 visit if subjects missed Week 24 visit or the response status cannot be determined due to missing SJC66 or TJC68 value at Week 24, then efficacy data collected after nominal Week 30 visit will be excluded from the primary analysis

   If subjects missed both visits (Week 24 and Week 30) or the response status cannot be determined at both visits, all the efficacy data collected will be included for analysis.

2) All available data (including data collected under standard of care).

<u>Selection of Efficacy Data in the Event of Multiple Records in an Analysis Visit Window</u>

If multiple valid and nonmissing efficacy measurements exist in an analysis window, then records will be chosen based on the following rules if a single value is needed:

- The record closest to the nominal day for that visit will be selected

- If there are 2 records that are equidistant from the nominal day, or more than 1 record (with time known) on the selected day, the latest record will be taken

- If chronological order cannot be determined (eg, more than 1 record on the same day with time missing), for any given subject, the worst outcome will be selected.

For the calculation of composite endpoints including DAS28(CRP), ACR20/50/70, ACR-N, SDAI, and CDAI, we use the following steps unless otherwise specified:

- Step 1: Assign individual components to analysis visit windows defined in Section 3.8.2

- Step 2: Within each analysis visit window, select the component-level data based on the rules for selecting efficacy data as above

- Step 3: Calculate the composite endpoint based on the selected component-level data in Step 2.

Missing Data Imputation

Below are the descriptions for the imputation methods that will be used throughout the efficacy analyses:

- Observed case (OC): Missing values remain missing. For the categorical composite endpoints, in the case that some components are missing, the composite endpoint assessment will be derived based on the non-missing components. If non-missing components are not sufficient to determine final composite endpoint, then the composite endpoint will be set as missing. For continuous composite endpoints (including ACR-N), if any components are missing, the composite endpoints will be set as missing.
- Last observation carried forward (LOCF): Baseline measurements will not be carried forward to postbaseline. Only postbaseline measurements will be LOCF. For the composite endpoints, the last nonmissing postbaseline observation will be carried forward to subsequent visits for each individual component, and then calculate the composite endpoints using individual components imputed by LOCF as described above. If a subject does not have a nonmissing observed record for a postbaseline visit, the last postbaseline record prior to the missed visit will be used for this postbaseline visit. If the last nonmissing observation prior to the missing visits cannot be determined due to multiple measurements occurring at the same time or the time not available within the same day, the worst outcome will be used for LOCF. If missing components still exist after LOCF, the composite endpoints will be calculated using the same rules as described in OC.
- Non-responder imputation (NRI): For all binary response measurements, starting from OC, all missings will be set as non-responders.

If subject only had baseline measurements, LOCF and OC analyses will not include this subject. But this subject will be treated as non-responder in NRI analyses.

6.1.1. Tender/Swollen Joint Counts (TJC/SJC)

The TJC68 and SJC66 will be collected during the course of the study. The assessment for each joint will be from the following selections: Tender Only, Swollen Only, Tender and Swollen, Joint Non-Evaluable or Missing, or Not Tender or Swollen.

Individual joint with missing assessment will not be imputed. If at least half of the joints are assessed at a given visit, the prorated tender and swollen joint counts will be calculated using the following formula:

$$\text{TJC68} = \frac{\text{Total number of tender joints}}{\text{68 - (Number of nonevaluable or missing joints out of 68 joints)}} \times 68$$

$$\text{SJC66} = \frac{\text{Total number of swollen joints}}{\text{66 - (Number of nonevaluable or missing joints out of 66 joints)}} \times 66$$

If less than half of joints are assessed at a given visit, joint counts are treated as missing for that visit.

A more abbreviated assessment considering 28 joints as listed in Table 6-1 for both tenderness and swelling will also be conducted (as part of the TJC68 and SJC66 assessment), denoted as TJC28 and SJC28, respectively.

Table 6-1. Composition of the 28 Joints

Joints	Number
Shoulder Joints (Left and Right)	2
Elbow Joints (Left and Right)	2
Wrist Joints (Left and Right)	2
Metacarpophalangeal Joints I-V (Left and Right)	10
Proximal Interphalangeal Joints I-V (Left and Right) – Hands only	10
Knee Joints (Left and Right)	2

If there exist non-evaluable or missing joints among the 28 joints, similar prorated tender and swollen joint counts will be calculated as follows:

$$\text{TJC28} = \frac{\text{Total number of tender joints}}{\text{28 - (Number of nonevaluable or missing joints out of 28 joints)}} \times 28$$

$$\text{SJC28} = \frac{\text{Total number of swollen joints}}{\text{28 - (Number of nonevaluable or missing joints out of 28 joints)}} \times 28$$

If less than half of the 28 joints are assessed at a given visit, TJC28 and SJC28 are treated as missing for that visit.

6.1.2. Global Assessment of Disease Activity

Subject’s Global Assessment of Disease Activity (SGA) and Physician’s Global Assessment of Disease Activity (PGA) based on a 0-100 visual analog scale (VAS) will be recorded during the study, with 0 indicating “no disease activity” and 100 indicating “maximum disease activity” (or similar description of disease severity).

6.1.3. Health Assessment Questionnaire Disability Index (HAQ-DI)

The HAQ-DI score is defined as the average of the scores of eight functional categories (dressing and grooming, arising, eating, walking, hygiene, reach, grip, and other activities), administered by the patient. Responses in each functional category are collected as: without any difficulty; with some difficulty; with much difficulty; unable to do a task in that area and with or without aids or devices. The HAQ-DI score ranges from 0 (no disability) to 3 (completely disabled), when 6 or more categories are non-missing. Detailed algorithm for calculating HAQ-DI score is described in Appendix 2.

HAQ-DI also includes a separate pain assessment and subject will be requested to mark the severity of the pain in the past week on a 0-100 VAS, with 0 indicating “no pain” and 100 indicating “severe pain”.

6.2. Primary Efficacy Endpoint

The primary efficacy endpoint is the proportion of subjects who achieve ACR20 response at Week 24.

6.2.1. Definition of the Primary Efficacy Endpoint

A subject achieves ACR20 response when this subject has

- ≥ 20% improvement from baseline in TJC68, AND
- ≥ 20% improvement from baseline in SJC66, AND
- ≥ 20% improvement from baseline in at least 3 of the following 5 items:
  a) PGA
  b) SGA
  c) Subject's pain assessment
  d) HAQ-DI score
  e) hsCRP

Percent improvement from baseline at a postbaseline visit is calculated as follows for all 7 components mentioned above:

$$\%\text{ improvement} = \frac{\text{baseline value - postbaseline value}}{\text{baseline value}} \times 100$$

If the baseline value is 0, then the percent improvement from baseline is set to missing.

In the case that some ACR20 components are missing other than TJC68 or SJC66, the ACR20 assessment will be based on the non-missing components. If non-missing components are not sufficient to determine ACR20 response, then the ACR20 response will be considered as missing.

6.2.2. Statistical Hypothesis for the Primary Efficacy Endpoint

In the primary analysis, the ACR20 response rate at Week 24 in the filgotinib 200 mg in combination with MTX group will be compared to MTX alone group for a superiority test at the 2-sided 0.05-level. If we denote the ACR20 response rate at Week 24 in the filgotinib 200 mg in combination with MTX group and MTX alone group as P_1 and P_2, respectively, the null and alternative hypotheses for the superiority test on the primary efficacy endpoint are as follows:

H_0: $P_1 = P_2$;

vs

H_1: $P_1 \neq P_2$.

6.2.3. Primary Analysis of the Primary Efficacy Endpoint

To test for superiority of filgotinib 200 mg in combination with MTX group versus MTX alone group in proportion of subjects who achieve ACR20 at Week 24, a logistic regression analysis with treatment groups and stratification factors in the model will be used. Subjects who do not have sufficient measurements to establish efficacy at Week 24 will be considered as failures (ie, NRI). The p-value from the logistic regression model for testing the superiority of filgotinib 200 mg in combination with MTX as compared to MTX alone will be provided. The 2-sided 95% confidence interval (CI) of the ACR20 response rate at Week 24 based on normal approximation for each treatment group will be provided for each treatment group. In addition, non-stratified ACR20 response rate difference between treatment groups with its 95% CI based on the normal approximation will be provided.

The comparisons of other filgotinib dose groups versus MTX alone group in proportion of subjects achieving ACR20 at Week 24 will also be conducted with the similar logistic regression model with NRI.

The number and percentage of ACR20 nonresponders observed and nonresponders due to missing at Week 24 will be summarized respectively by treatment groups.

For subjects with observed ACR20 outcomes (ie, responders and observed nonresponders) at Week 24, actual values and change from Baseline at Week 24 in individual components, including TJC68, SJC66, SGA, PGA, Subject's Pain Assessment, and hsCRP, will be summarized by using descriptive statistics (sample size, mean, SD, median, Q1, Q3, minimum, and maximum) by treatment groups.

The proportions of subjects achieving ACR20 (using NRI) will be plotted over time by treatment.

6.2.4. Sensitivity Analysis of the Primary Endpoint

Sensitivity analyses of the primary efficacy endpoint will be performed.

6.2.4.1. Per-Protocol Analysis

To evaluate the impact of study conduct on the primary analysis, the proportion of subjects who achieve ACR20 at Week 24 will be analyzed based on the PP Analysis Set as defined in Section 3.1.3 using the logistic regression analysis with NRI for the comparisons of filgotinib dose groups versus MTX alone group. The 2-sided 95% CI of the ACR20 response rate at Week 24 based on normal approximation for each treatment group will be provided.

6.2.4.2. Missing Data Imputation Analyses

To evaluate the impact of missing data on the ACR20 response rate at Week 24, the following missing value imputation methods will be used:

- The analysis specified in Section 6.2.3 will be performed by using OC and LOCF methods as described in Section 6.1.

