# Statistical Analysis Plan for Study M19-051

# A Multicenter, Randomized, Double-Blind, Placebo-Controlled Dose-Ranging Study to Evaluate the Safety and Efficacy of Upadacitinib in Subjects with Non-Segmental Vitiligo

## Date: 27 December 2022

## Version 2.0

## 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 upadacitinib Study M19-051 'A Multicenter, Randomized, Double-Blind, Placebo-Controlled Dose-Ranging Study to Evaluate the Safety and Efficacy of Upadacitinib in Subjects with Non-Segmental Vitiligo.'

Study M19-051 examines the efficacy and safety of upadacitinib for the treatment of adult subjects with non-segmental vitiligo (NSV).

The analyses of pharmacokinetic endpoints and biomarkers 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, Hypotheses and Estimands

The primary objective of this study is to evaluate the safety and efficacy of upadacitinib for the treatment of adult subjects with NSV. The primary efficacy objective is based on the percent change from Baseline in Facial Vitiligo Area Scoring Index (F-VASI) at Week 24 with upadacitinib treatment compared to placebo in the Intent-to-Treat (ITT) Population, which consists of all randomized subjects.

The hypothesis corresponding to the primary endpoint is:

Percent change from Baseline in F-VASI with upadacitinib is greater than that with placebo at Week 24.

The estimand corresponding to the primary endpoint is defined as:

Difference in the percent change from Baseline in F-VASI at Week 24, regardless of treatment discontinuation, between each of the upadacitinib dose groups compared with placebo in the adult subjects with NSV.

## 2.2 Study Design Overview

This is a Phase 2, multicenter, randomized, double-blinded, parallel-group, placebo-controlled dose-ranging study that will evaluate the safety and efficacy of upadacitinib in adult subjects ≥ 18 to 65 years of age with NSV. The study is comprised of a 35-day Screening Period, a 24-week double-blind treatment period (Period 1), a 28-week blinded long-term extension (Period 2), and a 30-day Follow-up Period.

Subjects who meet eligibility criteria at Baseline will be randomized in a 2:2:2:1:1 ratio to one of five treatment groups:

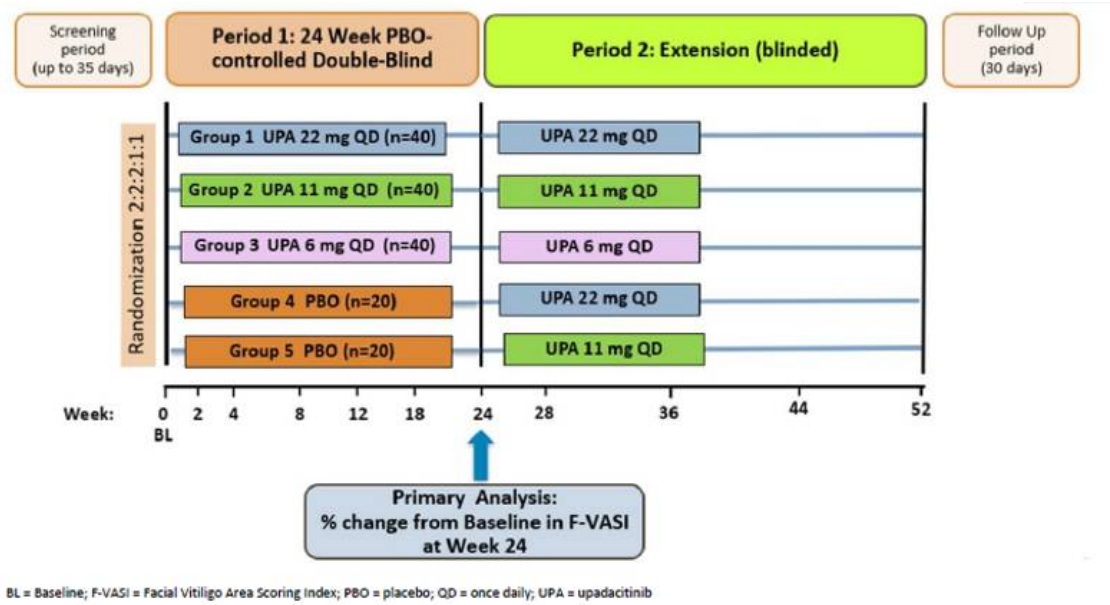
- Group 1: upadacitinib 22 mg once daily (QD) (N = 40) (Period 1) → upadacitinib 22 mg QD (Period 2)
- Group 2: upadacitinib 11 mg QD (N = 40) (Period 1) → upadacitinib 11mg QD (Period 2)
- Group 3: upadacitinib 6 mg QD (N = 40) (Period 1) → upadacitinib 6 mg QD (Period 2)
- Group 4: placebo (N = 20) (Period 1) → upadacitinib 22 mg QD (Period 2)
- Group 5: placebo (N = 20) (Period 1) → upadacitinib 11 mg QD (Period 2)

At Week 24, subjects who were randomized to placebo at Baseline will be switched to either upadacitinib 22 mg (Group 4) or upadacitinib 11 mg (Group 5) in a blinded fashion per pre-specified randomization assignments.

The AbbVie study team will be unblinded to perform the Week 24 primary analysis. The unblinding will take place after all subjects have completed the Week 24 visit or have prematurely discontinued prior to Week 24. Sites and subjects will remain blinded throughout the study.

The schematic of the study is shown in Figure 1.

**Figure 1. Study Schematic**




BL = Baseline; F-VASI = Facial Vitiligo Area Scoring Index; PBO = placebo; QD = once daily; UPA = upadacitinib

## 2.3 Treatment Assignment and Blinding

All subjects will be assigned a unique identification number by the IRT at the screening visit. For subjects who rescreen, the screening number assigned by the IRT at the initial screening visit should be used. The IRT will assign a randomization number that will encode the subject's treatment group assignment according to the randomization schedule.

Subjects who meet eligibility criteria at Baseline will be randomized in a 2:2:2:1:1 ratio to one of the five treatment groups:

- Group 1: upadacitinib 22 mg QD (N = 40) (Period 1) → upadacitinib 22 mg QD (Period 2)
- Group 2: upadacitinib 11 mg QD (N = 40) (Period 1) → upadacitinib 11mg QD (Period 2)

- Group 3: upadacitinib 6 mg QD (N = 40) (Period 1) → upadacitinib 6 mg QD (Period 2)
- Group 4: placebo (N = 20) (Period 1) → upadacitinib 22 mg QD (Period 2)
- Group 5: placebo (N = 20) (Period 1) → upadacitinib 11 mg QD (Period 2)

Randomization will be stratified by the age group (≤ 50 and > 50), Baseline disease severity (T-VASI < 15 and ≥ 15), and status of active vitiligo (Yes/No) defined as follows:

- Active: Showing new lesions or progression (enlargement) of existing lesions within the last 6 months and/or presenting clinical subtypes indicative of progressing vitiligo (i.e., confetti-like depigmentation, trichrome pattern, or Koebner phenomenon).

At Week 24, subjects who were randomized to placebo at Baseline will be switched to either 22 mg (Group 4) or 11 mg (Group 5) upadacitinib in a blinded fashion per pre-specified randomization assignments.

### 2.4 Sample Size Determination

Assuming a Week 24 percent change from Baseline in F-VASI of 0% in the placebo arm, a conservative estimate based on a prior study in vitiligo,[1] the planned total sample size of 160 subjects (40 subjects in each upadacitinib group, 40 subjects in total in the placebo groups) will provide more than 90% power to detect the treatment difference of 40% reduction (assuming a standard deviation of 54.8%) in at least 1 upadacitinib group versus placebo using a 2-sided significance level of 0.1 based on the two-sample t-test.

## 3.0 Endpoints

### 3.1 Primary Endpoint

The primary endpoint is the percent change from Baseline in F-VASI at Week 24.

## 3.2 Secondary Endpoints

- Achievement of F-VASI 75 (≥ 75% improvement in F-VASI from Baseline) at Week 24;
- Achievement of F-VASI 50 (≥ 50% improvement in F-VASI from Baseline) at Week 24;
- Achievement of total Vitiligo Area Scoring Index (T-VASI) 50 (≥ 50% improvement in T-VASI from Baseline) at Week 24;
- Percent change from Baseline in T-VASI at Week 24;
- Change from Baseline in the vitiligo quality-of-life (VitiQoL) total score at Week 24.

