Protocol Number: SRA-MMB-301

Official Title: A Randomized, Double-blind, Phase 3 Study to Evaluate the Activity of Momelotinib (MMB) versus Danazol (DAN) in Symptomatic, Anemic Subjects with Primary Myelofibrosis (PMF), Post-polycythemia Vera (PV) Myelofibrosis, or Post-essential Thrombocythemia (ET) Myelofibrosis who were Previously Treated with JAK Inhibitor Therapy

NCT Number: NCT04173494

Document Date: 30 DEC 2021

STATISTICAL ANALYSIS PLAN

Study Title:	A Randomized, Double-Blind, Phase 3 Study to Evaluate the Activity of Momelotinib (MMB) versus Danazol (DAN) in Symptomatic, Anemic Subjects with Primary Myelofibrosis (PMF), Post-Polycythemia Vera (PV) Myelofibrosis, or Post Essential Thrombocythemia (ET) Myelofibrosis who were Previously Treated with JAK Inhibitor Therapy
Protocol Version:	Version 2.0, 18 December 2020
Study Code:	SRA-MMB-301
Investigational Product:	Momelotinib (MMB)
SAP Version:	V2.0
SAP Date:	30 December 2021
SAP Author:	[redacted] Senior Director, Biometrics

Compliance Statement:	This trial will be conducted in accordance with Protocol SRA-MMB-301, the International Council for Harmonisation (ICH), Guideline for Good Clinical Practice (GCP), and the applicable country and regional (local) regulatory requirements.

TABLE OF CONTENTS

LIST OF TABLES

APPENDIX LIST OF TABLES

1. LIST OF ABBREVIATIONS AND DEFINITION OF TERMS

Abbreviation or Specialist Term	Explanation
AE	Adverse Event
ALP	Alkaline Phosphatase
ALT/SGPT	Alanine Aminotransferase
AST/SGOT	Aspartate Transaminase
ATC	Anatomical and Therapeutic Chemical
BID	Twice daily
CI	Confidence interval
CMH	Cochran-Mantel-Haenszel
CTC	Common Toxicity Criteria
DAN	Danazol
DIPSS	Dynamic International Prognostic Scoring System
EC	Early Cross-over
ECDF	Empirical Cumulative Distribution Function
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic case report form
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire
ePRO	Electronic patient reported outcome
EQ-5D	EuroQoL Five Dimension
ET	Essential Thrombocythemia
FED	Fedratinib
GP	General Practice Physician
Hgb	Hemoglobin

Abbreviation or Specialist Term	Explanation
HR	Hazard ratio
IMP	Investigational Medicinal Product
IPW	Inverse probability weighting
ITT	Intention-To-Treat
LCM	Left coastal margin
LDH	Lactate dehydrogenase
LFS	Leukemia-Free Survival
LOCF	Last Observation Carried Forward
MAR	Missing at random
MCT	Meaningful change threshold
MedDRA	Medical Dictionary for Regulatory Activities
MF	Myelofibrosis
MF-8D	Myelofibrosis-8 Dimension
MFSAF	Myelofibrosis Symptom Assessment Form
MMB	Momelotinib
MMRM	Mixed model for repeated measures
NCI-CTCAE	National Cancer Institute Common Terminology Criteria for Adverse Events
NRI	Non-responder imputation
OC	Observed case
OLE	Open-Label Extension
OS	Overall Survival
PGIC	Patient Global Impression of Change
PGIS	Patient Global Impression of Severity
PMF	Primary Myelofibrosis

Abbreviation or Specialist Term	Explanation
PP	Per Protocol
PRO	Patient-Reported Outcomes
PROMIS	Patient-Reported Outcomes Measurement Information System
PS	Performance Status
PT	Preferred term
PV	Polycythemia Vera
RBC	Red Blood Cells
RPFST	Rank-preserving structural failure time
RT	Randomized Treatment
RUX	Ruxolitinib
SAE	Serious Adverse Event
SAF	Safety
SOC	System organ class
SRR	Splenic Response Rate
TD	Transfusion Dependent
TEAE	Treatment-emergent adverse event
TI	Transfusion Independent
TR	Transfusion Requiring
TSS	Total Symptom Score
ZINB	Zero-Inflated Negative Binomial

2. INTRODUCTION

This Statistical Analysis Plan was written for the clinical trial SRA-MMB-301 conducted in subjects with Primary Myelofibrosis (PMF), Post-Polycythemia Vera (PV) Myelofibrosis, or Post Essential Thrombocythemia (ET) Myelofibrosis who were previously treated with JAK Inhibitor therapy. The ICH guideline E3 "Structure and Content of Clinical Study Reports" was used as a guide to the writing of the plan.

3. STUDY DESIGN AND OBJECTIVES

3.1. Study Objectives and Endpoints

3.1.1. Primary Objective and Endpoint

Primary Objective	Primary Endpoint
To determine the efficacy of MMB versus DAN assessed by improvement of MFSAF TSS in subjects with PMF, post-PV MF, or post-ET MF who were previously treated with approved JAK inhibitor therapy	Proportion of subjects with MFSAF TSS response at Week 24. TSS response is defined as a ≥ 50% reduction in mean TSS over the 28 days immediately prior to the end of Week 24 compared to baseline

3.1.2. Key Secondary Objectives and Endpoints (Type I error protected hierarchy)

Key Secondary Objectives	Key Secondary Endpoint(s)
To compare the effect of MMB versus DAN on TI status at Week 24	Proportion of subjects with TI status at the end of Week 24; TI status defined as not requiring RBC or whole blood transfusion (except in the case of clinically overt bleeding) for ≥ 12 weeks, with all Hgb levels during the ≥ 12-week interval of ≥ 8 g/dL (except in the case of clinically overt bleeding)
To compare SRR for subjects treated with MMB versus DAN	Proportion of subjects who have splenic response (reduction in spleen volume of ≥ 25% from baseline and also reduction of ≥ 35% from baseline) at the end of Week 24
To compare change from baseline MFSAF TSS at Week 24 in subjects treated with MMB versus DAN	Change from baseline of mean TSS over the 28 days immediately prior to the end of week 24
To compare RBC transfusion requirements in subjects treated with MMB versus DAN	Proportion of subjects with zero RBC or whole blood units transfused during the 24-week Randomized Treatment Period

3.1.3. Other Secondary Objectives and Endpoints

Other Secondary Objectives	Other Secondary Endpoint(s)
To assess the duration of MFSAF TSS response	Duration of the end of Week 24 MFSAF TSS response (assessed until the end of Week 48); For subjects who achieve a Week 24 TSS response, the duration of response is defined as the number of days from the start of the initial 28-day period (during the 24-Week Randomized Treatment Period) in which the subject has a ≥ 50% reduction from baseline TSS to the first day of the 7-day assessment that determines the mean TSS for the 28-day period during which the subject's TSS equals or exceeds their baseline value. TSS will be assessed during the last 7 days (± 7 days) of each month during the open label extended treatment period until Week 48.
To assess duration of TI status at Week 24	For subjects who achieve TI status at Week 24, duration of TI is defined as the number of days from the first day of a period of at least 12 weeks, during which a subject received no transfusions and had no Hgb < 8 g/dL (except in the case of clinically overt bleeding), to the first RBC or whole blood transfusion or Hgb level < 8 g/dL (again, except in the case of clinically overt bleeding) (assessed until the end of Week 48)
To compare the benefit of MMB versus DAN on anemia response and transfusion requirements, and to estimate the duration of response	Cumulative transfusion risk for MMB versus DAN at the end of Week 24, measured by a proportional hazards recurrent events model Proportion of subjects with TD status at the end of Week 24, defined as requirement of ≥ 4 RBC or whole blood units in an 8-week period immediately prior to the end of Week 24 (and Week 48 for MMB arm). Assessed in all subjects and in the subset of subjects who were TI at baseline. Proportion of hemoglobin responses. Hemoglobin responses are defined as increases of ≥ 1, ≥ 1.5, or ≥ 2 g/dL from baseline in Hgb over the 24-week randomized treatment period and the last 12 weeks of the period with any Hgb values within 4 weeks after a transfusion excluded. Assessed in all subjects and in the subset of subjects who were TI at baseline.
To compare the effect of MMB versus DAN on TI status at Week 24	Proportion of baseline TD subjects with TI status at the end of Week 24 Duration of TI status from Week 24 in baseline TD subjects
To characterize the safety of MMB	Safety assessments including the type, frequency, severity per CTC grading system (CTCAE v5.0, 2017), timing of onset, duration, and relationship to study drug of any AEs or abnormalities of laboratory tests, as well as SAEs or AEs leading to discontinuation of study drug

Other Secondary Objectives	Other Secondary Endpoint(s)
To compare OS and leukemia-free survival (LFS) of subjects treated with MMB versus DAN	OS, defined as the interval from the first study drug dosing date to death from any cause LFS, defined as the interval from the first study drug dosing date to any evidence of leukemic transformation and/or death
To compare patient-reported fatigue and physical function for MMB versus DAN	Mean change from baseline in disease-related fatigue (assessed as "Fatigue (tiredness, weariness)" by the MFSAF) in MMB versus DAN subjects from baseline to each evaluation timepoint Mean change from baseline in cancer-related fatigue (assessed by the EORTC QLQ-C30 fatigue domain) in MMB versus DAN subjects from baseline to each evaluation timepoint Mean change from baseline in physical function score (assessed by Patient-Reported Outcomes Measurement Information System [PROMIS]) in MMB versus DAN subjects from baseline to each evaluation timepoint

3.1.4. Exploratory Objectives and Endpoints

Exploratory Objectives	Exploratory Endpoint(s)
To compare patient-reported health status and health-related QoL for MMB versus DAN	Changes from baseline in EQ-5D index and VAS scores at each evaluation timepoint Changes from baseline in MF-8D classification, derived from responses to MFSAF and EORTC-QLQ-C30, at each evaluation timepoint
To assess association of MMB exposure (PK) with outcome	Correlation of plasma concentration of MMB and results of efficacy assessment
To determine the efficacy of MMB versus DAN on improvement in MFSAF TSS in subsets defined by baseline transfusion requirements	Assessed in baseline TD, TI and non-TD subsets; MFSAF TSS response rate, to the end of Week 24
To assess time to symptomatic splenic progression for subjects treated with MMB versus DAN	Time from first dose to symptomatic splenic progression
To explore potential correlates with response including but not limited to mutational analysis	Measures of symptom and anemia response and exploratory analyses including but not limited to mutational analysis
To explore health care utilization requirements for MMB versus DAN	Hospitalization rates, transfusion rates, and utilization of other medical care during the 24-Week Randomized Treatment Period, and during the study as compared to baseline based on data captured from patient records for the 12 weeks prior to randomization and recorded throughout the study

Exploratory Objectives	Exploratory Endpoint(s)
To assess baseline ferritin level as predictive biomarker for MMB vs DAN treatment effect measured by transfusion independence response at week 24.	Transfusion independence status at Week 24 by baseline ferritin level

The study protocol presents a list of objectives and endpoints different from that shown above, with some entries in different categories and some removed from the list. In particular, rolling 12-week Transfusion Independent (TI) endpoints and their duration, along with duration of anemia response, rate of RBC or whole blood transfusion, some duplicative hemoglobin (Hgb) and anemia response measures, Patient Global Impression of Severity (PGIS) and Patient Global Impression of Change (PGIC), any Total Symptom Score (TSS) response (as opposed to strictly Week 24) and duration of same, and time to TSS deterioration have been removed.

3.2. Study Design

This is a randomized, double-blind study intended to confirm the differentiated clinical benefit of momelotinib (MMB) versus Danazol (DAN) in subjects who have previously received approved JAK inhibitor therapy for myelofibrosis (MF) for a minimum of 90 days, or a minimum of 28 days if JAK inhibitor therapy is complicated by red blood cell (RBC) transfusion requirement of at least 4 units in 8 weeks, or Grade 3/4 adverse events (AEs) of thrombocytopenia, anemia, or hematoma.

Subjects will be randomized in a 2:1 ratio to receive MMB (plus DAN-matching placebo) or DAN (plus MMB-matching placebo). A non-deterministic biased coin minimization procedure (Pocock and Simon (1975); Han, 2009) will be used to reduce imbalances between treatment arms for the following baseline potential prognostic factors: Myelofibrosis Symptom Assessment Form (MFSAF) TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the left coastal margin (LCM) (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+), and investigational site. Details of the randomization method are in Section 4.