- Multiple imputation (MI): The MI procedure replaces each missing binary ACR20 value with a set of plausible values that represent the uncertainty about the right value to impute. Multiple imputed datasets will be generated based on logistic regression models. These multiple imputed data sets are then analyzed by using the method for the primary analysis for complete data as specified in Section 6.2.3. The result of each set of imputed data will be combined using Rubin's rule {Rubin 1987} and provided. The stratification factors will be included in the imputation model as covariates, and all available data at post-baseline visits up to the time point of interest will be included in the longitudinal model.

- Tipping point analysis: To assess the robustness of analysis results under MNAR (missing not at random) assumption, a delta-adjusting pattern-mixture approach for tipping point analysis {Ratitch 2013} will be conducted for the primary efficacy endpoint. The impact from missing data on the comparisons between filgotinib dose groups versus MTX alone group in ACR20 response rate at Week 24 will be evaluated. Specifically, the proposed method will perform a series of analyses with a range of different values of the shift parameter δ applied to the imputed datasets at which the conclusion about the statisitical significance of the estimated treatment effect is altered. Each δ value is classified as either 'altering the study's conclusion' or tips, 'keeping the study's conclusion unchanged'. The tipping points that alter the statistical conclusion will be provided. For each δ value, multiple imputed datasets will be generated. The same analysis method for the primary analysis for complete data as specified in Section 6.2.3 will be applied when analyzing adjusted data generated under the different δ values.

6.2.5. Subgroup Analysis of the Primary Endpoint

Subgroup analysis comparing each filgotinib dose groups to the MTX alone group will be performed at Week 24 for the primary endpoint in the subgroups specified in Section 3.4.

The proportion of subjects who achieve ACR20 will be analyzed using the Fisher's exact test based on the NRI method for comparison between treatment groups. The number and percentage of subjects with ACR20 will also be provided for each treatment group within the subgroups.

6.3. Key Secondary Efficacy Endpoints

The key secondary efficacy endpoints are:

- Change from Baseline in the HAQ-DI score at Week 24

- Change from Baseline in the mTSS at Week 24

- Proportion of subjects who achieve DAS28(CRP) < 2.6 at Week 24

- Change from Baseline in SF-36 PCS at Week 24

- Change from Baseline in FACIT-Fatigue at Week 24

6.3.1. Definition of Key Secondary Efficacy Endpoints

6.3.1.1. Modified Total Sharp Score (mTSS)

Subject's radiographs of bilateral hands, wrists and feet will be taken at screening and protocol specified visits. The radiographs will be evaluated through central review by independent joint assessors using the mTSS method. The mTSS (range [0, 448]) is defined as the erosion score (range [0, 280]) plus the joint space narrowing (JSN) score (range [0, 168]). The change in erosion score, JSN score and mTSS will be computed for each patient at scheduled visits as the following:

$$\text{change in score} = \text{postbaseline score - baseline score}$$

For the missing joint score (regardless of reasons of missing), if at least half of the joints are assessed for hand joints erosion, foot joints erosion, and JSN, respectively, at a given visit, the total scores for the afore mentioned joints will be prorated using the following formula:

$$\text{Hand joints erosion score}=\frac{\text{total score among available hand joints}}{\text{sum of weights among the available hand joints}}\times 160$$

$$\text{Foot joints erosion score}=\frac{\text{total score among available foot joints}}{\text{sum of weights among the available foot joints}}\times 120$$

$$\text{JSN score}=\frac{\text{total score among available joints}}{\text{sum of weights among the available joints}}\times 168$$

The weight is defined as the possible maximum score of each joint.

If any of the hand joints erosion, foot joints erosion, or JSN has more than half of the joints not evaluable at a given visit, the mTSS will be treated as missing for that visit.

Detailed algorithm for calculating mTSS is described in Appendix 4.

6.3.1.2. Disease Activity Score for 28 Joint Count using CRP

The DAS28(CRP) score is calculated as follows:

$$\text{DAS28(CRP)} = 0.56\sqrt{\text{TJC28}} + 0.28\sqrt{\text{SJC28}} + 0.36\ln(\text{CRP} + 1) + 0.014 \times \text{SGA} + 0.96,$$

where
CRP = hsCRP measurement (mg/L)
SGA = subject's global assessment of disease activity on a 0-100mm VAS

Higher DAS28(CRP) value indicates more severe disease activity.

No component-level imputation will be performed for the calculation of DAS28(CRP) for the primary analyses. If any components are missing, the DAS28(CRP) will be set as missing.

6.3.1.3. 36-Item Short-form Health Survey

The SF-36 Version 2 is a 36-item, self-reported, generic, comprehensive, and health-related quality of life questionnaire that yields an 8 health domains (physical functioning, role-physical, bodily pain, general health, vitality, social functioning, role-emotional, and mental health). Each domain is scored by summing the individual items and transforming the scores into a 0 to 100 scale with higher scores indicting better health status or functioning. In addition, 2 summary scores, the physical component score (PCS) and the mental component score (MCS) will be evaluated based on the 8 SF-36 domains.

6.3.1.4. FACIT-Fatigue

The FACIT-Fatigue scale is a brief, 13-item, symptom-specific questionnaire that specifically assesses the self-reported severity of fatigue and its impact upon daily activities and functioning in the past 7 days. The FACIT-Fatigue uses 0 (“not at all”) to 4 (“very much”) numeric rating scales. Negatively stated items are reversed by subtracting the response from “4” before being added to obtain a total score.

6.3.2. Analysis Methods for Key Secondary Efficacy Endpoints

Hypothesis testing on the key secondary efficacy endpoints will commence after the testing on the primary efficacy endpoint reaches statistical significance, and will be performed according to the hierarchical testing principle at the 2-sided 0.05 level as described in Section 3.5. If a null hypothesis is not rejected, formal sequential testing will be stopped and only nominal significance will be reported for the remaining hypotheses.

For proportion of subjects who achieve DAS28(CRP) < 2.6 at Week 24, the similar logistic regression analysis with the same model specification as the primary endpoint analysis described in Section 6.2.3 will be adopted for the comparisons between treatment groups. The NRI method will be used to impute missing values. Comparisons will be made between each filgotinib dose group and the MTX alone group. The p-value from the logistic regression model will be reported for statistical inference. The 2-sided 95% CI of the response rate of DAS28(CRP) < 2.6 at Week 24 based on normal approximation will be provided for each treatment group. In addition, non-stratified response rate difference between treatment groups with its 95% CI based on the normal approximation will be provided.

The change from Baseline in HAQ-DI at Week 24 will be analyzed using the mixed-effects model for repeated measures (MMRM) that include all available data at postbaseline visits. Subjects that have a baseline value and at least one postbaseline value are included in the analysis. The MMRM model will be used to evaluate treatment effect on change score from Baseline, with baseline value, stratification factors, treatment, visit, and treatment by visit interaction included as fixed effects and subject being the random effect. An unstructured variance-covariance matrix will be used. The Kenward-Roger method will be used to estimate the degrees of freedom. Missing change scores due to missing study visits or early withdrawal will not be otherwise imputed using the MMRM approach. The least squares (LS) means along with the two-sided 95% CIs and the p-value for the difference in mean change from Baseline in HAQ-DI at Week 24 between each filgotinib dose group and MTX alone group from MMRM will be provided.

The change from Baseline in SF-36 PCS, FACIT-Fatigue, and mTSS at Week 24 will be analyzed using similar MMRM method as HAQ-DI, by including all available data at postbaseline visits. To test for a treatment difference between each filgotinib dose group and MTX alone group, LS mean difference along with 95% CI and p-values from the MMRM model will be presented.

The proportions of subjects achieving DAS28(CRP) < 2.6 (using NRI) will be plotted over time by treatment. Plots of mean ± SD of changes from Baseline in HAQ-DI, SF-36 PCS, FACIT-Fatigue scale score, and mTSS over time will be presented, respectively.

6.3.3. Sensitivity Analysis for Key Secondary Efficacy Endpoints

6.3.3.1. Per-Protocol Analysis

The analyses of the key secondary efficacy endpoints will be repeated using the PP Analysis Set specified in Section 3.1.3.

6.3.3.2. Missing Data Imputation Analyses

The following imputation methods will be explored:

- Change from Baseline in the HAQ-DI score at Week 24

  1) MI: All subjects with baseline measurement will be included. The MI procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Multiple imputed datasets will be generated based on linear regression models on observed HAQ-DI scores. These multiple imputed data sets are then analyzed by using the analysis method specified in Section 6.3.2 for complete data. The result of each set of imputed data will be combined using Rubin’s rule {Rubin 1987} and provided. Stratification factors and baseline values will be included in the imputation model as covariates, and all available data at baseline and postbaseline visits up to the time point of interest will be included in the longitudinal model.

  2) Tipping point analysis: All subjects with baseline measurement will be included. To access the robustness of analysis results for HAQ-DI change from Baseline at Week 24 under MNAR assumption, a set of shift parameter that adjust the imputed values for observation will be examined {Yuan 2014}. The shift parameter that alters the conclusion of the hypothesis testing for HAQ-DI change from Baseline at Week 24 under the MAR assumption will be reported as the tipping point and provided. The tipping point analysis will be conducted by iteratively assigning plausible outcomes to missing values for subjects in different treatment groups independently until the analyses conclusion is reversed. The same analysis method as specified in Section 6.3.2 will be applied when analyzing adjusted data generated under each plausible shift parameter.

- Change from Baseline in mTSS at Week 24

  1) MI: Similar MI procedure used for analyzing the change from Baseline in the HAQ-DI score at Week 24 will be performed for change from Baseline in the mTSS at Week 24.

  2) Tipping point analysis: Similar tipping point analysis method used for analyzing the change from Baseline in the HAQ-DI score at Week 24 will be performed.

