Statistical Analysis Plan for Study M16-191

A Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study of Navitoclax in Combination with Ruxolitinib Versus Ruxolitinib in Subjects with Myelofibrosis (TRANSFORM-1)

Date: 01 April 2023

Version 3.0

Table of Contents

List of Tables

List of Figures

List of Appendices

1.0 Introduction

This Statistical Analysis Plan (SAP) describes the statistical analyses for Navitoclax Study M16-191 titled "A Randomized, Double-Blind, Placebo-Controlled, Phase 3 Study of Navitoclax in Combination with Ruxolitinib Versus Ruxolitinib in Subjects with Myelofibrosis (TRANSFORM-1)" Protocol Version 7.0.

Study M16-191 examines the efficacy and safety of navitoclax in combination with ruxolitinib in adult subjects with primary or secondary myelofibrosis who have not previously received a JAK2 inhibitor.

The analyses of pharmacokinetic endpoints, PRO research endpoints and biomarker research endpoints will not be covered in this SAP.

The SAP will not be updated in case of administrative changes or amendments to the protocol unless the changes impact the analysis.

Unless noted otherwise, all analyses will be performed using SAS Version 9.4 (SAS Institute Inc., Cary, NC 27513) or later under the UNIX operating system.

2.0 Study Design and Objectives

2.1 Objectives and Hypotheses

The primary objective of the study is to evaluate the effect of navitoclax in combination with ruxolitinib on splenomegaly response when compared to ruxolitinib in subjects with myelofibrosis.

The secondary objectives of the study are:

1. To evaluate the effect of navitoclax in combination with ruxolitinib on the onset, magnitude, and duration of disease response, including Total Symptom Score (TSS), effects on spleen, bone marrow fibrosis, and anemia.

2. To evaluate the effect of navitoclax in combination with ruxolitinib on measures of health-related quality of life (HRQoL) including fatigue, and physical functioning.
3. To evaluate the effect of navitoclax in combination with ruxolitinib on overall survival (OS) and leukemia-free survival (LFS).

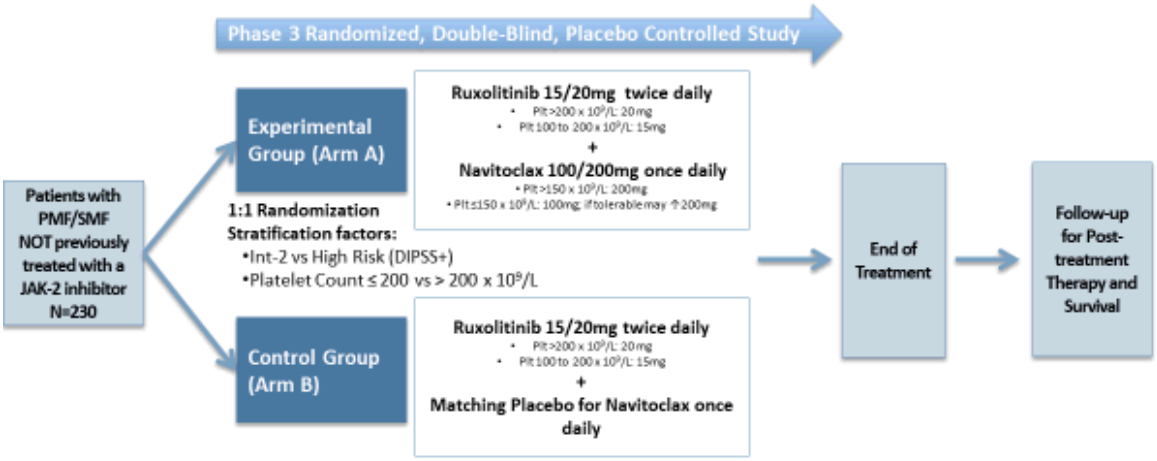
The exploratory objectives of the study are:

1. To evaluate responses to navitoclax and ruxolitinib in subjects with high molecular risk (HMR) mutations.
2. To evaluate the effect of navitoclax in combination with ruxolitinib on progression-free survival (PFS).
3. To evaluate the effect of navitoclax in combination with ruxolitinib on the frequency of mutated alleles.
4. Exploration of biomarkers predictive of navitoclax activity and response may be performed. Potential analysis may include, but will not be limited to, the evaluation of:
   - BCL-2 family profiling
   - Inflammatory cytokine reduction
   - Mutational status.

2.2 Study Design Overview

The schematic of the study is shown in Figure 1.

Figure 1. Study Schematic



DIPSS+ = Dynamic International Prognostic Scoring System Plus; Int = intermediate; Plt = platelet; PMF = primary myelofibrosis; SMF = secondary myelofibrosis

2.3 Treatment Assignment and Blinding

Subjects will be randomized 1:1 to one of the following treatment arms:

- Arm A: navitoclax + ruxolitinib
- Arm B: placebo + ruxolitinib

Randomization will be stratified by:

- Intermediate-2 versus high risk (DIPSS+)
- Platelet count $\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L

2.4 Sample Size Determination

The primary endpoint of the study is at least 35% reduction in spleen volume at Week 24 (SVR_{35W24}) from baseline as measured by magnetic resonance imaging (MRI) or computed tomography (CT) scan, per International Working Group (IWG) criteria. Approximately 230 subjects will be enrolled using a 1:1 randomization ratio to navitoclax

+ ruxolitinib and placebo + ruxolitinib by the stratified factors (DIPSS+: intermediate-2 versus high risk, and platelet count: $\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L).

Based on published results, the SVR35W24 rate in subjects receiving ruxolitinib is expected to be approximately 40%.[1] Improvement in SVR_{35W24} rate from 40% to [redacted] is considered clinically meaningful. With 230 subjects, the study will have approximately 90% power if true SVR_{35W24} rates in the control (placebo + ruxolitinib) and experimental (navitoclax + ruxolitinib) arms are 40% and [redacted], respectively.

With a total of 230 subjects, the study will also have approximately [redacted] power to detect a statistically significant overall survival (OS) improvement in navitoclax + ruxolitinib over placebo + ruxolitinib by collecting [redacted] death events, assuming a true hazard ratio of [redacted]. A total of two OS analyses are planned: the first will be performed when [redacted] death events (70% OS events) are observed; the final OS analysis will be performed when a total of [redacted] death events (100% OS events) have been observed. The Lan-DeMets alpha spending function with O'Brien-Fleming boundary will be used to determine the efficacy boundaries for these planned OS analyses.

3.0 Endpoints

3.1 Primary Endpoint(s)

The primary endpoint of the study is at least 35% reduction in spleen volume at Week 24 (SVR_{35W24}) from baseline as measured by magnetic resonance imaging (MRI) or computed tomography (CT) scan, per International Working Group (IWG) criteria.

3.2 Secondary Endpoint(s)

The secondary endpoints are:

- Change in total symptom score (TSS) at Week 24 from baseline as measured by Myelofibrosis Symptom Assessment Form (MFSAF) v4.0
- At least 35% reduction in spleen volume from baseline (SVR_{35}) as measured by MRI or CT scan, per IWG criteria

- Duration of SVR_{35}
- Change in fatigue from baseline at Week 24, as measured by the PROMIS Fatigue SF 7a
- Change in physical functioning at Week 24 from baseline, as measured by the physical functioning domain of the European Organisation for Research and Treatment of Cancer (EORTC) quality of life questionnaire (QLQ)-C30
- Anemia response per IWG criteria
- Overall survival
- Leukemia-free survival
- Reduction in grade of bone marrow fibrosis from baseline as measured by the European consensus grading system.

3.3 Other Efficacy Endpoint(s)

The other endpoints are:

- At least 50% reduction in palpable splenomegaly from baseline per IWG criteria
- Red blood cell (RBC) transfusion during study drug treatment
- Change in quality of life from baseline as measured by the global health status/quality of life domain of the EORTC-QLQ-C30
- Change from baseline in the summary score for the EORTC QLQ-C30
- Progression-free survival (PFS) per IWG criteria
- Change in frequency of allelic mutations from baseline
- Change in EQ-5D-5L from baseline
- Change in fatigue-related symptoms from baseline as assessed by the PROMIS Fatigue 7a symptom items
- Change in impacts associated with fatigue from baseline as assessed by the PROMIS Fatigue 7a impact items
- Overall response of clinical improvement per IWG criteria
- At least 50% reduction in TSS at Week 24 from baseline (TSS_{50W24})

- At least 50% reduction in TSS at any time from baseline (TSS_{50})
- Time to first TSS_{50}
- Duration of TSS_{50}
- TSS response rate at any time based on meaningful change threshold (MCT)
- Time to first TSS response based on MCT
- Duration of TSS response based on MCT.

