Statistical Analysis Plan for Study M20-040

A Phase 2, Multicenter, Randomized, Placebo-Controlled, Double-Blind Study to Evaluate Upadacitinib in Adult Subjects with Moderate to Severe Hidradenitis Suppurativa

Date: 30 April 2021

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 M20-040 'A Phase 2, Multicenter, Randomized, Placebo-Controlled, Double-Blind Study to Evaluate Upadacitinib in Adult Subjects with Moderate to Severe Hidradenitis Suppurativa.'

Study M20-040 examines the efficacy and safety of upadacitinib 30 mg versus placebo for the treatment of signs and symptoms of moderate to severe hidradenitis suppurativa (HS) in adult subjects.

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 assess the efficacy and safety of upadacitinib 30 mg QD in adult subjects with moderate to severe HS. The primary efficacy objective will be assessed based on the achievement of HiSCR after 12 weeks of treatment with upadacitinib 30 mg when compared to placebo in the Intent-to-Treat (ITT) population, which consists of all randomized subjects.

Hypothesis corresponding to the primary endpoint is:

- The proportion of subjects achieving HiSCR with upadacitinib is greater than that with placebo at Week 12.

The estimand corresponding to the primary endpoint is defined using composite variable strategy:

- Achievement of HiSCR at Week 12 without the initiation of antibiotics for HS-related infections and without premature discontinuation of study drug due to lack of efficacy prior to Week 12.

2.2 Study Design Overview

This is a Phase 2, multicenter, randomized, double-blinded, parallel-group, placebo-controlled study to evaluate the efficacy and safety of upadacitinib in adult subjects with moderate to severe HS.

The duration of the study will be up to 57 weeks and will include an approximately 35-day screening period followed by a 48-week double-blinded treatment period and a 30-day follow up visit after the last dose of study drug.

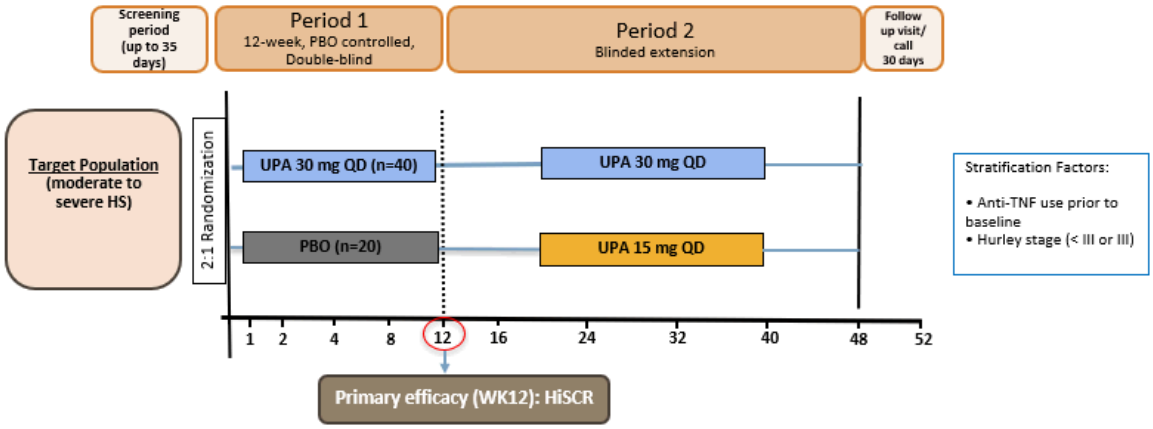
Subjects who meet eligibility criteria will be randomized in a 2:1 ratio to 1 of the 2 arms as shown below:

- Upadacitinib 30 mg once daily (QD) (N = 40): Daily oral doses of upadacitinib 30 mg from the Baseline visit through Period 1 and Period 2.
- Placebo (N = 20): Placebo for upadacitinib from the Baseline visit up to the Week 12 visit (Period 1). At Week 12, subjects will be switched to blinded upadacitinib 15 mg QD through Period 2.

The schematic of the study is shown in Figure 1. Further details regarding study procedures are included in the Operations Manual (Appendix E).

The primary efficacy analysis will be conducted when all ongoing subjects have completed their Week 12 visit.

Figure 1. Study Schematic



HiSCR = ≥50% reduction from baseline in the total AN count, with no increase in abscess or draining fistula counts

PBO = placebo; UPA = upadacitinib

2.3 Treatment Assignment and Blinding

Subjects will be randomized to upadacitinib 30 mg QD or placebo in a 2:1 ratio. Randomization will be stratified by anti-TNF use prior to baseline (yes or no) and Hurley stage (< III or III).

All AbbVie personnel with direct oversight of the conduct and management of the trial (with the exception of AbbVie Drug Supply Management Team), the investigator, study site personnel, and the subject will remain blinded to each subject's treatment through Week 12. Sites and subjects will remain blinded for the duration of the study.

2.4 Sample Size Determination

The primary endpoint is the achievement of HiSCR at Week 12. The primary analysis compares the proportion of subjects achieving HiSCR response to that of historical placebo (25% HiSCR response). Assuming the Week 12 HiSCR response rate is 57% for the upadacitinib 30 mg arm, 40 subjects on upadacitinib will provide more than 95% power with a one-sample Chi-square test and one-sided alpha = 0.05 significance level.

The historical placebo rate of 25% is assumed based on the response rate of placebo subjects satisfying the same eligibility criteria from the two Humira studies (Study M11-313 and Study M11-810).

3.0 Endpoints

3.1 Primary Endpoint(s)

The primary endpoint is the achievement of HiSCR at Week 12. HiSCR is defined as at least a 50% reduction in the total abscess and inflammatory nodule count (AN count) with no increase in abscess count and no increase in draining fistula count relative to Baseline.

3.2 Secondary Endpoint(s)

The secondary endpoint is the achievement of NRS30, defined as at least 30% reduction and at least 1 unit reduction from Baseline in Patient's Global Assessment (PGA) of Skin Pain (numeric rating scale [NRS]) – at worst at Week 12 among subjects with baseline NRS $\geq$ 3.

3.3 Other Efficacy Endpoint(s)

The primary and secondary endpoints will also be analyzed for all visits where measurements are taken other than Week 12. In addition, the following endpoints will be analyzed at all visits noted in the Study Activities Table (Protocol Appendix D).