- Proportion of subjects who achieve DAS28(CRP) < 2.6 at Week 24

  1) The analysis using the same logistic regression model specified in Section 6.3.2 will be performed by using OC and LOCF methods as described in Section 6.1.

  2) MI: All subjects with baseline measurement will be included. The MI procedure replaces each missing composite DAS28(CRP) value with a set of plausible values that represent the uncertainty about the right value to impute. Baseline value and stratification factors will be included in the imputation model as covariates, and all available data at baseline and all post-baseline visits through the last available visits will be included in the longitudinal model. The multiple imputed datasets will be generated. These multiple imputed data sets are then used to identify subjects with DAS28(CRP) ≤ 3.2 at Week 12, and analyzed by using the analysis method specified in Section 6.3.2 for complete data. The results from each set of imputed data sets will then be combined using Rubin's rule {Rubin 1987}.

  3) Tipping point analysis: Similar tipping point method as described in Section 6.2.4.2 will be applied to DAS28(CRP) < 2.6 response rate at Week 24.

- Change from Baseline in SF-36 PCS at Week 24

  1) MI: Similar MI procedure used for analyzing the change from Baseline in the HAQ-DI score at Week 24 will be performed.

  2) Tipping point analysis: Similar tipping point analysis method usedfor analyzing the change from Baseline in the HAQ-DI score at Week 24 will be performed.

- Change from Baseline in FACIT-Fatigue at Week 24

  1) MI: Similar MI procedure used for analyzing the change from Baseline in the HAQ-DI score at Week 24 will be performed.

  2) Tipping point analysis: Similar tipping point analysis method used for analyzing the change from Baseline in the HAQ-DI score at Week 24 will be performed.

6.3.4. Subgroup Analysis of Key Secondary Endpoints

Subgroup analysis comparing each filgotinib dose groups to the MTX alone group will be performed at Week 24 for the key secondary efficacy endpoint in the subgroups specified in Section 3.4.

The subgroup analysis for the proportion of subjects with DAS28(CRP) < 2.6 will be performed using the Fisher's exact test based on the NRI method. The number and percentage of subjects with DAS28(CRP) < 2.6 will also be provided for each treatment group within the subgroups

The change from Baseline in HAQ-DI will be analyzed using the MMRM method with baseline value, treatment, visit, and treatment by visit interaction included as fixed effects and subjects being the random effect. The LS mean, LS mean difference, SE, and 95% CI will be presented. The change from Baseline in mTSS, change from Baseline in SF-36 PCS and change from Baseline in FACIT-Fatigue will be analyzed similarly using the MMRM model, respectively. Descriptive statistics for actual values and change from baseline will also be presented for each treatment group within the subgroups.

6.4. Other Secondary Efficacy Endpoints

Other secondary efficacy endpoints include:

- Change from Baseline in the mTSS at Week 52
- The proportion of subjects who achieve ACR50 and ACR70 at Weeks 4, 12, 24, and 52, ACR20 at Weeks 4, 12, and 52, and ACR20/50/70 over time from Day 1 through Week 52
- Change from Baseline in individual components of the ACR response at Weeks 4, 12 and 24, and 52 and over time from Day 1 through Week 52
- The proportion of subjects who achieve change (ie, decrease) in HAQ-DI of ≥ 0.22 at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52
- Change from Baseline in DAS28(CRP) at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52
- The proportion of subjects who achieve DAS28(CRP) ≤ 3.2 at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52
- The proportion of subjects who achieve DAS28(CRP) < 2.6 at Weeks 4, 12, and 52, and over time from Day 1 through Week 52
- ACR-N and European League Against Rheumatism (EULAR) response at Weeks 4, 12, 24 and 52, and over time from Day 1 through Week 52
- Change from Baseline in CDAI at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52

- Change from Baseline in SDAI at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52
- The proportion of subjects who had no radiographic progression from Baseline (applying threshold of 0.5 on change from baseline in mTSS) at Week 24 and 52
- Absolute value and change from Baseline in SF-36, FACIT-Fatigue, and the EQ-5D over time at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52
- Absolute value and change from Baseline in WPAI-RA at Weeks 4, 12, 24, and 52, and over time from Day 1 through Week 52

6.4.1. Definition of Other Secondary Efficacy Endpoints

6.4.1.1. ACR50 and ACR70

ACR50 and ACR70 are similarly defined as ACR20 (see Section 6.2.1), except the improvement threshold from Baseline is 50% and 70%, respectively.

6.4.1.2. SDAI and CDAI

Simplifed Disease Acitivity Index (SDAI) is a composite measure that sums the TJC28, SJC28, the SGA on a 0-10 scale, the PGA on a 0-10 scale, and the CRP (in mg/dL). SDAI is scored as follows {Aletaha 2005}:

$$\text{SDAI} = \text{TJC28} + \text{SJC28} + \text{SGA} + \text{PGA} + \text{CRP}$$

Higher SDAI score indicates more severe disease activity status.

Clinical Disease Activity Index (CDAI) is a further simplification of the SDAI that excludes the CRP, which is calculated using the following formula {Aletaha 2005}:

$$\text{CDAI} = \text{TJC28} + \text{SJC28} + \text{SGA} + \text{PGA}$$

CDAI can range from 0 to 76, with higher score indicating more severe disease activity status.

No component-level imputation will be performed for the calculation of both SDAI and CDAI. If any components are missing, the SDAI and CDAI will be set as missing.

6.4.1.3. ACR-N

ACR-N is defined as the smallest percentage improvement from baseline in swollen joints, tender joints and the median of the following 5 items (PGA, SGA, subject’s pain assessment, HAQ-DI and CRP). It has a range between 0 and 100%. In particular, ACR-N = min { improvement in TJC68 (%), improvement in SJC66 (%), median [improvement in SGA (%), improvement in PGA (%), improvement in pain assessment (%), improvement in HAQ-DI (%), improvement in hsCRP (%)] }.

If this calculation results in a negative value, then the ACR-N is set to 0. If any components are missing, the ACR-N will be set as missing.

6.4.1.4. EULAR Response

Subject's response will be categorized according to the following table based on the DAS28(CRP).

Table 6-2. EULAR Response Criteria

	DAS28(CRP) Improvement from Baseline		
DAS28(CRP) at Visit	**> 1.2**	**> 0.6 and ≤ 1.2**	**≤ 0.6**
≤ 3.2	Good Response	Moderate Response	No Response
> 3.2 and ≤ 5.1	Moderate Response	Moderate Response	No Response
> 5.1	Moderate Response	No Response	No Response

6.4.1.5. European Quality of Life 5 Dimensions – 5 Levels (EQ-5D-5L)

The EQ-5D-5L is a standardized measure of health status of the subject at the visit (same day) that provides a simple, generic measure of health for clinical and economic appraisal. The EQ-5D-5L consists of 2 components: a descriptive system of the subject's health and a rating of his or her current health state using a 0 to 100 VAS. The descriptive system comprises the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The subject is asked to indicate his/her health state by ticking (or placing a cross) in the box associated with the most appropriate statement in each of the 5 dimensions. The VAS records the subject's self-rated health on a vertical VAS in which the endpoints are labeled "best imaginable health state" on the top and "worst imaginable health state" on the bottom. Higher EQ VAS indicates better health.The EQ-5D-5L will be scored according to the developer's instructions (scoring guidelines).

6.4.1.6. Work Productivity and Activity Impairment Questionnaire (WPAI) – RA

The WPAI is a questionnaire developed to measure impairments in work activities in subjects with RA. The questionnaire consists of 6 questions:

Q1: currently employed,

Q2: work time missed due to RA,

Q3: work time missed due to other reasons, Q4: hours actually worked,

Q5: degree RA affected productivity while working (0-10 VAS; with 0 indicating no effect and 10 indicating RA completely prevented the subject from working),

Q6: degree RA affected productivity in regular unpaid activities (0-10 VAS; with 0 indicating no effect and 10 indicating RA completely prevented the subject's daily activities).

The recall period for questions 2 to 6 is 7 days.

WPAI-RA outcomes are expressed as impairment percentages, with higher numbers indicating greater impairment and less productivity, that is, worse outcomes, as follows:

- Absenteeism (work time missed) due to RA: 100 × {Q2/(Q2+Q4)}

- Presenteeism (impairment while working) due to RA: 100 × {Q5/10}

- Work productivity loss (overall work impairment) due to RA: 100 × {Q2/(Q2+Q4) + [(1-Q2/(Q2+Q4))x(Q5/10)]}

- Activity impairment due to RA: 100 × {Q6/10}.

If Question 1 (Are you currently employed?) is 'NO', then only the activity impairment score can be determined.

6.4.1.7. Healthcare Resource Utilization Questionnaire

The Healthcare Resource Utilization Questionnaire (HRUQ) is designed to assess healthcare usage during the previous three months across a number of direct medical cost domains.

6.4.1.8. Treatment Satisfaction Questionnaire for Medication (TSQM)

TSQM Scale scores are computed by adding the items loading on each factor. The lowest possible score is subtracted from this composite score and divided by the greatest possible score minus the lowest possible score. This provides a transformed score between 0 and 1 that should be multiplied by 100. Note that only one item may be missing from each scale before the subscale should be considered invalid for that respondent. Detailed scoring method is available in the Quintiles TSQM Scoring Manual v1.4.

CCI [redacted]

[redacted]

6.4.2. Analysis Methods for Other Secondary Efficacy Endpoints

The FAS will be used for all summaries and analyses of other secondary efficacy endpoints. Nominal p-values will be presented, if applicable.