## 3.3 Other Efficacy Endpoints

Additional Efficacy Endpoints include the primary and all secondary endpoints assessed at visits as noted in the Study Activities table (Protocol Appendix D), other than Week 24. Additional Efficacy Endpoints also include the following measurements assessed at visits as specified in the Study Activities table (Protocol Appendix D):

- Achievement of F-VASI 90 (≥ 90% improvement in F-VASI from Baseline);
- Achievement of T-VASI 75 (≥ 75% improvement in T-VASI from Baseline);
- Achievement of T-VASI 90 (≥ 90% improvement in T-VASI from Baseline);
- Mean and percent change from Baseline in the vitiligo extent score (VES);
- Achievement of vitiligo noticeability scale (VNS) score of "A lot less noticeable (4)" or "No longer noticeable (5)";
- Achievement of Dermatology Life Quality Index (DLQI) total score of "0" or "1";
- Change from Baseline in the Hospital Anxiety and Depression Scale (HADS) anxiety and depression scores;
- Proportion of subjects selecting each of the response categories of VitiQoL 16 questions;

- Achievement of Physician's Global Impression of Change-Vitiligo (PhGIC-V) of "Much better (1)" or "A little better (2)";
- Achievement of Patient's Global Impression of Change-Vitiligo (PaGIC-V) of "Much better (1)" or "A little better (2)";
- Achievement of Face – Physician Global Vitiligo Assessment (F-PhGVA) of "No depigmentation (0)" or "Limited extent of depigmentation (1)";
- Achievement of Total – Physician Global Vitiligo Assessment (T-PhGVA) of "No depigmentation (0)" or "Limited extent of depigmentation (1)";
- Achievement of Face – Patient Global Vitiligo Assessment (F-PaGVA) of "No depigmentation (0)" or "Limited extent of depigmentation (1)";
- Achievement of Total – Patient Global Vitiligo Assessment (T-PaGVA) of "No depigmentation (0)" or "Limited extent of depigmentation (1)."

### 3.4 Safety Endpoints

Safety evaluations include AEs, serious adverse events (SAEs), adverse events of special interest (AESIs), AEs leading to discontinuation, vital signs, laboratory tests (hematology, chemistry, liver function tests), and physical examination.

## 4.0 Analysis Populations

The following Intent-to-Treat (ITT) populations will be used for the analyses.

- The ITT Population in Period 1 (ITT_1) includes all randomized subjects in Period 1. Subjects will be included in the analysis according to the treatment groups that they are randomized to. Analyses will include the following 4 groups: upadacitinib (UPA) 22 mg QD, UPA 11 mg QD, UPA 6 mg QD and placebo.
- The ITT Population in Period 2 (ITT_2) includes all subjects entered Period 2. Subjects will be included in the analysis according to the 5 treatment groups that they are randomized to at baseline: UPA 22 mg QD-> UPA 22 mg QD, UPA 11 mg QD-> UPA 11 mg QD, UPA 6 mg QD-> UPA 6 mg QD, Placebo-> UPA 22 mg QD and Placebo-> UPA 11 mg QD.

- The ITT_1/ITT_2 Populations will be used for all efficacy analyses across Period 1/Period 2.

The following three Safety Populations will be used for the safety analysis:

- The Safety Population in Period 1 (Safety_1) is defined as all subjects who are randomized and received at least one dose of study drug in Period 1. Subjects will be summarized in the following 4 groups: UPA 22 mg QD, UPA 11 mg QD, UPA 6 mg QD and placebo, based on the treatment actually received (defined as the most frequent dose regimen received in Period 1).
- The Safety Population in Period 2 (Safety_2) is defined as all subjects received at least one dose of study drug in Period 2. Subjects will be included in the 5 groups same as which they were randomized to at baseline: UPA 22 mg QD-> UPA 22 mg QD, UPA 11 mg QD-> UPA 11 mg QD, UPA 6 mg QD-> UPA 6 mg QD, Placebo-> UPA 22 mg QD and Placebo-> UPA 11 mg QD.
- The all upadacitinib treated (ALL_UPA) Population is defined as subjects who received at least one dose of upadacitinib in the study. Subjects will be analyzed based on the treatment actually received: UPA 22 mg QD, UPA 11 mg QD and UPA 6 mg QD.

# 5.0 Subject Disposition

The number of subjects for each of the following categories will be summarized in Period 1, for overall and for each treatment group in ITT_1 Population:

- Subjects randomized;
- Subjects who took at least one dose of study drug in Period 1;
- Subjects who completed Period 1;
- Subjects who completed study drug in Period 1;
- Subjects who prematurely discontinued from study in Period 1;
- Subjects who prematurely discontinued study drug in Period 1;

The summary will also be provided for Period 2 for each of the following categories and by treatment groups in the ITT_2 Population:

- Subjects who entered Period 2;
- Subjects who took at least one dose of study drug in Period 2;
- Subjects who completed study;
- Subjects who completed study drug in Period 2;
- Subjects who prematurely discontinued from study in Period 2;
- Subjects who prematurely discontinued study drug in Period 2.

Number and percentage of subjects who discontinue study drug or who discontinue from study will be summarized by primary reason for each treatment group and overall.

In addition, the above summaries (except for the reason for premature discontinuation) will also be summarized by center in the accountability table.

# 6.0 Study Drug Duration and Compliance

For Safety_1/Safety_2 Population, duration of treatment will be summarized for each treatment group in Period 1 and Period 2. For ALL_UPA Population, duration of treatment will be summarized overall and for each upadacitinib dose group. Duration of treatment is defined as follows:

**Period 1:**

For subjects who did not continue into Period 2:

- Date of last dose of study drug in Period 1 – Date of first dose of study drug in Period 1 + 1.

For subjects who continued into Period 2:

- Minimum of (Date of first dose of study drug in Period 2 – Date of first dose of study drug in Period 1, Date of last dose of study drug in Period 1 – Date of first dose of study drug in Period 1 + 1.)

**Period 2:**

For subjects who continued into Period 2:

- Date of last dose of study drug in Period 2 – Date of first dose of study drug in Period 2 + 1.

**ALL_UPA:**

Study drug duration during the administration of study drug in ALL_UPA Population is defined as:

- Date of last dose of upadacitinib – Date of first dose of upadacitinib + 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 each treatment duration interval (≥ 4 weeks, ≥ 12 weeks, ≥ 24 weeks, ≥ 36 weeks, ≥ 48 weeks) will be summarized for ALL_UPA Population.

Treatment compliance for Period 1 and 2 will be summarized by treatment groups for Safety Populations (Safety_1 and Safety_2). For ALL_UPA Population, treatment compliance will be summarized overall and for each upadacitinib dose group. Treatment compliance is defined as the number of tablets actually taken divided by the number of tablets that should have been taken. For subjects who prematurely discontinued the study drug, the planned tablets will only be counted prior to that scheduled visit of discontinuation. Percent compliance will be summarized.

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

Demographics, baseline or disease characteristics, and medical history will be summarized for the ITT_1 Population overall and by treatment groups. Prior medications will be summarized for the Safety_1 Population overall and by treatment groups. Concomitant medications will be summarized for the Safety_1 and Safety_2 Populations overall and by treatment groups. Categorical variables will be summarized with the number and percentage of subjects; percentages will be calculated based on the number of non-missing observations. 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

Continuous demographic variables include age, weight, height, and body mass index (BMI). Categorical demographic variables include sex, ethnicity, race, country, age ($\leq 50$ and $> 50$), weight ($< 60$ or $\geq 60$ kg), BMI ($< 25$, $\geq 25$ - $< 30$ or $\geq 30$ kg/$m^2$), tobacco user (current, former, never, unknown), and alcohol user (current, former, never, unknown).

Disease characteristics include the following Baseline F-VASI, Baseline T-VASI, Baseline VitiQoL total score and each domain score, Baseline VES, Baseline HADS, vitiligo disease diagnosis duration in years as continuous variables and Baseline VNS, Baseline DLQI, Baseline F-PhGVA, Baseline T-PhGVA, Baseline F-PaGVA, Baseline T-PaGVA, subject Fitzpatrick skin type, history of vitiligo-related symptoms (increased sensitivity, itch, and pain), status of active vitiligo, Baseline T-VASI ($\leq 10$ and $> 10$)as categorical variables.

## 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 (by MedDRA system organ class and preferred term) will be summarized overall and by treatment groups in the Intent-to-Treat Population in Period 1 (ITT_1). 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 will be summarized by generic name in the Safety population. A prior medication is defined as any medication taken prior to the date of the first dose of study drug and will be summarized in Period 1. For Period 1 and 2, a concomitant medication in each period is defined as any medication that started prior to the date of the first dose of study drug in each period and continued to be taken after the first dose of study drug in each period or any medication that started on or after the date of the first dose of study drug in each period, but not after the date of the last dose +1 of study drug in each period. 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.

In addition, subjects' prior exposure to topical treatment (topical corticosteroid [TCS], topical calcineurin inhibitor [TCI]), systemic corticosteroids, phototherapy, procedure, and other will be summarized for each treatment group.

## 8.0 Handling of Potential Intercurrent Events for the Primary Endpoint

The primary efficacy endpoint of percent change in F-VASI at Week 24 (defined in Section 3.1) will be analyzed based on the ITT_1 Populations. For treatment policy strategy, the occurrence of the intercurrent event (treatment discontinuation) is considered as irrelevant in defining the treatment effect of interest. Under this strategy, the value for the variable of interest is used regardless of whether or not the intercurrent event occurs.

# 9.0 Efficacy Analyses

## 9.1 General Considerations

All efficacy analyses will be conducted in the ITT Population in each period (ITT_1 and ITT_2). All tests will be 2-sided at an alpha level of 0.1.

The Primary Analysis will be performed after all ongoing subjects have completed Week 24 or have prematurely discontinued prior to Week 24, data pertaining to Period 1 has been cleaned, and the database has been locked. This will be the only and final analysis for the efficacy endpoints in Period 1.

For analysis in Period 1, continuous endpoints and categorical endpoints will be analyzed using Mixed-Effect Model Repeat Measurement (MMRM) and Cochran-Mantel-Haenszel (CMH), respectively.