- Experience of flare, defined as an increase in AN count of at least a 25% with a minimum increase of 2 relative to Baseline (flare);
- Change from Baseline in Dermatology Life Quality Index (DLQI);
- Change from baseline in Hidradenitis Suppurativa Symptom Assessment (HSSA);
- Change from baseline in HS-related swelling, assessed based on the HSSA;
- Change from baseline in HS-related odor, assessed based on the HSSA;
- Change from baseline in HS-related worst drainage, assessed based on the HSSA;

- Change from baseline in Hidradenitis Suppurativa Impact Assessment (HSIA).

3.4 Safety Endpoint(s)

The following endpoints will be included in the safety analyses:

- Treatment emergent adverse events (TEAEs);
- Serious adverse events (SAEs);
- Adverse events of special interest (AESIs);
- Adverse events (AEs) leading to discontinuation;
- Vital signs and laboratory tests.

4.0 Analysis Populations

The following populations will be used for the analyses.

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

The following 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.
- The all upadacitinib treated (ALL_UPA) Population is defined as subjects who received at least one dose of upadacitinib in the study. This population will be used to provide a comprehensive summary of safety.

For safety analyses, subjects will be analyzed based on the treatment actually received.

5.0 Subject Disposition

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

- 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 in ITT 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 reason (primary reason and all reasons) for each treatment group and overall. Subjects with multiple reasons for premature discontinuation will be counted once in the calculation of the number and percentage of total.

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

The number of subjects and percentage of screen failure will be summarized by primary reason.

6.0 Study Drug Duration and Compliance

For each safety population, duration of treatment will be summarized for each treatment group and for all dose groups combined. 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 each treatment duration interval (≥ 4 weeks, ≥ 12 weeks, ≥ 24 weeks, ≥ 36 weeks, ≥ 48 weeks) will be summarized for ALL_UPA Population.

Treatment compliance will be summarized by treatment groups for Safety_1 Population and upadacitinib dose groups for ALL_UPA Population. Treatment compliance is defined as the number of tablets actually taken divided by the number of tablets that should have been taken. Percent compliance will be summarized.

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 overall and by treatment group. 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

Demographics and baseline disease characteristics will be summarized for ITT Population.

Continuous demographic variables include age, weight, height, and body mass index (BMI). Categorical demographic variables include sex, ethnicity, race, age (<40, 40 – 65, ≥ 65 years), weight (≤ 100 or > 100 kg), BMI (< 25, ≥ 25 - < 30 or ≥ 30 kg/m^2), tobacco user (current, former, never, unknown), and alcohol user (current, former, never, unknown).

Disease characteristics include HS disease duration in years, Baseline lesion counts (by lesion type), Baseline skin pain NRS (among all subjects and among subjects with baseline value ≥ 3), Baseline DLQI, Baseline HSSA, Baseline HSIA and Baseline hsCRP as continuous variables, prior exposure to TNF antagonists (yes, no) and baseline Hurley stage (< III, III), HS history (anatomic regions, types of lesions) 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 group for ITT Population. 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 Safety_1 Population. A prior medication is defined as any medication taken prior to the date of the first dose of 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. Concomitant medications will be summarized for each safety population.

In addition, subjects' prior exposure to antibiotics (systemic and topical), biologics, Jak-inhibitor, and other will be summarized for each treatment group and by reason of discontinuation.

8.0 Handling of Potential Intercurrent Events for the Primary Endpoint

The primary efficacy endpoint, achievement of HiCSR at Week 12 (defined in Section 3.1) will be analyzed in the ITT Population and the following method will be used to address potential intercurrent events:

- Intercurrent events of initiation of antibiotics for HS-related infections and discontinuation of study drug due to lack of efficacy before Week 12 are included in the definition of primary endpoint; subjects who have either intercurrents above will be considered not achieving the response. No other intercurrent events are considered.

9.0 Efficacy Analyses

9.1 General Considerations

All efficacy analyses will be conducted in ITT Population in each period. All tests will be 1-sided at an alpha level of 0.05.

The Primary Analysis will be conducted after all ongoing subjects have completed Week 12, data pertaining to Period 1 has been cleaned, and a database lock has occurred for the purpose of efficacy and safety analysis in Period 1. This is the only and final efficacy analysis for Period 1.

In Period 1, all categorical variables at Week 12 and experience of flare in Period 1 will be analyzed using one sample Chi-square test comparing with historical placebo rate. Supplemental analyses will be conducted to compare in-trial data in upadacitinib group vs in-trial placebo data combined with historical placebo data for primary and secondary endpoints, and compare in-trial data in upadacitinib group vs in-trial placebo data for all categorical endpoints.

Continuous variables will be analyzed using Mixed-Effect Model Repeat Measurement (MMRM) (for endpoints other than DLQI) and Analysis of Covariance (ANCOVA) (for

DLQI) comparing with historical placebo mean change from Baseline at Week 12. Supplemental analyses will be conducted to compare in-trial data in upadacitinib group vs in-trial placebo data.

Long-term efficacy in the Period 2 will be summarized using the observed case approach.

Any subject who is randomized based on a wrong stratum will be analyzed according to the actual stratum the subject belongs to.

For variables other than PGA Skin Pain NRS, "Baseline" refers to the last non-missing observation on or before the date of the first administration of study drug or the date of randomization if no study drug is given. The "Baseline" of PGA Skin Pain NRS is defined as the last non-missing weekly average score, calculated based on the daily scores from the past 7 days (with non-missing values in 4 or more days of the 7 day period), before the date of the first administration of study drug or on or before the date of randomization if no study drug is given.

9.2 Handling of Missing Data

Missing data could occur due to various reasons, including missing visits/assessments, early withdrawal from the study, or missing due to COVID-19 infection or logistic restriction.

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. Sensitivity analyses will be performed to assess the impact of missing data and the robustness of the conclusion.

Handling of intercurrent events and missing data for the efficacy analyses is described below.

9.2.1 Categorical Endpoints in Period 1

- The primary approach for handling missing data in the analysis of categorical endpoints (including the primary endpoint) will use **N**on-**R**esponder **I**mputation while incorporating Multiple Imputation (MI) to handle missing data due to **C**OVID-19 (NRI-C).