Subjects randomized to receive MMB who complete the Randomized Treatment (RT) Period to the end of Week 24 may continue to receive MMB in the Open-Label Extension (OLE) Treatment Period and potentially in an extended access protocol.

Subjects randomized to receive DAN may cross-over to MMB open-label treatment in the following circumstances: a) at the end of Week 24 if they complete the RT Period; b) at the end of Week 24 if they discontinue treatment with DAN but continue study assessments and do not receive prohibited medications; c) at any time during the RT Period if they meet the protocol-defined criteria for confirmed symptomatic splenic progression.

Subjects randomized to receive DAN who are receiving clinical benefit at the end of Week 24 may continue open-label DAN therapy up to Week 48. The decision whether to remain on DAN or cross-over to MMB must be made at the end of Week 24.

Analysis of the primary efficacy endpoint will occur when the outcome of the primary endpoint is determinable for all subjects ie, when all subjects have either completed the RT Period or dropped out.

Details of the study design can be found in the clinical trial protocol.

3.3. Sample Size Justification

3.3.1. Summary and Base Assumptions

A sample size of 180 subjects was determined based on power considerations to detect a statistically significant treatment difference in the proportion of subjects with TSS response (primary endpoint), as well as in the proportion of subjects with TI status and in Splenic Response Rate (SRR, secondary endpoints).

With a sample size of 180 subjects who will be randomized to MMB or DAN in a 2:1 ratio, using a 2-sided significance level of 0.05, the study will have a 98.8% power to detect a true difference of 21% in TSS (23% with MMB versus 2% with DAN), or a 90% power to detect a true difference of 15% in TSS (17% with MMB versus 2% with DAN) based on the method in Fleiss et al (1980). The study will also have a 90% power to detect a true difference of 24% in the proportion of subjects with TI status (45% versus 21%) and a true difference of 14% in SRR (15% versus 1%).

Table 1: Power Computations for 180 Subjects

Endpoint	Assumed true proportions			Power to detect difference between treatments (superiority)
	MMB %	DAN %	True proportion difference	
TSS24	23	2	21%	98.8%
TSS24	17	2	15%	90%
TI24	45	21	24%	90%
SRR24 (25% reduction)	15	1	14%	90%
Proportion with no transfusions during first 24 weeks	70	45	35%	90%

*All power computations made with East software, version 6.5 (Cytel), or SAS, Version 9.4.

4. RANDOMIZATION

Subjects will be randomized in a 2:1 ratio to MMB arm or DAN arm. A non-deterministic biased coin minimization method (Pocock, 1975; Han, 2009) will be used to reduce imbalance between treatment arms for the following baseline potential prognostic factors: MFSAF TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+), and investigational site. Allocation probability of 0.9 and 0.8 will

be used in randomizing a patient to MMB arm and DAN arm each when it is the preferred treatment arm per the imbalance score. The unbiased randomization ratio under these allocation probabilities is 2:1 (Han, 2009).

Weighted sum of the marginal imbalance across the 4 factors will be used as the imbalance score to minimize.

5. GENERAL ANALYSIS DEFINITIONS

Data will be analyzed using SAS (Version 9.4 or higher).

No tests of significance will be carried out to compare treatment arms on baseline data because any observed differences between them must be attributed to chance.

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

Descriptive statistics will be tabulated as follows:

- Categorical data will be summarized in contingency tables presenting frequencies and percentages.
- Continuous data will be summarized using number of non-missing values (n), mean, standard deviation, median, first and third quartiles, minimum and maximum values. In addition, Empirical Cumulative Distribution Function (ECDF) plots may be provided by treatment arm as necessary.

The tables will be created by treatment arm. Efficacy results will be summarized under the treatment arms to which patients were randomized and safety results will be summarized under the treatment arms which patients received unless specified otherwise.

5.1. Baseline, Study Period and Visit Window Definitions

5.1.1. Baseline

Unless otherwise specified, the baseline is defined as the last assessment done before or on the day of first dose date. Assessments done on the date of 1st dose date are assumed to take place before the administration, unless specified otherwise. Baseline definitions for specific endpoints are defined as below:

1. MFSAF TSS

   Baseline MFSAF TSS will be defined relative to the date on which the baseline period for TSS was “triggered” on the handheld electronic patient reported outcome (ePRO) device; this may also be referred to as the “ePRO visit date”. This date may or may not be the same as the date of the baseline clinic visit as recorded in the eCRF.

   The baseline TSS will be computed as the mean of the TSS values generated on the date of the baseline period triggering per the handheld ePRO device and on the six days immediately following that triggering date, with the following caveats:

- If there is more than one TSS value generated on the date of baseline triggering, the last TSS value on that date will be used and other values from that date will be ignored.
- Any values that are generated on or after the date of first dosing will not be used in the computation of the baseline TSS.

2. Transfusion and Hgb:

   Baseline RBC or whole blood transfusion rate per subject is defined as the number of units of RBC or whole blood transfusion required per month and determined from number of RBC or whole blood transfusions given in the 8-week period prior to the first day of dosing.

   Baseline Hgb level is computed from the Hgb reading on and prior to the first day of dosing, depending upon when transfusions were given before dosing:

   - If no RBC or whole blood transfusions were given prior to the first day of dosing, the baseline Hgb level will be the last central Hgb level on or prior to the first day of dosing.
   - If at least one RBC or whole blood transfusion was given prior to the first day of dosing but they all were given more than 28 days prior to the last central Hgb prior to first day of dosing, then the baseline Hgb level will be the last central Hgb level on or prior to the first day of dosing.
   - If at least one RBC or whole blood transfusion was given prior to the first day of dosing and the last such transfusion was given within 28 days prior to the last central Hgb prior to first day of dosing, then the baseline Hgb level will be the last central or local Hgb level prior to the last such transfusion; if the last central and local Hgb levels occur on the same day, then the central Hgb level should be used.
   - If the preceding conditions are not met due to no such central Hgb levels satisfying the conditions, then local Hgb levels should be added to the collection of potential Hgb levels used to determine baseline Hgb level.
   - If none of the current conditions are met then the baseline Hgb level is the last Hgb assessment (regardless of local or central) prior to the most recent transfusion prior to the first date of first dose.
   - If no central or local Hgb levels meet the preceding conditions, then baseline Hgb should be set to missing.

Based on transfusion and Hgb data collected before dosing, patients will be classified into the following categories:

- Transfusion dependent (TD) status at baseline is defined as requiring ≥ 4 units of RBC or whole blood transfusions in the 8 weeks prior to first day of dosing. Only RBC or whole blood transfusions given when Hgb levels are ≤ 9.5 g/dL (as evidenced on the transfusion record) are counted towards TD. RBC or whole blood transfusions

given for clinically overt bleeding (as assessed by the investigator) are not counted towards TD.

- Transfusion independent status at baseline is defined as not receiving RBC or whole blood transfusion (except in the case of clinically overt bleeding) in the 12 weeks prior to first day of dosing, with all (central and local) Hgb levels collected prior to first day of dosing ≥ 8 g/dL.
- Transfusion requiring (TR) status at baseline is defined as not meeting TD or TI criteria.

5.1.2. Study Periods

The trial includes the following:

- Screening period: within 6 weeks prior to randomization
- Baseline period: 7 consecutive days (BL1 to BL7) immediately prior to randomization
- Randomization
- Day 1 (first dose of study treatment)
- Randomized Treatment Period: 24 weeks with visits at Weeks 2 and 4 (± 2 days), and every 4 weeks (± 3 days) until the end of Week 24, or until cross-over to open-label treatment.
- Open-Label Extended Treatment Period: Begins following the completion of Week 24, or following the Cross-Over Visit for early cross-over (EC) in the event of confirmed symptomatic splenic progression. Visits will occur every 4 weeks (± 3 days) to the end of Week 48 and thereafter every 12 weeks (± 7 days) to the end of the OLE Treatment Period or until the Treatment Discontinuation Visit, whichever occurs first. Transition to an MMB extension study, if available, may occur once a subject has completed at least Week 48 (or in the event of EC, 24 weeks after cross-over) on-study.
  - OLE Treatment Period (MMB); for subjects randomized to MMB, open-label treatment with MMB will continue until the subject withdraws from the study or enrolls in an extended access protocol.
  - OLE Treatment Period (DAN); subjects who do not wish to receive OL MMB and are receiving clinical benefit from DAN in the Randomized Treatment Period may continue to the OLE Treatment Period with DAN for 24 weeks, to the end of Week 48.
  - OLE Treatment Period (cross-over from DAN to MMB); for subjects randomized to receive DAN, cross-over to treatment with open-label MMB may occur at the end of Week 24. Early cross-over may occur in the event of confirmed symptomatic splenic progression. Open-label treatment with MMB may continue until the subject withdraws from the study or enrolls in an extended access protocol.

- Safety Follow-Up Visit: 30 days (± 7 days) after the last dose, at which time the subject is to enter the Survival Follow-Up Period
- Survival Follow-Up: subjects will be assessed every 3 months (± 7 days) post-last dose (end of Week 204 / Week EC180) to 7 years post-first dose (Day 1)

Transition to an MMB extension study, if available, may occur once a subject has completed at least Week 48 (or in the event of EC, 24 weeks after cross-over) on-study.

For the purpose of results presentation, the following periods are defined:

Period	Start	End
Study Treatment Period	Randomization	Study treatment discontinuation
Randomized Treatment Period	Randomization	Last day of dosing with randomized treatment or end of Week 24, whichever is earlier
OLE Treatment Period (MMB)	Initiation of open-label MMB treatment at the end of Week 24	Study treatment discontinuation
OLE Treatment Period (DAN)	Initiation of open-label DAN treatment at the end of Week 24	Study treatment discontinuation
OLE Treatment Period (cross-over from DAN to MMB)	Initiation of open-label MMB at the end of Week 24, or following the Cross-Over Visit for early cross-over	Study treatment discontinuation
Follow-up Period	Study treatment discontinuation	Final survival follow-up visit, 7 years after first dose of study treatment

For the purpose of assigning observations/events (eg, AEs) to specific study periods, actual dates (eg, onset dates) will be used.

5.2. Planned Analyses

The primary analysis of the primary efficacy endpoint will occur when the outcome of the primary endpoint is determinable for all subjects, ie, when each subject has completed the RT Period, crossed over early or dropped out from the randomized treatment. The data cutoff will be determined as when the aforementioned event has occurred. At that time, the study will be unblinded and all other study endpoints will also be analyzed as well. All the data prior to the data cutoff, including the data from the OLE Treatment Period, will be included in the primary analysis.

Additional unplanned analyses may be performed at other timepoints, subsequent to the primary Week 24 analysis, for regulatory or publication purposes.

The analysis of Overall Survival (OS) and Leukemia-Free Survival (LFS) will be also performed at a later stage than the primary analysis timing for the primary and other secondary efficacy

endpoints, as appropriate to satisfy requirements for long-term safety and OS, with a final analysis at completion of the follow-up period.

The derivation of meaningful change (MCT) in MFSAF will be described in a separate document.

5.3. Definition of Analysis Sets

5.3.1. Intention-To-Treat (ITT) Analysis Set

The Intention-To-Treat (ITT) analysis set includes all randomized subjects.

The ITT analysis set will be used as the primary analysis set for efficacy analyses except the TI Status at Week 24 endpoint.

5.3.2. Per Protocol (PP) Analysis Set

The Per Protocol (PP) analysis set consists of randomized subjects who received at least one dose of study medication and who do not have any major protocol violation. Study treatment assignment for the purpose of analysis will be designated according to the actual treatment received (see Section 5.5 for more details).

A list of major protocol violations is listed in Section 6.2.

The PP analysis set will be used for sensitivity analysis of selected efficacy endpoints including TI24 NI.

5.3.3. Safety (SAF) Analysis Set

The Safety (SAF) analysis set will include all subjects in the ITT Analysis set who received at least one dose of study drug.

The SAF analysis set will be used for safety analyses.