The analysis models and imputation methods for categorical and continuous endpoints in Period 1 and Period 2 are summarized in the following table:

## Table 1. Analysis Models and Imputation Methods

| Period | Type of Variables | Imputation Method | Intercurrent Events Handling | Analysis Model |
|---|---|---|---|---|
| Period 1 | Continuous | MMRM[a] | The treatment policy strategy is used in this study. Under this strategy, all values for the variable of interest are used regardless of the intercurrent event. | MMRM model includes treatment, visit and treatment by visit interaction, stratification factors (age group [≤ 50 and > 50][e], baseline disease severity [T-VASI < 15 and ≥ 15][d], and status of active vitiligo [Yes/No]), and baseline measurement |
| | Categorical | NRI-MI[b] | The treatment policy strategy is used in this study. Under this strategy, all values for the variable of interest are used regardless of the intercurrent event. | CMH test adjusting for stratification factors (age group [≤ 50 and > 50], baseline disease severity [T-VASI < 15 and ≥ 15], and status of active vitiligo [Yes/No]) |
| Period 2 | Continuous | AO[c] | The treatment policy strategy is used in this study. Under this strategy, all values for the variable of interest are used regardless of the intercurrent event. | ANCOVA includes treatment, stratification factors (age group [≤ 50 and > 50], baseline disease severity [T-VASI < 15 and ≥ 15][d], and status of active vitiligo [Yes/No]), and baseline measurement |
| | Categorical | AO[c] | The treatment policy strategy is used in this study. Under this strategy, all values for the variable of interest are used regardless of the intercurrent event. | Descriptive summary |

CMH = Cochran-Mantel-Haenszel; ANCOVA = Analysis of Covariance

a. Mixed-Effect Model Repeat Measurement.

b. Non-Responder Imputation while incorporating Multiple Imputation (MI) to handle missing data due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random.

c. As observed (AO): The AO analysis will not impute values for missing evaluations, and thus a subject who does not have an evaluation at a scheduled visit will not be included in the AO analysis for that visit.

d. The stratification factor of baseline disease severity (T-VASI <15 and ≥ 15) will not be included in model for percent change from Baseline in T-VASI.

e. If the MMRM model cannot converge, the stratification factor of age will not be included in the model.

To evaluate long-term efficacy, all endpoints (categorical and continuous) will be summarized descriptively at each visit by treatment groups overall and by stratification factors.

**Analysis of Continuous Variables**

In Period 1, the Baseline means and visit means will be presented for each treatment group who have both Baseline and post-baseline visit values. Change (and/or percent change) from Baseline in the treatment groups will be compared using MMRM as described in Table 1. Point estimates, SE, and 95% confidence interval (CI) of LS mean change (and/or percent change) from Baseline within treatment groups, along with p-values between each upadacitinib treatment group and placebo will be provided. MMRM will be the primary approach to handle missing data for continuous endpoints (Section 9.2) in Period 1. In addition, the primary efficacy endpoint will be analyzed using MI defined in Section 9.2 as the sensitivity analysis.

For long-term efficacy, the raw value at Baseline and each visit will be summarized using mean for each of the 5 treatment groups as observed. For change (and/or percent change) from Baseline, LS mean, SE, and 95% CIs generated by ANCOVA by visit as described in Table 1. Descriptive summary statistics including median, Q1, Q3, minimum and maximum of change (and/or percent change) from Baseline will also be provided.

**Analysis of Categorical Variables**

In Period 1, frequencies and percentages will be provided along with 95% CIs. Point estimates, 95% CIs and p-values from the CMH test for the difference in proportions between each upadacitinib group and placebo will be provided. Breslow-Day test will be performed to test the homogeneity between strata.

Non-Responder Imputation while incorporating Multiple Imputation (MI), NRI-MI, will be the primary approach to handle missing data for categorical endpoints in Period 1 (Section 9.2 and Appendix E).

- If there are missing data due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random, a total of 30 datasets will be generated by PROC MI, and subjects will be characterized as responders or non-responders based on MI imputed values.
  - If not all the imputed datasets are identical, analysis will be performed as described in Appendix E.
  - If all the 30 imputed datasets are identical, all statistical inferences will be calculated using the first imputed dataset: 1) The corresponding 95% CIs will be constructed based on normal approximation to the binomial distribution. 2) Adjusted differences, 95% CIs and p-values for the comparisons between each upadacitinib group and placebo will be calculated based on CMH test adjusting for stratification factors.
- If there are no missing data due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random, the NRI-MI method will be the same as traditional NRI method. All statistical inferences will be calculated using the NRI imputed datasets.

For long-term efficacy, descriptive summary statistics including frequencies and percentages for categorical endpoints, along with 95% CIs based on normal approximation, will be provided based on as observed data.

Unless otherwise specified, any subject who is randomized based on a wrong stratum will be analyzed according to the actual stratum the subject belongs to.

"Baseline" refers to the last non-missing observation before the first administration of study drug or randomization if no study drug is given.

## 9.2 Handling of Missing Data

In Period 1, missing data will be imputed using the following methods for the efficacy analyses:

- Mixed-Effect Model Repeat Measurement (MMRM): The repeated measures analysis will be conducted using a mixed model including observed

measurements at all visits. The mixed model includes treatment, visit and treatment-by-visit interaction, and stratification factors (age group [≤ 50 and > 50], Baseline disease severity [T-VASI < 15 and ≥ 15], and status of active vitiligo [Yes/No]) derived from actual values as fixed factors, and Baseline value as a covariate. An unstructured variance covariance matrix will be used. If the model cannot converge, an appropriate covariance structure matrix (e.g., autoregressive (1) or compound symmetry) will be used, and/or the stratification factor of age will not be included in the model. Parameter estimation is based on the assumption of data being missing at random and using the method of restrictive maximum likelihood (REML). MMRM will be the primary approach in the analysis of continuous variables in Period 1.

- Non-Responder Imputation while incorporating Multiple Imputation (MI) to handle missing data due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random (NRI-MI): The NRI-MI will categorize any subject who does not have an evaluation during a pre-specified visit window (either due to missing assessment or due to early withdrawal from the study) as a non-responder for the visit. The only exceptions are:
  1) when the subject is a responder both before and after the visit window in Period 1, the subject will be categorized as a responder for the visit.
  2) missing data due to COVID-19 infection or logistical restriction or any other missing data that can be reasonably assumed to be Missing at Random will be handled by MI. The NRI-MI will be the primary approach in the analyses of categorical variables in Period 1. Random seeds are provided in Appendix D and more details are provided in Appendix E. NRI-MI is the primary approach to handle missing data for categorical endpoints in Period 1.
- Multiple Imputation (MI): The MI approach will be used as a sensitivity analysis for the primary efficacy endpoint. Markov Chain Monte Carlo (MCMC) will be first applied to augment data into monotonic missing pattern and PROC MI will be used to generate 30 datasets using the regression method. The variables to be included in the imputation model are: treatment group, stratification factors (age group [≤ 50 and > 50], Baseline disease severity [T-VASI < 15 and ≥ 15], and status of active vitiligo [Yes/No]) derived from actual values, Baseline, and measurements at each visit up to the end of the Period 1. The random seed for MCMC and the random seed for

PROC MI are specified in Appendix D. The imputed post-baseline measurements will be rounded to the same precision as the observed data. The imputed endpoint will be analyzed using each of the 30 datasets using ANCOVA for the primary endpoint. SAS PROC MIANALYZE will be used to generate the final inferences of the difference between upadacitinib group and placebo.

In Period 2, all values collected in the study will be used to summarize the long-term efficacy as observed (AO) in Period 2:

- The AO analysis will not impute values for missing evaluations, and thus a subject who does not have an evaluation at a scheduled visit will not be included in the AO analysis for that visit.

## 9.3 Primary Efficacy Endpoint and Analyses

### 9.3.1 Primary Efficacy Endpoint

The primary endpoint is the percent change from Baseline in F-VASI at Week 24.

The corresponding statistical null hypothesis to the primary endpoint is that there is no difference between any upadacitinib treatment group and placebo, in the LS mean of percent change from Baseline in F-VASI at Week 24.

### 9.3.2 Main Analysis of Primary Efficacy Endpoint

The primary analysis of the primary efficacy endpoint will be conducted in the ITT_1 Population based on treatment as randomized. The primary analysis is the comparison between upadacitinib and placebo at Week 24 using MMRM, and the corresponding analyses are specified in Section 9.1.

The attributes of the estimands corresponding to the primary efficacy endpoint are summarized in Table 2.

**Table 2. Summary of the Estimand Attributes of the Primary Efficacy Endpoint**

| | Attributes of the Estimand | | | | |
|---|---|---|---|---|---|
| **Estimand Label** | **Treatment** | **Endpoint** | **Population** | **Intercurrent Events** | **Statistical Summary** |
| Percent change from Baseline in F-VASI at Week 24 | Upadacitinib 22 mg QD/ Upadacitinib 11 mg QD/ Upadacitinib 6 mg QD/ Placebo | Percent change from Baseline in F-VASI at Week 24 | adult subjects with NSV | For treatment policy strategy, the occurrence of the intercurrent event (treatment discontinuation) is considered irrelevant in defining the treatment effect of interest. Under this strategy, all values for the variable of interest are used regardless of the intercurrent event. | Difference in the percent change from Baseline in F-VASI at Week 24 between each upadacitinib dose and placebo |

The second analysis of the primary efficacy endpoint is to test a pre-specified set of dose-response models among upadacitinib dose groups and placebo group at Week 24 using Multiple Comparison Procedure – Modeling (MCP-Mod).