The proportion of subjects who achieve ACR20/50/70 and proportion of subjects with change in HAQ-DI ≥ 0.22 (ie, reduction in HAQ-DI ≥ 0.22), DAS28(CRP) ≤ 3.2, and DAS28(CRP) < 2.6 will be analyzed using the the same logistic regression method with NRI as the primary endpoint analysis described in Section 6.2.3. The 2-sided 95% CIs for the proportion based on normal approximation will be provided for each treatment group. The OC and LOCF values will be analyzed as a sensitivity approach to the primary method that uses NRI.

The proportion of subjects having no radiographic progression at Week 24 and 52 will be analyzed using the same logistic regression method based on the OC. Comparison will be made between each filgotinib dose group and MTX alone group. The 2-sided 95% CIs for the proportion based on normal approximation will be provided for each treatment group. In addition, similar analysis will be prerformed with thresholds of 0 and the smallest detectable change (SDC). The change from Baseline in mTSS, DAS28(CRP), CDAI and SDAI will be analyzed using MMRM method that include all available data at postbaseline visits through Week 52. The MMRM models will include baseline value, stratification factors, treatment, visit, and treatment by visit interaction included as fixed effects and subject being the random effect. An unstructured variance-covariance matrix will be used. The Kenward-Roger method will be used to estimate the degrees of freedom. Missing change scores due to missing study visits or early withdrawal will not be otherwise imputed using the MMRM approach. The LS means along with 95% CIs of the difference in mean change scores from Baseline between each filgotinib dose group and MTX alone group from the MMRM model will be provided for each postbaseline visit. Descriptive statistics of actual and change in DAS28(CRP), CDAI, SDAI, and mTSS from baseline to postbaseline visits will also be performed by treatment and visit.

Actual values and change from Baseline in individual components of the ACR response (including TJC68, SJC66, SGA, PGA, Subject's pain assessment, and hsCRP), TJC 28, SJC28, and ACR-N will be summarized using descriptive statistics (sample size, mean, SD, median, Q1, Q3, minimum, and maximum) by treatment and visit. In addition, the MMRM model with baseline value, stratification factors, treatment, visit, and treatment and visit interaction as fixed effects and subject being the random effect will be performed to compare each filgotinib dose group and MTX alone group. The LS means and 95% CI of the difference between each filgotinib dose group and MTX alone group will be provided.

Actual values and change from Baseline in ACR-N will be summarized using descriptive statistics (sample size, mean, SD, median, Q1, Q3, minimum, and maximum) by treatment and visit.

Number and percentage of subjects for EULAR responses will be presented for each visit through Week 52 by treatment.

For HRQoL endpoints, the analysis methods are detailed below:

- **SF-36**

  The transformed scores and change from baseline in the 8 domains and 2 summary component scores (PCS and MCS) of the SF-36 will be summarized at each visits through Week 52 by treatment group using descriptive statistics (sample size, mean, SD, median, Q1, Q3, minimum and maximum).

  The change from Baseline in PCS and MCS at each postbaseline visit through Week 52 will be analyzed using MMRM method that includes all available data at postbaseline visits, without imputation for missing data. The MMRM models will be used to evaluate treatment effect on change score from Baseline, with baseline value, stratification factors, treatment, visit, and treatment by visit interaction included as fixed effects and subject being the random effect. An unstructured variance-covariance matrix will be used. The Kenward-Roger method will be used to estimate the degrees of freedom. To test for a treatment difference between each filgotinib dose group and MTX alone group, the LS means along with the 95% CIs of the difference and p-value from MMRM will be provided for each postbaseline visit.

- **FACIT-Fatigue**

  The change from baseline in FACIT-Fatigue scale score at each postbaseline visit through Week 52 will be analyzed using similar MMRM models as for SF-36. To test for a treatment difference between each filgotinib dose group and placebo, LS mean difference along with 95% CI and p-values will be presented for each postbaseline visit. Missing change scores due to missing study visits or early withdrawal will not be otherwise imputed using the MMRM approach. Summary statistics of the actual score and change from baseline in FACIT-Fatigue scale score will also be displayed by treatment and visit.

- **EQ-5D-5L**

  Summary statistics of two outcomes of the EQ-5D-5L described as the following will be provided:

  — A health profile: the number and percentage of subjects at each categorical response for the 5 dimensions (mobility, self-care, usual activity, pain/discomfort, and anxiety/depression) will be provided by treatment and visit.

  — A self-perceived current health score: calculated from patient level EQ VAS responses (continuous variable). Descriptive statistics (sample size, mean, SD, min, median, and max) will be provided by treatment and visit.

  The change from baseline in EQ VAS score through Week 52 will be analyzed using similar MMRM models as for SF-36. To test for a treatment difference between each filgotinib dose group and placebo, LS mean difference, 95% CI and p-values will be presented. Missing change scores due to missing study visits or early withdrawal will not be otherwise imputed using the MMRM approach.

- **WPAI – RA**

  Summary statistics of the actual scores and change from baseline in each of the four types of scores of WPAI-RA (absenteeism, presenteeism, work productivity loss, and activity impairment) will be summarized by treatment and visit using descriptive statistics (sample size, mean, SD, median, Q1, Q3, minimum and maximum).

  The analyses of absenteeism, presenteeism, and work productivity loss will be based on subjects employed at both baseline and postbaseline visits.

- **HRUQ**

  Summary statistics will be provided for all the following parameters by treatment group and visit.

  Healthcare resource utilization – Outpatient visits:

  — Number and percentage of subjects with any medical visits, and the number of visits related to RA will be provided for the categories:

  - ■ Outpatient Healthcare Provider
  - ■ Emergency Room
  - ■ Chiropractor, Physical or Occupational Therapist

  Healthcare resource utilization – Inpatient hospitalization:

  — Number and percentage of subject with hospital stays, and number of stays related to RA

  — Descriptive statistics for the total number of days in hospital related to RA

  — Number and percentage of subject with nursing home or rehabilitation center stays, and number of stays related to RA

  — Descriptive statistics for the total number of days in nursing home or rehabilitation center related to RA

- **TSQM**

  Descriptive statistics of the scores from each of the 4 domains (Effectiveness, Side-Effects, Convenience, and Global Satisfaction) will be presented by treatment group and by visit.

CCI [redacted] [redacted]

[redacted]

Plots of proportions of subjects for categorical endpoints and mean ± SD for continuous endpoints (including each individual component of the ACR response) will be presented over time by treatment.

In addition, all available data including those collected under the standard of care will be summarized by treatment group for the proportion of subjects who achieve ACR20, DAS28(CRP) ≤ 3.2, and DAS28(CRP) < 2.6, and the actual value and change from Baseline in HAQ-DI, mTSS, SF-36 PCS score, and FACIT-Fatigue. The observed values will be analyzed, and no formal statistical testing will be performed.

6.5. Changes From Protocol-Specified Efficacy Analyses

The following changes were made from the protocol-specified efficacy analyses:

- It has been observed from the blinded data that the enrollment rate is low in some combinations of stratification factors , which will result in a small cell counts for the Cochran- Mantel-Haenszel test adjusted for the stratification factors. The logistic regression analysis with treatment and stratification factors in the model will be used instead to analyze binary endpoints.

- The following two endpoints have been updated as key secondary efficacy endpoints in Section 6.3 and included in the hierarchy testing in Section 3.5, in order to support inclusion in the label.

  — Change from Baseline in SF-36 PCS at Week 24

  — Change from Baseline in FACIT-Fatigue at Week 24

- Given the importance of determining the lowest effective dose for filgotinib in RA, the hierarchical ordering has been updated in Section 3.5 to test filgotinib 100 mg in combination with MTX immediately after each test for filgotinib 200 mg in combination with MTX (including the primary endpoint and each of the key secondary endpoints). In addition, the testing for ACR20, HAQ-DI, and mTSS is moved up earlier in the list to accommodate the changes aforementioned as well as because of the importance of these endpoints in determining drug efficacy in RA. These changes together caused a relative move down for testing of DAS28(CRP) < 2.6.

7. SAFETY ANALYSES

7.1. Adverse Events and Deaths

7.1.1. Adverse Event Dictionary

Clinical and laboratory adverse events (AEs) will be coded using the MedDRA 21.0. System organ class (SOC), high-level group term (HLGT), high-level term (HLT), preferred term (PT), and lower-level term (LLT) will be provided in the AE dataset.

7.1.2. Adverse Event Severity

Adverse events are graded by the investigator as Grade 1, 2, 3, 4, or 5 according to toxicity criteria specified in the protocol. The severity grade of events for which the investigator did not record severity will be categorized as “missing” for tabular summaries and data listings. The missing category will be listed last in summary presentation.

7.1.3. Relationship of Adverse Events to Study Drug

Related AEs are those for which the investigator selected “Related” on the AE CRF to the question of “Related to Study Treatment.” Relatedness will always default to the investigator’s choice, not that of the Medical Monitor. Events for which the investigator did not record relationship to study drug will be considered related to study drug for summary purposes. However, by-subject data listings will show the relationship as missing.

7.1.4. Serious Adverse Events

Serious adverse events (SAEs) will be identified and captured as SAEs if AEs met the definitions of SAE that were specified in the study protocol. SAEs captured and stored in the clinical database will be reconciled with the SAE database from the Gilead Pharmacovigilance & Epidemiology Department before data finalization.

7.1.5. Treatment-Emergent Adverse Events

7.1.5.1. Definition of Treatment-Emergent Adverse Events

Treatment-emergent adverse events (TEAEs) are defined as one or both of the following:

- Any AEs with an onset date on or after the study drug start date and no later than 30 days after permanent discontinuation of study drug

- Any AEs leading to premature discontinuation of study drug

7.1.5.2. Incomplete Dates

If the onset date of the AE is incomplete and the AE stop date is not prior to the first dosing date of study drug, then the month and year (or year alone if month is not recorded) of onset determine whether an AE is treatment emergent. The event is considered treatment emergent if both of the following 2 criteria are met:

- The AE onset is the same as or after the month and year (or year) of the first dosing date of study drug, and

- The AE onset date is the same as or before the month and year (or year) of the date corresponding to 30 days after the date of the last dose of study drug

An AE with completely missing onset and stop dates, or with the onset date missing and a stop date later than the first dosing date of study drug, will be considered to be treatment emergent. In addition, an AE with the onset date missing and incomplete stop date with the same or later month and year (or year alone if month is not recorded) as the first dosing date of study drug will be considered treatment emergent.