5.4. Subgroup Definitions

Subgroup analyses for the following factors are planned for primary and selected secondary efficacy endpoints (see Section 5.1.1 for definitions):

- Subgroup of subjects defined by transfusion status (TI/TR/TD) at baseline
- Subgroup of subjects define by transfusion status (TI/non-TI) at baseline
- Subgroups specific to gender (male, female)
- Subgroups specific to age (< 65 years, ≥ 65 years)
- Subgroups specific to race (using options from CRF)
- Subgroups specific to baseline platelets count (< 50, ≥ 50 but ≤ 150, > 150 but ≤ 300, > 300 × 10^9/L)
- Subgroups specific to baseline platelets count (≤ 150, > 150 × 10^9/L)
- Subgroups specific to baseline platelets count (≤ 200, > 200 × 10^9/L)

- Subgroups based on baseline MFSAF TSS (< 22, ≥ 22)
- Subgroups based on baseline spleen volume median
- Subgroups based on RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+)
- Subgroups based on baseline Hgb (< 8 g/dL, ≥ 8 g/dL)
- Subgroups based on baseline glomerular filtration rate (30-60 / 60+)
- Subgroups based on Dynamic International Prognostic Scoring System (DIPSS) prognostic category (low, intermediate-1, intermediate-2, high risk)
- Subgroups based on Myelofibrosis diagnosis (PMF, post-PV MF, post-ET MF)
- Subgroups based on JAK2V617F mutation status (positive, negative, unknown)
- Subgroups based on prior JAKI total daily dose received immediately prior to enrollment (three groups: 0, < 20 mg twice daily (BID) of ruxolitinib (RUX) or ≤ 200mg of fedratinib (FED), ≥ 20 mg BID of RUX or > 200 mg of FED)
- Subgroups based on Geographical Region (Asia, Australian Asia, Europe, North America)
- Subgroups based on duration of JAKI treatment received prior to randomization (< 12 weeks, ≥ 12 weeks)
- Subgroup of subjects receiving ongoing JAKI at screening

Subjects with missing subgroup determining variables will not be included in a subgroup. Subgroups with too few subjects may be combined as necessary.

To graphically display treatment effect changes across subgroups, a forest plot of proportion differences (TSS response, TI status, SRR) or treatment group least-square mean differences (TSS change from baseline) by subgroup will be produced.

Additionally, analyses of duration of response/improvement will be restricted to those subjects qualifying for the response/improvement of interest. This will be clarified in the appropriate sections below.

5.5. Treatment Assignment and Treatment Arms

All analyses performed on the ITT analysis set will consider subjects in the treatment arm to which they were allocated at randomization, regardless of any accidental or intentional receipt of other treatment or any treatment cross-over during the treatment period of the study.

All analyses performed on the PP analysis set or the SAF analysis set will consider subjects in the treatment arm that corresponds to the treatment actually received, as long as the same treatment was received during the entire course of the RT Period considered. In case of accidental receipt of a wrong treatment for a limited period of time, the subject will be considered in treatment arm corresponding to the assigned treatment at randomization. Subjects assigned to DAN who crossover to MMB will still be considered in the DAN arm in the RT Period analysis.

More specific details are provided in the safety section regarding the presentation of safety data collected during the OLE Treatment Period.

5.6. Calculated Variables

Baseline specific definitions are provided in Section 5.1.1. Efficacy endpoints derivation is described further in appropriate sections. This section described general conventions used in the calculation of variables.

- Change from baseline is defined as the post-baseline value – baseline value.
- Percent change from baseline is defined as 100*(post-baseline value – baseline value)/baseline value. If the baseline value is 0 and the post-baseline value is 0, the change from baseline and the percent change from baseline are both defined to be 0. If the baseline value is 0 and the post-baseline value is not 0, the change from baseline is the same as the post-baseline value and the percent change from baseline will be missing.
- Duration of exposure (months): (date of last dose of study drug – date of first dose of study drug + 1) / 30.4375.
- Actual cumulative dose (mg): sum of all actual doses administered.
- Planned cumulative dose (mg): [200 mg/daily for MMB and 600 mg/daily for DAN] x [number of days in the treatment period].
- Dose intensity (mg/day): ratio of actual cumulative dose received and actual duration of exposure in days.
- Relative dose intensity (%): 100 * ratio of actual cumulative dose received and planned cumulative dose.
- Rate of RBC or whole blood transfusion (per subject-month): total number of units transfused in the period / duration of period (months)

5.7. Partial Dates

Partial or missing dates in general will not be imputed.

For the assignment to prior or concomitant medication, the general rule will be that medications should be considered concomitant unless demonstrated otherwise.

Thus, for purposes of assignment to prior or concomitant medication, the following rules will be applied in case of incomplete or missing dates:

- If end date is missing, the end date should be considered the last date of dosing or start date whichever is later.
- If end date is presented as year only, the end date should be considered the last day of dosing or start date whichever is later if dosing ended in that year or the last day of the year of dosing did not end in that year.

- If end date is presented as year and month only, the end date should be considered the last day of dosing if dosing ended in that month or the last day of the month if dosing did not end in that month.
- If start date is missing, the start date should be considered the first date of dosing or end date whichever is earlier.
- If start date is presented as year only, the start date should be considered the first day of dosing if dosing started in that year or the first day of the year if dosing did not start in that year, or end date whichever is earlier.
- If start date is presented as year and month only, the start date should be considered the first day of dosing if dosing started in that month or the first day of the month if dosing did not start in that month, or end date whichever is earlier.

The imputed dates will only be used for the assignment to prior or concomitant and will not be used in any other calculation and will not be listed.

For the assignment of AE to treatment-emergent category the general rule to apply is that AEs should be considered treatment-emergent unless shown otherwise. Thus, the assignment of AEs to either pre-treatment-emergent or treatment-emergent should follow the same rules as those applied to medications for determining prior or concomitant medications.

In case of partial or missing onset dates not allowing comparison with the start date of OLE Treatment Period (therefore not allowing immediate allocation to either RT Period or OLE Treatment Period), the AE will by default be reported under the RT Period.

Similar rule will apply for allocating an AE to either the OLE Treatment Period or the Follow-Up Period: in case of doubt due to partial/missing onset date, the AE will be reported under the OLE Treatment Period.

The imputed dates will only be used for determining whether AEs are treatment-emergent or which period AEs should be allocated to, and will not be used in any other calculation. There will not be listed as such (only original partial date will be listed).

5.8. Methods to be Used for Handling Missing Data

5.8.1. Derivation of the primary variable

For calculation of baseline mean TSS, if more than 3 daily TSS results are missing from the 7-day baseline assessment period, the score will be considered missing. The definition of baseline assessment period is in Section 5.1.1 and Section 9.3 .

For calculation of mean TSS at Weeks 4, 8, 12, 16, 20, 24, if fewer than 20 daily measurements out of 28 are available, TSS will be set to missing for the time point considered. The definition of the 28 day period at Weeks 4, 8, 12, 16, 20, 24 is in Section 9.3 .

For calculation of mean TSS during the OLE Treatment Period, TSS will only be considered missing if all of the 7 daily TSS results are missing.

5.8.2. Missing data adjustment strategies

For each efficacy endpoint, the specific strategy to be used for handling missing data will be specified further in the appropriate sections. This section provides definitions of the main methods used in the analysis and referred to in the endpoint-specific sections of this SAP.

Missing data for response binary efficacy outcomes (primary and key secondary efficacy endpoints) will be handled, in the primary analysis, using the non-responder imputation (NRI) approach, as described below. A summary of randomized treatment discontinuation reasons (including death, disease progression, worsening splenomegaly, leukemic transformation) will be presented to further describe subjects with a missing Week 24 evaluation who are imputed as non-responders in the primary endpoint analysis. For sensitivity purpose, alternative methods can be considered, which are also described below.

1. Non-responder imputation (NRI) approach:

   For response variables, subjects with a missing evaluation will be considered as a non-responder. Therefore, the denominator will always be the number of subjects by treatment arm in the analysis set being analyzed.

2. Observed case (OC) approach:

   Method consists of using measurements as available, without any imputation of missing data. As a result, only subjects with available data will be included in the analysis.

3. Multiple imputation (MI):

   For sensitivity purposes, a multiple imputation approach can be considered to impute missing data. More details can be found in the applicable sections of this SAP.

4. Censoring:

   Missing data for time-to-event variables will be handled by censoring subjects with unobserved events. For continuous outcomes collected at several post-baseline timepoints, missing data will be handled through direct-likelihood approach by using a mixed model for repeated measurements (MMRM).

5. Last observation carried forward (LOCF):

   A LOCF approach will be used as sensitivity analysis for continuous and binary outcomes to be evaluated at one specific timepoint (eg, Week 24).

6. Maximum likelihood (ML) approach:

   When the data are examined via MMRM (repeated measures mixed model), the missing data are handled by ML approach under the missing at random (MAR) assumption. The parameter of interest is estimated such that it maximizes the overall likely likelihood including both subjects with missing and non-missing outcome.

5.9. Changes to Protocol

Not applicable.

6. STUDY PATIENTS

6.1. Disposition of Patients

The number of subjects who were screened but did not meet inclusion criteria will be presented by inclusion/exclusion criteria.

The number of subjects in each population will be tabulated by treatment arm and overall.

The numbers and percentages of subjects described below will be presented for the ITT population. For each column (treatment group), the denominator for the percentage calculation will be the total number of subjects randomized for that column with exception of that the denominator for the percentage calculation of study drug completion status and reason of study drug discontinuation in the OLE Treatment Period will be the total number of subjects who are treated in the OLE Treatment Period for each column.

- Treated (Safety Analysis Set) for that study phase (by RT and OLE Treatment Periods)
- Completed study drug for that study phase (by RT and OLE Treatment Periods)
- Did not complete study drug with reasons for premature discontinuation of study drug for that study phase (by RT and OLE Treatment Periods) (or completion of study period for such subjects)
- Completed the RT Period
- Did not complete the RT Period with reasons for premature discontinuation of RT Period
- Continuing study drug in the OLE Treatment Period
- Completed the protocol-planned duration of the study
- Did not complete the study with reasons for premature discontinuation of study
- Continuing the study
- Number of subjects who reached 24 weeks in the OLE Treatment Period
- Number of subjects who discontinued the OLE Treatment Period prior to 24 weeks and the reason for discontinuation

The primary reason for discontinuation of any of the study periods and terminating the study will be summarized. The details of the ‘other reason’ will be included in the listing of individual data.

In addition, the total number of subjects who were enrolled in each study phase, and the number and percentage of subjects in each of the disposition categories listed above will be displayed in a flowchart.

Changes in study visit schedules, missed visits/assessments, or patient discontinuation due to COVID-19 will be summarized by treatment arm. A listing of all subjects affected by COVID-19 related study disruption will be generated. Number of COVID-19 infected subjects and the

serious adverse events (SAEs) in those subjects will be summarized by treatment arm. (FDA, Guidance for Industry 2021).

6.2. Major (aka Important) Protocol Deviations

Protocol deviations occurring after subjects entered the study are documented during routine monitoring.

The number and percentage of subjects with major protocol deviations will be summarized by treatment arm and overall for the ITT analysis set. The details will be listed by patient. Major protocol deviations include the following:

- Subjects that received the wrong treatment
- Subjects that violated inclusion/exclusion criteria
- Subjects that received prohibited concomitant medication
- Subjects with missing baseline TSS
- Subjects with informed consent form violation
- Subjects with unplanned unblinding

7. DEMOGRAPHIC AND OTHER BASELINE CHARACTERISTICS

Descriptive statistics with respect to subject characteristics at baseline will be displayed for the ITT and the SAF analysis set, both by treatment group and overall. A summary of key demographic data and also a listing presenting demographic and baseline data per subject will be presented.

The following parameters will be summarized:

- Demographics (Age, sex)
- Discrepancies between the minimization/stratification factors recorded in the randomization system and those recorded in the eCRF
- Disease history
- Treatment history
- Transfusion and hemoglobin history
- Medical history
- Hospital, GP / family doctor, and urgent care history
- Physical exam
- Vital signs
- ECG data
- ECOG performance status (PS)

- Spleen length (palpation/ultrasound) and spleen volume (scan)
- Laboratory assessments
- PROs: MFSAF, EORTC QLQ-C30, PROMIS, PGIS, EQ-5D
- DIPSS
- JAK2, MPL, CALR mutational status

Summaries for selected parameters will also be made for those subjects entering the OLE Treatment Period of the study.

8. (PRIOR AND) CONCOMITANT TREATMENT

(Prior and) Concomitant medications will be classified according to World Health Organization Drug Dictionary.

The number and percentage of participants receiving a (prior or) concomitant medication will be displayed by anatomical main group (1st level of the Anatomical Therapeutical Chemical –ATC-classification) and chemical subgroup (4th level of the ATC classification) for the safety population.

Medications will be reported as prior when they start before the first day of study treatment.