  The NRI-C 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, the subject will be categorized as a responder for the visit. 2) missing data due to COVID-19 infection or logistical restriction will be handled by Multiple Imputation. Of note, the intercurrent events of initiation of antibiotics for HS-related infections or discontinuation of study drug due to lack of efficacy will lead the subjects being counted as non-responders at later visits. Subjects whose change/percent change from Baseline cannot be calculated because of a missing Baseline will be considered as non-responders at all post-baseline visits. When determining the HiSCR response status, a subject with missing Baseline value will be counted as a non-responder at any post-baseline visit unless a subject's abscess, inflammatory nodule, and draining fistula counts are all zero at a specific visit, where the subject will be a HiSCR responder at that visit. More details are provided in Appendix E.
- A sensitivity analysis for categorical endpoints will use **NRI** with **N**o special data handling for missing due to **C**OVID-19 (NRI-NC).

NRI-NC will be performed in the same way as NRI-C without the exception #2 above. That is, missing due to COVID-19 infection or logistical restriction will also be counted as non-responders.

The NRI-C and NRI-NC will not be applicable to experience of flare in Period 1 since it is event-driven.

- Multiple Imputation (MI), a sensitivity analysis for the primary 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. Taking abscess count as an example, the variables to be included in the imputation model are: major stratum (prior exposure to anti-TNF [yes, no], Hurley Stage [< III, III]), 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 same imputation strategy will be applied to the inflammatory nodule count and the draining fistula count. HiSCR status will be determined accordingly. Using the one sample Chi-square test, the imputed endpoint will be analyzed using each of the 30 datasets. A pooled test statistic from the chi-square distributed statistics after analyzing each of the 30 imputed datasets will be used to test the difference between upadacitinib group and historical placebo rate with Rubin's strategy. Note that measurements will be considered as missing after receiving antibiotics for HS-related infections or discontinued study drug due to lack of efficacy before MI. Regardless of MI imputed values, subjects receiving antibiotics for HS-related infections or discontinued study drug due to lack of efficacy will be counted as non-responders at later visits.

9.2.2 Continuous Endpoints in Period 1

For continuous endpoints except for DLQI, missing data will be handled using Mixed-Effect Model Repeat Measurement (MMRM).

- MMRM will be conducted using a mixed effect model including observed measurements at all visits. For the comparison with historical placebo change

from Baseline, the dependent variable is change from Baseline, and the mixed model for Period 1 includes visit, main stratification factors at randomization (prior exposure to anti-TNF [yes, no], Hurley Stage [<III, III]), and the corresponding Baseline as a continuous variable. For the comparison with in-trial placebo data, the dependent variable is change from Baseline, and the mixed effect model for Period 1 includes treatment, visit, treatment by visit interaction, main stratification factors at randomization, and the corresponding Baseline as a continuous variable. Subjects' observations after receiving antibiotics for HS-related infections or discontinued study drug due to lack of efficacy will be excluded from the analysis. An unstructured variance covariance matrix (UN) will be used. If the model cannot converge, an appropriate covariance structure matrix (e.g., autoregressive (1) or compound symmetry) will be used. The parameter estimations are based on the method of restrictive maximum likelihood (REML). The fixed effects will be used to report model-based means at corresponding visits.

No missing data handling will be used for DLQI in Period 1 since there is only one post-baseline assessment in Period 1. An ANCOVA model will be applied. For the comparison with historical placebo change from Baseline, the dependent variable is change from Baseline in DLQI, and the model includes main stratification factors and the corresponding Baseline as a continuous variable. For the comparison with in-trial placebo data, the dependent variable is change from Baseline in DLQI, and the model includes treatment, main stratification factors and the corresponding Baseline as a continuous variable.

9.2.3 Summary of Long-Term Efficacy

Long-term efficacy in the Period 2 will be summarized using the observed case approach.

- Observed Case (OC) while on study drug: The OC analysis will be used for the summaries of long-term efficacy, which will not impute values for missing evaluations, and thus a subject who does not have an evaluation on a scheduled visit will not be included in the OC analysis for that visit. The OC analysis will be performed for all variables and will not include values after subjects

receiving antibiotics for HS-related infections or discontinued study drug due to lack of efficacy.

9.3 Primary Efficacy Endpoint and Analyses

9.3.1 Primary Efficacy Endpoint

The primary endpoint is the achievement of HiSCR at Week 12. HiSCR is defined as at least a 50% reduction in the total AN count with no increase in abscess count and no increase in draining fistula count relative to Baseline.

The corresponding statistical null hypothesis to the primary endpoint is that there is no difference between upadacitinib 30 mg and historical placebo rate, in the proportion of subjects achieving HiSCR at Week 12.

9.3.2 Main Analysis of Primary Efficacy Endpoint

Analysis of the primary endpoint will be conducted in the ITT Population based on treatment as randomized. Comparison of the primary endpoint will be made between upadacitinib 30 mg group and historical placebo rate (25%) using a one sample Chi-square test at Week 12. Point estimates, p-value and 95% CIs for the difference in proportions between upadacitinib 30 mg and the historical placebo rate (25%) will be provided.

The attributes of the estimands corresponding to the primary efficacy endpoint are summarized in Table 1.

Table 1. Summary of the Estimand Attributes of the Primary Efficacy Endpoint

	Attributes of the Estimand				
Estimand	**Treatment**	**Endpoint**	**Population**	**Intercurrent Events**	**Statistical Summary**
Primary efficacy endpoint of HiSCR	Upadacitinib 30 mg QD	Achievement of HiSCR at Week 12 without the initiation of antibiotics for HS-related infections or discontinuation of study drug due to lack of efficacy	All randomized subjects (ITT Population)	The intercurrent of initiation of antibiotics for HS-related infections or discontinuation of study drug due to lack of efficacy are included in the endpoint attribute using composite variable strategy. No other intercurrent event is considered.	Difference between upadacitinib 30 mg and placebo in the proportion of subjects achieving HiSCR at Week 12

9.3.3 Sensitivity Analyses of the Primary Efficacy Endpoint

Historical placebo data pre-selected by propensity score matching combined with in-trial placebo data will be used to estimate treatment effect of upadacitinib against placebo. Subjects with historical placebo data of HiSCR were pre-selected using propensity score matching from two Humira Phase 3 studies (M11-313 and M11-810) and the Risankizumab study (M16-833), whose study populations, entry criteria and study designs are similar to this study.