7.1.6. Summaries of Adverse Events and Deaths

Treatment-emergent AEs will be summarized based on the Safety Analysis Set.

7.1.6.1. Summaries of AE Incidence in Combined Severity Grade Subsets

The number and percentage of subjects who experienced at least 1 TEAE will be provided and summarized by SOC, HLT, PT, and treatment group. For other AEs described below, summaries will be provided by SOC, PT, and treatment group:

- TEAEs of Grade 3 or higher (by maximum severity)

- TEAEs of Grade 2 or higher

- All TE treatment-related AEs

- TE Treatment-related AEs of Grade 3 or higher (by maximum severity)

- TE Treatment-related AEs of Grade 2 or higher

- All TE SAEs

- All TE treatment-related SAEs

- All TEAEs leading to premature discontinuation of any study drug

- All TEAEs leading to premature discontinuation of study

- All TE SAEs leading to death (ie, outcome of death)
- All TEAEs leading to temporary interruption of any study drug

A brief, high-level summary of AEs described above will be provided by treatment group and by the number and percentage of subjects who experienced the above AEs. All deaths observed in the study will be also included in this summary.

Multiple events will be counted only once per subject in each summary. Adverse events will be summarized and listed first in alphabetic order of SOC and HLT within each SOC (if applicable), and then by PT in descending order of total frequency within each SOC. For summaries by severity grade, the most severe grade will be used for those AEs that occurred more than once in an individual subject during the study.

In addition to the above summary tables, all TEAEs and TE treatment-related AEs will be summarized by PT only, in descending order of total frequency.

In addition, data listings will be provided for the following:

- All AEs, indicating whether the event is treatment emergent
- All AEs of Grade 3 or higher
- All AEs of Grade 2 or higher
- SAEs
- Deaths
- All SAEs leading to death (ie, outcome of death)
- AEs leading to premature discontinuation of any study drug
- AEs leading to premature discontinuation of study
- AEs leading to temporary interruption of any study drug

7.1.7. Adverse Events of Special Interest

Events of interest will be identified by the use of either SMQs or MSTs. However, should additional cases not detected by the predefined search term listings be identified during the clinical review process, these cases will be reported by the respective category.

7.1.7.1. Adjudication Committee for MACE

An independent cardiovascular safety endpoint adjudication committee (CVEAC) will be formed to periodically review and adjudicate all potential major cardiovascular adverse events (MACE). MACE isdefined as cardiovascular death, nonfatal myocardial infarction and nonfatal stroke.

To identify the MACE, the following potential cases identified using MedDRA search listing (MST) searches will be adjudicationed:

- All deaths
- Cardiovascular events (meeting serious criteria)
- Myocardial infarction
- Hospitalization for unstable angina
- Transient ischemic attack
- Stroke
- Hospitalization for cardiac failure
- Percutaneous coronary intervention

The CVEAC will review those potential MACE and related clinical data to determine whether a MACE has developed. The CVEAC's role and responsibilities and the data to be provided to the CVEAC are described in a mutually agreed upon CVEAC charter. The CVEAC charter defines the CVEAC membership, adjudication process, meeting logistics, and meeting frequency. The number and percentage of subjects with positively adjudicated MACE will be summarized by treatment group using the PT term.

A by-subject listing for all patients who have a potential MACE or a positively adjudicated MACE at any time will be provided.

7.1.7.2. Other Adverse Events of Special Interest

In addition to general safety parameters and MACE, safety information on other adverse events of special interest (AESIs) will also be analyzed. AESIs will be identified by either laboratory results, standardized MedDRA queries (SMQs), or sponsor defined MSTs, or a combination of these methods as indicated below.

- All infections (defined as all PTs in the Infections and Infestations SOC)
- Serious infections (defined as all PTs in the Infections and Infestations SOC that are SAEs)
- Infections of special interest as defined below (identified by using MST)
  a) Herpes zoster
  b) Active tuberculosis
  c) Opportunistic infections
  d) Hepatitis B or C infections

- Deep vein thrombosis (DVT) and pulmonary embolism (PE) (identified by using MST)
- Malignancy (including lymphoma; not including nonmelanoma skin cancer) (identified by using MST)
- Nonmelanoma skin cancer (identified by MST)
- GI (Gastrointestinal) perforations (identified by MST)

The number and percentage of subjects with aforementioned events of special interest will be provided by the PT term for each AE of special interests.

A by-subject listing for all subjects having AE of special interests at any time will be provided for each AE of special interest.

7.2. Laboratory Evaluations

Laboratory data collected during the study will be analyzed and summarized using both quantitative and qualitative methods. Summaries of laboratory data will be provided for the Safety Analysis Set and will include data collected up to the last dose of any study drug plus 30 days for subjects who have permanently discontinued study drug, or all available data at the time of the database snapshot for subjects who were ongoing at the time of an interim analysis. The analysis will be based on values reported in conventional units. When values are below the LOQ, they will be listed as such, and the closest imputed value will be used for the purpose of calculating summary statistics as specified in Section 3.7.

A by-subject listing for laboratory test results will be provided by subject ID number and visit in chronological order for hematology, serum chemistry, and urinalysis separately. Values falling outside of the relevant reference range and/or having a severity grade of 1 or higher on the CTCAE severity grade will be flagged in the data listings, as appropriate.

No formal statistical testing is planned.

7.2.1. Summaries of Numeric Laboratory Results

Descriptive statistics of Baseline values, values at each postbaseline visit and change from Baseline at each postbaseline visit will be provided by treatment group for the following laboratory tests:

- Hematology
  - Hematocrit
  - Hemoglobin
  - Platelet count
  - Red blood cell count

- Mean corpuscular volume
- White blood cell (WBC) count
- Lymphocytes
- Monocytes
- Neutrophils
- Eosinophils
- Basophils

- Chemistry
  - Alanine aminotransferase (ALT)
  - Aspartate aminotransferase (AST)
  - Alkaline phosphatase (ALP)
  - Total bilirubin
  - Serum creatinine
  - Creatinine clearance by Cockcroft-Gault formula
  - Creatinine phosphokinase (CPK)
  - Glucose
- Lipid
  - Triglycerides
  - Total cholesterol
  - HDL
  - LDL
  - LDL/HDL ratio

A baseline laboratory value will be defined as the last measurement obtained on or prior to the date/time of first dose of any study drug. Change from baseline to a postbaseline visit will be defined as the visit value minus the baseline value. The mean, median, Q1, Q3, minimum, and maximum values will be displayed to the reported number of digits; SD values will be displayed to the reported number of digits plus 1.

Median (Q1, Q3) of the observed values for the laboratory tests specified above will be plotted using a line plot by treatment group and visit.

In the case of multiple values in an analysis window, data will be selected for analysis as described in Section 3.8.3.

7.2.2. Graded Laboratory Value

The CTCAE Version 4.03 (with one modification specified in protocol) will be used to assign toxicity grades (0 to 4) to laboratory results for analysis. Grade 0 includes all values that do not meet the criteria for an abnormality of at least Grade 1. For laboratory tests with criteria for both increased and decreased levels, analyses for each direction (ie, increased, decreased) will be presented separately.

7.2.2.1. Treatment-Emergent Laboratory Abnormalities

Treatment-emergent laboratory abnormalities are defined as values that increase at least 1 toxicity grade from baseline at any postbaseline time point, up to and including the date of last dose of any study drug plus 30 days for subjects who permanently discontinued study drug, or the last available date in the database snapshot for subjects who were still on treatment at the time of an interim analysis.

If the relevant baseline laboratory value is missing, any abnormality of at least Grade 1 observed within the time frame specified above will be considered treatment emergent.

7.2.2.2. Treatment-Emergent Marked Laboratory Abnormalities

Treatment-emergent marked laboratory abnormalities are defined as values that increase from baseline by at least 3 toxicity grades at any postbaseline time point, up to and including the date of the last dose of any study drug plus 30 days for subjects who permanently discontinued study drug or the last available date in the database snapshot for subjects who were still on treatment at the time of an interim analysis.

If the relevant baseline laboratory value is missing, any Grade 3 or higher values observed within the timeframe specified above will be considered treatment-emergent marked abnormalities.

7.2.2.3. Summaries of Laboratory Abnormalities

Laboratory data that are categorical will be summarized using the number and percentage of subjects in the study with the given response at baseline and each scheduled postbaseline visit.

The following summaries (number and percentage of subjects) for treatment-emergent laboratory abnormalities will be provided by laboratory test and treatment group; subjects will be categorized according to the most severe postbaseline abnormality grade for a given laboratory test:

- Graded laboratory abnormalities

- Grade 3 or higher laboratory abnormalities
- Marked laboratory abnormalities

For all summaries of laboratory abnormalities, the denominator is the number of subjects with nonmissing postbaseline values up to 30 days after last dosing date.

A by-subject listing of treatment-emergent Grade 3 or higher laboratory abnormalities and marked laboratory abnormalities will be provided separately by subject ID number and visit in chronological order. These listings will include all test results that were collected throughout the study for the laboratory test of interest, with all applicable severity grades or abnormal flags displayed.