3.4 Safety Endpoint(s)

Safety and tolerability will be assessed by evaluating adverse events (AEs), physical examinations, vital signs, electrocardiogram, and clinical laboratory data for the entire study treatment duration.

4.0 Analysis Populations

The following population sets will be used for the analyses.

The Intent-to-Treat (ITT) Population includes all randomized subjects. Unless otherwise specified, the ITT Population will be used for all efficacy analyses. Subjects will be included in the analysis according to the treatment arms that they are randomized to.

The TSS Sensitivity Analysis Population includes all randomized subjects who meet the following criteria:

- subjects with baseline TSS score ≥ 12; or
- subjects with at least 2 symptoms with a baseline symptom score ≥ 3.

The Safety Analysis Set consists of all subjects who received at least 1 dose of any component of the study drug. Subjects will be included in the analysis according to the study drug that they actually received.

The pharmacokinetic (PK) analysis population consists of all subjects who received at least 1 dose of navitoclax. The PK analysis population will be used for all PK analyses.

The study drug is defined as any component of navitoclax + ruxolitinib for subjects in Arm A and any component of placebo + ruxolitinib for subjects in Arm B.

5.0 Subject Disposition

The total number of subjects who were screened, randomized, and treated will be summarized. Reasons for exclusion, including screen failure, will be summarized.

A summary of subject accountability will be provided where the number of subjects in each of the following categories will be summarized for each treatment arm:

- Subjects randomized in the study;
- Subjects who took at least one dose of study drug;
- Subjects who discontinued study drug (all reasons and primary reason);
- Subjects who discontinued study (all reasons and primary reason).

6.0 Study Drug Duration and Compliance

For the safety analysis set, duration of treatment of navitoclax/placebo and ruxolitinib will be summarized for each treatment arm. Duration of treatment is defined for each subject as last dose date minus first dose date + 1. Duration of treatment will be summarized using the number of subjects treated, mean, standard deviation, median, minimum and maximum. In addition, the number and percentage of subjects in the following treatment duration intervals will be summarized:

- < 4 weeks (0 to 28 days),
- ≥ 4 weeks to < 12 weeks (28 to 84 days),
- ≥ 12 weeks to < 24 weeks (84 to 168 days),
- ≥ 24 weeks to < 36 weeks (168 to 252 days),
- ≥ 36 weeks to < 52 weeks (252 to 364 days),
- ≥ 52 weeks to < 78 weeks (364 to 546 days),
- ≥ 78 weeks to < 104 weeks (546 to 728 days),

- ≥ 104 weeks (≥ 728 days).

Subject-years of exposure of navitoclax/placebo and ruxolitinib will be summarized as duration of treatment from all subjects (days) divided by 365.25 for each treatment arm.

The relative dose intensity (RDI) will be summarized for navitoclax/placebo and ruxolitinib for each treatment arm. The RDI is defined as the ratio of the observed dose intensity (ODI) to the planned dose intensity (PDI) and expressed in a percentage. The ODI is defined as:

$$\text{ODI} = \frac{\text{Actual total dose}}{\text{Actual total duration of treatment exposure (days)}}.$$

The PDI is defined as

$$\text{PDI} = \frac{\text{Planned total dose over actual duration of treatment}}{\text{Actual total duration of treatment exposure (days)}}.$$

The actual total duration of treatment exposure in the above equations is defined as last dose date minus first dose date + 1 for navitoclax/placebo and ruxolitinib, respectively. The RDI will be summarized using the number of subjects treated, mean, standard deviation, median, quartiles, minimum and maximum.

7.0 Demographics, Baseline Characteristics, Medical History, and Prior/Concomitant Medications

Demographics, baseline or disease characteristics, medical history, and prior and concomitant medications will be summarized for the ITT population overall and by treatment arm. Categorical variables will be summarized with the number and percentage of subjects; percentages will be calculated based on all subjects, unless otherwise specified. Continuous variables will be summarized with descriptive statistics (number of

non-missing observations, mean and standard deviation, median, minimum and maximum).

7.1 Demographics and Baseline Characteristics

All baseline characteristic summary statistics and analyses are based on characteristics prior to the first dose of study drug or date of randomization for untreated subjects.

Distributions of the continuous demographic and baseline characteristic variables will be summarized by treatment arms with the number of non-missing observations, mean, standard deviation, and median, as well as the minimum and maximum values.

For the categorical demographic and baseline characteristic variables, the frequency and percentages of subjects within each category will be summarized by treatment arms. The number of subjects with missing information will also be summarized.

The following demographic and baseline characteristics will be summarized:

Demographics

- Age (years) and Age Categories (18 - < 65, 65 - < 75, ≥ 75 years)
- Gender (Male, Female)
- Race (White, Black or African American, Asian, Other)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino)
- Region (US, Europe, Japan, Other Regions)
- Height (cm)
- Weight (kg)
- Body Mass Index (kg/m^2)

Baseline and Disease-Related Characteristics

- DIPSS+ Risk Group Reported in IRT (Intermediate-2, High)
- DIPSS+ Risk Group Reported in EDC (Intermediate-1, Intermediate-2, High)

- Platelet Count Reported in IRT ($\leq 200 \times 10^9$/L, $> 200 \times 10^9$/L)
- Baseline Platelet Count ($\leq 200 \times 10^9$/L, $> 200 \times 10^9$/L)
- Type of Myelofibrosis (Primary Myelofibrosis, Secondary Myelofibrosis)
  - Type of Secondary Myelofibrosis (Post-Polycythemia Vera Myelofibrosis, Post-Essential Thrombocythemia Myelofibrosis)
- JAK2 V617F Mutation (Detected, Not Detected)
- CALR Frameshift Mutation (Detected, Not Detected)
- MPL W515 Mutation (Detected, Not Detected)
- Negative of JAK2, CALR and MPL (Yes, No)
- High Molecular Risk (Yes, No)
- Time Since Myelofibrosis Diagnosis (months)
- Number of Prior Lines of Therapies for Myelofibrosis
- Reasons for Change of Therapy (Splenomegaly, Worsening Symptoms, Increased Need for Transfusion, Anemia, Thrombocytopenia, Increasing WBC, Increasing LDH, Other)
- Baseline ECOG Performance Status (0, 1, 2)
- Baseline Spleen Length by Palpation (cm)
- Baseline Spleen Volume (cm^3)
- Baseline TSS score
- Baseline Bone Marrow Fibrosis Grade (MF-0, MF-1, MF-2, MF-3)
- Baseline Transfusion Dependent (Yes, No)
- Baseline Hemoglobin (g/dL)
- Baseline CTCAE Grade of Anemia
- Baseline Platelet Count (10^9/L)
- Baseline CTCAE Grade of Thrombocytopenia
- Baseline Neutrophil Count (10^9/L)
- Baseline CTCAE Grade of Neutropenia
- Baseline White Blood Cell Count (10^9/L)
- Baseline Lactate Dehydrogenase (U/L)

- Baseline Renal Function (Normal, Mild Impairment, Moderate Impairment, Severe Impairment)
- Baseline Hepatic Function (Normal, Mild Impairment, Moderate Impairment, Severe Impairment)
- Alcohol Use (Unknown, Never, Current, Former)

7.2 Medical History

Medical history data will be coded using the Medical Dictionary for Regulatory Activities (MedDRA). The actual version of the MedDRA coding dictionary will be noted in the statistical tables and clinical study report. The number and percentage of subjects in each medical history category, including myelofibrosis related conditions and symptoms, by MedDRA system organ class and preferred term will be summarized overall and by treatment arm. The system organ class (SOC) will be presented in alphabetical order, and the preferred terms will be presented in alphabetical order within each SOC. Subjects reporting more than one condition/diagnosis will be counted only once in each row (SOC or preferred term).

7.3 Prior and Concomitant Medications

Prior and concomitant medications, including prior treatment for myelofibrosis and preceding treatment for polycythemia vera or essential thrombocythemia prior to developing secondary myelofibrosis, will be summarized by generic name. A prior medication is defined as any medication taken prior to the date of the first dose of study drug. All medications are considered prior medications for subjects who did not receive any study drug. A concomitant medication is defined as any medication that started prior to the date of the first dose of study drug and continued to be taken after the first dose of study drug or any medication that started on or after the date of the first dose of study drug, but not after the date of the last dose of study drug.

The number and percentage of subjects taking medications will be summarized by generic drug name based on the World Health Organization (WHO) Drug Dictionary for both

prior and concomitant medications. A subject who reports the use of two or more medications will be counted only once in the summary of any prior or concomitant medication. Subject reporting two or more uses of the same medication will be counted only once in the total for the associated generic drug name.

If an incomplete or missing start date was collected for a medication, the medication will be assumed to be a concomitant medication, or prior medication for subjects who did not receive any study drug, unless there is evidence to the contrary.