The propensity scores were estimated for all randomized subjects in this study and historical placebo subjects with selected common baseline demographics and characteristics variables:

- Age
- Gender
- Race
- Smoking status
- Baseline BMI

- Baseline AN count (abscess count and inflammatory nodule count)
- Duration of HS since diagnosis
- Baseline Hurley Stage
- Baseline draining fistula

The "nearest neighbor matching" were performed based on a distance measure, caliper, with the propensity score for each subject calculated from the logistic regression. The number of synthetic placebo subjects was determined by the simulation results, for each in-trial placebo rate assumption, with pre-determined criteria that the number of synthetic placebo subjects not to exceed 68 (approximately 3 times the number of in-trial placebo subjects) and the one-sided false positive rate was controlled at 10% (simulation details can be found in Appendix F and the list of matching subjects can be found in Appendix G:

In-Trial Placebo Rate	Number of Synthetic Placebo Subjects
≤ 33%	68
>33% and ≤ 34%	23*
>34%	0

* The 23 subjects with the highest rank in Appendix G Table A will be used in the analysis.

After database lock, comparison between the in-trial data of the upadacitinib group and the in-trial placebo data combined with the pre-selected historical placebo subjects' data will be performed. Frequencies and percentages will be summarized along with 95% confidence interval (CI). Pairwise comparisons of upadacitinib vs. placebo will be made using CMH test. Point estimates, *p*-value and 95% CIs for the difference in proportions between each upadacitinib group and placebo will be provided. The stratification factors (prior exposure to anti-TNF [yes, no], Hurley Stage [<III, III]) will be adjusted in the model.

Comparison between the in-trial data of the upadacitinib group and the in-trial placebo subjects will also be performed. The analysis method will be the same as described above.

The NRI-C will be the primary approach for missing data handling in the analysis of the primary efficacy endpoint. NRI-NC and MI will be used as sensitivity analysis.

9.4 Secondary Efficacy Analyses

9.4.1 Main Analyses of Secondary Efficacy Endpoints

The secondary endpoint NRS30, defined as the achievement of at least 30% reduction and at least 1 unit reduction from Baseline in Patient's Global Assessment (PGA) of Skin Pain (numeric rating scale [NRS]) – at worst at Week 12 among subjects with baseline NRS ≥ 3.

NRS30 will be analyzed using the same method as for the primary endpoint and comparing with the historical placebo rate (22.5%) at Week 12. The historical placebo rate of 22.5% is assumed based on the response rate of placebo subjects from the two Humira studies (Study M11-313 and Study M11-810), who satisfied the same eligibility criteria.

The same supplemental analysis to compare upadacitinib 30 mg group vs in-trial placebo group combined with pre-selected placebo subjects' data from the three existing trials will also be performed as described in Section 9.3.3. The actual number of synthetic placebo subjects is determined as below (simulation details can be found in Appendix F and the list of matching subjects can be found in Appendix G):

In-Trial Placebo Rate	Number of Synthetic Placebo Subjects
≤ 36%	68
>36% and ≤ 37%	23*
>37	0

* The 23 subjects with the highest rank in Appendix G Table B will be used in the analysis.

Comparison between the in-trial data of the upadacitinib group and the in-trial placebo subjects will also be performed. The analysis method will be the same as described in Section 9.3.3.

In addition, analgesic use will be handled as follows:

- Prohibited analgesic (including opioids): A subject will be counted as a non-responder from the day that the subject initiates prohibited analgesic to 5 days after the end of such analgesic use.
- Protocol-allowed analgesic use:
  - A subject who enter the study without concomitant analgesic will be counted as a non-responder from the day that the subject initiates a protocol-allowed analgesic to 2 days after the end of such analgesic use.
  - A subject who entered the study on a stable dose of a protocol-allowed concomitant analgesic will be counted as a non-responder from the first day of an analgesic dose increase to 2 days after the end of the dose increase.

The NRI-C will be the primary approach for missing data handling in the analysis of the secondary efficacy endpoint. NRI-NC will be used as sensitivity analysis.

9.5 Additional Efficacy Analyses

9.5.1 Additional Categorical Endpoints in Period 1

The analysis of flare during Period 1 will be performed using the same method as for the primary endpoint, comparing with historical placebo rate of the occurrence of flare in 12 weeks (34%). The historical placebo rate is assumed based on the response rate of placebo subjects from the two Humira studies (Study M11-313 and Study M11-810), who satisfied the same eligibility criteria. A supplemental analysis will be performed to compare upadacitinib 30 mg and in-trial placebo group.

9.5.2 Additional Continuous Endpoints in Period 1

Continuous endpoints except for DLQI will be analyzed using MMRM including observed measurements at all visits, as described in Section 9.2. For HSSA total score, HS-related swelling, odor, worst drainage, and HSIA at Week 12, the comparison will be made between upadacitinib 30 mg group and the historical change from baseline at Week

12 values of -0.72, -0.8, -0.72, -0.89 or -0.83 (HSIA historical value is the average of Week 8 and Week 16 since there is no Week 12 measurement), respectively, using MMRM. The historical change from baseline at Week 12 values are obtained based on the placebo subjects from Risankizumab Study M16-833, who satisfied the same eligibility criteria.

For change from baseline in DLQI at Week 12, the comparison will be made between upadacitinib 30 mg group and historical change from baseline (-2.5) using an ANCOVA model. The historical change from baseline of -2.5 is assumed based on the mean change from Baseline of placebo subjects from the two Humira studies (Study M11-313 and Study M11-810), who satisfied the same eligibility criteria.

Continuous endpoints will also be analyzed using MMRM for endpoints other than DLQI and ANCOVA for DLQI to compare upadacitinib 30 mg QD with in-trial placebo group.

9.5.3 Summary of Long-Term Efficacy

Long-term efficacy in Period 2 will be summarized by visit for ITT Population using the observed case approach, as described in Section 9.2.3.

9.6 Efficacy Subgroup Analyses

To evaluate the consistency of the efficacy over demographic and other baseline characteristics, summaries will be provided for the following subgroups for the primary efficacy endpoint.

- Age group (< 40 years, ≥ 40 – < 65 years, ≥ 65 years)
- Sex (male, female)
- Race (white, non-white)
- Smoking (current, ex or never)
- Body mass index (normal: < 25, overweight: ≥ 25 – < 30, obese: ≥ 30)
- Duration of HS since diagnosis (by median)
- Prior exposure to TNF antagonists (yes, no)

- Baseline Hurley stage (< III, III)

For age and BMI subgroups with fewer than 10% subjects will be combined with their adjacent subgroup.