The dose-response relationship among upadacitinib dose groups and the placebo will be characterized for the primary endpoint using MCP-Mod[2]. The percent change from Baseline and standard deviations in each upadacitinib dose groups and placebo based on the primary analysis will be used, and ADDPLAN DF software Version 4.0 will be used to perform the MCP-Mod analyses.

A set of 6 pre-specified standardized candidate dose-response models, as described in Table 3, will be utilized to examine the dose-response relationship. For the primary endpoint, a statistically significant dose-response relationship will be declared if at least one model is identified by the MCP-Mod method to be statistically significant at 2-sided significance level of 0.1. Instead of choosing an optimal dose-response relationship model, a weighted average of the minimum effective dose (MED) (i.e., the recommended MED) across all significant models will be calculated for the MED selection.

1. The MED of each statistically significant model will be determined based on the pre-specified clinical meaningful target of a 40% target difference from placebo.
2. The weighted average of MED across all significant models will be calculated, with the weight $W_i$ of model $i$ calculated using Akaike information criterion (AIC) as:

$$W_i = \frac{exp(-0.5(AIC_i - \overline{AIC}))}{\sum_{j=1}^{J} exp(-0.5(AIC_j - \overline{AIC}))},\ \overline{AIC} = \frac{1}{J}\sum_{j=1}^{J} AIC_j, i = 1,\ldots,J,$$

$$\text{J } is\ the\ number\ of\ significant\ models.$$

The rationale for using a weighted average of MED is because selecting a single model as the 'best' representation of observed data and subsequently using it to estimate the MED does not account for uncertainty in model selection. The uncertainty in model selection can be accounted for by averaging estimates of MED over all significant models.

**Table 3. Candidate Models**

| Model | $f(d,\theta)$<br>$d = dose,$<br>$\theta = Model\ Parameters$ | $f^0(d,\theta)$<br>Standardized Model | Initial Value(s) for Parameter(s) |
|---|---|---|---|
| **Linear** | $E_0 + \delta d$ | $d$ | NA |
| **Exponential** | $E_0 + E_1\left[exp\left(\frac{d}{\delta}\right) - 1\right]$ | $exp\left(\frac{d}{\delta}\right) - 1$ | $\delta$ =44 |
| **Logistic** | $E_0 + \frac{E_{max}}{1 + exp\left(\frac{ED_{50} - d}{\delta}\right)}$ | $\frac{1}{1 + exp\left(\frac{ED_{50} - d}{\delta}\right)}$ | $ED_{50} = 4.45, \delta = 5.96$ |
| **Emax** | $E_0 + \frac{E_{max} d}{ED_{50} + d}$ | $\frac{d}{ED_{50} + d}$ | $ED_{50} = 2.41$ |
| **SigEMax** | $E_0 + \frac{E_{max} d^h}{ED_{50} + d^h}$ | $\frac{d^h}{ED_{50} + d^h}$ | $ED_{50} = 7.28, h = 2.66$ |
| **Quadratic** | $E_0 + \beta_1 d + \beta_2 d^2$ | $d + \frac{p_2}{\lvert\beta_1\rvert} d^2$ | $\delta$ =-0.02 |

Note: For logistic, Emax, SigEMax, Quadratic models, the initial values for parameters are based on the assumptions that upadacitinib 11 mg will achieve 75% of the maximum efficacy effect, and upadacitinib 22 mg will achieve 95% of the maximum efficacy effect. For exponential model, the initial values for the parameter is solely based on the assumption that upadacitinib 22 mg will achieve 95% of the maximum efficacy effect.

The fitted dose-response curves will be presented graphically for all statistically significant models along with confidence bands.

### 9.3.3 Sensitivity Analyses of the Primary Efficacy Endpoint

A sensitivity analysis using MI to handle missing data for the primary efficacy endpoint will be performed, as described in Section 9.2.

## 9.4 Secondary Efficacy Analyses

### 9.4.1 Main Analyses of Secondary Efficacy Endpoints

Secondary efficacy endpoints in Period 1 will be analyzed by comparing each upadacitinib treatment group with placebo.

Comparisons of the categorical endpoints will be made between each upadacitinib treatment group and placebo using the CMH test, adjusting for the stratification factors derived from actual values under a two-sided significance level of 0.1 for each upadacitinib group, and the corresponding analyses are specified in Section 9.1.

The continuous endpoints will be analyzed using MMRM at a two-sided significance level of 0.1 for each upadacitinib group, and the corresponding analyses are specified in Section 9.1 and Section 9.2.

## 9.5 Additional Efficacy Analyses

In Period 1, unless otherwise specified, additional efficacy endpoints will be compared between the each upadacitinib treatment group and placebo. The categorical endpoints and continuous endpoints will be analyzed by CMH and MMRM, respectively, and the analyses are specified in Section 9.1 and Section 9.2. Proportion of subjects selecting each of the 16 VitiQoL questions will be summarized for each of the 16 questions by response categories at each visit as observed.

In Period 2, the categorical endpoints will be summarized by treatment groups as observed, and the continuous endpoints will be analyzed as observed using ANCOVA as described in Table 1. The analyses are specified in Section 9.1 and Section 9.2.

In Addition, at the time of the Primary Analysis, additional long-term analysis up to Week 52 will be provided for the below variables among all subjects who may potentially reach to Week 52 by the data cutoff using NRI-MI (categorical variables) or MMRM (continuous variables):

- F-VASI 50, F-VASI 75,
- T-VASI 50,
- Percent change from Baseline in F-VASI,
- Percent change from Baseline in T-VASI

At the time of the final analysis, the above variables will be evaluated up to Week 52 using NRI-MI (categorical variables) or MMRM for the continuous variables.

In addition, for VitiQoL, VNS, DLQI and HADS, analyses will be performed, for up to Week 52, for the following:

- Change from Baseline in VitiQoL question 16 regarding the severity of skin condition.
- Proportion of subjects selecting each of the response categories of VNS
- Change from Baseline in Dermatology Life Quality Index (DLQI) total score
- Achievement of a clinically meaningful improvement in DLQI total score defined as $\Delta$ DLQI $\geq 4$ (for subjects with DLQI $\geq 4$ at baseline)
- Achievement of a clinically meaningful improvement in HADS anxiety and depression scores defined as $\Delta$ HADS $\geq 1.5$ (for subjects with HADS $\geq 1.5$ at baseline)

## 9.6 Efficacy Subgroup Analyses

To evaluate the consistency of the efficacy over demographic and other Baseline characteristics, the primary efficacy endpoint will be analyzed in the following subgroups:

- Age group ($\leq 50$ and $> 50$)
- Age group (($\geq 18$ - $< 28$, $\geq 28$ - $< 40$ and $\geq 40$)
- Sex (male, female)
- BMI (normal: $< 25$, overweight: $\geq 25$ - $< 30$, obese: $\geq 30$)
- Race (White, Asian, Black, and Other)
- Weight ($<$ median, $\geq$ median)

- Baseline disease severity (T-VASI < 15 and ≥ 15)
- Baseline disease severity (T-VASI ≤ 10 and > 10),
- Status of active vitiligo (Yes/No)
- Baseline F-VASI (< median, ≥ median)
- Fitzpatrick skin types (≤ II, III, > III)
- Duration of vitiligo since diagnosis (< median, ≥ median)
- Prior topical therapy (with and without)
- Prior phototherapy (with and without)

'BMI ≥ 30' or 'BMI ≥ 25 - < 30' subgroups will be combined when either of the subgroups have fewer than 10% subjects. Any Race subgroups with fewer than 10% subjects will be combined with 'Other' subgroup for analyses.

# 10.0 Safety Analyses

## 10.1 General Considerations

Safety data will be summarized for each safety population.

## 10.2 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.

### 10.2.1 Treatment-Emergent Adverse Events

For Period 1, treatment-emergent adverse events (TEAEs) are defined as any adverse events that begin or worsen in severity after initiation of study drug through 30 days following the last dose of study drug in Period 1 or prior to first dose in Period 2, whichever is earlier. For Period 2, TEAEs are defined as any adverse events that begin or worsen in severity after initiation of study drug in Period 2 through 30 days following the last dose of study drug in Period 2. A TEAE during the administration of upadacitinib (i.e., among the ALL_UPA Population) is defined as any AE with an onset date on or after the first dose of upadacitinib and within 30 days after the last dose of upadacitinib. Events where the onset date is the same as the study drug start date are assumed to be treatment emergent.

All TEAEs 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 alphabetical order within each SOC. For the Safety_1 and ALL_UPA Populations, the number and percentage of subjects experiencing TEAEs and the number of events per 100 patient-years of exposure will be summarized.