Medications will be reported as concomitant in the following cases:

- when they (1) start before, on or after first day of study treatment and (2) continue afterwards (beyond first day of study treatment) or stop after the first day of study treatment;
- when they start and stop on the first day of study treatment.

Medications started before the first day of study treatment and continuing afterwards will be reported both as prior and concomitant.

Separate summaries will be presented for prior medications and concomitant medications.

An additional summary will be presented for concomitant medication taken for myelofibrosis indication during the RT Period. This summary would exclude any medication with a start date falling after the end of Week 24.

(Prior and) Concomitant medication summaries will be sorted alphabetically by generic term within ATC class.

9. EFFICACY EVALUATION

9.1. Formal Statistical Comparisons

The overall type I error for this trial is controlled at 5% (2-sided) for the primary endpoint and key secondary endpoints, by using a hierarchical testing procedure.

Only in the case the primary endpoint meets statistical significance at the primary analysis, the key secondary endpoints will be tested sequentially.

Formal comparisons of the primary and key secondary endpoints will be performed according to the following hierarchy:

Hierarchical Testing					
Test Order	**Endpoint**	**Testing**	**Criterion for significance**	**Testing***	**Criterion for significance**
1	MFSAF TSS 24 response (primary)	Superiority	$P \leq 0.05$		
2	TI 24 status	Non-inferiority	lower limit of 95% confidence interval on (MMB TI proportion) – 0.80*(DAN TI proportion) > 0	Superiority	$P \leq 0.05$*
3	SRR 24 (based on 25% reduction criterion)	Superiority	$P \leq 0.05$		
4	MFSAF TSS 24 change from baseline	Superiority	$P \leq 0.05$		
5	SRR 24 (based on 35% reduction criterion)	Superiority	$P \leq 0.05$		
6	Rate of no transfusion at Week 24	Superiority	$P \leq 0.05$		

* If non-inferiority is concluded for TI status, then the p-value associated with the test of superiority will also be calculated.

Additional secondary endpoints will also be examined, but not included in the hierarchical testing:

- Proportion of subjects with ≤ 4 RBC or whole blood units transfused during the 24-week RT Period estimated as crude proportion/test by a Cochran-Mantel-Haenszel (CMH) test or estimated via Kaplan-Meier methods and tested via a Wald test
- Cumulative transfusion risk for MMB versus DAN through the end of Week 24, measured by a proportional hazard recurrent events analysis
- Mean change in cancer-related fatigue (assessed by the EORTC QLQ-C30 fatigue domain) from baseline to each evaluation timepoint by Week 24 as tested via MMRM model.
- Mean change in physical function score (assessed by PROMIS) from baseline by Week 24 as tested via MMRM model.

The duration of MFSAF TSS response, the duration of TI response at Week 24, the characterization of safety, PRO and QoL endpoints, and the comparison of OS and LFS of subjects treated with MMB versus DAN will also be examined.

9.1.1. TI Non-inferiority Margin Rationale

While Danazol, the active control, is recommended as treatment for myelofibrosis-associated anemia by NCCN, only case series are available in the literature that describes MF-associated anemia benefit. This endpoint incorporates change in both hemoglobin levels and transfusion requirements [(1) transfusion cessation, in transfusion-dependent patients, and (2) an Hb increase >2 g/dl, in patients without transfusion dependency, both lasting for a minimum of 12 weeks at any time during therapy]. Clinical improvement was documented in 15 out of 50 patients; 30% response proportion, including 5 of 27 transfusion dependent patients achieving TI (18.5 %); Cervantes (2015).

This MMB verses DAN study, however, examines a TI endpoint, evaluating the ability to maintain or convert to transfusion independence for 12 weeks immediately prior to week 24. In the absence of more historical data on the Danazol treatment effect on this TI endpoint, a highly conservative value of 80% was chosen as the fraction of the control arm treatment effect that MMB arm must maintain in order to declare a non-inferiority on the TI endpoint.

9.2. General Statistical Methods

All efficacy analyses will be performed on the ITT population unless noted otherwise. Some sensitivity analyses will be performed on the PP population.

Binary outcomes will be described by proportions by treatment arm and compared with a CMH test stratified by baseline MFSAF TSS (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+), as recorded in the randomization system. For binary endpoints in the alpha-control hierarchy, the 95% exact binomial confidence interval (CI) will be generated for the per-arm proportion estimate. The difference between the two proportions will be estimated by Mantel-Haenszel common risk difference with the stratified Newcombe confidence interval unless specified otherwise (Xin, 2010). For subgroups and for endpoints not in the hierarchy, a non-stratified exact CI will be generated for the difference in treatment proportions. Discrepancies between the stratification values used in the randomization and the actual values will be summarized and sensitivity analysis using the actual values may be performed if more than 10% of randomized subjects have discrepancies.

Time to event outcomes (“survival times”) will be described by treatment arm using the Kaplan-Meier method. Subjects who have not had the event of interest at the time of the analysis will be censored at the time of the last follow up. Summary statistics will be provided by treatment arm in terms of the number of events, median and 95% CI and survival probabilities at specific time points (such as 1 year, 2 years, etc.). When comparing the survival curves of the two arms, a log-rank test stratified by baseline MFSAF TSS, baseline spleen length, and baseline RBC or whole blood units transfused, as recorded in the randomization system will be used. A stratified Cox regression model will be used to estimate the hazard ratio and its 95% CI. Follow-up

duration for OS and LFS endpoints will be summarized by reverse Kaplan-Meier method (Schemper, 1996).

Continuous outcomes collected at several post-baseline timepoints will be analyzed using a MMRM model, including factors for treatment arm and baseline MFSAF TSS, baseline spleen length, and baseline RBC or whole blood units transfused, as recorded in the randomization system.

Stratification factors will be used in the analysis as they were collected in the randomization system at allocation unless specified otherwise.

Any changes or adjustments to these general methods for particular endpoints will be specified below.

9.3. Primary Endpoint Analysis

MFSAF TSS response rate at Week 24 is the primary study endpoint. The MFSAF TSS response rate is defined as the proportion of subjects who achieve a ≥ 50% reduction in mean MFSAF TSS at Week 24, compared to baseline.

Baseline MFSAF TSS, as specified previously, will be defined relative to the date on which the baseline period for TSS was "triggered" on the ePRO handheld device. This date may or may not be the same as the date of the baseline clinic visit.

The baseline TSS will be computed as the mean of the TSS values generated on the date of the baseline period triggering and on the six days immediately following that triggering date, with the following caveats:

- If there is more than one TSS value generated on the date of baseline triggering, the last TSS value on that date will be used and other values from that date will be ignored.
- Any values that are generated on or after the date of first dosing will not be used in the computation of the baseline TSS.
- If fewer than four assessments are available for computing baseline for a subject, that subject's baseline will be considered missing.

The Week 24 TSS is defined as the average of the daily TSS from a consecutive 28-day period prior to Week 24. The consecutive 28-day period at Week 24 is defined as the latest eligible period of 28 consecutive days that has ≥ 20 available daily TSS; the last day of that 28-day period must be prior to or on the last RT Period participation date, have nonmissing daily TSS, and fall between Days 161 and 168, inclusive. If no such consecutive 28-day period with ≥ 20 available daily TSS is available or the last RT Period participation day was prior to Day 161, the Week 24 TSS will be considered missing.

The 7 domains of the MFSAF represent the seven symptoms of MF identified through existing patient- and clinician-based evidence to be the most relevant: fatigue (weariness, tiredness), night sweats, pruritus, abdominal discomfort, pain under the left ribs, early satiety, and bone pain. Each symptom domain is to be assessed on an 11 point numeric rating scale ranging from 0 to 10, with the TSS representing the sum of the scores across these seven domains, thus

representing a range of scores from 0 to 70, with a higher score corresponding to more severe symptoms.

For purposes of computing the proportion of subjects with at least a 50% reduction in TSS at Week 24, if the Week 24 TSS is missing, such a subject will be considered to be a non-responder (NRI approach).

In accordance with the prohibition of non-study active anti-MF therapy, TSS scores after receiving other active MF therapy (as defined in protocol section 5.3.3, followed by medical review) during the treatment periods will be excluded in determining MFSAF TSS response at Week 24.

The primary analysis of MFSAF TSS response will be performed on the ITT analysis set using a CMH test, stratified by: MFSAF TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+), as recorded for randomization. Primary inference will be based on the asymptotic p-value based on the Wald statistic from this CMH test.

The analysis will be repeated as a sensitivity analysis on the population of subjects excluding those with major GCP violation in their PRO data.

P-value from re-randomization test: P-value for the primary endpoint of TSS24 will be also calculated by re-randomization test as a sensitivity analysis. Ten thousand simulated trials will be generated such that the patients randomized in the study are re-randomized into two arms under the same study randomization method, so as to produce an empirical distribution of the test statistic (CMH test as described above) under the null hypothesis. Patients will be re-randomized in the same order they were originally randomized, taking into account stratification factors based on baseline TSS, spleen size, number of transfusion units, and investigative site. The empirical p-value is the frequency, calculated as total number of times out of 10,000, that a simulated test statistic is strictly larger (ie, more extreme) than the test statistic on the observed data using the original randomization allocation. For a given analysis, the summary table will display the asymptotic parameter estimate and confidence interval, plus the empirically estimated p-value (“re-randomization p-value”).

In order to evaluate the impact of the stratification on results, and more specifically the impact of empty strata on the power of the test, a sensitivity analysis of the primary comparison will be performed using Pearson chi-square test (unstratified).

In case more than 10% of subjects cross over from DAN to MMB treatment prior to the end of Week 24, a sensitivity analysis will be conducted on data collected prior to that treatment switch, ie, carrying forward the MFSAF TSS value from latest timepoint under DAN treatment for those subjects crossing over.

In order to assess the results’ robustness to missing data handling method, the following sensitivity analyses will be performed:

- The analysis described above for the primary analysis will be repeated on all available data without any imputation for missing values ("observed cases" approach).

- The analysis described above will be performed using the LOCF approach, ie, carrying forward the latest non-missing MFSAF TSS value.
- A MI procedure will be used to handle missing scores (MI step performed by treatment group and strata used at randomization), and a logistic regression model will be applied to the derived response status (after MI step), including fixed categorical effects for MFSAF TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+).

MFSAF TSS response rate at Week 24 will also be summarized by treatment arm in all subgroups defined in Section 5.4.

TSS at each of the visit timepoints is defined as the mean of the readings taken on the 28 days on and prior to the index date of the visit, under the condition that at least 20 of the 28 days must possess a TSS value, lest the value be considered missing.

9.4. Secondary Efficacy Endpoints Analysis

9.4.1. Week 24 TI Status (key secondary endpoint)

The first key secondary endpoint is the proportion of subjects with TI status in the terminal 12 weeks of the 24-week RT Period (ie, TI Status at Week 24).

Transfusion independent status at week 24 is defined as:

- no RBC or whole blood transfusions (except in the case of clinically overt bleeding) for ≥ 12 weeks (immediately prior to the Week 24 visit), and
- no central or local Hgb level < 8 g/dL during the same interval (except in the case of clinically overt bleeding) and
- at least 2 Hgb assessments in the 12 week (84 day) period and
- the time between the earliest and latest Hgb assessments in the 12 week period is at least 42 days and
- the Week 24 visit falls no later than day 176 and
- the Week 24 visit falls no earlier than day 161.

The Week 24 TI status will be analyzed on the ITT analysis set, as well as on the PP analysis set for sensitivity purposes.

In accordance with the prohibition of non-study active anti-MF therapy, patients receiving other active MF therapy during the RT Period (as defined in protocol section 5.3.3, followed by medical review) will be set to “Not TI” for TI status at Week 24. Subjects without TI-status at Week 24 (including missing TI status) will be set to “Not TI”.

Non-inferiority of MMB will be based on synthesis approach (FDA Guidance, 2016) where the treatment effect of the active control (DAN here) is not pre-specified, but the percentage of the

active control effect to be preserved is specified. 80% of the response rate in the DAN arm should be preserved in the MMB arm to declare non-inferiority.

A stratum-adjusted 2-sided 95% CI based on Koch et al (1989), will be calculated for the difference:

Non-inferiority difference = p(MMB) – 0.80 x p(DAN)

where p(MMB) is the proportion of subjects with TI status in the MMB arm and p(DAN) is the proportion of subjects with TI status in the DAN arm.

If the lower bound of the CI is greater than 0, MMB will be declared to be non-inferior to DAN.

The method above employs, in its computation, the sample size in each treatment less one in the denominator for each stratum cell.