10.0 Safety Analyses

10.1 General Considerations

Safety data will be summarized for each safety population. Safety summaries will be presented by treatment group. For ALL_UPA Population, upadacitinib 30 mg, 15 mg and a total group for all subjects on upadacitinib will be summarized. For the safety analysis, subjects are assigned to a treatment group based on the treatment actually received, regardless of the treatment randomized.

A subject's actual treatment sequence will be determined by the most frequent dose regimen received in Period 1.

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 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 through 30 days following the last dose of study drug in Period 2. 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.

The number and percentage of subjects experiencing TEAEs 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:

- 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 4.03)
- Any serious TEAE (SAE)
- Any treatment-emergent SAE related to study drug according to the investigator
- Any TEAE leading to discontinuation of study drug
- TEAEs of Special Interest
- Any TEAE leading to death
- Any deaths
- Deaths occurring ≤ 30 days after last dose of study drug
- Deaths occurring > 30 days after last dose of study drug.

The overview of AEs per 100 patient-years of exposure will also be provided for the above categories.

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

TEAEs 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 severity and SOC and PT; and by subject number and 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, TEAEs will be summarized by PT and sorted by decreasing frequency for upadacitinib 30 mg group for Safety_1 Population, and by the total active group for ALL_UPA Population.

10.2.4 Treatment-Emergent Adverse Events per Patient-Years of Exposure

Exposure-adjusted AEs per 100 patient-years will be provided by SOC and PT, where AEs per 100 patient-years of exposure are defined as the number of AEs divided by the total exposure in 100 patient-years. The study drug exposure is defined as (last dose – first dose) + 1.

10.2.5 SAEs (Including Deaths) and Adverse Events 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.

In addition, the event rate per 100 patient-years of exposure will also be provided by SOC and PT for each treatment emergent SAE and TEAE leading to study drug discontinuation.

10.2.6 Adverse Events of Special Interest

Adverse events of special interest 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. 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 are provided in Appendix B.

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

In addition, the event rate per 100 patient-years of exposure will also be provided by SOC and PT for each AESI.

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.

Mean change from baseline to each applicable post-baseline visit will be summarized for selected laboratory variables, with the number of observations, baseline mean, and visit mean for each safety population. The change from baseline mean, standard error, and 95% confidence interval will be presented for the mean change from baseline within each treatment group and difference between treatment groups (active vs. placebo). Percent change from baseline in LDL-C, HDL-C, total cholesterol and ratio of LDL-C to HDL-C will also be summarized similarly.

Changes in laboratory parameters will be tabulated using shift tables by NCI CTC criteria (Version 4.03) for each safety population. A shift table from baseline to the worse value (based on NCI CTC criteria) during treatment will be created. A similar shift table will be provided to summarize shifts from baseline to the final post-baseline value.

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 at least one laboratory value meeting the criteria will be summarized for each safety population. 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 the following criteria of potential clinical interest should be presented:

- ALT $\geq 3 \times$ ULN
- ALT $\geq 5 \times$ ULN

- ALT ≥ 10 × ULN
- ALT ≥ 20 × ULN
- AST ≥ 3 × ULN
- AST ≥ 5 × ULN
- AST ≥ 10 × ULN
- AST ≥ 20 × ULN
- TBL ≥ 2 × ULN
- Alkaline phosphatase ≥ 1.5 × ULN
- ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 1.5 × ULN
- ALT and/or AST ≥ 3 × ULN and concurrent TBL ≥ 2 × ULN

10.4 Analysis of Vital Signs

Vital sign measurements of systolic and diastolic blood pressure, pulse rate, and body temperature will be summarized for each safety population.

Mean 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. The change from baseline mean, standard error, and 95% confidence interval will be presented for the mean change from baseline within each treatment group and difference between treatment groups (active vs. placebo).

Vital sign variables will be evaluated based on PCI criteria (Appendix C). For each vital sign PCI criterion, the number and percentage of subjects who have at least one 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

Not applicable.

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 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 nature of the phase 2 proof of concept study, all the endpoints will be tested at alpha = 0.05 (1-sided) and no alpha-spending is needed.

14.0 Version History

Table 2. SAP Version History Summary

Version	Date	Summary
1.0	25 Sep 2020	Original version
2.0	30 Apr 2021	• Added the supplemental analyses of the comparison between in-trial upadacitinib data with in-trial placebo data combined with historical placebo data for primary and secondary endpoints. • Added the supplemental analyses of the comparison between in-trial upadacitinib data with in-trial placebo data for all categorical endpoints. • Provided details for the MMRM and ANCOVA for continuous endpoints. • Added the comparisons of upadacitinib with single placebo change from baseline for HSSA total score and components and HSIA based on the estimates from Study M16-833. • Clarified the population for the summary of exposure and concomitant medications. • Adjust the calculations of the study drug exposure in the TEAE per patient-years of exposure.

15.0 References

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

5. 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.
6. 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.
7. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30(4):377-99.
8. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985;39(1):33-8.
9. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res. 2011;46(3):399-424.

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	
CPK	U/L		> 5.0 × ULN
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 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-C

Endpoints	Random Seed	
	MCMC procedure	**PROC MI**
Abscess count	1001	9001
Draining fistula count	1002	9002
Inflammatory nodule count	1003	9003
NRS30	1004	9004
≥25% AN increase	1005	9005

B. Random Seeds for MI

Endpoints	Random Seed	
	MCMC procedure	**PROC MI**
Abscess count	1006	9006
Draining fistula count	1007	9007
Inflammatory nodule count	1008	9008

Appendix E. Non-Responder Imputation Incorporating Multiple Imputation to Handle Missing Data Due to COVID-19 Pandemic 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 guidance in March 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 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 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-C)

2.1 Overall Description of the Method

For a dichotomized outcome variable with missing data, the NRI-C 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 the particular Study Period, 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, the information will be captured in the database and the subject's response status will be imputed using multiple imputation.

Of note, later visits of subjects receiving antibiotics for HS-related infections or discontinued study drug due to lack of efficacy will be set as missing before imputation. As a result, these assessments will not contribute to the imputation and the subjects will be counted as non-responders for the analysis at later visits. When determining the HiSCR response status, a subject with missing Baseline value will be counted as a non-responder at any post-baseline visit unless a subject's abscess, inflammatory nodule, and draining fistula counts are all zero at a specific visit, then the subject will be a HiSCR responder at that visit. Non-responder imputation incorporating multiple imputation (NRI-C) for missing due to COVID-19 will be implemented as follows.