8.0 Efficacy Analyses

8.1 General Considerations

Unless otherwise specified, all randomized subjects in the ITT population will be included in the efficacy analyses. All statistical tests will be 2-sided at an alpha level of 0.05.

The primary analysis will be performed after all ongoing subjects have completed 24 weeks of follow-up and the database has been locked. Unless otherwise specified, data after the cutoff date will be excluded from statistical analyses.

Unless otherwise specified, all subjects will be analyzed according to the baseline platelet count and DIPSS+ score collected in Electronic Data Capture (EDC) system. For subjects with missing karyotype information, the DIPSS+ score will be calculated by assuming these subjects do not have unfavorable karyotype. The DIPSS+ score will be categorized by intermediate (including both intermediate-1 and intermediate-2) and high risk.

Unless otherwise specified, binary variables (e.g., response rates) will be analyzed using Cochran-Mantel-Haenszel (CMH) test, stratified by DIPSS+ risk group (intermediate versus high risk) and platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L). The 95% confidence interval for the response rate will be provided for each treatment arm using the exact binomial distribution. The strata-adjusted difference in the response rates between the treatment arms and the 95% confidence interval for the difference will also be provided.

Unless otherwise specified, continuous data from patient-reported outcome will be analyzed using linear mixed effects model with treatment arm, visit, treatment by visit interaction, DIPSS+ risk group (intermediate versus high risk), platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L) and baseline score as fixed effects. An unstructured variance covariance matrix will be used. The compound symmetry variance covariance matrix will be used if the model does not converge with the unstructured matrix. Change in score from baseline will be compared between the treatment arms for the specified post-baseline visit. The least-squares mean change from baseline and the corresponding 95% confidence interval will be provided for each treatment arm. In addition, the least-squares mean of the difference between the treatment arms and the corresponding 95% confidence interval will be provided.

Unless otherwise specified, time-to-event endpoints will be analyzed using Kaplan-Meier methodology and compared between treatment arms using the log-rank test, stratified by DIPSS+ risk group (intermediate versus high risk) and platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L). The hazard ratio between treatment arms and the corresponding 95% confidence interval will be estimated using the Cox proportional hazards model, stratified by DIPSS+ risk group (intermediate versus high risk) and platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L).

Unless otherwise specified, if a subject prematurely breaks the blind, data after the blind break will be included in the statistical analysis as appropriate based on the definition and analysis of the efficacy endpoints.

Unless otherwise specified, baseline refers to the last non-missing observation before the first administration of study drug or randomization if no study drug is given.

Unless otherwise specified, a confirmed disease progression is defined as any of the following:

- Two consecutive disease responses of progressive disease or relapse, in which case the date of the confirmed disease progression will be the date of the first progressive disease or relapse

- A progressive disease due to leukemic transformation
- A progressive disease or relapse that is the last disease response of a subject.

At the time this SAP is written, the evidence-based MCT for change from baseline in TSS at Week 24 is determined to be –10.

8.2 Handling of Missing Data

Unless otherwise specified, missing data will be analyzed using the following methods for the efficacy analyses:

- Binary variables: If a subject does not have evaluation during a specific visit window, the subject will be considered a non-responder for that visit.
- Continuous variables: Mixed-effect Model Repeat Measurement (MMRM) will be used as the primary approach for handing missing data. Parameter estimation in linear mixed effects models will be based on the assumption of data being missing at random and the method of restrictive maximum likelihood (REML).
- Time-to-event variables: The time-to-event analysis will be based on observed data and will not impute values for missing evaluations. Subjects who do not have the event of interest will be censored at the censoring date defined for the respective endpoint in Section 8.3, Section 8.4 and Section 8.5.

Tipping Point Analysis for Binary Variables

The tipping point analyses will be performed to assess the robustness of the primary analysis results for SVR_{35W24} and TSS_{50W24} under missing not at random assumptions. The tipping point analyses are two-dimensional (i.e., assumptions about the missing outcomes on the treatment arm and the placebo arm can vary independently). The response rate among those subjects with a missing response is assumed to be p_0 for the control arm and p_1 for the treatment arm, and the response rate p_0 and p_1 systematically vary from 0% to 100% by every 10%, respectively. Given a set of (p_0, p_1), the subjects with a missing response will be randomly assigned as responders or non-responders using

binomial distribution to generate 30 imputed datasets, and the same CMH method used for the primary analysis will be performed on each of the multiple imputed datasets to obtain the results for each comparison of the treatment arm versus the control arm. These results will then be aggregated using Rubin's method.

If one pair of shift parameters is found to just reverse the study conclusion (two-sided *p*-value > 0.05), then the shift parameters are identified as the tipping point. The results for a grid of shift parameter combinations are tabulated.

Tipping Point Analysis for Continuous Variables

The tipping point analyses will be conducted as a two-dimensional sensitivity analysis to check the departure from the MAR assumption on the treatment arm and the control arm independently. The focused scenarios are when missing outcomes on the treatment arm are worse than the imputed values using the MAR assumption on the treatment arm, while missing outcomes on the control arm are better than the imputed values using the MAR assumption on the control arm. In each treatment group, missing values will be first imputed via MI under MAR assumption using observed data, and then a shift parameter will be added to the imputed values. A different shift parameter may be specified for each treatment group and be implemented by PROC MI using the missing not at random (MNAR) statement. The imputation uses a two-step approach, an augmentation step using Markov Chain Monte Carlo (MCMC) and an imputation step using Monotone Regression. The MNAR statement is applied in the imputation step. The number of imputed datasets is 30. For TSS at Week 24, the imputation model will include DIPSS+ risk group (intermediate versus high risk), platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L), and TSS scores at baseline, Week 8 and Week 12. For PROMIS Fatigue score at Week 24 and physical functioning score at Week 24, the imputation model will include DIPSS+ risk group (intermediate versus high risk), platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L), and TSS scores at baseline, Week 4 and Week 12.

In cases where the shifted values have exceeded the maximum value of the endpoint (i.e., out of range), the minimum or maximum value of the endpoint will be used in

further analysis steps. For each pair of shift values, ANCOVA model using PROC MIXED will be performed on each imputed dataset for a pairwise comparison between treatment and control arms. ANCOVA model will include the fixed effects of treatment, stratification factors, and the baseline score. Analysis results from each imputed dataset will be aggregated using Rubin's method.

If one pair of shift parameters is found to reverse the study conclusion (two-sided p-value > 0.05), then the shift parameters are identified as the tipping point. The results for a grid of shift parameter combinations will be tabulated.

8.3 Primary Efficacy Endpoint(s) and Analyses

8.3.1 Primary Efficacy Endpoint(s)

The primary efficacy endpoint is at least 35% reduction in spleen volume at Week 24 (SVR_{35W24}) from baseline as measured by magnetic resonance imaging (MRI) or computed tomography (CT) scan and assessed centrally, per International Working Group (IWG) criteria.[2]

8.3.2 Handling of Missing Data for the Primary Efficacy Endpoint(s)

The analysis window for Week 24 is defined as from [redacted] per study activity schedule with consideration of COVID-19 related interruptions. Subjects who do not have spleen volume assessment in the analysis window will be treated as non-responders. Subjects who do not receive any study drug will be treated as non-responders. Subjects without baseline spleen volume assessment or with unevaluable baseline spleen volume will be treated as non-responders. Spleen volume assessments after confirmed disease progression or start of post-study treatment will not be included in the analysis.

8.3.3 Primary Efficacy Analysis

The proportion of subjects who achieve SVR_{35W24} will be compared between the treatment arms using Cochran-Mantel-Haenszel (CMH) test, stratified by DIPSS+ risk group (intermediate versus high risk) and platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L). The DIPSS+ risk group collected in EDC and the baseline platelet count collected in EDC will be used in the primary efficacy analysis as described in Section 8.1. The 95% confidence interval for the SVR_{35W24} rate will be provided for each treatment arm using the exact binomial distribution.

8.3.4 Additional Analyses of the Primary Efficacy Endpoint(s)

The following additional analyses will be performed for SVR_{35W24}:

- Percent change from baseline in spleen volume at Week 24 will be summarized for each treatment arm and presented graphically in waterfall plots.
- Sensitivity analysis of SVR_{35W24} will be performed by excluding subjects whose DIPSS+ risk group is intermediate-1.
- Spleen volume reduction at Week 24 will be descriptively summarized by the following categories: $\geq 10\%$ to $< 25\%$, $\geq 25\%$ to $< 35\%$, $\geq 35\%$ to $< 50\%$, and $\geq 50\%$.
- CMH test stratified by DIPSS+ risk group (intermediate-2 versus high risk) and platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L) collected in IRT.
- Tipping point analysis as described in Section 8.2.

Logistic regression with model selection procedure using relevant demographic and baseline characteristics in Section 7.1 may be performed to evaluate the robustness of the primary analysis.