If there is any stratum treatment cell with fewer than two subjects, the strata for the number of baseline transfusions will be collapsed from three strata to two strata by combining the 0 units and 1-to-4 units strata or the 1-to-4 units and 5+ units strata, depending on which combination provides a more uniform spread of subjects across strata.

If this strata-adjustment procedure still fails to produce at least 2 subjects in each stratum, then all the strata will be combined into one stratum and the method will be applied with the single stratum.

If non-inferiority is concluded, then the p-value associated with the test of superiority will also be calculated, using a CMH test stratified by baseline MFSAF TSS, baseline spleen length, and baseline RBC or whole blood units transfused.

The 80% of DAN response threshold represents a conservative margin of approximately 4 percentage points under the expected DAN response proportion of 21%. The expected response proportion for DAN is based on available clinical literature for DAN treatment in MF with consideration of the patient population to be enrolled in this study.

In addition, the proportion of subjects with TI conversion status at Week 24 (defined, for subjects who were TD at baseline, as a switch to TI status at Week 24, TI status being defined earlier) will be computed and compared between treatment arms using same methodology as described above (CMH test) in the subgroup of subjects with transfusion dependent (TD) status at baseline.

The proportion of subjects with TI status at Week 24 will also be summarized by treatment arm in all subgroups defined in Section 5.4. Shift tables showing baseline and Week 24 TI status will be generated.

A supplemental analysis of TI status will be performed considering subjects who crossover from DAN to MMB before Week 24 TI status evaluation as non-responders (“Not TI”) at Week 24.

9.4.2. Splenic Response Rate at week 24 (key secondary endpoint)

Splenic response rate is defined as the proportion of subjects who have splenic response (reduction in spleen volume of ≥ 25% from baseline) at the end of Week 24. Scans taken no more than 10 days after the beginning of OL treatment may be considered as valid scans for

assessment. Scans taken more than 10 days after the beginning of OL treatment will coerce the splenic response outcome to non-responder.

The primary analysis of splenic response will be performed on the ITT analysis set, as well as on the PP analysis set for sensitivity purposes, using a CMH test, stratified by baseline MFSAF TSS, baseline spleen length, and baseline RBC or whole blood units transfused.

In accordance with the prohibition of non-study active anti-MF therapy, subjects receiving other active MF therapy during the RT Period (as defined in protocol section 5.3.3, followed by medical review) will be set to non-responder for SRR response at Week 24.

Subjects with a missing evaluation at baseline or Week 24 and subjects with differing modalities of spleen scanning at baseline and Week 24 (ie, CT at baseline but MRI at Week 24 or vice versa) will be considered as non-responders for SRR (NRI approach).

SRR at Week 24 will also be summarized by treatment arm in all subgroups defined in Section 5.4.

Furthermore, an additional splenic response rate at Week 24 will also be computed with a reduction criterion of 35% replacing the above 25% reduction criterion.

Spleen volume measurements at Week 48, and their corresponding changes from baseline, will be summarized for subjects with Week 24 spleen response. Two sets of summary statistics, one for the 35% reduction criterion and one for the 25% reduction criterion, will be produced.

Palpation-based spleen size measurements will be summarized at baseline and Week 24 and the proportion of responders, based on this methodology, will be computed for each treatment group.

9.4.3. MFSAF TSS Change from Baseline

TSS change from baseline at week 24 is a key secondary endpoint.

Individual symptom scores, TSS, as well as their change and percent change from baseline, will be summarized using descriptive statistic by treatment arm at each evaluation timepoint. Mean TSS and mean change from baseline in TSS will be displayed graphically over time by treatment arm.

Change from baseline and percent change from baseline in TSS at Week 24 will be presented in an ECDF plot, by treatment arm, with the TSS value on the horizontal axis and the proportion of score less than or equal to each TSS value on the vertical axis. Separate curves will denote the two treatment groups. Week 24 TSS and change-from-baseline TSS will also be plotted as scatter plots with baseline TSS on the horizontal axis and Week 24 value on the vertical axis, using side-by-side plots to show the two treatment groups.

Changes from baseline in TSS at Weeks 4, 8, 12, 16, 20 and 24 will be analyzed using a MMRM model, using all available subject-level derived scores (daily data summarized for each 4-week period) on the ITT analysis set. In addition to terms for treatment arm, timepoint (week) and treatment-by-week interaction, the model will include (as fixed effects) the same factors as those used in the primary endpoint analysis: MFSAF TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+). LSMEANS

(SE) will be presented by treatment arm at each timepoint. Confidence intervals (95%) for the treatment differences will be derived from the model at each analysis timepoint.

The p-value for the LS mean difference between the two arms at week 24 will be used for the inference of this key secondary endpoint.

Example SAS code:

```
Proc Mixed Data=TSS;
Class SUBJECT VISIT TRT
      BSL_STRATUM SPLEEN_STRATUM RBC_STRATUM;
Model TSSCFB = TRT VISIT TRT*VISIT
               BSL_STRATUM SPLEEN_STRATUM RBC_STRATUM
               / DDFM=KenwardRoger;
Repeated VISIT / Subject= SUBJECT Type=UN R RCorr;
Run;
```

MFSAF TSS change from baseline to Week 24 will also be summarized by treatment arm in all subgroups defined in Section 5.4.

Validity of Missing at Random Assumption: A control-based multiple imputation under a missing not at random assumption will assess of the extent to which the MAR assumption in the above estimand is robust. Control-based multiple imputation uses a pattern mixture model framework and uses the distributions of responses from the control arm to impute missing for both treatment arms. In this analysis, missing observations in the treatment arm are assumed to have the statistical behavior of the control arm after dropout, that is, the treatment effect becomes equivalent to control subjects (Cro, 2020).

Validity of the multivariate normality assumption will also be investigated.

9.4.4. Duration of Week 24 MFSAF TSS Response

The duration of the Week 24 TSS response will be assessed to the end of Week 48.

For subjects who achieve a Week 24 TSS response, the duration of response is defined as the number of days from the start of the initial 28-day period (during the 24-Week RT Period) in which the subject has a ≥ 50% reduction from baseline TSS to the first day of the 7-day assessment that determines the mean TSS for the 28-day period during which the subject's TSS equals or exceeds their baseline value. Subjects (TSS24 responders) will be censored at the first day of the last 7 day period if there is no mean TSS of the 7 day assessments which equals or exceeds the baseline value. TSS will be assessed during the last 7 days (± 7 days) of each month during the open label extended treatment period until Week 48. During the OLE Treatment Period of the study, averages will be computed for any TSS data available; there is no non-zero minimum number of days required to compute a weekly TSS value during the OLE Treatment Period.

By definition, the duration of TSS response will be analyzed in the subgroup of subjects with TSS response status at Week 24.

Kaplan-Meier methods will be used to estimate the median duration of TSS response, as well as its 1st and 3rd quartiles, in each treatment arm. Results will be displayed in a Kaplan-Meier curve.

A supplemental analysis of duration of the Week 24 TSS response may also be performed:

- Subjects crossing over to another treatment before loss of response will be censored on the first day of the last seven-day period during which TSS was collected before crossover.
- Subjects crossing over before achieving response will not be considered in this analysis.

9.4.5. Duration of Week 24 TI Status

Duration of TI status will be assessed to the end of Week 48 in all subjects with TI status at the end of Week 24.

For subjects who achieve TI status at Week 24, duration of TI is defined as the number of days from (a) the first day of a 12-week period that satisfies the 12-week TI status definition, to (b) the first RBC or whole blood transfusion or Hgb level < 8 g/dL (except in the case of clinically overt bleeding) (assessed until end of Week 48). Duration of TI24 will be censored at date of the last Hgb assessment not less than 8 g/dL. By definition, the duration of TI will be analyzed in subjects with TI status at Week 24 (subset of ITT data set).

Kaplan-Meier methods will be used to estimate the median duration of TI, as well as its 1st and 3rd quartiles, in each treatment arm. Results will be displayed in a Kaplan-Meier curve.

A supplemental analysis of duration of Week 24 TI status will be performed:

- Subjects crossing over to another treatment (DAN to MMB) before loss of response will be censored on the date of the last Hgb assessment not less than 8 g/dL prior to the crossover date.
- Subjects crossing over before achieving response will not be considered in this analysis.

Duration of TI will also be evaluated in the subgroup of subjects who were TD or TR, or TD at baseline (and who were also TI at Week 24).

9.4.6. TD Status

The proportion of subjects with TD status at the end of Week 24 will be compared between groups using a CMH test stratified by baseline MFSAF TSS, baseline spleen length, and baseline RBC or whole blood units transfused, in both the ITT analysis set and the subset of subjects who were TI at baseline.

TD status at Week 24 is defined as requirement of ≥ 4 RBC or whole blood units (note units and not simply transfusions) in an 8-week period immediately prior to the end of Week 24. Only RBC or whole blood transfusions given when Hgb levels are ≤ 9.5 g/dL are counted towards TD but that RBC or whole blood transfusions given for clinically overt bleeding or accident/injury are not counted towards TD.

Transfusions with unknown status relative to clinically overt bleeding should be considered to be not due to clinically overt bleeding. Transfusions without Hgb levels should be considered to be due to Hgb level of ≤ 9.5 g/dL.

Transfusion dependent status at a given timepoint of interest is defined as:

- four or more units of RBC or whole blood transfusions (except in the case of clinically overt bleeding) in the 8 weeks (immediately prior to the timepoint of interest) with each such transfusion in response to a Hgb assessment of ≤ 9.5 g/dL, and
- at least 2 Hgb assessments in the 8 week (56 day) period and
- the time between the earliest and latest Hgb assessments in the 8 week period is at least 28 days.

If a subject's TI status is determined to be "TI" at a given timepoint, then that subject will have TD status set to "not TD".

A subject with last RT Period participation date (the latest of recorded lab and procedure dates and the Week 24 visit date prior to OL start) prior to day 155 and who does not present with evidence of being TD will be classified as "Unknown".

A subject who is determined to be neither TI nor TD but has sufficient evidence to preclude being classified as Unknown will be classified as "TR" (transfusion requiring).

Shift tables demonstrating shifts from baseline transfusion status (TI/TR/TD) to Week 24 transfusion status (TI/TR/TD/Unknown) will be produced for each treatment group.

9.4.7. Hemoglobin Responses

Hemoglobin responses are defined as increases of (a) ≥ 1, (b) ≥ 1.5, or (c) ≥ 2 g/dL from baseline in Hgb, as measured at any point in time during the observation period. Two observation periods are being considered: (1) the period during the last 12 weeks of the RT Period and (2) the entire 24-week RT Period. For each of the two observation periods, a subject will be classified as having an Hgb response (a, b, or c) if there exists at least one central Hgb reading sufficiently higher than the baseline Hgb reading during that observation period. Hemoglobin values that occur within 4 weeks after a transfusion will be excluded.

The proportion of responders will be presented by treatment arm, on the ITT analysis set and on the subgroups of subjects who were TI at baseline, using the NRI approach.

Comparisons will be performed using a CMH test, as described for the primary and key secondary analyses.

9.4.8. RBC or Whole Blood Units Transfused

The proportion of subjects with zero RBC or whole blood units transfused (key secondary endpoint) and the proportion of subjects with ≤ 4 RBC units transfused during the 24-week RT Period will be evaluated as secondary endpoints.

Comparison of two arms will be performed using a stratified CMH test, on the ITT analysis set, using the NRI approach for each of these two proportions. MFSAF TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+) will be used as strata in the analysis.

Time-to-first RBC or whole blood transfusion will be estimated by treatment group using Kaplan-Meier methods, as well as time-to-third and time-to-fifth transfusions. Kaplan-Meier methods applied to time-to-first transfusion will be used to estimate the proportion of subjects in each treatment group who have zero units transfused in the first 24 weeks following first day of dosing and the standard error of the proportion.

Comparison of the two treatment groups at specific timepoints (Week 12 and Week 24) will be carried out via Wald tests of the difference in proportions, scaled by the estimated standard error of the difference.

Comparison of the treatments with regard to time-to-first will also be carried out via a stratified log-rank test of the time-to-first unit transfused employing the randomization strata. Stratified Cox regression will be performed to estimate treatment effect measured as hazard ratio.

The cumulative RBC or whole blood transfusion risk (ie, accumulating all transfusions per subject as repeated events over time) will be evaluated and compared between treatment arms using a zero-inflated negative binomial (ZINB) model, adjusted for baseline MFSAF TSS, baseline spleen length, and baseline RBC or whole blood units transfused. This model will employ time on study as an offset in the model to account for disparate durations in the study.