7.2.3. Laboratory Evaluations of Special Interest

7.2.3.1. Liver-Related Laboratory Evaluations

Liver-related abnormalities after initial study drug dosing will be examined and summarized using the number and percentage of subjects who were reported to have the following laboratory test values for postbaseline measurements:

- AST: (a) > 3 times of the upper limit of reference range (ULN); (b) > 5 x ULN; (c) > 10 x ULN; (d) > 20 x ULN
- ALT: (a) > 3 x ULN; (b) > 5 x ULN; (c) > 10 x ULN; (d) > 20 x ULN
- AST or ALT > 3 x ULN and total bilirubin > 2 x ULN

The summary will include data from all postbaseline visits up to 30 days after the last dose of any study drug. For individual laboratory tests, subjects will be counted once based on the most severe postbaseline values. For both the composite endpoints of AST or ALT and total bilirubin, subjects will be counted once when the criteria are met at the same postbaseline visit date. The denominator is the number of subjects in the Safety Analysis Set who have nonmissing postbaseline values of all relevant tests at the same postbaseline visit date. A listing of subjects who met at least 1 of the above criteria will be provided.

7.2.3.2. Complete Blood Count-Related Laboratory Evaluations

Complete blood count (CBC)-related abnormalities such as anemia, leucopenia, neutropenia, lymphopenia, and thrombocytopenia after initial study drug dosing will be examined and summarized using the number and percentage of subjects who were reported to have the following laboratory test values for postbaseline measurements:

- Hemoglobin: (a) any postbaseline worsening CTCAE grade from baseline; (b) baseline value of less than Grade 3 and increase to Grade 3 or higher at worst postbaseline; (c) baseline value of less than Grade 3 and increase to Grade 4 or higher at worst postbaseline

- WBC count: (a) any postbaseline worsening CTCAE grade from baseline; (b) baseline value of less than Grade 3 and increase to Grade 3 or higher at worst postbaseline; (c) baseline value of less than Grade 3 and increase to Grade 4 or higher at worst postbaseline

- Absolute neutrophil count: (a) any postbaseline worsening CTCAE grade from baseline; (b) baseline value of less than Grade 3 and increase to Grade 3 or higher at worst postbaseline; (c) baseline value of less than Grade 3 and increase to Grade 4 or higher at worst postbaseline

- Lymphocyte count: (a) any postbaseline worsening CTCAE grade from baseline; (b) baseline value of less than Grade 3 and increase to Grade 3 or higher at worst postbaseline; (c) baseline value of less than Grade 3 and increase to Grade 4 or higher at worst postbaseline

- Platelet count: (a) any postbaseline worsening CTCAE grade from baseline; (b) baseline value of less than Grade 3 and increase to Grade 3 or higher at worst postbaseline; (c) baseline value of less than Grade 3 and increase to Grade 4 or higher at worst postbaseline

The summary will include data from all postbaseline visits up to 30 days after the last dose of any study drug.

7.3. Body Weight and Vital Signs

Descriptive statistics will be provided by treatment group for body weight, BMI and vital signs (systolic and diastolic blood pressures [mmHg], pulse [beats/min]) as follows:

- Baseline value

- Values at each postbaseline visit

- Change from baseline at each postbaseline visit

A baseline value will be defined as the last available value collected on or prior to the date/time of first dose of any study drug. Change from baseline to a postbaseline visit will be defined as the postbaseline value minus the baseline value. Body weight and vital signs measured at unscheduled visits will be included for the baseline value selection.

In the case of multiple values in an analysis window, data will be selected for analysis as described in Section 3.8.3. No formal statistical testing is planned.

A by-subject listing of vital signs (systolic and diastolic blood pressure [mmHg], pulse [beats/min], respiration [breaths/min], and body temperature [°C]) will be provided by subject ID number and visit in chronological order. In the same manner, a by-subject listing of body weight, height, and BMI will be provided separately

7.4. Prior and Concomitant Medications

Medications collected at screening and during the study will be coded using the World Health Organization (WHO) Drug dictionary version BSEP17

All the analyses in this section will be performed for general prior /concomitant medications and RA-specific prior/concomitant medications separately, unless otherwise specified.

7.4.1. Prior Medications

Prior medications are defined as any medication taken before a subject took the first study drug.

Prior medications will be summarized by Anatomical Therapeutic Chemical (ATC) drug class preferred name using the number and percentage of subjects for each treatment group and overall. A subject reporting the same medication more than once will be counted only once when calculating the number and percentage of subjects who received that medication. The summary will be provided by preferred name in order of descending overall frequency. For drugs with the same frequency, sorting will be done alphabetically.

For the purposes of analysis, any medication with a start date prior to the first dosing date of any study drug will be included in the prior medication summary regardless of when the stop date is. If a partial start date is entered the medication will be considered prior unless the month and year (if day is missing) or year (if day and month are missing) of the start date are after the first dosing date. Medications with a completely missing start date will be included in the prior medication summary, unless otherwise specified.

Summaries will be based on the Safety Analysis Set. No formal statistical testing is planned.

7.4.2. Concomitant Medications

Concomitant medications are defined as medications taken while a subject took study drug. Use of concomitant medications will be summarized by ATC drug class preferred name using the number and percentage of subjects for each treatment group. A subject reporting the same medication more than once will be counted only once when calculating the number and percentage of subjects who received that medication. The summary will be provided by preferred term in descending overall frequency within each ATC medical class. For drugs with the same frequency, sorting will be done alphabetically.

For the purposes of analysis, any medication with a start date prior to or on the first dosing date of any study drug and continued to take after the first dosing date, or started after the first dosing date but prior to or on the last dosing date of study drug will be considered concomitant medications. Medications started and stopped on the same day as the first dosing date or the last dosing date of any study drug will also be considered concomitant. Medications stopped on the same day as the first dosing date will be considered concomitant. Medications with a stop date prior to the date of first dosing date of any study drug or a start date after the last dosing date of any study drug will be excluded from the concomitant medication summary. If a partial stop date is entered, any medication with the month and year (if day is missing) or year (if day and month are missing) prior to the date of first study drug administration will be excluded from the concomitant medication summary. If a partial start date is entered, any medication with the month and year (if day is missing) or year (if day and month are missing) after the study drug stop date will be excluded from the concomitant medication summary. Medications with completely missing start and stop dates will be included in the concomitant medication summary, unless otherwise specified.

Summaries will be based on the Safety Analysis Set. No formal statistical testing is planned.

All prior and concomitant medications (other than per-protocol study drugs) will be provided in a by-subject listing sorted by subject ID number and administration date in chronological order.

7.5. Electrocardiogram Results

7.5.1. Investigator Electrocardiogram Assessment

A shift table of the investigators' assessment of ECG results at each visit compared with baseline values will be presented by treatment group using the following categories: normal; abnormal (not clinically significant); abnormal (clinically significant); or missing. The number and percentage of subjects in each cross-classification group of the shift table will be presented. Subjects with a missing value at baseline or postbaseline will not be included in the denominator for percentage calculation.

No formal statistical testing is planned.

A by-subject listing for ECG assessment results will be provided by subject ID number and visit in chronological order.

7.6. Other Safety Measures

A data listing will be provided for subjects who become pregnant during the study.

7.7. Changes From Protocol-Specified Safety Analyses

There are no deviations from the protocol-specified safety analyses.

8. PHARMACOKINETIC ANALYSES

Concentrations of filgotinib and GS-829845 in plasma will be determined using validated bioanalytical assays. Plasma concentrations of filgotinib and the active metabolite of filgotinib, GS-829845, will be listed and summarized using descriptive statistics (eg, sample size, arithmetic mean, geometric mean, % coefficient of variation, standard deviation, median, minimum and maximum).

8.1. PK Analyses Related to Intensive PK Sampling

CCI

8.1.1. Estimation of Pharmacokinetic Parameters

PK parameters will be estimated using Phoenix WinNonlin® software using standard noncompartmental methods. The linear/log trapezoidal rule will be used in conjunction with the appropriate noncompartmental model, with input values for dose level, dosing time, plasma concentration, and corresponding real-time values, based on drug dosing times whenever possible.

All predose sample times before time-zero will be converted to 0. Predose samples may also serve as the 24-hr post dose sample if appropriate.

For area under the curve (AUC), samples BLQ of the bioanalytical assays occurring prior to the achievement of the first quantifiable concentration will be assigned a concentration value of 0 to prevent overestimation of the initial AUC. Samples that are BLQ at all other time points will be treated as missing data in WinNonlin. The nominal time point for a key event or dosing interval (τ) may be used to permit direct calculation of AUC over specific time intervals. The appropriateness of this approach will be assessed by the PK scientist on a profile-by-profile basis.

Pharmacokinetic parameters such as AUC_{tau}, λ_z and $t_{1/2}$ are dependent on an accurate estimation of the terminal elimination phase of drug. The appropriateness of calculating these parameters will be evaluated upon inspection of PK data on a profile-by-profile basis by the PK scientist.

8.1.2. Pharmacokinetic Parameters

CCI

CCI

The analytes and parameters presented in Table 8-2 will be used to evaluate the PK objectives of the study. The PK parameters to be estimated in this study are listed and defined in the Pharmacokinetic Abbreviations section.

Table 8-2. Pharmacokinetic Parameters for Each Analyte

Analyte	Parameters
Filgotinib	AUC_{last}, AUC_{tau}, C_{tau}, C_{max}, T_{max}, C_{last}, T_{last}, λ_z, CL_{ss}/F, V_z/F, and $t_{1/2}$, if appropriate
GS-829845	AUC_{last}, AUC_{tau}, C_{tau}, C_{max}, T_{max}, C_{last}, T_{last}, λ_z, and $t_{1/2}$, if appropriate

8.1.3. Statistical Analysis Methods

Individual subject concentration data and individual subject PK parameters for filgotinib and GS-829845 will be listed and summarized using descriptive statistics by treatment. Summary statistics (number of subjects [n], mean, SD, coefficient of variation [%CV], median, minimum, maximum, Q1, and Q3) will be presented for both individual subject concentration data by time point and treatment; and individual subject PK parameters by treatment. Moreover, the geometric mean, 95% CI, and the mean and SD of the natural log-transformed values will be presented for individual subject PK parameter data.