8.4 Secondary Efficacy Analyses

8.4.1 Key Secondary Efficacy Analyses

8.4.1.1 Change in Total Symptom Score (TSS) at Week 24 from Baseline

The total symptom score will be measured by Myelofibrosis Symptom Assessment Form (MFSAF) v4.0.[3] The daily TSS scores will be calculated by summing up scores from all 7 questions from the questionnaire. All 7 questions need to be answered to calculate a daily score. [redacted]

TSS score at Week 24 is defined as [redacted]

TSS scores will be analyzed using a linear mixed effects model with treatment arm, visit, treatment by visit interaction, DIPSS+ risk group (intermediate versus high risk), platelet count ($\leq 200 \times 10^9$/L versus $> 200 \times 10^9$/L) and baseline score as fixed effects, as described in Section 8.1. An unstructured variance covariance matrix will be used. The compound symmetry variance covariance matrix will be used if the model does not converge with the unstructured matrix. Weekly scores up to Week 24 will be included in the mixed model. Parameter estimation is based on the method of restrictive maximum

likelihood. Within the framework of the mixed model, change in TSS score from baseline at Week 24 will be compared between the treatment and control arm.

The tipping point analysis described in Section 8.2 will be performed as a sensitivity analysis to assess departure from the MAR assumption. For subjects who do not have Week 24 TSS score, reasons for missing Week 24 TSS score will be summarized for each treatment arm and individual scores prior to Week 24 will be plotted to explore the validity of the missing at random assumption.

The following supplementary analyses will be performed:

- Change in TSS score at Week 24 from baseline by excluding the fatigue item
- TSS_{50W24} rate
- TSS_{50} rate at any time
- Time to first TSS_{50}
- Duration of TSS_{50}
- TSS response rate at any time based on meaningful change threshold (MCT)
- TSS response rate at Week 24 based on meaningful change threshold (MCT)
- Time to first TSS response based on MCT
- Duration of TSS response based on MCT.

TSS_{50W24} rate is defined as the proportion of subjects who achieve at least 50% reduction in total symptom score at Week 24 from baseline. TSS scores obtained after confirmed disease progression or start of post-study treatment will not be included in the analysis. Subjects who do not have TSS score at Week 24 will be treated as non-responders. Subjects who do not receive any study drug will be treated as non-responders. TSS_{50W24} rate will be analyzed using the statistical method described for binary variables in Section 8.1.

TSS_{50} rate any time is defined as the proportion of subjects who achieve at least 50% reduction in total symptom score from baseline at any time during the study. TSS scores obtained after confirmed disease progression or start of post-study treatment will not be

included in the analysis of TSS_{50} rate. Subjects who do not have any TSS score after randomization will be treated as non-responders. Subjects who do not receive any study drug will be treated as non-responders. TSS_{50} rate will be analyzed using the statistical method described for binary variables in Section 8.1.

Time to first TSS_{50} is defined as the time from randomization to the first date of achieving TSS_{50}. Time to first TSS_{50} will be summarized descriptively for each treatment arm.

Duration of TSS_{50} is defined as the time from the first date of achieving TSS_{50} to the first date of loss of TSS_{50}. Loss of TSS_{50} is defined as not maintaining TSS_{50} and at least 50% increase in TSS from nadir (the lowest score in the previous weeks). Subjects who do not lose TSS_{50} response will be censored at the last TSS assessment. Subjects with ≥ 2 consecutive missing TSS scores prior to losing TSS_{50} response will be censored at the last non-missing TSS assessment prior to losing TSS_{50} response. The distribution of duration of TSS_{50} will be estimated for each treatment arm using Kaplan-Meier methodology. Median duration of TSS_{50} and the corresponding 95% confidence interval will be provided for each treatment arm. Only subjects who achieve TSS_{50} at any time will be included in the analysis of duration of TSS_{50}.

TSS response rate at any time during the study is defined as proportion of subjects who achieved change in TSS score from baseline $\leq$ MCT at any time during the study. The MCT, derived using anchor-based analyses, has been determined to be at least 10 points for improvement in TSS from baseline. The validation of MCT of –10 is described in the Psychometric Analysis Report. TSS scores obtained after confirmed disease progression or start of post-study treatment will not be included in the analysis of TSS response rate. Subjects who do not have any TSS score after randomization will be treated as non-responders. Subjects who do not receive any study drug will be treated as non-responders. TSS response rate will be analyzed using the statistical method described for binary variables in Section 8.1. Similar analyses will be performed for TSS response rate at Week 24. The maximum reduction in TSS from baseline at any time will be summarized.

Time to TSS response is defined as the time from randomization to the first date of achieving TSS response based on MCT of at least 10 points. Time to first TSS response will be summarized descriptively for each treatment arm.

Duration of TSS response is defined as the time from the first date of achieving TSS response to the first date of loss of TSS response. Loss of TSS response is defined as diminishment of TSS below MCT improvement of 10 points. Subjects who do not lose TSS response will be censored at the last TSS assessment. Subjects with ≥ 2 consecutive missing TSS scores prior to losing TSS response will be censored at the last non-missing TSS assessment prior to losing TSS response. The distribution of duration of TSS response will be estimated for each treatment arm using Kaplan-Meier methodology. Median duration of TSS response and the corresponding 95% confidence interval will be provided for each treatment arm. Only subjects who achieve TSS response at any time will be included in the analysis of duration of TSS response. Sensitivity analysis of duration of TSS response may be performed by accounting for disease progression or death.

In addition, descriptive summaries of weekly TSS scores and individual symptom scores, change from baseline and percent change from baseline will be provided for each treatment arm. Change and percent change in TSS scores and individual symptom scores from baseline at Week 24 may also be presented graphically for each treatment arm.

TSS analyses will be performed on the ITT population. Sensitivity analyses will be performed using the TSS Sensitivity Analysis Population described in Section 4.0.

8.4.1.2 At Least 35% Reduction in Spleen Volume from Baseline (SVR_{35})

SVR_{35} rate is defined as the proportion of subjects who achieve at least 35% reduction in spleen volume from baseline at any time during the study. Subjects who do not have any post-baseline spleen volume assessment or who have had confirmed disease progression or started post-study treatment prior to achieving SVR_{35} will be treated as non-responders. Subjects who do not receive any study drug will be treated as non-responders. Subjects

without baseline spleen volume assessment or with unevaluable baseline spleen volume will be treated as non-responders.

SVR_{35} rate will be analyzed using the statistical method described for binary variables in Section 8.1. In addition, descriptive summaries of spleen volume and percent change from baseline at each visit will be provided for each treatment arm and will be presented graphically. Maximum percent reduction from baseline will also be presented graphically for each treatment arm. Time to first SVR_{35} will be descriptively summarized for each treatment arm. The proportion of subjects who achieve at least 35% reduction in spleen volume from baseline at Week 12 will be descriptively summarized for each treatment arm.

8.4.1.3 Change in Fatigue from Baseline at Week 24

Patient reported outcome on fatigue will be assessed using the global fatigue score as measured by PROMIS Cancer Fatigue SF 7a.[4] Scores will be computed according to the PROMIS Cancer Fatigue SF 7a scoring manual. All questions must be answered to calculate the total score. Change in scores from baseline at Week 24 will be compared between the treatment arms using the linear mixed effects model described in Section 8.1. The tipping point analysis described in Section 8.2 will be performed as sensitivity analyses. T-scores will be used in the analysis.

8.4.1.4 Anemia Response

Definition of Baseline Transfusion Dependence

Baseline transfusion dependence (TD) is defined as transfusion of at least 6 units of packed red blood cells (PRBC) within 12 weeks on or prior to the date of the first dose of study drug (or randomization for untreated subjects). In addition, the most recent transfusion episode must have occurred within 4 weeks on or prior to the date of the first dose of study drug (or randomization for untreated subjects). Subjects with baseline hemoglobin level of < 10 g/dL who do not meet the criteria for baseline transfusion dependence are considered baseline transfusion independent (TI).

Definition of Baseline Hemoglobin

For subjects who receive PRBC transfusion within 4 weeks on or prior to the date of the first dose of study drug (or randomization for untreated subjects), the baseline hemoglobin value is defined as the minimum of the hemoglobin value collected prior to the transfusion, if available, and the last non-missing hemoglobin value on or prior to the date of the first dose of study drug (or randomization for untreated subjects).

For subjects who do not receive PRBC transfusion within 4 weeks on or prior to the date of the first dose of study drug (or randomization for untreated subjects), the baseline hemoglobin value is defined as the last non-missing hemoglobin value on or prior to the date of the first dose of study drug (or randomization for untreated subjects).