2.2 Multiple Imputation (MI) and MAR Assumption

When a dichotomized variable is derived from continuous scales, for example, HiSCR (as at least a 50% reduction in the total abscess and inflammatory nodule (AN) count with no increase in abscess count and no increase in draining fistula count relative to Baseline), the multiple imputation will be applied to the original scales of abscess, inflammatory nodule and draining fistula counts, assuming multivariate normal distributions. Then the

dichotomized variable will be derived from the imputed values. For demonstration purposes, we use HiSCR as an example in this Appendix E, and provide sample code for the imputation of abscess count for one sample Chi-square test and two sample CMH test in Appendix E Section 3.0. The same imputation strategy will be applied to the inflammatory nodule count and the draining fistula count. HiSCR status will be determined accordingly.

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., abscess count 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., abscess count 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 group, prior

exposure to anti-TNF category, Baseline Hurley Stage, Baseline abscess count, all post-baseline visits of abscess count up to the end of the Study Period. 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 group, prior exposure to anti-TNF category, Baseline Hurley Stage, Baseline abscess count, all post-baseline visits of abscess count up to the end of the Study Period. 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 (Graham, Olchowski, and Gilreath 2007).[5] 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[5] suggested that the number of repetitions (K) should be at least equal to the percentage of missing (White et al., 2011).[7]

2.4 Derivation of Response Status and Non-Responder Imputation

Same as the abscess count, we also impute to obtain complete datasets for the inflammatory nodule count and the draining fistula count. For each 'complete' dataset, the imputed post-baseline values will be rounded to the same precision as the observed data. HiSCR status will be determined accordingly.

The imputed response status for missing due to reasons other than COVID-19 will be overridden by non-responder imputation (Section 2.1) to ensure that multiple imputation is only applied to missing due to COVID-19:

- Using NRI-C approach, all missing due to reasons other than COVID-19 will be categorized as non-responders, including visits after a subject receives antibiotics for HS-related infections or discontinued study drug due to lack of efficacy. When determining the HiSCR response status, a subject with missing Baseline value will be counted as a non-responder at any post-baseline visit unless a subject's abscess, inflammatory nodule, and draining fistula counts are all zero at a specific visit, then the subject will be a HiSCR responder at that visit.

- 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 SPP-specified visit window in the particular Study Period.

2.5 Analysis

The statistical analysis will use the Chi-square test or CMH test adjusted by the actual stratification factors.

2.5.1 Analysis of Each Dataset

For each of the K 'complete' datasets, the Chi-square test or 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

For the comparison between upadacitinib 30 mg and historical placebo rate, the chi-square distributed statistics from each of the K datasets will be pooled using the procedure of Rubin (1987) and Li et al. (1991). Denote by χ_k^2 chi-square distributed statistics with l degrees of freedom estimated in each of the K imputed datasets. A pooled test statistics can be obtained as follows:

$$D_\chi = \frac{\frac{\bar{\chi}^2}{l} - \frac{K+1}{K-1}\bar{r}_\chi}{1+\bar{r}_\chi}$$

where $\bar{\chi}^2 = \frac{1}{K}\sum_{k=1}^{K}\chi_k^2;\ \bar{r}_\chi = \left(1+\frac{1}{K}\right)\frac{1}{K-1}\sum_{k=1}^{K}\left(\sqrt{\chi_k^2} - K^{-1}\sum_{k=1}^{K}\sqrt{\chi_k^2}\right)^2$

The pooled p-value for the hypothesis test based on D_χ can be obtained using F distribution with l and V_x as numerator and denominator degrees of freedom, respectively, as follows:

$$P_\chi = Pr\left[F_{l,V_\chi} > D_\chi\right];\quad V_\chi = l^{-\frac{3}{K}}(K-1)\left(1+\frac{1}{\bar{r}_\chi}\right)^2$$

For the comparison between in-trial data from upadacitinib group and the in-trial placebo data combined with subject-level historical data, the results from the K 'complete' datasets will be synthesized using the SAS procedure PROC MIANALYZE, following Rubin's formula (Rubin, 1987),[6] to derive the MI estimator of the treatment difference for the final inferences.

We fit the analysis model to the kth 'complete' dataset, denoting the estimate of the treatment difference q by $\tilde{\theta}_k$ from the kth '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:

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

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

It has been shown[6] that the statistic

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

has an approximate t_v distribution where $v = (K-1)\left(1+\frac{W}{B}\right)^2$. 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 COUNT_2; SET COUNT;
  /*THE MCMC STATMENT BELOW ASSUMES MULTI-VARIATE NORMAL*/
  /*  USE ALL DATA IN THE IMPUTATION MODEL*/
  IF HURLEY=1 THEN HURLEY1 = 1 ELSE HURLEY1 = 0; /* CREATE DUMMY
VARIABLE FOR HURLEY STAGE 1*/
  IF HURLEY=2 THEN HURLEY2 = 1 ELSE HURLEY2 = 0; /* CREATE DUMMY
VARIABLE FOR HURLEY STAGE 2*/
RUN;

/*AUGMENTATION STEP -- TO HAVE 30 MONOTONE MISSING DATASETS*/
PROC MI DATA= COUNT_2 OUT= COUNT_MONO NIMPUTE=30 SEED= 1001 /*RANDOM
SEED PRE-DEFINED*/
  ROUND=. . . . 1 1 1 1 1 /*VALUE ROUND TO INTEGER*/
  MIN=. . . . 0 0 0 0 0  /*MINIMUM VALUE OF COUNT IS 0*/
  MAX=. . . . . . . . .   /*MAXIMUM VALUE*/
MCMC IMPUTE=MONOTONE ;
/*NOTE: CATEGORICAL VARIABLES SUCH AS HURLEY1 HURLEY2 ARE DUMMY, CREATED
ABOVE*/
/*NOTE: ALL OTHER NON-DUMMIED COVARIATES MUST BE CONTINUOUS*/
VAR TRT01PN TNFN HURLEY1 HURLEY2 BASE WK2 WK4 WK8 WK12;
/*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= COUNT_MONO OUT= COUNT_FULL NIMPUTE=1 SEED= 9001  /*RANDOM
SEED PRE-DEFINED*/
```