Individual concentration data listings and summaries will include all subjects with concentration data. The sample size for each time point will be based on the number of subjects with nonmissing concentration data at that time point. The number of subjects with concentration BLQ will be presented for each time point. For summary statistics, BLQ values will be treated as 0 at predose and one-half of the lower limits of quantitation (LLOQ) for postdose time points.

Individual PK parameter data listings and summaries will include all subjects for whom PK parameter(s) can be derived. The sample size for each PK parameter will be based on the number of subjects with nonmissing data for that PK parameter.

The following tables may be provided for each analyte by treatment:

- Individual subject concentration data and summary statistics
- Individual subject plasma PK parameters and summary statistics.

CCI

[redacted]

Individual, mean, and median postdose concentration values that are ≤ LLOQ will not be displayed in the figures and remaining points connected.

The following listings will be provided:

- PK sampling details (and PK concentrations) by subject, including procedures, differences in scheduled and actual draw times, and sample age
- Individual data on determination of plasma half-life and corresponding regression correlation coefficient.

9. BIOMARKER ANALYSIS

A separate biomarker analysis plan (BAP) will document methods to analyze biomarker assessments.

10. REFERENCES

Aletaha D, Nell VP, Stamm T, Uffmann M, Pflugbeil S, Machold K, et al. Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score. Arthritis research & therapy 2005;7 (4):R796-806.

Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. Pharm Stat 2013;12 (6):337-47.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley & Sons, Inc, 1987:

Yuan Y. Sensitivity Analysis in Multiple Imputation for Missing Data. 2014:

11. SOFTWARE

SAS® Software Version 9.X. SAS Institute Inc., Cary, NC, USA.

nQuery Advisor(R) Version X.0. Statistical Solutions, Cork, Ireland.

12. SAP REVISION

Revision Date (DD MMM YYYY)	Section	Summary of Revision	Reason for Revision

13. APPENDICES

Appendix 1. Lists of RA Medications

1) List of Oral Corticosteroid (WHO Preferred Terms)

- BETAMETHASONE
- BETAMETHASONE DIPROPIONATE
- BETAMETHASONE SODIUM PHOSPHATE
- CORTISONE
- DEXAMETHASONE
- DEXAMETHASONE PALMITATE
- DEXAMETHASONE SODIUM PHOSPHATE
- HYDROCORTISONE
- MEPREDNISONE
- METHYLPREDNISOLONE
- METHYLPREDNISOLONE ACETATE
- METHYLPREDNISOLONE SODIUM SUCCINATE
- MOMETASONE FUROATE
- PREDNISOLONE
- PREDNISOLONE FARNESYLATE
- PREDNISOLONE SODIUM SUCCINATE
- PREDNISONE
- TRIAMCINOLONE

2) List of DMARDs and investigational bDMARDs (WHO Preferred Terms)

- ABATACEPT
- ADALIMUMAB
- ANAKINRA

- BAMINERCEPT
- BELIMUMAB
- BIMEKIZUMAB
- CABIRALIZUMAB
- CERTOLIZUMAB
- CERTOLIZUMAB PEGOL
- CHLOROQUINE
- CHLOROQUINE PHOSPHATE
- CHLOROQUINE SULFATE
- CICLOSPORIN
- CLAZAKIZUMAB
- CORTISONE
- DENOSUMAB
- DEXAMETHASONE
- DEXAMETHASONE PALMITATE
- DEXAMETHASONE PHOSPHATE
- DEXAMETHASONE SODIUM PHOSPHATE
- EPRATUZUMAB
- ETANERCEPT
- GOLIMUMAB
- HYDROXYCHLOROQUINE
- HYDROXYCHLOROQUINE SULFATE
- INFLIXIMAB

- INTERLEUKIN INHIBITORS
- INTERLEUKIN-2
- LEFLUNOMIDE
- METHOTREXATE
- METHOTREXATE SODIUM
- OCRELIZUMAB
- PENICILLAMINE
- REMTOLUMAB
- RITUXIMAB
- SARILUMAB
- SECUKINUMAB
- SIRUKUMAB
- SULFASALAZINE
- TABALUMAB
- TILDRAKIZUMAB
- TOCILIZUMAB
- TREGALIZUMAB

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

The HAQ-DI score is defined as the average of the scores of eight functional categories (dressing and grooming, arising, eating, walking, hygiene, reach, grip, and other activities), usually administered by the patient. Responses in each functional category are collected as 0 (without any difficulty) to 3 (unable to do a task in that area), with or without aids or devices.

The highest score for questions in each category (range 0 to 3) determines the score for the category, unless aids or devices are required. Dependence on equipment or physical assistance increases a lower score (ie, scores of 0 or 1) to the level of 2 to more accurately represent underlying disability. The eight category scores are averaged into an overall HAQ-DI score on a scale from 0 (no disability) to 3 (completely disabled) when 6 or more categories are non-missing. If more than 2 categories are missing, the HAQ-DI score is set to missing. The HAQ-DI can be treated as a continuous measure.

The HAQ-DI score using aids (and/or devices) is computed by taking the maximum score of the questions in each category (range: [0, 3]) and whether or not aids/devices are used (0 or 1):

A = max(dressing & grooming category questions, 2*aids indicator) +
max(rising category questions, 2*aids indicator) +
max(eating category questions, 2*aids indicator) +
max(walking category questions, 2*aids indicator) +
max(hygiene category questions, 2*aids indicator) +
max(reach category questions, 2*aids indicator) +
max(grip category questions, 2*aids indicator) +
max(usual activities category questions, 2*aids indicator)

HAQ-DI = A/(total number of categories with at least 6 non-missing)

The following table shows the contribution of the 43 questions used to calculate the HAQ-DI:

HAQ-DI Category:	Category questions: At least 6 categories must have scores to compute the HAQ-DI.		HAQ-DI Category Score with Aids/Devices Calculation:
	Category Questions	Aids/Devices Indicators	
Dressing / Grooming	HAQ0101, HAQ0102 (DRESS, HAIR)	HAQ0114, HAQ0119 (DRSG, GROOM)	Using each question with a scale of 0-3, calculate the category score as the maximum of the category questions. If the Aids/Devices indicator is "No", no need to adjust the category score. If the Aids/Devices indicator is "Yes" and the category score is <2, then the category score with the Aids/Devices is set to 2. If the Aids/Devices indicator is "Yes" and the category score is ≥2, then the category score with Aids/Devices is the calculated category score without adjustment. For example: The Dressing/Grooming category score is 2 if subject answered 1 for both questions 1 and 2 and "Yes" for both question 14 and 18. In the HAQ-DI score calculation, questions on other device/aids will not be used.
Arising	HAQ0103, HAQ0104 (STAND, BED)	HAQ0116, HAQ0120 (CHAIR, ARISING)	
Eating	HAQ0105,HAQ0106, HAQ0107 (MEAT, LIFT,MILK)	HAQ0115, HAQ0121 (UTENSIL, EAT)	
Walking	HAQ0108, HAQ0109 (WALK, STEPS)	HAQ0110, HAQ0111, HAQ0112, HAQ0113, HAQ0122 (CANE, WALKER, CRUTCH, WHEEL, WALKING)	
Hygiene	HAQ0123, HAQ0124, HAQ0125 (WASH, BATH, TOILET)	HAQ0134, HAQ0135, HAQ0137, HAQ0139, HAQ0142 (RAISEAT, BATHBAR, BATHSEAT, LONGBATH, HYGIENE)	
Reach	HAQ0126, HAQ0127 (REACH, BEND)	HAQ0138, HAQ0143 (LONGRCH, REACH)	
Grip	HAQ0128, HAQ0129, HAQ0130 (OPENCAR, JAR, FAUCET)	HAQ0136, HAQ0144 (JAROPEN, GRIP)	
Activity	HAQ0131, HAQ0132, HAQ0133 (SHOP, INCAR, CHORES)	HAQ0145 (ERRAND)	

Handling Missing Data: If no more than 2 categories have missing category scores, then the HAQ-DI is the mean of the non-missing category scores. Otherwise, the HAQ-DI score is set to missing.

If any of the category questions are missing, but the aids/device indicator is non-missing, the category score can still be computed. However, if all category questions and its aids/device indicators are missing, then the category score is considered missing.

Appendix 3. Treatment Satisfaction Questionnaire for Medication (TSQM)

TSQM scale scores are computed by adding the items loading on each factor. The lowest possible score is subtracted from this composite score and divided by the greatest possible score minus the lowest possible score. This provides a transformed score between 0 and 1 that should be multiplied by 100. Note that only one item may be missing from each scale before the subscale should be considered invalid for that respondent.

EFFECTIVENESS:

([Sum(Item 1 + Item 2 + Item 3) – 3] divided by 18) * 100

If one item is missing
([Sum(Item 1? + Item 2? + Item 3?) – 2] divided by 12) * 100

SIDE-EFFECTS:

If Question 4 is answered 'No', then score = 100

Else,

([Sum(Item 5 to Item 8) – 4] divided by 16) * 100

If one item is missing
([Sum(Item 5? to Item 8?) – 3] divided by 12) * 100

CONVENIENCE:

([Sum(Item 9 to Item 11) – 3] divided by 18) * 100

If one item is missing
([Sum(Item 9? to Item 11?) – 2] divided by 12) * 100

GLOBAL SATISFACTION:

([Sum(Item 12 to Item 14) – 3] divided by 14) * 100

If either Item 12 or 13 is missing
([Sum(Item 12? to Item 14?) – 2] divided by 10) * 100

If Item 14 is missing
([Sum(Item 12 and Item 13) – 2] divided by 8) * 100

Appendix 4. Modified Total Sharp Score

Modified Sharp Method

The joint erosion score is a summary of erosion severity in 32 joints of the hands and 12 joints in the feet, as shown below for the joints assessed for erosions.