Definition of Anemia Response

For subjects who are baseline transfusion independent and whose baseline hemoglobin is < 10 g/dL, anemia response is defined as ≥ 2 g/dL increase in hemoglobin from baseline without receiving any PRBC transfusion within 2 weeks or erythropoietin supplements within 4 weeks. Hemoglobin values after the last dose + 30 days, confirmed disease progression or start of post-study treatment will not be considered in the analysis of anemia response.

For subjects who are baseline transfusion dependent, anemia response is defined as a period of at least 12 consecutive weeks without PRBC transfusion after the first dose of study drug until the last dose + 30 days, start of post-study treatment, confirmed disease progression or death, which occurs earlier.

Subjects who do not receive any study drug will be treated as non-responders. Subjects who are baseline transfusion independent and whose baseline hemoglobin is ≥ 10 g/dL will not be included in the analysis since these subjects are not applicable to the evaluation of anemia response.

Anemia response rate is defined as the proportion of subjects who achieve anemia response and will be analyzed using the statistical method described for binary variables in Section 8.1. The main analysis of anemia response will include both TD and TI subjects, with separate summaries provided for TD and TI subjects.

8.4.1.5 Change in Physical Functioning at Week 24 from Baseline

Physical functioning will be assessed using the physical functioning domain of EORTC QLQ-C30[5]. Scores will be computed according to the EORTC QLQ-C30 scoring manual. Change in scores from baseline at Week 24 will be compared between the treatment arms using the linear mixed effects model described in Section 8.1. The tipping point analysis described in Section 8.2 will be performed as sensitivity analyses.

8.4.1.6 Reduction in Grade of Bone Marrow Fibrosis from Baseline

Grade of bone marrow fibrosis will be measured by the European consensus grading system[6] and assessed centrally. The proportion of subjects who achieve a reduction of at least 1 grade in bone marrow fibrosis from baseline will be analyzed using the statistical method described for binary variables in Section 8.1. Data after subjects have confirmed disease progression or started post-study treatment will be excluded from the analysis.

Subjects without any post-baseline grade of bone marrow fibrosis measurement will be treated as non-responders. Subjects who do not receive any study drug will be treated as non-responders. Subjects without baseline bone marrow fibrosis assessment will be treated as non-responders.

The analysis of reduction in grade of bone marrow fibrosis will be performed on subjects who have a bone marrow fibrosis grade determined at baseline and have at least one post-baseline assessment.

Time to first reduction in grade of bone marrow fibrosis from baseline will be descriptively summarized for each treatment arm. Reduction in grade of bone marrow fibrosis at Week 24 will be descriptively summarized.

8.4.1.7 Overall Survival (OS)

Overall survival is defined as the time from randomization to death from any cause. Subjects who are alive will be censored at the last known alive date. The last known alive date will be determined by selecting the last available date of the following study procedures: randomization, start date of adverse event, bone marrow collection, disease assessment, spleen assessment, vital signs assessment, electrocardiogram assessment, clinical laboratory collection, study drug administration, start date of concomitant or post-treatment medicine or procedure, survival follow-up, biospecimen sample collection, quality of life assessments, and performance status.

Overall survival will be analyzed using the statistical method described for time-to-event variables in Section 8.1.

8.4.1.8 Leukemia-Free Survival (LFS)

Leukemia-free survival is defined as the time from randomization to leukemic transformation (≥ 20% blasts) by bone marrow blasts or peripheral blood blasts per IWG criteria, or death from any cause, whichever occurs earlier, on or prior to the start of post-study treatment. Subjects who do not have leukemic transformation and are alive will be censored at the date of the last adequate peripheral blood blast assessment or bone marrow assessment on or prior to the start of post-study treatment. Subjects who are alive and without any hematology laboratory assessment for blast or bone marrow assessment will be censored at the date of randomization.

Leukemia-free survival will be analyzed using the statistical method described for time-to-event variables in Section 8.1.

8.4.1.9 Duration of SVR_{35}

Duration of SVR_{35} is defined as the time from the first date of SVR_{35} to the first assessment where SVR_{35} is not maintained and the spleen volume is ≥ 25% increase from nadir (the lowest spleen volume in the previous assessments), confirmed relapse or leukemic transformation per IWG criteria, whichever is earlier. Subjects without the

defined events will be censored at the date of the last spleen volume assessment, the last disease assessment, the last bone marrow assessment, or the last hematology lab test, whichever occurs the latest and on or prior to the start of the post-study treatment.

The distribution of duration of SVR_{35} will be estimated for each treatment arm using Kaplan-Meier methodology. Median duration of SVR_{35} and the corresponding 95% confidence interval will be provided for each treatment arm. There will be no statistical comparison for duration of SVR_{35}.

Only subjects who achieve SVR_{35} will be included in the analysis.

8.5 Additional Efficacy Analyses

8.5.1 At Least 50% Reduction in Palpable Splenomegaly from Baseline

The proportion of subjects who achieved at least 50% reduction in palpable splenomegaly from baseline will be analyzed using the statistical method described for binary variables in Section 8.1.

Subjects who do not have any post-baseline assessment will be treated as non-responders. Data after subjects have confirmed disease progression or start post-study treatment will be excluded from the analysis. Subjects who do not receive any study drug will be treated as non-responders. Subjects without baseline assessment will be treated as non-responders.

8.5.2 Red Blood Cell Transfusion

Post-baseline RBC transfusion evaluation period is from after the first dose of study drug to earliest day of on or before the last dose of study drug + 30 days, or the start of post-study treatment, confirmed disease progression or death, which occurs earlier.

Post-baseline transfusion independence rate will be calculated as the proportion of subjects who achieved transfusion independence during the evaluation period.

Transfusion independence is defined as a period of at least 12 weeks (≥ 84 days) with no PRBC transfusions during the evaluation period. Subjects who do not receive any study drug will be considered not achieving transfusion independence.

The duration of transfusion independence will be summarized for each treatment arm. The duration of transfusion independence is defined as the duration of the first time period in which a subject received no PRBC transfusions for at least 12 weeks (≥ 84 days). The descriptive statistics (median and range) will be provided for the duration of transfusion independence.

Incidence rate of PRBC transfusions/subject/month, defined as total number of PRBC transfusions from all subjects divided by total duration of treatment from all subjects (months), will be calculated for each treatment arm.

Time from first dose of study drug to the first time receiving 2 or more units of PRBC transfusion will be summarized for each treatment arm. The total number and amount of PRBC transfusions from each subject will be summarized for each treatment arm.

8.5.3 EORTC QLQ-C30

Scores for the subscales/items from the EORTC QLQC30 will be calculated according to the scoring manuals. Change in scores from baseline will be analyzed using the linear mixed effects model described in Section 8.1.

8.5.4 Progression-Free Survival (PFS)

Progression-free survival is defined as the time from randomization to the date of confirmed relapse or progressive disease, or death from any cause, whichever occurs earlier, on or prior to the start of post-study treatment. Subjects without any defined PFS event will be censored at the date of last adequate disease assessment on or prior to the start of post-study treatment. Subjects who are alive and without any adequate disease assessment will be censored at the date of randomization.

Progression-free survival will be analyzed using the statistical method described for time-to-event variables in Section 8.1.

8.5.5 Change in Frequency of Allelic Mutations from Baseline

Absolute change and percent change from baseline in variant allele frequency any time during the post baseline period will be derived by specific driver gene, which is mutated-JAK2, -CALR, and -MPL, if baseline allele frequency and post-baseline allele frequency are available based on the Next generation sequencing (NGS) method in the peripheral blood.

Descriptive analysis will be performed to analyze the change from baseline in variant in driver gene allele frequency at different time points. Waterfall and line plots will be used for visualization. The proportion of subjects who achieved ≥ 20% and ≥ 50% reduction in driver gene allele frequency at Week 24 and at any time will be summarized for each treatment arm.

8.5.6 EQ-5D-5L

Scores for EQ-5D-5L[7] and its visual analog scale will be calculated according to the scoring manuals. Change in scores from baseline will be analyzed using the linear mixed effects model described in Section 8.1.

8.5.7 PROMIS Fatigue 7a

Scores for PROMIS Cancer Fatigue 7a symptom items and impact items will be calculated according to the scoring manuals. All questions of a subdomain must be answered to calculate a score for that subdomain. The score for a subdomain will be calculated as the average of the item scores in that subdomain. Change in scores from baseline will be analyzed using the linear mixed effects model described in Section 8.1. Raw scores will be used in the analysis.

8.5.8 Overall Response of Clinical Improvement

Rate of overall response of clinical improvement is defined as the proportion of subjects who achieve best response of clinical improvement (CI) per IWG criteria. Subjects who do not have any disease assessment or who have started post-study treatment prior to achieving best response of CI will be treated as non-responders. Responses achieved after confirmed relapse or progressive disease will not be included in the analysis.