Additionally, a proportional rate/mean model will be used as an exploratory analysis, treating each transfusion as a recurrent event, including follow-up time after the last transfusion, censoring patients at their last follow-up date. To account for the number of units used in a transfusion, a weighting based on number of units transferred will apply to the transfusion event.

9.4.9. Overall Survival

Overall survival is defined as the interval from the first study drug dosing date (or randomization date for subjects who did not receive treatment) to death from any cause. Subjects without a documented death at the time of analysis will be censored on the last date known to be alive.

The survival functions will be estimated by the Kaplan-Meier method. Treatment difference will be assessed by the stratified log-rank test, and its magnitude will be assessed by the Hazard Ratio (HR) from a stratified Cox proportional hazard model. Stratification factors will be: MFSAF TSS baseline score (≥ 22 versus < 22), baseline palpable spleen length below the LCM (≥ 12 cm versus < 12 cm), and baseline RBC or whole blood units transfused in the 8-week period prior to randomization (0, 1-4, and 5+).

Survival rates will be estimated by treatment arm at selected time points such as Week 24, Week 48, Week 96 and Week 204.

OS will be analyzed on the ITT analysis set. The final analysis of OS will be performed at completion of follow-up data collection (end of the study). However, OS will also be evaluated at each data cut-off point used for the various analyses, ie, at the time of Week 24 primary analysis, and at the time all subjects have reached end of Week 96. These interim OS results, obtained from immature data, will have to be interpreted with caution.

9.4.10. Leukemia-Free Survival

Leukemia-free survival is defined as the interval from first study drug dosing date (or randomization date for subjects who did not receive treatment) to any evidence of leukemic transformation and/or death (from any cause). Patient without evidence of leukemia or death at the time of analysis will be censored at the date of last assessment of their disease.

Leukemia-free survival will be analyzed using the same method as described for OS, on the ITT analysis set. Leukemia-free survival will be evaluated at same timepoints as described above for OS.

9.4.11. Disease-Related Fatigue (MFSAF)

The analysis of this item score will be performed using the same methods employed on the TSS score (Section 9.4.3) including the same descriptive statistics, ECDF plots and MMRM model. The ITT analysis set will be used for these analyses; a subset of the ITT analysis set made up of those subjects with non-missing baseline assessments may be used for further exploratory analyses.

9.4.12. Cancer-Related Fatigue (EORTC QLQ-C30)

The analysis of this item score will be performed using the same methods employed on the TSS score (Section 9.4.3) including the same descriptive statistics, ECDF plots and MMRM model. The ITT analysis set will be used for these analyses; a subset of the ITT analysis set made up of those subjects with non-missing baseline assessments may be used for further exploratory analyses.

9.4.13. Physical Function (PROMIS)

The analysis of this item score will be performed using the same methods employed on the TSS score (Section 9.4.3) including the same descriptive statistics, ECDF plots and MMRM model. The ITT analysis set will be used for these analyses; a subset of the ITT analysis set made up of those subjects with non-missing baseline assessments may be used for further exploratory analyses.

9.4.14. EQ-5D

The analysis of this item score will be performed using the same methods employed on the TSS score (Section 9.4.3) including the same descriptive statistics, ECDF plots and MMRM model. The ITT analysis set will be used for these analyses; a subset of the ITT analysis set made up of those subjects with non-missing baseline assessments may be used for further exploratory analyses.

9.4.15. MF-8D

The analysis of this item score will be performed using the same methods employed on the TSS score (Section 9.4.3) including the same descriptive statistics, ECDF plots and MMRM model. The ITT analysis set will be used for these analyses; a subset of the ITT analysis set made up of those subjects with non-missing baseline assessments may be used for further exploratory analyses.

9.5. Exploratory Endpoints Analysis

9.5.1. MFSAF TSS in Subgroups

The MFSAF TSS response proportions at Weeks 4, 8, 12, 16, 20 and 24 will be summarized by treatment arm, on the subgroups of TD subjects, TI subjects, and non-TD subjects.

9.5.2. Joint Distribution of TSS, TI, and Spleen Response at Week 24

The joint distribution of subjects relative to TSS, TI, and Spleen response at week 24 will be computed for each treatment group.

9.5.3. Symptomatic Splenic Progression

Confirmed splenic progression is defined as meeting either of the following criteria:

a. Increase in spleen volume ≥ 25% from baseline, or
b. Symptomatic splenic progression defined as meeting both of the following criteria:
   - Worsening of early satiety with weight loss ≥ 5% from baseline, or worsening of sustained splenic pain following either:
   - For subjects not previously receiving narcotic pain medication, initiation of new narcotic pain medication use for ≥ 5 days, or
   - ≥ 50% increase from baseline in the daily dose of narcotic pain medication for ≥ 5 days.
   - Increase in spleen volume ≥ 15% from baseline

Confirmed splenic progression is recorded as a check box upon exit from the RT Period of the study.

A comparison of the proportion of subjects in the two treatment groups with confirmed splenic progression will be produced using CMH methods.

Time to spleen volume increase of at least 15% during the first 24 weeks of the study will be determined for each such subject. Subjects without such an increase will be censored at the date of the spleen scan that confirms the absence of such an increase. Kaplan-Meier methods will be used to determine the time to 15% spleen volume increase for the two treatment groups.

9.5.4. Correlation between Responses and Exploratory Endpoints

The prognostic and predictive potential of JAK2, MPL, CALR mutational status and other somatic mutations will be evaluated using, as clinical response variables, MFSAF TSS change from baseline to Week 24, TSS response at Week 24, TI responses at Week 24, and SRR at Week 24, as applicable.

Mutation statuses as collected at baseline will be presented by treatment arm.

An ANOVA model using covariates for treatment arm, stratification factors as used at randomization, and baseline mutation status will be fit to the change from baseline in MFSAF TSS to evaluate the prognostic potential of each mutational status. Additionally, a similar model

will be used which will include an interaction between treatment and the mutational status, in order to assess the predictive potential of each mutational status.

For TSS response and TI responses, a similar approach will be used with a logistic regression.

Baseline disease characteristics (DIPSS, etc.) and other baseline characteristics may also be explored within an ANOVA framework.

9.5.5. Healthcare Utilization Requirements

The following patient healthcare resource utilization at each visit will be summarized with the number and percentage of subjects for the ITT population and by treatment arm:

- Transfusions; and
- Healthcare category and healthcare encounter (eg, hospital, GP/family doctor, and urgent care visits).

These endpoints will also be summarized cumulatively for the RT Period and for the OLE Treatment Period.

9.5.6. Other PRO Endpoint Analysis

Descriptive responder analyses, longitudinal responder analyses, and time-to-event and duration analyses for MFSAF, EORTC QLQ-C30, EQ-5D, and PROMIS Physical Function endpoints will be also performed. The details of these analyses are described in a separate document (see appendix).

9.5.7. Baseline Ferritin as Potential Biomarker

The baseline ferritin level will be used to create subgroups within and across which transfusion independence response will be examined to assess the potential of baseline ferritin to be used as a potential biomarker for anemia benefit. Other analysis to examine the baseline ferritin level as predictive or prognostic biomarker for transfusion independence rate also may be applied.

10. SAFETY EVALUATION

10.1. Extent of Exposure

Duration of exposure, number of doses prescribed and taken, number of treatment interruptions, dose intensity and relative dose intensity over the 24-week RT Period, as well as over the OLE Treatment Period, will be summarized using descriptive statistics by treatment arm.

Treatment modifications and interruptions will be presented in detail in listings of individual data.

10.2. Adverse Events

Adverse events will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) and will be graded according to the National Center Institute Common Terminology Criteria for

AEs (NCI-CTCAE criteria v5.0, 2017). Adverse events will be analyzed according to their type, incidence, severity and relationship to the study treatment (as assessed by investigator).

Adverse events will be tabulated if they are treatment-emergent. Treatment-emergent Adverse Events (TEAEs) are defined as AEs occurring or worsening on or after the first dose of study treatment, and no later than 30 days after last dose of study drug received. Missing or partial AE start date will be estimated in order to include events in summary tables in case of doubt (see Section 5.7 for more details). Adverse events that are not categorized as treatment-emergent will only be displayed in listings of individual data.

Treatment-emergent AEs will also be presented according to the study period in which the onset date of the event falls. Separate tabulations and listings will be presented for:

- The 24-week RT Period (tabulation by treatment arm MMB versus DAN)
- The complete OLE Treatment Period

In the OLE Treatment Period, tabulations will be presented by treatment arm from RT Period (MMB versus DAN), with an additional distinction made in the DAN treatment arm between those subjects continuing with DAN treatment up to Week 48 and those switching to MMB treatment (DAN to DAN versus DAN to MMB).

Tabulations in the OLE treatment will therefore be presented by the following treatment groups:

- MMB - MMB
- DAN - DAN assuming any subject elects to continue DAN in the OLE Treatment Period
- DAN - MMB (switchers)

A summary table will present by treatment arm the number and percentage of patients with at least one:

- TEAE
- TEAE related to the study treatment
- TEAE with grade of at least 3 that are related to the study treatment
- TEAE leading to permanent discontinuation of the study treatment
- TEAE leading to study treatment interruption and/or dose modification
- Serious TEAE
- Serious TEAE that are at least grade 3
- Serious TEAE that are related to study treatment
- Non-serious TEAE
- Grade 3 or higher TEAE
- Grade 3 or 4 (severe) TEAE
- Fatal TEAE

- Fatal TEAE that are related to study treatment

Presentations for Serious and Non-serious TEAEs will include proportion of subjects in each arm, total number of occurrences in each arm, total number of occurrences that are related to each treatment, number of deaths in each arm, and the number of deaths related to each treatment.

In addition, tabulations of the number of patients who experienced TEAEs as well as severity of the events will be presented by system organ class (SOC) and preferred term (PT). Patients will only be counted once for each preferred term. In case a patient experienced the same event more than once, the worst severity will be presented.

The following tabulations will be presented:

- All TEAEs
- TEAEs leading to permanent discontinuation of the study treatment
- TEAEs related to the study treatment

Tabulations by SOC and PT will be presented by decreasing total frequencies (across treatment arms).

Tabulations on the complete OLE Treatment Period will be repeated replacing number of subjects with subject incidence per 100 patient-years of exposure. Listings of all AEs by subject will be provided, flagging the ones that are treatment-emergent, including the patient identifier, verbatim, preferred term, duration of the event, severity, relationship, period and date of onset.

10.3. Deaths and Serious Adverse Events

Serious adverse events, fatal AEs and NCI/CTC grade 3/4 toxicities will be provided in listings of individual data, irrespectively of the fact that they are treatment-emergent or not.

The number of deaths will be tabulated together with the primary cause of death.

10.4. Clinical Laboratory Determination

Worst-by-subject CTC grade over the RT Period will be presented by treatment arm for:

- hematological parameters (RBC, Hgb, hematocrit, platelets, white blood cells, neutrophils, lymphocytes, blast count)
- hepatic parameters (AST, ALT, ALP, total bilirubin, Gamma-GT, LDH) and creatinine
- other chemistry parameters (sodium, potassium, chloride, calcium, phosphorus, albumin, alkaline reserve, prothrombin rate)

Shift tables of grade at baseline versus worst grade per patient during RT Period will be created for all these parameters, by treatment arm.

Similar tables will be presented over the OLE Treatment Period.

Additionally, a tabulation of the number and percentage of subjects with any liver function abnormalities over the entire treatment period will be presented by treatment arm.

Liver function abnormalities will be defined by the following criteria:

- ALT/SGPT > 3 × ULN and concurrent (or subsequent within 21 days) total bilirubin > 2 × ULN
- AST/SGOT > 3 × ULN and concurrent (or subsequent within 21 days) total bilirubin > 2 × ULN
- ALT/SGPT and/or AST/SGOT > 3 × ULN and concurrent (or subsequent within 21 days) total bilirubin > 2 × ULN
- ALT/SGPT and/or AST/SGOT > 3 × ULN and concurrent (or subsequent within 21 days) total bilirubin > 2 × ULN and ALP < 1.5 × ULN

The grading of the laboratory parameters will only be determined based on laboratory values and not on any symptoms or concomitant medications.

Laboratory measurements collected after end of treatment (follow-up period) will only be displayed in listings of individual data.

ANCOVA models will be used to estimate differences in the two treatments in mean Hgb and platelet levels at Week 24, using baseline values as covariates.