Hands/Fingers MCP: Metacarpophalangeal joints I - V(n = 10)	Hands/Wrist: Multangular bones(trapezium & trapezoid) (n = 2)
Hands/Fingers PIP: Proximal interphalangeal joints II - V (n = 8)	Hands/Wrist: Scaphoid(Navicular) bones (n = 2)
Hands/Fingers IP (thumbs): Interphalangeal joints (n = 2)	Hands/Wrist: Lunate bones (n=2)
Hands/Fingers (thumbs): Proximal Metacarpal joints (n = 2)	Feet/Toes MTP: Metatarsophalangeal joints I – V (n = 10)
Hands/Wrist: Distal Radius (n = 2)	Feet/Toes IP (big toes): Interphalangeal joints (n = 2)
Hands/Wrist: Distal Ulna (n = 2)	

An erosion score of 0 to 5 is given to each joint in the hands and wrists, and a score of 0 to 10 is given to each joint in the feet. Each hand and wrist joint is scored, according to the surface area involved, from 0 to 5, with 5 indicating extensive loss of bone from more than one half of the articulating bone (0 indicates no erosion). Because each side of a foot joint is graded on this scale, the maximum erosion score for a foot joint is 10. Thus, the maximum erosion score for hands and wrists is 32*5 + 12*10 = 280. A joint may not be evaluable due to surgery (i.e., joint replacement) or may be radiographically insufficient for reading.

The joint space narrowing (JSN) score summarizes the severity of JSN in 30 joints of the hands/wrists and 12 joints of the feet, as shown below for the joints assessed for JSN.

Hand/Fingers MCP: Metacarpophalangeal joints I – V (n = 10)	Hand/Wrist: Capitate-scaphoid Joint (n = 2)
Hand/Fingers PIP: Proximal interphalangeal joints II - V (n = 8)	Hand/Wrist: Scaphoid-radius Joint (n=2)
Hand/Fingers CMC: Carpometacarpal joints III, IV, V (n = 6)	Foot/Toes MTP: Metatarsophalangeal joints (n = 10)
Hand/Wrist: Scaphoid-Trapezium Joint (n = 2)	Foot/Toes IP: Interphalangeal joints (n = 2)

Assessment of JSN for each hand (15 joints per hand) and foot (6 joints per foot), including subluxation, is scored from 0 to 4, with 0 indicating no/normal JSN and 4 indicating complete loss of joint space, bony ankylosis, or luxation. Thus, the maximal JSN score is (30*4) + (12*4) = 168. A joint may not be evaluable due to subluxation, luxations, surgery (i.e., joint replacement) or may be radiographically insufficient for reading.

The mTSS is defined as the erosion score plus the JSN score, with a range of 0 to 448.

The maximum scores (adding up both hands/wrists and feet) are:

Radiography	**Maximum score of hands/wrists**	**Maximum score of feet**	**Maximum score (add up hands/wrists and feet)**
ES (erosion score)	32 joints * 5 score = 160	12 joints * 10 score = 120	280
JSN (joint space narrowing) score	30 joints * 4 score = 120	12 joints * 4 score = 48	168
TSS (Total Sharp score)	280	168	448

Appendix 5. Cortocosteroids

The following table will be used for converting non-prednisone medications to prednisone equivalent:

*Example: Patient is taking 8 mg of Methylprednisolone orally daily. To get the equivalent dose of prednisone: 8 mg Methylprednisolone = (5*8)/ 4 = 10 mg prednisone.*

Corticosteroids Name	Equivalent Dose (mg) to 5 mg Prednisone
Betamethasone	0.75
Betamethasone Dipropionate	0.75
Betamethasone Sodium Phosphate	0.75
Cortisone	20
Dexamethasone	0.75
Dexamethasone Palmitate	0.75
Dexamethasone Phosphate	0.75
Dexamethasone Sodium Phosphate	0.75
Hydrocortisone	20
Meprednisone	4
Methylprednisolone	4
Methylprednisolone Acetate	4
Methylprednisolone Sodium Succinate	4
Prednisone	5
Prednisolone	5
Prednisolone Farnesylate	5
Prednisolone Sodium Succinate	5
Triamcinolone	4

Appendix 6. SAS Programming for Tipping Point Analysis for Binary Endpoint

The following **%tp_binary** macro generates multiple imputed data and a set of the shift parameters that adjust the imputed values will be examined.

```
/*---------------------------------------------------------------------------------------------------------------------*/
/*--- Delta-Adjustment Method for Tipping Point Analysis for Binary Endpoint                          ---*/
/*---                                                                                                  ---*/
/*--- Generate imputed data set for specified shift parameters                                         ---*/
/*--- data= input data set                                                                             ---*/
/*--- smin= min shift parameter for active drug                                                        ---*/
/*--- smax= max shift parameter for active drug                                                        ---*/
/*--- sinc= increment of the shift parameter for active drug                                           ---*/
/*--- pmin= min shift parameter for placebo drug                                                       ---*/
/*--- pmax= max shift parameter for placebo drug                                                       ---*/
/*--- pinc= increment of the shift parameter for placebo drug                                          ---*/
/*--- trt= treatment group indicator                                                                   ---*/
/*--- out= output imputed data set                                                                     ---*/
/*---------------------------------------------------------------------------------------------------------------------*/

%macro tp  binary( data=, smin=, smax=, sinc=, pmin=, pmax=, pinc=, trt=, out=);
        data &out;
        set _null_;
        run;

        %let ncase_pbo = %sysevalf( (&pmax-&pmin)/&pinc, ceil );

        %do pc=0 %to &ncase_pbo;

                %let pj= %sysevalf( &pmin + &pc * &pinc);

                /*------------ # of shift values ------------*/
                %let ncase= %sysevalf( (&smax-&smin)/&sinc, ceil );

                /*------- Imputed data for each shift -------*/
                        %do jc=0 %to &ncase;

                        %let sj= %sysevalf( &smin + &jc * &sinc);

                                proc mi data=&data seed=14823 nimpute=20 out=outmi;
                                        var trt01pn strat1V strat2V strat3V das28n;
                                        class trt01pn strat1V strat2V strat3V das28n;
                                        monotone logistic (das28n / link=glogit);
                                        mnar adjust( das28n(event='1') / adjustobs=(trt01pn="&trt") shift= &sj)
                                                adjust( das28n(event='1') / adjustobs=(trt01pn='3') shift= &pj);

                                run;

                                data outmi;
                                        set outmi;
                                        Shift  Trt= &sj;
                                        Shift_Pbo= &pj;
                                run;
```

```
                    data &out;
                              set &out outmi;
                    run;
          %end;
%end;
%mend tp_binary;
```

Appendix 7. SAS Programming for Tipping Point Analysis for Continuous Endpoint

The following **%tp_conti** macro generates multiple imputed data sets and a set of the shift parameters that adjust the imputed values will be examined.

```
/***********************************************************************************/
/*--- Modified Tipping Point Analysis from Yuan's method                         ---*/
/*---                                                                            ---*/
/*--- Generate imputed data set for specified shift parameters                   ---*/
/*--- data= input data set                                                       ---*/
/*--- smin= min shift parameter for active drug                                  ---*/
/*--- smax= max shift parameter for active drug                                  ---*/
/*--- sinc= increment of the shift parameter for active drug                     ---*/
/*--- pmin= min shift parameter for placebo drug                                 ---*/
/*--- pmax= max shift parameter for placebo drug                                 ---*/
/*--- pinc= increment of the shift parameter for placebo drug                    ---*/
/*--- trt= treatment group indicator (eg, 1= filgotinib 200 mg)                  ---*/
/*--- out= output imputed data set                                               ---*/
/***********************************************************************************/

%macro tp  conti( data=, smin=, smax=, sinc=, pmin=, pmax=, pinc=, trt=, out=);
          data &out;
          set _null_;
          run;

          %let ncase_pbo = %sysevalf( (&pmax-&pmin)/&pinc, ceil );

   %do pc=0 %to &ncase_pbo;

          %let pj= %sysevalf( &pmin + &pc * &pinc);

                    /*------------ # of shift values ------------*/
                    %let ncase= %sysevalf( (&smax-&smin)/&sinc, ceil );

                    /*------- Imputed data for each shift -------*/
                    %do jc=0 %to &ncase;

                    %let sj= %sysevalf( &smin + &jc * &sinc);

                    proc mi data=&data out=all_mono nimpute=15 seed=123;
                              var  v  1 v  2 v  3 v  4 v  5 v  6 v  7 v  8 v_9 ;
                              mcmc chain=multiple impute=monotone;
                    run;

                    proc sort data=all_mono; by _Imputation_ trt01pn; run;

                    proc mi data=all  mono out=outmi seed=465 nimpute=1;
                              by _Imputation_;
                              var trt01pn strat1V strat2V strat3V v_1 v_2 v_3 v_4 v_5 v_6 v_7 v_8 v_9;
                              class trt01pn strat1V strat2V strat3V;
                              monotone regression;
                              mnar adjust( v_5 / adjustobs=(trt01pn="&trt") shift=&sj)
                                        adjust( v_5 / adjustobs=(trt01pn='3') shift= &pj);
```

```
                run;

                data outmi;
                        set outmi;
                        Shift_Trt= &sj;
                        Shift_Pbo= &pj;
                run;

                data &out;
                        set &out outmi;
                run;
                %end;
        %end;
%mend tp_conti;
```