```
  ROUND=. . . 1 1 1 1 1 /*VALUE ROUND TO INTEGER*/
  MIN=. . . 0 0 0 0 0   /*MINIMUM VALUE OF COUNT IS 0*/
  MAX=. . . . . . . .    /*MAXIMUM VALUE*/
  MINMAXITER=1000;
  /*CLASS CATEGORICAL VARIABLES*/
  CLASS TRT01P TNFN HURLEY;
  VAR TRT01P TNFN HURLEY BASE WK2 WK4 WK8 WK12;
  MONOTONE REG (WK2 WK4 WK8 WK12);  /* IMPUTED SEQUENTIALLY, FROM WK 1
TO 16, WITH COVARIATES CONSTRUCTED FROM THE CORRESPONDING PRECEDING
VARIABLES*/
  BY _IMPUTATION_;                    /*FOR EACH OF THE 30 MONOTONE
MISSING DATASETS, IMPUTE A 'COMPLETE' DATASET*/
RUN;


/*We will get the imputed "complete" datasets for abscess, inflammatory
nodule, and draining fistula counts. For sample code below, we denote
the following variables: ABS12, INF12 and DRA12 as the corresponding
lesion count at Week 12; and AN12 as the sum of ABS12 and INF12.
Determine the HiSCR status at Week 12.  */
DATA ALL; SET COUNT_FULL;
    IF 0<= AN12 <=0.5*AN_BASE AND ABS12<=ABS_BASE AND DRA12 <= DRA_BASE
THEN HISCR_12=1;
    ELSE HISCR_12=0;
RUN;


/*************************************************************************
*/
/*         DATA HANDLING STEPS TO MERGE COVID-19 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 WITH TRADITIONAL
NRI*/
DATA ALLF; SET ALL;
  /*COVID19_XX='Y' IF MISSING AT WEEK XX IS DUE TO COVID-19; IF NOT,
OVERRIDE WITH TRADITIONAL NRI*/
  /*VARIABLE HISCR_NRI_12: TRADITIONAL NRI DATA AT WEEK 12, WHICH COVERS
THE SPECIAL HANDLING SUCH AS THE BEFORE-AND-AFTER EXCEPTION IN THE
PARTICULAR STUDY PERIOD*/
  IF COVID19_MISS NE 'Y' THEN HISCR_12 = HISCR_NRI_12;
RUN;
PROC SORT DATA=ALLF; BY _IMPUTATION_ SUBJID; RUN;
```

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

/*KEY CODE: ANALYZING EACH 'COMPLETE' DATASET*/

/*********************************/
/*******************************/
/*FOR ONE SAMPLE CHI-SQUARE TEST*/
/*******************************/

ODS OUTPUT ONEWAYCHISQ=CHI;
PROC FREQ DATA=ADEFF;
  BY _IMPUTATION_;
  TABLES CRIT2FN/CHISQ TESTP=(0.75 0.25) ;
RUN;

DATA CHI;
  SET CHI;
  KEEP _IMPUTATION_ NVALUE1 LABEL1;
RUN;

/*GET CHI-SQUARE STATISTICS*/
DATA CHISQ_VALUE;
  SET CHI(WHERE = (LABEL1 = "Chi-Square"));
  VALUE = NVALUE1;
  DROP NVALUE1;
RUN;

/*GET DEGREE OF FREEDOM*/
DATA CHISQ_DF;
  SET CHI(WHERE = (LABEL1 = "DF"));
  DF = NVALUE1;
  DROP NVALUE1;
RUN;

DATA CHISQ;
  MERGE CHISQ_VALUE CHISQ_DF;
  BY _IMPUTATION_;
  A=1;
RUN;

/*CALCULATE MEAN OF CHI-SQUARE STATISTICS*/
PROC MEANS DATA = CHISQ MEAN NOPRINT;
  VAR VALUE;
  OUTPUT OUT = M_CHISQ MEAN = M_VALUE N = COUNT;
RUN;

DATA M_CHISQ;
```

```
  SET M_CHISQ;
  A=1;
RUN;

DATA CHISQ1;
  MERGE CHISQ M_CHISQ;
  BY A;
  MEAN_VALUE = M_VALUE;
  M = COUNT;
  DROP _FREQ_ _TYPE_ M_VALUE COUNT;
RUN;

DATA TEMP;
  SET CHISQ1;
  SQRT_VALUE=SQRT(VALUE);
RUN;

/*CALCULATE MEAN OF SQUARE ROOT OF CHI-SQUARE STATISTICS*/
PROC MEANS DATA = TEMP MEAN NOPRINT;
  VAR SQRT_VALUE;
  OUTPUT OUT=M_CHISQ1 MEAN=M_SQRT_VALUE;
RUN;

DATA M_CHISQ1;
  SET M_CHISQ1;
  A=1;
RUN;

DATA CHISQ2;
  MERGE CHISQ1 M_CHISQ1;
  BY A;
  MEAN_SQRT=M_SQRT_VALUE;
  DROP _FREQ_ _TYPE_ M_SQRT_VALUE;
RUN;

DATA CHISQ3;
  SET CHISQ2;
  RX = (1+1/M)*(((VALUE**0.5) - MEAN_SQRT)**2)/(M-1);
RUN;

/*CALCULATE MEAN OF RX*/
PROC MEANS DATA = CHISQ3 MEAN NOPRINT;
  VAR RX;
  OUTPUT OUT = T_RX SUM = T_RX;
RUN;

DATA T_RX;
  SET T_RX;
  A=1;
```

```
RUN;

DATA CHISQ4;
  MERGE CHISQ3 T_RX;
  BY A;
  RX_RESULT = T_RX;
  DROP _FREQ_ _TYPE_ T_RX;
RUN;

PROC SORT DATA=CHISQ4 NODUPKEY;
  BY MEAN_VALUE DF RX_RESULT M MEAN_SQRT;
RUN;

/*CALCULATE P VALUE*/
DATA P_CHISQ;
  SET CHISQ4;
  DROP RX;
  DX = ((MEAN_VALUE/DF) - (M+1)*RX_RESULT/(M-1))/(1+RX_RESULT);
  VX = (DF **(-3/M)) * (M-1) * ((1 + 1/RX_RESULT) ** 2);
  PVAL = 1 - CDF("F", DX, DF, VX);
RUN;

/*************************/
/*FOR TWO SAMPLE CMH TEST*/
/*************************/

/*INDIVIDUAL-LEVEL DATA --> # OF RESPONDERS & # OF SUBJECTS, TO BE READ-
IN TO PROC STDRATE */
PROC FREQ DATA=ALLF;
  BY _IMPUTATION_;
  TABLES TRT01PN*STRATA*HISCR_12/LIST NOCUM NOPRINT OUT=COUNT_TABLE;
RUN;
DATA COUNT_TABLE; SET COUNT_TABLE;
  DROP PERCENT;
RUN;
PROC TRANSPOSE DATA=COUNT_TABLE OUT=FREQ_TABLE PREFIX=RESP;
ID HISCR_12;
BY _IMPUTATION_ TRT01PN STRATA;
VAR COUNT;
RUN;
DATA FREQ_TABLE1; SET FREQ_TABLE;
  CASE=RESP1;
  SIZE=SUM(RESP0, RESP1);
  KEEP _IMPUTATION_ TRT01PN STRATA CASE SIZE;
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 STRATA / 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;
```