### 10.2.2 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 for the Safety_1 and ALL_UPA Populations:

- Any TEAE
- Any TEAE related to study drug according to the investigator
- Any severe TEAE (Grade 3 and above according to National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) version 5)
- Any serious TEAE (SAE)
- Any TEAE leading to discontinuation of study drug
- Any TEAE leading to death
- TEAEs of special interest

- All deaths
- Deaths occurring ≤ 30 days after last dose of study drug
- Deaths occurring > 30 days after last dose of study drug.

The overview of number of TEAEs/deaths per 100 patient-years of exposure will also be provided for the above categories for the Safety_1 and ALL_UPA Populations.

TEAEs per 100 patient-years of exposure are defined as the number of TEAEs divided by the total exposure in 100 patient-years. The exposure-adjusted TEAE rate per 100 patient-years is calculated as:

$$100 \times \frac{\text{Number of TEAEs}}{\text{Total Patient Years}}$$

where total patient years is defined as the sum of the study drug duration (defined in Section 6.0) +1 day (not to exceed the start of subsequent period) of all subjects normalized by 365.25, and rounded to 1 decimal place.

### 10.2.3 Treatment-Emergent Adverse Events by SOC and/or PT

Treatment-emergent adverse events will be summarized by SOC and PT; by maximum relationship to study drug as assessed by the investigator (e.g., reasonable possibility or no reasonable possibility) and SOC and PT; by maximum toxicity and SOC and PT; by relationship to COVID-19 and SOC and PT; by ≥5% subjects in any treatment group by SOC and PT. The number and percentage of subjects experiencing TEAEs will be provided for these summaries in the Safety_1 and ALL_UPA Populations. Number of TEAEs per 100 patient-years of exposure will be summarized by SOC and PT for the Safety 1 and All_UPA Populations. 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 toxicity 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 total active treatment group.

### 10.2.4 SAEs (Including Deaths) and TEAE Leading to Study Drug Discontinuation

SAEs (including deaths) and TEAEs leading to study drug discontinuation will be summarized by SOC and PT and in listing format. The number and percentage of subjects experiencing SAEs and TEAE leading to study drug discontinuation as well as the number of events per 100 patient-years of exposure will be summarized for the Safety_1 and ALL_UPA Populations.

### 10.2.5 Adverse Events of Special Interest

Treatment emergent Adverse events of special interest (AESI) will be summarized by SOC and PT and will be based on standardized or company MedDRA queries (SMQs or CMQs) or based on adjudication results. The number and percentage of subjects experiencing treatment emergent AESIs and the number of events per 100 patient-years of exposure of the AESI overview table will be summarized for the Safety_1 and ALL_UPA Populations. For ALL_UPA, number of events per 100 patient-years of exposure of the AESI by SOC and PT will be summarized.

Adverse events of special interest are categorized as follows:

- Serious infections
- Opportunistic infection excluding tuberculosis and herpes zoster
- Possible malignancy
- Malignancy
- Non-melanoma skin cancer (NMSC)
- Malignancy excluding NMSC
- Lymphoma
- Hepatic disorder
- Adjudicated gastrointestinal perforation

- Anemia
- Neutropenia
- Lymphopenia
- Herpes zoster
- Renal dysfunction
- Active tuberculosis
- Adjudicated MACE
- Adjudicated VTE

Detailed information about the search criteria is provided in Appendix B.

Tabular listings of adverse events of special interest will be provided.

### 10.3 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, except for Baseline where SAE-related laboratory assessments on or before the first dose of study drug will be excluded. The clinical laboratory tests defined in the protocol operations manual (e.g., hematology and clinical chemistry) will be summarized.

Raw values and change from Baseline (percent change if applicable) to each applicable post-baseline visit will be summarized for selected laboratory variables in the Safety_1 and Safety_2 Populations. The following descriptive statistics will be presented by treatment groups: number of observations, Baseline mean (Standard deviation, SD), visit mean (SD), medium, change from Baseline mean (SD), standard error, and the 95% confidence interval of the mean change from Baseline.

Percent change from Baseline in LDL-C, HDL-C, total cholesterol and ratio of LDL-C to HDL-C will also be summarized similarly in the Safety_1 and Safety_2 Populations.

For the Safety_1 and ALL_UPA Populations:

A shift table from baseline to the worse post-baseline value (based on NCI CTC criteria) during treatment will be created. A similar shift table will be provided to summarized shifts from baseline to the final post-baseline value. The number and percentage of subjects will be summarized.

Laboratory abnormalities will be evaluated based on Potentially Clinically Important (PCI) criteria (Appendix C). For each laboratory PCI criterion, the number and percentage of subjects who have a laboratory value meeting the criteria will be summarized. Listings will be provided to summarize subject-level laboratory data for subjects meeting PCI criteria.

For the purpose of assessing for potential Hy's law cases, the frequencies and percentages of subjects with post-baseline liver specific function test values that meet PCI criteria (Appendix C) of potential clinical interest should be presented.

## 10.4 Analysis of Vital Signs

Vital sign measurements of systolic and diastolic blood pressure, heart rate, respiratory rate, and body temperature will be summarized.

Each vital sign variable will be summarized for all time points (starting with Baseline) with the number of non-missing observations, mean and standard deviation, median, minimum and maximum. Change from Baseline to each applicable post-baseline visit will be summarized for each vital sign variable, with the number of observations, Baseline mean, and visit mean. Mean, standard error, and 95% confidence interval will be presented for the change from Baseline within each treatment group in Safety_1 and Safety_2 Populations.

Vital sign variables will be evaluated based on PCI criteria (Appendix C) for the Safety_1 and ALL_UPA Populations. For each vital sign PCI criterion, the number and percentage of subjects who have a vital sign value meeting the criteria will be summarized. Listings

will be provided to summarize subject-level vital sign data for subjects meeting PCI criteria.

### 10.5 Safety Subgroup Analyses

Not applicable.

### 10.6 Other Safety Analyses

A post-Baseline treatment ECG listing will be provided.

## 11.0 Other Analyses

Not applicable.

## 12.0 Interim Analyses

No formal interim analysis of efficacy is planned for this study. Routine safety reviews will be performed by an external DMC.

### 12.1 Data Monitoring Committee

An external data monitoring committee (DMC) composed of persons independent of AbbVie and with relevant expertise in their field will periodically review unblinded safety data from the ongoing study. The primary responsibility of the DMC will be to protect the safety of the subjects participating in this study.

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

Since there are no efficacy analyses for early stopping, no alpha adjustment is needed.

## 13.0 Overall Type-I Error Control

Due to the Phase 2 nature of the phase 2 study, all the endpoints will be tested at a two-sided significance level of 0.1 and overall type-1 error control is not considered necessary.

# 14.0 Version History

**Table 4. SAP Version History Summary**

| Version | Date | Summary |
|---|---|---|
| 1.0 | 30 July 2021 | Original version |
| 2.0 | 27 December 2022 | • Modify the ITT and Safety Population to ITT_1/ITT_2, Safety_1/Safety_2 Population.<br>• Removed Baseline PhGIC-V and T-PaGVA from the disease characteristic table in Section 7.1<br>• Added NRI-MI approach for subject who have opportunity to reach Week 52 for F-VASI 50, F-VASI 75, T-VASI 50. In additional, percent change from Baseline in F-VASI, T-VSI and change from Baseline in VitiQoL question 16 are added in Section 9.5<br>• Added additional efficacy subgroup analyses in Section 9.6<br>• Update severity to toxicity in Treatment Emergent Adverse Events by SOC and/or PT in Section 10.2.3<br>• Added a post-Baseline treatment ECT listing in Section 10.6<br>• Clarify the efficacy and safety analyses for each analysis population |

# 15.0 References

1. Rosmarin D, Pandya AG, Lebwohl M, et al. Ruxolitinib cream for treatment of vitiligo: a randomised, controlled, phase 2 trial. Lancet. 2020;396(10244):110-20.
2. Pinheiro J, Bornkamp B, Bretz F. Design and analysis of dose-finding studies combining multiple comparisons and modeling procedures. J Biopharm Stat. 2006;16(5):639-56.