Rate of overall response of clinical improvement will be analyzed using the statistical method described for binary variables in Section 8.1.

8.6 Efficacy Subgroup Analyses

To evaluate the potential impact of demographics and baseline characteristics on efficacy, subgroup analyses may be performed on, but not limited to, the primary efficacy endpoint. For subgroup analyses of binary variables, the difference in the proportions and the corresponding 95% confidence interval (exact unconditional confidence limits) for the difference will also be provided. For subgroup analyses of time-to-event variables, the hazard ratio and the 95% confidence interval based on unstratified Cox proportional hazards model will be provided. For subgroup analyses of continuous variables, the mean difference and the corresponding 95% confidence interval will be provided.

Subgroup analyses will be performed for subgroups including, but not limited to, those defined below:

- Age (18 - < 65 years, ≥ 65 years, ≥ 75 years)
- Race (White, Black or African American, Asian, Other)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino)
- DIPSS+ Risk Group (Intermediate-1, Intermediate-2, High)
- Baseline Platelet Count ($\leq 200 \times 10^9$/L, $> 200 \times 10^9$/L)
- Type of Myelofibrosis (Primary, Secondary)
- JAK2 V617F Mutation
- CALR Frameshift Mutation

- MPL W515 Mutation
- High Molecular Risk (Yes, No)

High molecular risk (HMR) is defined as any mutation within ASXL1, SRSF2, EZH2, IDH1/2, and U2AF1. For molecular subgroups, analysis for mutation subtypes may also be performed. In addition, correlation among efficacy endpoints, including but not limited to the following, will be explored:

- Anemia response and spleen response
- Depth of spleen response and reduction in grade of bone marrow fibrosis
- Reduction in grade of bone marrow fibrosis and anemia response
- Reduction in variant allele frequency and spleen response.

9.0 Safety Analyses

9.1 General Considerations

Safety data will be summarized for the safety analysis set. Safety summaries will be presented by treatment arm and in total. For the safety analysis, subjects are assigned to a treatment arm based on the treatment actually received, regardless of the treatment randomized. Unless otherwise specified, data after the cutoff date will be excluded from the safety analysis.

9.2 Adverse Events

Adverse events (AEs) will be summarized and presented using primary MedDRA System Organ Classes (SOCs) and preferred terms (PTs) according to the version of the MedDRA coding dictionary used for the study at the time of database lock. The actual version of the MedDRA coding dictionary used will be noted in the AE tables and in the clinical study report. Specific adverse events will be counted once for each subject for calculating percentages, unless stated otherwise. In addition, if the same adverse event occurs multiple times within a subject, the highest severity and level of relationship to investigational product will be reported.

9.2.1 Treatment-Emergent Adverse Events

Treatment-emergent AEs are defined as any AE with the onset that is after the first dose of study drug until 30 days after the last dose of the study drug. Events where the onset date is the same as the study drug start date are assumed to be treatment-emergent. If an incomplete or missing onset date was collected for an AE, the AE will be assumed to be treatment-emergent unless there is evidence to the contrary (e.g., the AE end date was prior to the date of the first dose of study drug). All treatment-emergent AEs will be summarized overall, as well as by primary MedDRA SOC and PT. The SOCs will be presented in alphabetical order, and the PTs will be presented in descending frequency order in the navitoclax + ruxolitinib group within each SOC.

The number and percentage of subjects experiencing treatment-emergent AEs will be summarized.

9.2.2 Adverse Event Overview

An overview of AEs will be presented consisting of the number and percentage of subjects experiencing at least one event for each of the following AE categories:

- Any treatment-emergent AE
- Any treatment-emergent AE with NCI toxicity Grade ≥ 3
- Any treatment-emergent AE with NCI toxicity Grade 3 or 4
- Any treatment-emergent AE with NCI toxicity Grade 3
- Any treatment-emergent AE with NCI toxicity Grade 4
- Any treatment-emergent AE with reasonable possibility related to navitoclax/placebo as assessed by the investigator
- Any treatment-emergent AE with reasonable possibility related to ruxolitinib as assessed by the investigator
- Any serious treatment-emergent AE
- Any treatment-emergent AE leading to discontinuation of navitoclax/placebo
- Any treatment-emergent AE leading to discontinuation of ruxolitinib

- Any treatment-emergent AE leading to interruption of navitoclax/placebo
- Any treatment-emergent AE leading to interruption of ruxolitinib
- Any treatment-emergent AE leading to dose reduction of navitoclax/placebo
- Any treatment-emergent AE leading to dose reduction of ruxolitinib
- Any treatment-emergent AE leading to death
- Any treatment-emergent AE related to COVID-19 infection
- All deaths
- Deaths occurring within 30 days after the last dose of study drug
- Deaths occurring more than 30 days after the last dose of study drug
- Deaths related to COVID-19 infection

9.2.3 Treatment-Emergent Adverse Events by SOC and/or PT

Treatment-emergent adverse event summaries will be presented by SOC and PT. Specific adverse events will be counted once for each subject for calculating percentages, unless stated otherwise. In addition, if the same adverse event occurs multiple times within a subject, the highest severity and level of relationship to investigational product will be reported.

In addition, treatment-emergent adverse events will be summarized by PT and sorted by decreasing frequency for the navitoclax + ruxolitinib arm, and summarized by SOC, PT and maximum NCI grade.

9.2.4 SAEs (Including Deaths) and Adverse Events Leading to Death

SAEs (including deaths) and AEs leading to death will be summarized by SOC and PT and in listing format.

9.2.5 Safety Topics of Interest

The safety topics of interest identified using standard MedDRA queries (SMQs) and custom searches are provided in Appendix B.

The incidence of each safety topic of interest will be summarized by SOC and PT. The AEs overview described in Section 9.2.2 and a listing of subject data will be provided for each search.

In addition, for each search, the risk difference for the overall incidence and the corresponding 95% confidence interval (CI) will be computed and displayed in a forest plot and summary table. The risk difference will be defined as the incidence proportion of the navitoclax + ruxolitinib treatment arm minus the incidence proportion of the placebo + ruxolitinib treatment arm. The normal approximation to the binomial distribution will be used to compute the 95% CI.

9.3 Analysis of Laboratory Data

Data collected from central and local laboratories, including additional laboratory testing due to an SAE, will be used in all analyses. Change from baseline will be summarized and presented for key lab parameters at scheduled post-baseline visits. Baseline is defined as the last non-missing observation before the first dose of study drug or randomization if no study drug is given.

For shifts relative to National Cancer Institute Common Toxicity Criteria for Adverse Events, baseline and post-baseline laboratory observations will be categorized as Grade 0, Grade 1, Grade 2, Grade 3, or Grade 4 according to NCI CTCAE grading. The baseline grade is defined as the grade of the last measurement collected on or prior to the first dose of study drug (or randomization for non-treated subjects). In cases where multiple values are collected on the same day, the maximum grade value will be selected as the value for that day for post-baseline values and the minimum grade will be selected as the value for that day for baseline values. If a subject had missing baseline and non-missing post-baseline for a given lab, the baseline grade will be assumed to be Grade 0. The maximum NCI toxicity grade value is the value with highest NCI toxicity grade collected after the first dose of study drug.

For each variable, cross tabulation tables will be generated that cross tabulate the number of subjects with baseline values of Grade 0, Grade 1, Grade 2, Grade 3, or Grade 4 versus maximum post-baseline values of Grade 0, Grade 1, Grade 2, Grade 3, or Grade 4. Subjects with missing measurements will be counted in the "missing" category. All treated subjects will be included in the cross tabulation regardless of whether baseline or post-baseline measurements are collected.

The laboratory shifts tables based on the two criteria below will be generated for each laboratory tests related to CTCAE:

1. Shifts from Grade 0 (Normal) at baseline to Grade 1 - 4 post-baseline (maximum) and worsening from an abnormal baseline value of at least one grade up post-baseline (maximum).
2. Shifts from Grade 0 - 2 at baseline to Grade 3 or 4 post-baseline (maximum) and from Grade 3 at baseline value to Grade 4 post-baseline (maximum).

The shift table analyses will include post-baseline measurements within 30 days of the last dose of the study drug. Detailed listings of all data for subjects experiencing NCI CTCAE Grade 3 to 4 blood chemistry and hematology values will be provided.

For key lab parameters including platelet count, lymphocytes, neutrophils, hemoglobin, total bilirubin, ALT, AST, and alkaline phosphatase, boxplots of laboratory values and summary tables will be displayed at scheduled visits.

Assessment of Hepatotoxicity

Elevations relative to the upper limit of normal (ULN) in ALT, AST, total bilirubin, and alkaline phosphatase as outlined in the US FDA Guidance for Industry[8] pertaining to premarketing clinical evaluations for drug-induced liver injury (DILI) will be summarized using the maximum post-baseline values. Laboratory values collected no more than 30 days following the last date of study drug will be used in the analysis.