Plots of estimated mean values for each treatment over time will be generated to examine effects due to crossover from the RT to OLE Treatment Periods.

10.5. Body Weight

Body weight, and change from baseline, will be summarized per collection timepoint and by treatment group over the entire study duration.

10.6. Spleen Measurements

Spleen volume measurements at baseline, Week 24 and Week 48 will be summarized by treatment arm, as well as spleen volume changes and percent changes from baseline at Week 24 and Week 48.

Spleen volume percent change from baseline at Week 24 will be displayed in a waterfall plot, separately for each treatment arm, or in ECDF plots. These plots will be prepared using the observed case approach, ie, using all observed Week 24 measurements.

10.7. ECOG Performance Status

Shifts from baseline to worst measurement over treatment period in the PS grade will be presented by treatment arm. Performance status grades by subject over time will be displayed in a listing.

11. REFERENCES

Cervantes F, Isola IM, Alvarez-Larrán A, Hernández-Boluda JC, Correa JG, Pereira A. Danazol therapy for the anemia of myelofibrosis: assessment of efficacy with current criteria of response and long-term results. Ann Hematol. 2015 Nov;94(11):1791-6.

Cro S, Morris TP, Kenward MG, Carpenter JR. Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. Statistics in Medicine. 2020;39(21):2815-2842.

Fleiss JL, Tytun A, Ury HK. A Simple Approximation for Calculating Sample Sizes for Comparing Independent Proportions. Biometrics. 1980;36(2):343–346.

Food and Drug Administration. Guidance for Industry: Non-Inferiority Clinical Trials to Establish Effectiveness. 2016 Nov.

Han B, Enas NH, McEntegart D. Randomization by minimization for unbalanced treatment allocation. Statistics in Medicine. 2009;28(27):3329-3346.

Koch GG, Carr GJ, Amara IA, et al. Statistical Methodology in the Pharmaceutical Sciences; chapter in Categorical Data Analysis. CRC Press. 1989.

Pocock SJ and Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics. 1975 Mar;31(1):103-15.

Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. Controlled Clinical Trials.1996 Aug;17(4):343-346.

12. APPENDIX

ADDITIONAL PRO ANALYSIS PLAN

This document includes the plan for the exploratory categorical and time to event analyses of the following patient reported outcome (PRO) endpoints in addition to the analyses planned in the Statistical Analysis Plan V1.0 of the SRA-MMB-301 study, as referred in the Section 9.5.6 of the document.

- Total Symptom Scores (TSS) and 7 item scores of MFSAF
- EORTC QLQ-C30 5 functional scales, 3 symptom scales, 1 global health status / QoL scale, and 6 single items
- PROMIS short form 10b physical function score and 4 additional single items

This document also includes the detailed scoring derivation of the two study PRO endpoints, EQ-5D index score and MF-8D score. The analysis methods for these two endpoints are described in Section 9.4 of the Statistical Analysis Plan V1.0.

1. GENERAL ANALYSIS DEFINITIONS

1.1. Continuous Response Calculations

The following PRO questionnaires are collected in the study. The description and assessment schedule for the questionnaires are located in the SRA-MMB-301 Clinical Trial Protocol V2.0 (Section 8.2).

1.1.1. MFSAF

Each of the 7 symptom domains is to be assessed on an 11-point numeric rating scale ranging from 0 to 10, with the TSS representing the sum of the scores across these seven domains, thus representing a range of scores from 0 to 70, with a higher score corresponding to more severe symptoms. If any of the 7 items has a missing score, then TSS is set to missing.

1.1.2. EORTC QLQ-C30

Five functional scales, three symptom scales, a global health status / QoL scale, and six single items will be derived from the 30 component items of QLQ-C30 v3.0 per the scoring manual (https://www.eortc.org/app/uploads/sites/2/2018/02/SCmanual.pdf).

The derivation algorithm for each scale score is as below:

For all scales, the *RawScore*, *RS*, is the mean of the component items:

$$RawScore = RS = (I_1 + I_2 + ... + I_n)/n$$

Then for **Functional scales**:

$$Score = \left\{1 - \frac{(RS-1)}{range}\right\} \times 100$$

and for **Symptom scales / items** and **Global health status / QoL:**

$$Score = \{(RS-1)/range\} \times 100$$

where I_i is the i-th item score included in the scale (among the total n items) and range is the difference between the maximum and minimum possible values of RS. The following table includes the range for each scale ("Item range" column).

The following table includes the item(s) included for each scale:

Table 1: Scoring the QLQ-C30 version 3.0

	Scale	Number of items	Item range*	**Version 3.0** Item numbers	Function scales
Global health status / QoL					
Global health status/QoL (revised)†	QL2	2	6	29, 30	
Functional scales					
Physical functioning (revised)†	PF2	5	3	1 to 5	F
Role functioning (revised)†	RF2	2	3	6, 7	F
Emotional functioning	EF	4	3	21 to 24	F
Cognitive functioning	CF	2	3	20, 25	F
Social functioning	SF	2	3	26, 27	F
Symptom scales / items					
Fatigue	FA	3	3	10, 12, 18	
Nausea and vomiting	NV	2	3	14, 15	
Pain	PA	2	3	9, 19	
Dyspnoea	DY	1	3	8	
Insomnia	SL	1	3	11	
Appetite loss	AP	1	3	13	
Constipation	CO	1	3	16	
Diarrhoea	DI	1	3	17	
Financial difficulties	FI	1	3	28	

* *Item range* is the difference between the possible maximum and the minimum response to individual items; most items take values from 1 to 4, giving *range* = 3.
† (revised) scales are those that have been changed since version 1.0, and their short names are indicated in this manual by a suffix "2" – for example, PF2.

The items included in each scale with the value range of 6 or 3 (Table 1) are converted to the functional/symptom scales scores from 0 to 100. A high score represents a higher response level. Thus, a high score for a functional scale and the global health status/QoL represents a high / healthy level of functioning and global health status, but a high score for a symptom scale / item represents a high level of symptomatology / problems.

If at least half of the component items are non-missing, the above algorithm will be applied for the available item scores. Otherwise, the scale's score is set to missing.

1.1.3. Physical Function (PROMIS)

PROMIS Short Form v2.0 Physical Function 10b is the instrument implemented in this study with 10 questions, and includes an additional 4 questions relating to physical function from the PROMIS item bank. The overall Physical Function (PF) total raw score is the sum of the 10 scores. (https://sites.duke.edu/centerforaging/files/2019/02/PROMIS-Physical-Function-Scoring-Manual.pdf). Responses to the 4 additional questions will be analyzed individually as reported.

1.1.4. EQ-5D

EQ-5D VAS score and index score are exploratory endpoints of the study.

EQ-5D VAS score is the value entered for the last question on the questionnaire asking the health state of the evaluation date on the scale of 0 (worst health) and 100 (best health).

EQ-5D index score is calculated from the 5 domain scores (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) with the scale of 1 (best score) to 5 (worst score). The index score is calculated as 1 + sum of the "value set" corresponding to each domain scale score. The estimates of Model 1 from Table 2 (Pickard, 2019), will be used for the value sets. Domain are notated by acronym, i.e. MO for mobility, SC for self-care, UA for usual activities, PD for pain/discomfort, AD for anxiety/depression, and thus MO2 in the "Dimension/Level" column of Table 2 corresponds to a patient selecting "2". For example, if a patient selected 22222 for the 5 domains, the index score becomes 1 -0.096-0.089-0.068-0.060-0.057 = 0.63. The digits of the five dimensions (e.g., 22222 in this example) are referred to as the patient's "health state" per the questionnaire. Note that the value set does not include the value for scale=1 as an answer of 1 corresponds to perfect health for the domain. If 1 is selected for a domain, 0 is subtracted for that domain. For example, the index score for a selection of 21111 for the 5 domains becomes 0.904 = 1-0.096. A higher EQ-5D index score represents the better health status.

Table 2: Value Set Estimates to Generate EQ-5D Index Score

Dimension/level	Model 1: cTTO (Tobit with heteroscedasticity, censored at −1, RE) (preferred model)			Model 2: DCE (mixed logit, rescaled to censored cTTO mean values)			Model 3: hybrid (Tobit with heteroscedasticity, censored at −1, conditional logit)		
	Estimate	SE	*P* value	Estimate	SE	*P* value	Estimate	SE	*P* value
MO2	−0.096	0.015	<.0001	−0.092	0.011	<.0001	−0.077	0.009	<.0001
MO3	−0.122	0.016	<.0001	−0.090	0.015	<.0001	−0.102	0.012	<.0001
MO4	−0.237	0.018	<.0001	−0.232	0.016	<.0001	−0.247	0.012	<.0001
MO5	−0.322	0.016	<.0001	−0.331	0.021	<.0001	−0.364	0.012	<.0001
SC2	−0.089	0.014	<.0001	−0.079	0.011	<.0001	−0.068	0.009	<.0001
SC3	−0.107	0.017	<.0001	−0.071	0.013	<.0001	−0.08	0.012	<.0001
SC4	−0.220	0.018	<.0001	−0.251	0.019	<.0001	−0.225	0.012	<.0001
SC5	−0.261	0.016	<.0001	−0.299	0.023	<.0001	−0.288	0.011	<.0001
UA2	−0.068	0.015	<.0001	−0.044	0.011	<.0001	−0.051	0.009	<.0001
UA3	−0.101	0.016	<.0001	−0.055	0.013	<.0001	−0.068	0.011	<.0001
UA4	−0.255	0.013	<.0001	−0.166	0.016	<.0001	−0.205	0.012	<.0001
UA5	−0.255	0.013	<.0001	−0.207	0.016	<.0001	−0.236	0.011	<.0001
PD2	−0.060	0.013	<.0001	−0.094	0.013	<.0001	−0.065	0.008	<.0001
PD3	−0.098	0.017	<.0001	−0.151	0.017	<.0001	−0.108	0.012	<.0001
PD4	−0.318	0.015	<.0001	−0.393	0.027	<.0001	−0.367	0.013	<.0001
PD5	−0.414	0.017	<.0001	−0.399	0.026	<.0001	−0.441	0.013	<.0001
AD2	−0.057	0.014	<.0001	−0.076	0.015	<.0001	−0.057	0.008	<.0001
AD3	−0.123	0.018	<.0001	−0.150	0.018	<.0001	−0.133	0.012	<.0001
AD4	−0.299	0.016	<.0001	−0.310	0.025	<.0001	−0.329	0.012	<.0001
AD5	−0.321	0.015	<.0001	−0.369	0.027	<.0001	−0.371	0.012	<.0001

Dimension ranking	PD-MO-AD-SC-UA	PD-AD-MO-SC-UA	PD-AD-MO-SC-UA
Estimated utility values by health state			
21 111	0.904	0.908	0.923
12 111	0.911	0.921	0.932
11 211	0.932	0.956	0.949
11 121	0.940	0.906	0.935
11 112	0.943	0.924	0.943
55 555	−0.573	−0.605	−0.699
No. of health states WTD, n (%)	624 (20.0)	669 (21.4)	733 (23.5)

AD indicates anxiety/depression; cTTO, composite time trade-off; DCE, discrete choice experiment; MO, mobility; PD, pain/discomfort; RE, random effects; SC, self-care; SE, standard error; UA, usual activities; WTD, worse than dead.

1.1.5. MF-8D

MF-8D is an exploratory endpoint of the study. The MF-8D score is constructed from selected components from the EORTC QLQ-C30 and MFSAF (Mukira, 2015). A higher MF-8D score represents the better status:

Table 3: EORTC QLQ-C30 and MFSAF Source Components for MF-8D Score Generation

Source	Component
EORTC QLQ-C30	2 (Long Walk)
EORTC QLQ-C30	3 (Short Walk)
EORTC QLQ-C30	22 (Worry)
EORTC QLQ-C30	18 (Tired)
MFSAF	2 (Night Sweats)
MFSAF	3 (Itching)
MFSAF	4 (Abdominal Discomfort)
MFSAF	5 (Pain Under Ribs on Left Side)
MFSAF	7 (Bone Pain)

The result from each question is transformed into a numeric amount as follows.