## 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 Adverse Events of Special Interest

Adverse Events of Special Interest (AESI) will be identified using the following search criteria:

| AESI | Type of MedDRA Query | Broad or Narrow Search | SMQ/CMQ Search Criteria |
|---|---|---|---|
| Serious Infections | CMQ | | "Infections" – Subset for SAEs |
| Opportunistic Infection excluding tuberculosis and herpes zoster | CMQ | | "Opportunistic Infection excluding tuberculosis and herpes zoster" |
| Possible malignancy | SMQ | Narrow | "Malignancies" |
| Malignancy | SMQ | | "Malignant Tumours" |
| Non-Melanoma Skin Cancer (NMSC) | SMQ/CMQ | SMQ Narrow | Skin Malignant tumours (Narrow SMQ) removing Melanoma CMQ |
| Malignancy excluding NMSC | | | Malignancy Narrow SMQ and removing NMSC output |
| Lymphoma | SMQ | | "Malignant Lymphomas" |
| Hepatic Disorder | SMQ | Narrow | "Drug Related Hepatic Disorders" |
| Adjudicated Gastrointestinal Perforations | Medical review of events identified by the "Gastrointestinal Perforation" SMQ Narrow search | | |
| Anemia | CMQ | | "Non-Hemolytic and Non-Aplastic Anemias" |
| Neutropenia | CMQ | | "Hematological Toxicity – Neutropenia" |
| Lymphopenia | CMQ | | "Hematological Toxicity – Lymphopenia (Veliparib Product Specific)" |
| Herpes Zoster | CMQ | | "Herpes Zoster" |
| Renal Dysfunction | SMQ | Narrow | "Acute Renal Failure" |

| AESI | Type of MedDRA Query | Broad or Narrow Search | SMQ/CMQ Search Criteria |
|---|---|---|---|
| Active Tuberculosis | CMQ | | "Active Tuberculosis" |
| Adjudicated cardiovascular events[a] | Output from CAC | | |
| MACE* | | | |
| Cardiovascular Death | | | |
| Non-fatal Myocardial Infarction | | | |
| Non-fatal Stroke | | | |
| Other Adjudicated Cardiovascular Events | | | |
| Undetermined/Unknown Cause of Deaths | | | |
| Adjudicated Thrombotic Events | Output from CAC | | |
| VTE** | | | |
| Deep Vein Thrombosis | | | |
| Pulmonary Embolism | | | |
| Other Venous Thrombosis | | | |
| Arterial Thromboembolic Events (non-cardiac, non-neurologic) | | | |

CAC = Cardiovascular Adjudication Committee; CMQ = company MedDRA query; PT = preferred term; SMQ = standard MedDRA query

* MACE: Major Adverse Cardiovascular Events, defined as cardiovascular death, non-fatal myocardial infarction and non-fatal stroke.

** VTE: Venous thromboembolic events, defined as deep vein thrombosis (DVT) and pulmonary embolism (PE) (fatal and non-fatal).

a. Reviewed and adjudicated by an independent Cardiovascular Adjudication Committee in a blinded manner.

## Appendix C. Potentially Clinically Important Criteria for Safety Endpoints

The criteria for Potentially Clinically Important (PCI) laboratory findings are described in Table C-1 and Table C-2, and the PCI criteria for vital sign findings are described in Table C-3.

### Table C-1. Criteria for Potentially Clinically Important Hematology Values

| | | Definition of Potentially Clinically Important (NCI CTCAE Grade 3 or higher) |
|---|---|---|
| **Hematology Variables** | **Units** | **Very Low** |
| Hemoglobin | g/dL | < 8.0 |
| Platelets count | $10^9$/L | < 50.0 |
| WBC count | $10^9$/L | < 2.0 |
| Neutrophils | $10^9$/L | < 1.0 |
| Lymphocytes | $10^9$/L | < 0.5 |

Note: A post-baseline value must be more extreme than the Baseline value with at least one CTCAE grade of worsening to be considered a potentially clinically important finding.

## Table C-2. Criteria for Potentially Clinically Important Chemistry Values

| | | Definition of Potentially Clinically Important (NCI CTCAE Grade 3 or higher) | |
|---|---|---|---|
| **Chemistry Variables** | **Units** | **Very Low** | **Very High** |
| ALP | U/L | | > 5.0 × ULN |
| SGOT/AST | U/L | | > 5.0 × ULN |
| SGPT/ALT | U/L | | > 5.0 × ULN |
| Albumin | g/L | < 20 | |
| Glucose | mmol/L | < 2.2 | > 13.9 |
| Triglycerides | mmol/L | | > 5.7 |
| Creatinine | mcmol/L | | > 3.0 × ULN |
| Potassium | mmol/L | < 3.0 | > 6.0 |
| Calcium | mmol/L | < 1.75 | > 3.1 |
| Sodium | mmol/L | < 130 | > 155 |
| Phosphate | mmol/L | < 0.6 | |
| Total Cholesterol | mmol/L | | > 10.34 |

Note: A post-baseline value must be more extreme than the Baseline value with at least one CTCAE grade of worsening to be considered a potentially clinically important finding.

## Table C-3. Criteria for Potentially Clinically Important Liver Function Tests

| | | Definition of Potentially Clinically Important |
|---|---|---|
| **Variables** | **Units** | **Very High** |
| AST | U/L | ≥ 3 × ULN |
| | | ≥ 5 × ULN |
| | | ≥ 10 × ULN |
| | | ≥ 20 × ULN |
| ALT | U/L | ≥ 3 × ULN |
| | | ≥ 5 × ULN |
| | | ≥ 10 × ULN |
| | | ≥ 20 × ULN |
| Total Bilirubin | UMOL/L | ≥ 2 × ULN |
| Alkaline phosphatase | U/L | ≥ 1.5 × ULN |
| ALT (U/L) and/or AST (U/L) and concurrent TBL (UMOL/L) | | ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 1.5 × ULN<br>ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 2 × ULN |

## Table C-4. Criteria for Potentially Clinically Important Vital Sign Values

| Vital Sign | Category | Criteria for Potential Clinically Significant Vital Signs |
|---|---|---|
| Systolic blood pressure | Low | Value ≤ 90 mmHg and decrease ≥ 20 mmHg from Baseline |
| | High | Value ≥ 160 mmHg and increase ≥ 20 mmHg from Baseline |
| Diastolic blood pressure | Low | Value ≤ 50 mmHg and decrease ≥ 10 mmHg from Baseline |
| | High | Value ≥ 100 mmHg and increase ≥ 10 mmHg from Baseline |
| Pulse | Low | Value ≤ 50 bpm and decrease ≥ 15 bpm from Baseline |
| | High | Value ≥ 120 bpm and increase ≥ 15 bpm from Baseline |
| Weight | High | > 7% increase from Baseline |
| | Low | > 7% decrease from Baseline |

## Appendix D. Random Seeds

In case of non-convergence, the random seed will be updated by adding 10000 at each attempt until convergence of model happens.

A. Random Seeds for NRI-MI

| Endpoints | Random Seed | |
|---|---|---|
| | **MCMC procedure** | **PROC MI** |
| F-VASI 50 | 1001 | 9001 |
| F-VASI 75 | 1002 | 9002 |
| F-VASI 90 | 1003 | 9003 |
| T-VASI 50 | 1004 | 9004 |
| T-VASI 75 | 1005 | 9005 |
| T-VASI 90 | 1006 | 9006 |
| VNS 4/5 | 1007 | 9007 |
| DLQI 0/1 | 1008 | 9008 |
| PhGIC-V 1/2 | 1009 | 9009 |
| PaGIC-V 1/2 | 1010 | 9010 |
| F-PhGVA 0/1 | 1011 | 9011 |
| T-PhGVA 0/1 | 1012 | 9012 |
| F-PaGVA 0/1 | 1013 | 9013 |
| T-PaGVA 0/1 | 1014 | 9014 |
| DLQI 4 | 1015 | 9015 |

B. Random Seeds for MI

| Endpoints | Random Seed | |
|---|---|---|
| | **MCMC procedure** | **PROC MI** |
| Percent change from Baseline in F-VASI | 1015 | 9015 |

# Appendix E. Non-Responder Imputation Incorporating Multiple Imputation to Handle Missing Data Due to COVID-19 Pandemic or Any Other Missing Data that Can Be Reasonably Assumed to Be Missing at Random for Dichotomized Outcome Variables

## 1.0 Overview

### 1.1 Background and Justification for Missing at Random (MAR) Assumption

The COVID-19 pandemic is interfering with the conduct of many ongoing trials, with potential impacts on treatment duration and the collection, analysis and the interpretation of clinical trial data. Some protocol-specified visits in the clinical trials may be impacted due to COVID-19 infection or logistical restrictions during the pandemic. For example, some scheduled visits may be missed due to self-quarantine or local government restrictions on travel; some visits may also be delayed or canceled due to healthcare resource constraints during the pandemic. Impacted visits due to COVID-19 will be recorded in the database. The probability of having missed visits and missing data due to COVID-19 infection or logistical restrictions related to the COVID-19 pandemic can be reasonably assumed to be unrelated to the unobserved values. Therefore, for the purpose of statistical analysis, it is reasonable to assume that these missing data are missing at random (MAR) and the statistical models that require MAR assumption are appropriate. In some cases, sensitivity analyses may be performed to assess the impact of missing data and the robustness of the conclusion.

### 1.2 FDA Guidance

FDA provided two guidance documents[1,2] in March 2020 and June 2020 on the efficacy collection and possible changes in the statistical analysis plan:

- "With respect to efficacy assessments, FDA recommends consultation with the appropriate review division regarding protocol modifications for the collection of efficacy endpoints, such as use of virtual assessments, delays in assessments, and alternative collection of research-specific specimens, if feasible. For individual instances where efficacy endpoints are not collected,

the reasons for failing to obtain the efficacy assessment should be documented (e.g., identifying the specific limitation imposed by COVID-19 leading to the inability to perform the protocol-specified assessment)."

- "If changes in the protocol will lead to amending data management and/or statistical analysis plans, the sponsor should consider doing so in consultation with the applicable FDA review division. Prior to locking the database, sponsors should address in the statistical analysis plan how protocol deviations related to COVID-19 will be handled for the prespecified analyses."

## 1.3 EMA Guidance

EMA provided guidance[3] in March 2020:

- "At this point in time it is not possible to give general applicable advice on how the different aspects related to the pandemic should be handled, as implications on clinical trials are expected to be manifold. Impact on the data collection, analysis and interpretation of results for each trial will need a thorough case-by-case assessment."
- "As a general principle, there are strong scientific reasons to conduct trials as planned and implement changes only when there is a convincing scientific reason that it improves interpretability of results."