The number and percentage of subjects who have at least one observed post baseline value meeting the following criteria will be tabulated by treatment:

- ALT: > 3 ×, > 5 ×, > 10 ×, or > 20 × ULN
- AST: > 3 ×, > 5 ×, > 10 ×, or > 20 × ULN
- Total bilirubin: >1.5 ×, > 2 × ULN
- Alkaline phosphatase: > 1.5 × ULN
- ALT and/or AST > 3 × ULN and total bilirubin > 1.5 × ULN at any visit
- ALT and/or AST > 3 × ULN and total bilirubin > 2 × ULN at any visit
- ALT and/or AST > 3 × ULN and total bilirubin > 2 × ULN within 72 hours
- ALT and/or AST > 3 × ULN and total bilirubin > 2 × ULN at the same visit

An evaluation of Drug Induced Serious Hepatotoxicity (eDISH) plot of the maximum post-baseline ALT value (as a multiple of ULN) vs. the maximum post-baseline total bilirubin value (as a multiple of ULN), not necessarily concurrent, will be presented, with reference lines for 3 × ULN for ALT and 2 × ULN for total bilirubin. A similar eDISH plot will be presented for AST vs. total bilirubin.

A listing of liver function tests for subjects with observed ALT and/or AST (> 3 × ULN) accompanied by total bilirubin (> 2 × ULN) at any time on study will be provided.

9.4 Analysis of Vital Signs

Change from baseline will be summarized and presented for each scheduled post-baseline visit for the vital sign parameters. Baseline is defined as the last non-missing observation before the first administration of study drug or randomization if no study drug is given.

9.5 Safety Subgroup Analyses

The incidence of treatment-emergent AEs overview, treatment-emergent AEs by SOC and PT, and safety topics of interest overview will be assessed for the subgroups defined below:

- Age (18 - < 65 years, ≥ 65 years, ≥ 75 years)
- Gender (Male, Female)
- Race (White, Black or African American, Asian, Other)
- Region (US, Europe, Japan, Other Regions)
- DIPSS+ Risk Group (Intermediate-1, Intermediate-2, High)
- Baseline Platelet Count (≤ 200 × 10^9/L, > 200 × 10^9/L)
- Baseline Renal Function (Normal, Mild Impairment, Moderate Impairment, Severe Impairment)
- Baseline Hepatic Function (Normal, Mild Impairment, Moderate Impairment, Severe Impairment)
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino)

10.0 Other Analyses

Summary statistics of plasma concentration of navitoclax will be tabulated for each visit.

11.0 Interim Analyses

There is no planned interim analysis for efficacy. An independent data monitoring committee will review the unblinded safety data on a regular basis to protect the safety of the subjects enrolled in the study.

11.1 Data Monitoring Committee

An independent data monitoring committee (IDMC) will regularly review unblinded safety data from the ongoing study according to the schedule provided in the IDMC charter, including adverse events and laboratory results. The unblinded safety analyses will be performed by an independent statistical data analysis center (SDAC) external to AbbVie and reviewed by the IDMC. AbbVie personnel will remain blinded and will not have access to the unblinded analyses prepared for the IDMC. The IDMC will provide recommendations to AbbVie as per the IDMC charter.

A separate IDMC charter describes the roles and responsibilities of the DMC members, frequency of data reviews, relevant data to be assessed, and general operations.

12.0 Overall Type-I Error Control

The fixed sequence testing procedure will be performed with a significance level of 0.05 (two-sided) for the primary efficacy endpoint and key secondary efficacy endpoints sequentially. The ranking of the endpoints is described in Table 1. If the statistical test is not significant for the primary efficacy endpoint, then statistical significance will not to be declared for any of the key secondary endpoints.

Except for overall survival and leukemia-free survival, the statistical test of other ranked efficacy endpoints will be performed one-time at the primary analysis of SVR_{35W24} rate or at Week 48, based on the testing sequence in Table 1. The planned alpha spending for overall survival and leukemia-free survival will be based on Lan-Demets alpha spending function with O'Brien-Fleming boundaries and is described in Table 1. The actual stopping boundaries will be derived based on the observed number of events in the extracted database. After the primary analysis of SVR_{35W24}, descriptive OS analyses in addition to the pre-planned analyses at 70% OS and 100% OS may be performed to support regulatory submissions as necessary.

Since there are no efficacy analyses for early stopping planned for the IDMC review, no alpha spending is needed for the IDMC review.

Table 1. Testing Sequence and Alpha Spending (Two-Sided) for Primary and Key Secondary Endpoints

Testing Sequence	Endpoint	Timing of Analysis			
		SVR_{35W24}	Week 48[a]	70% OS ([redacted] Death Events)	100% OS [redacted] Death Events)
1	SVR_{35W24} rate	0.05	No Test	No Test	No Test
2	Change from baseline in TSS score at Week 24	0.05	No Test	No Test	No Test
3	SVR_{35} rate	0.05	No Test	No Test	No Test
4	Anemia Response	0.05	No Test	No Test	No Test
5	Reduction in grade of bone marrow fibrosis	No Test	0.05	No Test	No Test
6	Overall survival	No Test	No Test	0.0149	0.0455
7	Leukemia-free survival	No Test	No Test	0.0149[b]	0.0455[b, d]
8	Change in fatigue at Week 24 from baseline	0.05[c]	No Test	No Test	No Test
9	Change in physical functioning at Week 24 from baseline	0.05[c]	No Test	No Test	No Test

a. "Week 48" denotes when the last subject enrolled has been followed up for at least 48 weeks.

b. Information fraction of OS is used as reference for the planned alpha spending.

c. Statistical significance will not be claimed until all endpoints ranked above are tested and claimed statistical significance.

d. The actual stopping boundary will be recalibrated based on the actual number of observed LFS events.

13.0 Version History

Table 2. SAP Version History Summary

Version	Date	Summary
1.0	19 Nov 2020	Original version
2.0	21 June 2021	The following updates were made: 1. Clarified the handling of missing baseline in the analysis of efficacy endpoints. 2. Clarified the additional analyses to be performed for the primary efficacy endpoint. 3. Clarified the handling of baseline score of 0 in the analysis of total symptom scores. 4. Added erythropoietin supplements to the definition of anemia response. 5. Updated the ranking of the key secondary endpoints based on Protocol Amendment 4. 6. Added the summary of difference and 95% confidence interval for difference to the analysis of response rates. 7. Updated OS projection at the analysis of SVR_{35W24}. 8. Clarified analyses of safety topics of interest and safety subgroups. 9. Corrected typos throughout the document.

Version	Date	Summary
3.0 DRAFT	20 October 2021	The following updates were made: 1. Clarified the analysis window for SVR_{35W24}. 2. Clarified that subjects with unevaluable spleen volume at baseline will be considered non-responders in the analyses of SVR_{35W24} and SVR_{35}. 3. Clarified that additional descriptive summaries will be provided in the analyses of TSS scores and SVR_{35}. 4. Clarified the analysis window and criteria for anemia response. 5. Clarified that data after disease progression will not be included in the analyses of certain efficacy endpoints. 6. Clarified definition of duration of SVR_{35}. 7. Clarified that the statistical test for change in fatigue from baseline at Week 24 will be based on the data cutoff at the primary analysis of SVR_{35W24}. 8. Added alcohol use to baseline characteristics summary. 9. Clarified that the risk difference and 95% CI for overall incidence for each safety topic of interest will be provided in summary tables in addition to forest plots. 10. Clarified that summary tables will be provided for key laboratory parameters in addition to boxplots. 11. Clarified wording for safety topics of interest. 12. Other editorial edits throughout the document.

Version	Date	Summary
3.0 DRAFT	22 July 2022	The following updates were made based on [redacted]: 1. Updated the handling of subjects without baseline TSS score or with a baseline TSS score of 0 that those subjects will be treated as non-responders in the analysis of TSS_{50W24}. 2. Updated the main analysis method of change from baseline in PROMIS Fatigue at Week 24. 3. Removed overall response of clinical improvement from the ranked key secondary endpoints. The following additional updates were made: 1. Updated the ranking of the key secondary endpoints. 2. Updated the analysis convention that stratification factors collected in EDC will be used in the stratified analyses. 3. Updated the categorization of DIPSS+ risk group. 4. Added sensitivity analysis of SVR_{35W24}. 5. Updated the analysis window for TSS_{50W24}. 6. Added sensitivity analysis of TSS_{50W24}. 7. Added analysis of TSS_{50} at any time, time to first $TSS_{50,}$ time to first SVR_{35}, time to first reduction in grade of bone marrow fibrosis. 8. Updated the analysis for post-baseline RBC transfusion. 9. Clarified the analysis of LFS, PFS and allelic mutations. 10. Added age subgroup to the efficacy subgroup analysis. 11. Added additional variables to baseline characteristics. 12. Other editorial edits throughout the documents.