Table 4: Per-component Numeric Value for MF-8D Score Generation

Source	Component	Response	Numeric value	Special issues
EORTC QLQ-C30	2 (Long Walk)	1 (Not at all)	0	
		2 (A little)	0	
		3 (Quite a bit)	0.074	Use only if Short Walk is not equal to “Very much”. If Short Walk is equal to “Very much”, then this is 0.000.
		4 (Very much)	0.074	Use only if Short Walk is not equal to “Very much”. If Short Walk is equal to “Very much”, then this is 0.000.
EORTC QLQ-C30	3 (Short Walk)	1 (Not at all)	0	
		2 (A little)	0	
		3 (Quite a bit)	0	
		4 (Very much)	0.122	
	22 (Worry)	1 (Not at all)	0	

EORTC QLQ-C30		2 (A little)	0.031
		3 (Quite a bit)	0.048
		4 (Very much)	0.075
EORTC QLQ-C30	18 (Tired)	1 (Not at all)	0
		2 (A little)	0
		3 (Quite a bit)	0.013
		4 (Very much)	0.072
MFSAF	2 (Night Sweats)	0-10	(value/10)*0.080
MFSAF	3 (Itching)	0-10	(value/10)*0.093
MFSAF	4 (Abdominal Discomfort)	0-10	(value/10)*0.145
MFSAF	5 (Pain Under Ribs on Left Side)	0-10	(value/10)*0.139
MFSAF	7 (Bone Pain)	0-10	(value/10)*0.178
	Constant	All values above are zero	0.000
		Any value above is non-zero	0.007

Following the transformation above, the resultant numeric values are summed. This sum is "utility lost" and ranges from a minimum of 0.000 to a maximum of 0.911.

Utility is then computed as 1.000 minus the utility lost and, as such, ranges from 0.089 to 1.000.

As the MF-8D employs readings from the MFSAF and the EORTC QLQ-C30, the MF-8D can only be computed at timepoints where both these instruments are collected.

1.2. Responder Status Derivation

Change from baseline in a continuous scale will be derived at each timepoint up to week 24. Responder status at each timepoint will be determined by applying the meaningful change threshold (MCT) for improvement to the change from baseline. MCTs for each scale are described in Section 1.2.3. The meaningful change threshold for deterioration might be applied to select PRO questionnaires depending on the proportion of patients with deterioration.

Any scores collected after cross-over to MMB treatment for the patients randomized to DAN will not be included in any derivations.

1.2.1. Baseline Score

MFSAF

Baseline TSS will be derived as described in Section 5.1.1 of the SAP. Baseline of individual items is defined the same as TSS, i.e., the average of the item's last daily value reported on the date of the baseline period triggered per the handheld ePRO device and on the six days immediately following that triggering date and before first dosing date. If more than 3 out of 7 daily values are missing, then the item's baseline score is considered missing.

Other PRO Scores and Individual Items

Baseline scores for each of the other QoL questionnaires described in Section 1.1 are defined in the same way as other study measurements (Section 5.1.1 of the SAP), i.e., the last non-missing assessment on or before the day of first dose date. Assessments done on the date of first dose date are assumed to be measured prior to the first dose date.

1.2.2. Change from Baseline in the RT Period

MFSAF

Since MFSAF is scheduled daily during the 24 week RT period, the values for TSS and individual items at weeks 4, 8, 12, 16, and 20 will be derived for every 28 day period (4 weeks) ending at weeks 4, 8, 12, 16, 20 (i.e., hypothetical 'visit' for analysis purpose). For each visit, each item's score is the average of the item's daily score reported during the previous 28 days. If there is more than one score collected on a same day, the last value on the day will be used. If more than 8 out of 28 daily scores are missing, then the item's score for the period is considered missing. The derivation of Week 24's value will follow Section 9.3. of the SAP.

In the time-to-event analysis, the first day of each 4-weeks period is considered the 'date' of for the visit (e.g., study day 1 for week 4, study day 29 for week 8).

EORTC QLQ-C30

Change from baseline will be calculated for week 12 and week 24 scores.

PROMIS

Change from baseline will be calculated for weeks 2, 4, 8, 12, 16, 20 and 24.

For the PRO questionnaires scheduled at a particular visit, if there are multiple scores collected on the same visit, the last one for the visit will be used.

1.2.3. Responder Status Derivation in the RT period

Responder status will be derived by applying the MCT to the change from baseline in continuous scale. Following MCTs will be used for each questionnaire/scale/item.

MFSAF

The MCT for the TSS will be derived by the methods described in the Meaningful Change Threshold Analysis Plan. Analysis will be repeated for absolute change MCT and % change MCT.

Individual items which have 11 levels (0 to 10) will be considered to have a response if the change from baseline is at least 3.

EORTC QLQ-C30:

The MCT for each sub-scale will be based on Table 5 below (Cocks, 2012). The medium column under "Improvements" will be used. For example, the MCT for cognitive function (CF) sub-scale is >7 which means it is a response if the change from baseline is greater than 7.

Table 5: Guidelines for Interpretation of Longitudinal Difference: EORTC QLQ-C30

Sub-scale	Deteriorations			No difference	Improvements		
	Large	Medium[a]	Small	Trivial	Small	Medium[a]	Large
FI	NE	<–10	–10 to –2	–2 to 3	>3	NE	NE
CF	NE	<–7	–7 to –1	–1 to 3	3–7	>7	NE
PF	<–17	–17 to –10	–10 to –5	–5 to 2	2–7	>7	NE
QL	<–16	–16 to –10	–10 to –5	–5 to 5	5–8	>8	NE
SF	NE	<–11	–11 to –6	–6 to 3	3–8	>8	NE
EF	NE	<–12	–12 to –3	–3 to 6	6–9	>9	NE
NV	<–16	–16 to –11	–11 to –5	–5 to 3	3–9	>9	NE
DY	NE	<–11	–11 to –5	–5 to 2	2–9	>9	NE
FA	<–15	–15 to –10	–10 to –5	–5 to 4	4–9	>9	NE
SL	<–17	–17 to –9	–9 to –2	–2 to 5	5–9	>9	NE
PA	<–20	–20 to –11	–11 to –3	–3 to 5	5–9	9–14	>14
CO	NE	<–15	–15 to –5	–5 to 4	4–10	>10	NE
DI	NE	<–15	–15 to –5	–5 to 3	3–11	>11	NE
RF	<–22	–22 to –14	–14 to –7	–7 to 6	6–12	>12	NE
AP	<–26	–26 to –14	–14 to –2	–2 to 7	7–13	>13	NE

Abbreviations: NE, not evaluable (a guideline for that size class was unobtainable); AP, appetite loss; CF, cognitive functioning; CO, constipation; DI, diarrhoea; DY, dyspnoea; EF, emotional functioning; FA, fatigue; FI, financial difficulties; NV, nausea and vomiting; PA, pain; PF, physical functioning; QL, global quality of life; RF, role functioning; SF, social functioning; SL, insomnia.
[a] Upper limits for medium improvements could not generally be estimated.

PROMIS Physical Function

The MCT for PROMIS Physical Function total raw score will be based on Yost, 2011 where a change from baseline of 4~6 is considered a meaningful change when the 10-item short form is used in cancer patients. A conservative threshold of 6 will be used as the MCT for this analysis (i.e. response if change from baseline is greater than 6).

Individual items which have 5 levels (1 for the worst to 5 for the best) will be considered to have a response if the change from baseline is 2 or greater.

1.2.4. Missing Data Handling

1.2.4.1. Observed Case Approach

In this approach, missing data for original scores and their derived response status will not be imputed and will be summarized/analyzed as is.

1.2.4.2. Non-Responder Imputation (NRI) Approach

In this approach, missing response status will be categorized as non-response

1.2.4.3. Multiple Imputation (MI) Approach

Multiple imputation will be used to impute missing scores (in original scales) at planned visits during the RT period using the same method described in Section 9.4.3 of the SAP. The imputation model might include other variables related to missingness in addition to the variables included in the main GEE model.

MCTs (Section 1.2.3) will be applied to the absolute change or percent change from baseline to imputed missing scores to derive imputed response status (Bell, 2019).

Multiple imputation assumes missing at random (MAR). A control-based multiple imputation (Cro, 2020) under missing not at random assumption will assess the extent to which the MAR assumption in the above estimation is robust. Control-based multiple imputation will be performed for the missing MF-SAF TSS.

2. ANALYSIS SET

All analyses in this document will be performed on the ITT Analysis Set.

3. ANALYSIS METHODS

3.1. Descriptive Analysis

This analysis will be performed for MFSAF individual items, MFSAF TSS, EORTC QLQ-C30 derived scales and items, PROMIS PF total score and individual items. Only scheduled visits in the RT period will be presented.

The distribution of change from baseline will be summarized by descriptive statistics such as mean, median, min/max and inter quartile range as well as number of subjects with missing data. For the MFSAF individual items, the raw scores at each timepoint will be grouped into 5 ordinal categories; 0 (none), 1-3 (mild), 4-6 (moderate), 7-9 (severe), 10 (very severe) and the distribution of the 5 groups will be summarized at each timepoint (including baseline).

The number/percentage of responders (for improvement) per corresponding MCT will be summarized. The distribution of change from baseline for non-responders will be further summarized into improvement of < MCT and deterioration. All summaries will be performed by the randomized treatment arm.

CMH test will be performed comparing response rate at each visit between the two treatments by using NRI (i.e., missing response considered as non-response).

3.2. Longitudinal Analysis of Response Status

This analysis will be performed for MFSAF TSS and PROMIS PF total score per visits in the RT period, defined in Section 1.2.2.

Missing scores in original unit will be imputed by multiple imputation first and transformed into response status based on the corresponding MCT. A generalized estimating equation (GEE) model will be fit on the dataset included imputed missing responses, where a linear relationship

between the logit link of the response status and treatment across visits will be tested during the RT period, and accounting for correlation across these visits within each patient. The model will include treatment effect, timepoint (visit), treatment vs timepoint interaction, stratification variables (except the sites). Autoregressive at 1 degree (i.e. AR1) will be used for the variance covariance structure. Odds ratio for response, its 95% confidence interval, and p-value will be presented at each timepoint as the treatment effect (MMB vs. DAN).

Below is a sample SAS code for the GEE model:

```
proc genmod data=data;
  class subject visit trt(ref='DAN');
  model chgfrombl=trt visit trt*visit / dist=binomial link=logit;
  repeated subject=subject / corrw type=AR;
  slice trt*visit / sliceby(visit) diff exp cl;
run;
```

3.3. Time to Event Analysis

3.3.1. Time to Response Analysis

This analysis will be performed for MFSAF TSS and individual items, PROMIS PF total score and individual items, in the RT period.

For each subject, time to first response is defined as the duration from first dose (or randomization if not dosed) to the first visit day (defined in Section 1.2.2) with response.

Time to first response for subjects without a response will be censored at the last visit day with non-missing response status. If a subject doesn't have any post-baseline measurement, the subject will be censored at 1 day.

Time to first response will be described with quartiles and their 95% confidence intervals, response rate at selected timepoints (e.g., day 28, 56), along with a graphic presentation, by Kaplan-Meier method. Stratified Log-rank test will be performed to compare the two arms with randomization stratification factors (except sites). Hazard ratio (where higher ratio means better response) and its 95% confidence interval will be estimated from a stratified Cox regression model. Any subject with cross-over during the RT period will be censored at the last visit before cross-over.

3.3.2. Duration of MFSAF TSS Response and Fatigue Item Response

Duration of the TSS response observed in the RT period will be assessed beyond the RT period to the end of week 48.

For subjects who achieve a TSS response at any visit during the RT period (Sections 1.2.2 and 1.2.3), the duration of response is defined as the number of days from the start of the initial 28-day period in which the subject achieved the response to the first day of the 7-day assessment that, in the RT period determines the mean TSS for the 28-day period, or the first of the 7-day

assessment in the open label phase, during which the subject's TSS equals or exceeds their baseline value. The responders will be censored at the first day of the last 7 day period if there is no mean TSS of the 7 day assessments which equals or exceeds the baseline value.

Duration of fatigue item response is defined in the same way as duration of TSS response.

4. REFERENCES

Bell ML, Floden L, Rabe BA, et al. Analytical approaches and estimands to take account of missing patient-reported data in longitudinal studies. *Patient related outcome measures.* 2019;10:129.

Cocks K, King MT, Velikova G, et al. Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire. *European Journal of Cancer.* 2012; 48.

Pickard AS, Law EH, Jiang R, et al. United States Valuation of EQ-5D-5L Health States Using an International Protocol. *Value Health.* 2019.

Mukira C, Rowen D, Brazier EB, et al. Deriving a Preference-Based Measure for Mylofibrosis from the EORTC QLQ-C30 and the MF-SAF. *Value in Health.* 2015.