## 1.4 Missing Data Handling for Missing Due to COVID-19 for Dichotomized Variables

In this document, a missing data handling method is proposed to handle missing data due to COVID-19 infection or logistical restrictions related to the COVID-19 pandemic under the general MAR framework. In particular, we explain using multiple imputation (MI) to handle missing data due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random in dichotomized variables in conjunction with non-responder imputation (NRI) for missing data due to other reasons.

# 2.0 Non-responder Imputation Incorporating Multiple Imputation (NRI-MI)

## 2.1 Overall Description of the Method

For a dichotomized outcome variable with missing data, the NRI-MI will categorize any subject who does not have evaluation during a pre-specified visit window as a non-responder for the visit, with two exceptions:

- If the subject is a responder both before and after the pre-specified visit window in Period 1, the subject will be categorized as a responder for the visit.
- If the reason for missing (e.g., missed visits, incomplete visit, out-of-schedule visits, or discontinuations of study drug) is due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random, the information will be captured in the database and the subject's response status will be imputed using multiple imputation.

In addition, subjects whose change/percent change from Baseline cannot be calculated because of a missing Baseline will be considered as a non-responder at all post-baseline visits unless the post-baseline value is zero.

Non-responder imputation incorporating multiple imputation (NRI-MI) for missing due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random will be implemented as follows.

## 2.2 Multiple Imputation (MI) and MAR Assumption

When a dichotomized variable is derived from a continuous scale, for example, F-VASI 75 (at least a 75% reduction in F-VASI relative to Baseline), the multiple imputation will be applied to the original scale, F-VASI (ranges from 0 – 100) assuming multivariate normal distribution. Then the dichotomized variable will be derived from the imputed value.

The MI procedure assumes that the data are missing at random (MAR). That is, for an outcome variable Y, the probability that an observation is missing depends only on the observed values of other variables, not on the unobserved values of the outcome variable Y. Statistical inference from the MI procedure is valid under the MAR assumption.

## 2.3 Imputation Algorithm

It is reasonable to assume the missing values of the longitudinal data for an outcome variable (e.g., F-VASI, the original scale of F-VASI 75, at each post-baseline visit) follows a monotone missing pattern. In practice, the missing data of the outcome variable might have an arbitrary (non-monotone) missing data pattern. An extra step may be added accordingly, to augment data into a monotone missing pattern.

For the outcome variable (e.g., F-VASI at each visit), K 'complete' datasets can be generated in two steps: augmentation step and imputation step. K, the number of repetitions, is determined below.

**Augmentation Step**

For datasets with non-monotone missing data pattern, the augmentation step will first impute enough values to augment the data into a monotonic missing pattern:

Markov Chain Monte Carlo (MCMC) will be applied to augment the data using PROC MI with the MCMC IMPUTE=monotone statement, assuming a multivariate normal distribution. The augmented data will be used in the subsequent imputation step to generate 'complete' datasets. Covariates included in the model are treatment, stratification factors (the age group [≤ 50 and > 50], Baseline disease severity [T-VASI < 15 and ≥ 15], and status of active vitiligo [Yes/No]), Baseline, and all post-baseline visits of the outcome variable according to the pre-specified order. For T-VASI related endpoints, Baseline disease severity (T-VASI < 15 and ≥ 15) will not be included in the model. Of note, categorical variables are included using the form of dummy variables.

Repeat the imputation process K=30 times using the procedure described above to form K=30 monotone missing datasets, where K is determined as described in "Repetition of Imputations (K)."

**Imputation Step**

For missing data with monotone missing patterns, the choice of multiple imputation using a parametric regression model that assumes multivariate normality is appropriate.

The imputation step is described below:

- The imputation model for the missing data is a regression model, which controls for treatment, stratification factors (the age group [≤ 50 and > 50], Baseline disease severity [T-VASI < 15 and ≥ 15], and status of active vitiligo [Yes/No]), Baseline, and all post-baseline visits of the outcome variable. For T-VASI related endpoints, Baseline disease severity (T-VASI < 15 and ≥ 15) will not be included in the model. The covariates included in the model and the order of these variables are consistent with the augmentation step.
- For each monotone missing dataset, using SAS PROC MI with MONOTONE REG model statement, the outcome variable at each post-baseline visit with missing values will be imputed sequentially with covariates constructed from their corresponding sets of preceding variables.

A 'complete' dataset with imputed values for the missing data is generated after the augmentation and imputation steps are completed.

**Repetition of Imputations (K)**

Repetition of imputations, K, must be determined in advance. When estimating the overall variance of multiple imputation, the additional sampling variance is the between-imputation variance divided by K. This value represents the sampling error associated with the overall or average coefficient estimates. It is used as a correction factor for using a specific number of imputations. The more imputations (K) are conducted, the more precise the parameter estimates will be. For example, with a 1% power falloff tolerance

in multiple imputation, as compared to an infinite number of imputations, multiple imputation requires 20 repetitions of imputation for 30% missing information and 40 repetitions for 50% missing information.[4] In the usual clinical settings expecting less than 30% missing information, K = 30 repetitions are deemed sufficient. When missingness exceeds 30%, depending on the power falloff tolerance level, number of repetitions may need to be increased. Recent research[4] suggested that the number of repetitions (K) should be at least equal to the percentage of missing.[6]

## 2.4 Derivation of Response Status and Non-Responder Imputation

For each 'complete' dataset, the imputed post-baseline values will be rounded to the same precision as the observed data. Response status (e.g., F-VASI 75 at each visit) will be determined accordingly.

The imputed response status for missing due to reasons other than COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random will be overridden by non-responder imputation (Section 2.1) to ensure that multiple imputation is only applied to missing due to COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random:

- Using NRI approach, all missing due to reasons other than COVID-19 or any other missing data that can be reasonably assumed to be Missing at Random will be categorized as non-responders. In addition, subjects whose change/percent change from Baseline cannot be calculated because of a missing Baseline will be considered as a non-responder at all post-baseline visits unless the post-baseline value is zero.
- The only exception is that a subject will be categorized as a responder for the visit if the subject is a responder both before and after an SAP-specified visit window in Period 1.

## 2.5 Analysis

The statistical analysis will use the Cochran-Mantel-Haenszel (CMH) test adjusted by the stratification factors derived from actual values.

### 2.5.1 Analysis of Each Dataset

For each of the K 'complete' datasets, the CMH test will be used to estimate the treatment difference versus placebo and the corresponding standard error.

### 2.5.2 Synthesis of Results for Statistical Inference

The results from the K 'complete' datasets will be synthesized using the SAS procedure PROC MIANALYZE, following Rubin's formula (Rubin, 1987[5]), to derive the MI estimator of the treatment difference for the final inferences.

**Rubin's formula**

We fit the analysis model to the k$^{th}$ 'complete' dataset, denoting the estimate of the treatment difference q by $\tilde{\theta}_k$ from the k$^{th}$ 'complete' dataset, and denoting the corresponding estimate of the variance as $V_k$.

The MI estimator of q (point estimator obtained from PROC MIANALIZE), $\tilde{\theta}_{MI}$, is the average of the K individual estimators:

$$\tilde{\theta}_{MI} = \frac{1}{K}\sum_{k=1}^{K}\tilde{\theta}_k.$$

The estimated variance of $\tilde{\theta}_{MI}$, is a combination of the between- and within-imputation variability as follows:

$$V_{MI} = W + (1 + \frac{1}{K})B,$$

where $W = \frac{1}{K}\sum_{k=1}^{K} V_k$ is the within-imputation variability and $B = \frac{1}{K-1}\sum_{k=1}^{K}(\tilde{\theta}_k - \tilde{\theta}_{MI})^2$ is the between-imputation variance.

It has been shown[1] that the statistic

$$T = \frac{\tilde{\theta}_{MI} - \theta}{\sqrt{V_{MI}}}$$

has an approximate $t_v$ distribution where $v = (K-1)(1+\frac{W}{B})^2$. Statistical inference, including hypothesis testing and confidence intervals for the treatment effect, will be based on this T-statistic.