Version	Date	Summary
3.0 DRAFT	29 November 2022	The following updates were made [redacted] 1. Changed the TSS endpoint from TSS_{50W24} to change in TSS score from baseline at Week 24. 2. Defined main and sensitivity analyses for change in TSS score from baseline at Week 24. 3. Defined TSS_{50W24} as a supplementary analysis. 4. Defined TSS_{50} rate at any time, time to TSS_{50}, duration of TSS_{50}, TSS response rate based on MCT, time to TSS response, and duration of TSS response as supplementary analyses. 5. Defined TSS Analysis Population. 6. Changed the main analysis of PROMIS Fatigue to linear mixed effects model and defined sensitivity analysis. The following additional updates were made: 1. Removed RBC Transfusion at Baseline from the baseline characteristics. 2. Added graphical summary of SVR. 3. Added Grade 3 TEAE and Grade 4 TEAE to AE overview. 4. Other editorial edits.

Version	Date	Summary
3.0	01 April 2023	The following updates were made [redacted] 1. Changed the key secondary endpoint TSS_{50W24} to change in TSS score at Week 24 from baseline. 2. Changed the key secondary endpoint time to deterioration in physical functioning to change in physical functioning score at Week 24 from baseline. 3. Added tipping point analysis for major SVR_{35W24} and TSS_{50W24}. 4. Changed the sensitivity analysis for continuous PRO endpoints to tipping point analysis. 5. Changed the analysis population for TSS analyses to ITT population. The TSS analysis population will be used for sensitivity analysis. 6. Added analysis of change in TSS score from baseline at Week 24 by excluding the fatigue symptom. 7. Clarified that all 7 questions need to be answered to calculate the daily TSS score. 8. Clarified that T-score will be used in the analysis of PROMIS Fatigue total score and raw score will be used in the analysis of PROMIS Fatigue subitems. The following additional updates were made: 1. Updated the ranking and analysis timing of the key secondary endpoints. 2. Clarified the strata-adjusted difference between the treatment arms will be summarized for binary endpoints. 3. Added the definition of confirmed disease progression. 4. Added summary of SVR_{35} rate at Week 12. 5. Added analysis of TSS response based on MCT at Week 24. 6. Added summary of maximum reduction in TSS from baseline at any time. 7. Added the value of MCT for TSS analysis. 8. Updated the definitions of duration of TSS_{50} and duration of TSS response based on MCT. 9. Clarified that all questions need to be answered to calculate PROMIS Fatigue total score and the subdomain score. 10. Added summary of reduction in grade of bone marrow fibrosis at Week 24

Version	Date	Summary
		11. Clarified that additional descriptive OS analysis may be performed based on regulatory interactions. 12. Added summaries for number and amount of PRBC transfusions. 13. Added summaries of ≥ 20% and ≥ 50% reduction in driver gene allele frequency. 14. Updated the summaries of JAK2, CALR, and MPL mutations. 15. Updated the analysis of correlation among efficacy endpoints. 16. Removed "Skin Cancer or Second Malignancies" from the safety topic of interest. 17. Added Ethnicity to safety subgroup analyses. 18. Added Race and Ethnicity to efficacy subgroup analyses. 19. Other editorial edits.

14.0 References

1. Verstovsek S, Mesa RA, Gotlib J, et al. A Double-Blind, Placebo-Controlled Trial of Ruxolitinib for Myelofibrosis. N Engl J Med. 2012;366(9):799-807.
2. Tefferi A, Cervantes F, Mesa R, et al. Revised response criteria for myelofibrosis: International Working Group-Myeloproliferative Neoplasms Research and Treatment (IWG-MRT) and European LeukemiaNet (ELN) consensus report. Blood. 2013;122(8):1395-8.
3. Gwaltney C, Paty J, Kwitkowski VE, et al. Development of a harmonized patient-reported outcome questionnaire to assess myelofibrosis symptoms in clinical trials. Leuk Res. 2017;59:26-31.
4. Cella D, Yount S, Rothrock N, et al. The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. Med Care. 2007;45 (5 Suppl 1):S3-11.

5. Aaronson NK, Ahmedzai S, Bergman B, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials oncology. J Natl Cancer Inst. 1993;85(5):365-76.
6. Thiele J, Kvasnicka HM, Facchetti F, et al. European consensus on grading bone marrow fibrosis and assessment of cellularity. Haematologica. 2005;90(8):1128-32.
7. Oppe M, Devlin NJ, van Hout B, et al. A program of methodological research to arrive at the new international EQ-5D-5L valuation protocol. Value Health. 2014;17(4):445-53.
8. Guidance for Industry: Drug-Induced Liver Injury: Premarketing Clinical Evaluation. US Food and Drug Administration Center for Drug Evaluation and Research. 2009.
9. Ramalingam SS, Kummar S, Saranopoulos J, et al. Phase I study of vorinostat in patients with advanced solid tumors and hepatic dysfunction: a National Cancer Institute Organ Dysfunction Working Group Study. J Clin Oncol. 2010;28(29):4507-12.
10. Tang Y. An Efficient Multiple Imputation Algorithm for Control-Based and Delta-Adjusted Pattern Mixture Models using SAS. Statistics in Biopharmaceutical Research. 2017;21(3):116-25.

Appendix A. Protocol Deviations

The number and percentage of subjects who reported at least one of the following protocol deviation categories will be provided.

- Subject entered into the study even though s/he did not satisfy entry criteria.
- Subject developed withdrawal criteria during the study and was not withdrawn.
- Subject received wrong treatment or incorrect dose of study.
- Subject took prohibited concomitant medication.

Appendix B. Definition of Safety Topics of interest

Safety topics of interest will be identified using the following search criteria:

Safety Topics of Interest	Criteria for Identification of Events
Thrombocytopenia	PTs: 10043554 Thrombocytopenia 10035528 Platelet Count Decreased
Neutropenia	PTs: 10016288 Febrile Neutropenia 10029354 Neutropenia 10029366 Neutrophil Count Decreased
Hemorrhagic Events	SMQ: 20000038 Haemorrhages (Narrow)
Serious Hemorrhagic Events Concurrent[a] with Grade 3 or 4 Thrombocytopenia	SAEs in the Hemorrhagic Events search concurrent with Grade 3 or 4 AE in Thrombocytopenia search
Serious Infections	SAEs in the SOC of Infections and Infestations
Serious Infections with Concurrent[a] with Grade 3 or 4 Neutropenia	SAEs in the Serious Infections search concurrent with Grade 3 or 4 AE in Neutropenia search
Drug-Induced Liver Injury (DILI)[b]	SMQs: 20000006 Drug Related Hepatic Disorders - Comprehensive (Broad) 20000007 Drug Related Hepatic Disorders - Severe Events Only (Narrow)

a. An AE will be considered concurrent with cytopenia event if the onset date is no more than 7 days prior to the onset of the cytopenia and no more than 7 days after the end of the cytopenia event.

b. The Drug Related Hepatic Disorders – Comprehensive SMQ (20000006) includes all PTs in the Drug Related Hepatic Disorders - Severe Events Only SMQ (20000007). Summaries will be provided for each SMQ.

Appendix C. Renal and Hepatic Function

Renal function is classified by the baseline creatinine clearance (mL/min) as follows.

Renal Function	CrCL (mL/min)
Normal	≥ 90
Mild Impairment	60 ≤ CrCL < 90
Moderate Impairment	30 ≤ CrCL < 60
Severe Impairment	CrCL < 30

Creatinine clearance is calculated by the Cockcroft Gault formula.

$$\text{CrCL} = \frac{(140 - \text{Age}) \bullet (\text{Weight in kg}) \bullet [0.85 \text{ if Female}]}{72 \bullet \text{Serum Creatinine (mg/dL)}}$$

Or, if serum creatinine is in µmol/L:

$$\text{CrCL} = \frac{(140 - \text{Age}) \bullet (\text{Weight in kg}) \bullet [1.23 \text{ if Male, } 1.04 \text{ if Female}]}{\text{Serum Creatinine (µmol/L)}}$$

Hepatic function is classified according to the baseline bilirubin and AST values as follows based on the National Cancer Institute Organ Dysfunction Working Group (NCI-ODWG) definition.[9]

Hepatic Function	Bilirubin (mg/dL)	AST (IU/L)
Normal	≤ ULN	≤ ULN
Mild Impairment	≤ ULN > ULN and ≤ 1.5 × ULN	> ULN Any value
Moderate Impairment	> 1.5 × ULN and ≤ 3 × ULN	Any value
Severe Impairment	> 3 × ULN	Any value

ULN is based on reported values from the lab.