## 3.0 Sample SAS Code

```
/**********************/
/*IMPUTATION ALGORITHM*/
/**********************/
/*NOTE: THIS APPROACH REQUIRES NO MISSING IN CATEGORICAL COVARIATES AND
REQUIRES AT LEAST ONE OBSERVATION IN BASELIBE OR ONE OF THE POST-
BASELINE VISIT*/

/*PRE-AUGMENTATION - CREATE DUMMY FOR CATEGORICAL VARIABLES*/
/*****************************************/
DATA VASI_2; SET VASI;
  /*THE MCMC STATMENT BELOW ASSUMES MULTI-VARIATE NORMAL*/
  /*IF TREATMENT OR ANY COVARIATES ARE CATEGORICAL WITH L>2 LEVELS*/
  /*NEED TO CREATE L-1 DUMMY VARIABLES*/
  /*HERE, TREATMENT (TRT01PN) HAS 4 LEVELS, SO WE NEED 3 DUMMY
VARIABLES*/
  IF TRT01PN=1 THEN TRT1=1;
  ELSE TRT1=0;
  IF TRT01PN=2 THEN TRT2=1;
  ELSE TRT2=0;
  IF TRT01PN=3 THEN TRT3=1;
  ELSE TRT3=0;
RUN;

/*AUGMENTATION STEP -- TO HAVE 30 MONOTONE MISSING DATASETS*/
PROC MI DATA=VASI_2 OUT=VASI_MONO NIMPUTE=30 SEED= 1001 /*RANDOM SEED
PRE-DEFINED*/
  ROUND=. . . . . 0.1 0.1 0.1 0.1 0.1 0.1 0.1 /*VALUE ROUND TO 1ST
DECIMAL*/
  MIN=. . . . . 0 0 0 0 0 0 0            /*MINIMUM VALUE OF F-VASI IS
0*/
  MAX=. . . . . 100 100 100 100 100 100 100;   /*MAXIMUM VALUE OF F-
VASI IS 100*/
  MCMC IMPUTE=MONOTONE ;
```

```
/*NOTE: CATEGORICAL VARIABLES SUCH AS TRT1 TRT2 TRT3 ARE DUMMY, CREATED
ABOVE*/
/*NOTE: ALL OTHER NON-DUMMIED COVARIATES MUST BE CONTINUOUS*/
/*SUPPOSE STRATAN (NUMERIC VARIABLE FOR STRATA) HAS ONLY 2 LEVELS, NO
NEED TO CREATE DUMMY*/
  VAR TRT1 TRT2 TRT3 AGEGR1N TVAGR1N ACTVGR1N BASE WK2 WK4 WK8 WK12 WK18
WK24;
/*CAUTION TO USE THE "BY" STATEMENT IN MCMC: */
/*MVN MODEL IS FITTED WITHIN EACH 'BY' GROUP, INSTEAD OF ACROSS ALL
GROUPS*/
RUN;

/*IMPUTATION STEP - DETERMINE IMPUTATION DISTRIBUTION AND RANDOMLY
IMPUTE MISSING VALUE TO GENERATE 'COMPLETE' DATASETS*/
/************************************************/
PROC MI DATA=VASI_MONO OUT=VASI_FULL NIMPUTE=1 SEED= 9001  /*RANDOM SEED
PRE-DEFINED*/
  ROUND=. . . . . 0.1 0.1 0.1 0.1 0.1 0.1 0.1 /*VALUE ROUND TO 1ST
DECIMAL*/
  MIN=. . . . . 0 0 0 0 0 0 0            /*MINIMUM VALUE OF F-VASI IS
0*/
  MAX=. . . . . 100 100 100 100 100 100 100;   /*MAXIMUM VALUE OF F-
VASI IS 100*/
  MINMAXITER=1000;
  CLASS TRT01PN AGEGR1N TVAGR1N ACTVGR1N;
  VAR TRT01PN AGEGR1N TVAGR1N ACTVGR1N BASE WK2 WK4 WK8 WK12 WK18 WK24;
  MONOTONE REG (WK2 WK4 WK8 WK12 WK18 WK24);  /* IMPUTED SEQUENTIALLY,
FROM WK 2 TO 24, WITH COVARIATES CONSTRUCTED FROM THE CORRESPONDING
PRECEDING VARIABLES*/
  BY _IMPUTATION_;                    /*FOR EACH OF THE 30 MONOTONE
MISSING DATASETS, IMPUTE A 'COMPLETE' DATASET*/
RUN;


/*DETERMINE DICHOTOMOUS RESPONSE STATUS, VASI 50*/
DATA ALL; SET VASI _FULL;
  IF 0<=WK2<=0.5*BASE THEN VASI50_2=1;
    ELSE VASI50 2=0;
  IF 0<=WK4<=0.5*BASE THEN VASI50_4=1;
    ELSE VASI50 4=0;
  IF 0<=WK8<=0.5*BASE THEN VASI50_8=1;
    ELSE VASI50_8=0;
  IF 0<=WK12<=0.5*BASE THEN VASI50_12=1;
    ELSE VASI50 12=0;
  IF 0<=WK18<=0.5*BASE THEN VASI50_18=1;
    ELSE VASI50 18=0;
  IF 0<=WK24<=0.5*BASE THEN VASI50_24=1;
    ELSE VASI50_24=0;
RUN;
```

```
/*********************************************************************
*/
/*    DATA HANDLING STEPS TO MERGE COVID-19 or Other MAR STATUS OMITTED
*/
/*            PLACE TO ADD DATA HANDLING AND MERGING STEPS
*/
/*********************************************************************
*/

/*FOR MI, SKIP THE FOLLOWING CODE, PROCEED TO THE CODE AFTER ANALYSIS
MODEL *//*OVERRIDE MISSING VALUES NOT DUE TO COVID-19 or Other MAR WITH
TRADITIONAL NRI*/
DATA ALLF; SET ALL;
  /*COVID19 XX='Y' IF MISSING AT WEEK XX IS DUE TO COVID-19 or Other
MAR; IF NOT, OVERRIDE WITH TRADITIONAL NRI*/
  IF COVID19_2 NE 'Y' THEN VASI50_2=VASI50NRI_2;
  IF COVID19_4 NE 'Y' THEN VASI50_4=VASI50NRI_4;
  IF COVID19_8 NE 'Y' THEN VASI50_8=VASI50NRI_8;
  /*VARIABLE VASI50NRI_XX: TRADITIONAL NRI DATA AT WEEK XX, WHICH COVERS
THE SPECIAL HANDLING SUCH AS THE BEFORE-AND-AFTER EXCEPTION*/
  IF COVID19_12 NE 'Y' THEN VASI50_12=VASI50NRI_12;
  IF COVID19_18 NE 'Y' THEN VASI50_18=VASI50NRI_18;
  IF COVID19_24 NE 'Y' THEN VASI50_24=VASI50NRI_24;
RUN;
PROC SORT DATA=ALLF; BY _IMPUTATION_ SUBJID; RUN;

/***************/
/*ANALYSIS MODEL*/
/***************/

/*KEY CODE: ANALYZING EACH 'COMPLETE' DATASET*/
/*********************************/
/*COMPARE TREATMENT GROUPS 1 (PLACEBO) AND 2 (HIGH-DOSE) ONLY*/
DATA ALL; SET ALL;
  WHERE TRT01PN IN (1,2);
RUN;

/*INDIVIDUAL-LEVEL DATA --> # OF RESPONDERS & # OF SUBJECTS, TO BE READ-
IN TO PROC STDRATE*/
PROC FREQ DATA=ALL;
  BY _IMPUTATION_;
  TABLES TRT01PN*AGEGR1N*TVAGR1N*ACTVGR1N*VASI50_24/LIST NOCUM NOPRINT
OUT=COUNT_TABLE;
  /*WEEK 24 RESULTS AS AN EXAMPLE*/
RUN;
DATA COUNT_TABLE; SET COUNT_TABLE;
  DROP PERCENT;
```

```
RUN;
PROC TRANSPOSE DATA=COUNT_TABLE OUT=FREQ_TABLE PREFIX=RESP;
ID VASI50_24;
BY  IMPUTATION_ TRT01PN AGEGR1N TVAGR1N ACTVGR1N;
VAR COUNT;
RUN;
DATA FREQ TABLE1; SET FREQ_TABLE;
  CASE=RESP1;
  SIZE=SUM(RESP0, RESP1);
  KEEP _IMPUTATION_ TRT01PN AGEGR1N TVAGR1N ACTVGR1N CASE SIZE;
RUN;

/*RE-ORDER TO SET 1 (PLACEBO) AS THE REFERENCE GROUP*/
DATA FREQ_TABLE2; SET FREQ_TABLE1;
  IF TRT01PN=2 THEN TRT01PN=0;
RUN;

/*CALCULATE THE COMMON RISK DIFF FOR EACH COMPLETE DATASET*/
PROC STDRATE DATA=FREQ TABLE2
  METHOD=MH STAT=RISK EFFECT=DIFF;
  BY _IMPUTATION_;
  POPULATION GROUP=TRT01PN EVENT=CASE TOTAL=SIZE;
  STRATA AGEGR1N TVAGR1N ACTVGR1N / ORDER=DATA STATS(CL=NONE) EFFECT;
  ODS OUTPUT EFFECT=EFFECT;
RUN;


/*COMBINING RESULTS USING PROC MIANALYZE*/
/*************************************/
PROC MIANALYZE DATA=EFFECT;
  ODS OUTPUT PARAMETERESTIMATES=RISK_DIFF_MH;
  MODELEFFECTS RISKDIFF;
  STDERR STDERR;
RUN;
```

# 4.0 References

1. FDA Guidance on Conduct of Clinical Trials of Medical Products during COVID-19 Pandemic - Guidance for Industry, Investigators, and Institutional Review Boards. FDA. 2020.
2. Statistical Considerations for Clinical Trials During the COVID-19 Public Health Emergency - Guidance for Industry. FDA. 2020.
3. Points to consider on implications of Coronavirus disease (COVID-19) on methodological aspects of ongoing clinical trials. EMA. 2020.

4. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prev Sci. 2007;8(3):206-13.
5. Rubin DB, Schenker N. Interval estimation from multiply-imputed data: a case study using agriculture industry codes. J Am Stat Assoc. 1987;81:366-74.
6. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99.