Appendix F. Simulations for Combining Subject-level Historical Placebo Data with In-trial Placebo Data

1.0 Introduction

In randomized controlled trials, randomization ensures that the different arms of the trial tend to be balanced, both in terms of measured and unmeasured characteristics of the population. Thus, the only systematic difference between treated and untreated patients is the exposure to treatment. This allows one to obtain unbiased estimates of the average treatment effect within the context of the study design. In the studies such that the treatment effect will be estimated with the combination of in-trial placebo subjects and historical placebo subjects, the distribution of subject characteristics is likely to vary systematically between the in-trial subjects and historical subjects and is often related to patient prognosis, patient characteristics, and physician preference. Thus, confounding may occur.

Propensity score matching method (Rosenbaum and Rubin, 1985) matches in-trial patients with synthetic placebo subjects based on propensity scores such that matched in-trial subjects and placebo subjects are comparable in terms of covariates used in propensity score determination. This process mimics randomization to create two comparable groups with a caution of limitation that the matching is conditional on included covariates.

To provide the guidance on how many synthetic placebo subjects can be matched for the primary and secondary analysis (HiSCR and NRS30), simulations have been conducted to evaluate the power and false positive rate of combining the data from historical placebo subjects and in-trial subjects under different scenarios.

2.0 Methods

The simulation procedures are outlined in the following sections.

2.1 Simulation of Baseline Covariates and Treatment Groups

Within each simulation, the following baseline covariates were generated for the 68 in-trial subjects from multivariate normal distribution:

- Age
- Baseline AN count
- Baseline BMI
- Duration of HS since diagnosis
- Baseline draining fistula
- Gender
- Race
- Smoking status

The multivariate normal distribution with mean vector μ, standard deviation S and correlation matrix ∑ is defined as below (covariates in the order as listed above):

$$\mu' = (35.8, 21.8, 35.2,11,3.56,0,0,0)$$

$$S' = (11.91, 20.73, 8.65, 8.4, 5.14, 1,1,1)$$

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & 0.4 & 0.1 & 0 & 0 & 0 \\ 0 & 1 & 0.1 & 0 & 0.3 & 0 & 0 & 0 \\ 0 & 0.1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0.4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.3 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0.2 & 1 \end{pmatrix}$$

The means and standard deviations for the continuous variables were estimated based on the baseline information from the in-trial subjects. For the binary variables, random

samples were generated using standard normal distribution and the binary outcome of 0 or 1 was determined by the proportions estimated from the in-trial subjects.

Based on the baseline information from 68 in-trial subjects, the two stratification factors, baseline Hurley Stage (<III, III), and the indicator of prior exposure to anti-TNF treatment (Yes, No) were generated assuming that 5% of the subjects are Hurley Stage <III and have prior exposure to anti-TNF treatment, 10% of the subjects are Hurley Stage III and have prior exposure to anti-TNF treatment, 55% of the subjects are Hurley Stage <III and do not have prior exposure to anti-TNF treatment, and 35% of the subjects are Hurley Stage III and do not have prior exposure to anti-TNF treatment.

Treatment groups (upadacitinib 30 mg and placebo) were generated in a 2:1 ratio within each of the 4 strata. With a total of 68 subjects, it was assumed that 45 subjects received upadacitinib 30 mg, and 23 subjects received placebo.

2.2 Matching Historical Placebo Subjects Using Propensity Score

In the matching phase, placebo subjects from existing studies (Study M11-313, Study M11-810, Study M16-833) were matched in a fixed ratio with in-trial subjects using propensity score method with pre-selected baseline covariates. Nearest neighbor matching method was used.

Different numbers of matched subjects were evaluated:

Number of Matched Subjects	Matching Ratio (in-trial: matching)	Control Set
23	1:1	In-trial placebo subjects (N=23)
46	1:2	In-trial placebo subjects (N=23)
68	1:1	All in-trial subjects (N=68)

3.0 Scenarios and Simulation Results for Primary Endpoint (HiSCR) and Secondary Endpoint (NRS30)

3.1 HiSCR

Different scenarios with regard to the variations in the in-trial placebo rate relative to the historical placebo rate (28%) were considered: in-trial placebo subjects perform worse than historical placebo subjects by 10% or 5%, in-trial placebo subjects perform similarly as the historical placebo subjects, and in-trial placebo subjects perform better than historical placebo subjects by 5% or 10%, as summarized in the following table.

Scenario	Placebo Rate	Upadacitinib Rate*
In-trial PBO rate = Historical rate -10%	0.18	0.59
In-trial PBO rate = Historical rate - 5%	0.23	0.59
In-trial PBO rate = Historical rate	0.28	0.59
In-trial PBO rate = Historical rate + 5%	0.33	0.59
In-trial PBO rate = Historical rate + 10%	0.38	0.59

* Assume that upadacitinib performs at least as good as min TPP.

One-sided power and false positive rates for each scenario and different numbers of matched subjects were calculated, and the results from 10000 simulations (random seed: 12345) are summarized in the following table:

	Power				False Positive Rate			
	Number of Matched Subjects			In-Trial Only	Number of Matched Subjects			In-Trial Only
Scenario	23	46	68	UPA vs PBO: 45:23	23	46	68	UPA vs PBO: 45:23
In-trial PBO rate = Historical rate - 10%	0.978	0.984	0.984	0.977	0.013	0.004	0.002	0.082
In-trial PBO rate = Historical rate - 5%	0.957	0.973	0.977	0.926	0.027	0.014	0.009	0.075
In-trial PBO rate = Historical rate	0.925	0.954	0.966	0.837	0.051	0.04	0.033	0.072
In-trial PBO rate = Historical rate + 5%	0.877	0.933	0.951	0.701	0.087	0.093	0.091	0.07
In-trial PBO rate = Historical rate + 6%*	0.865	0.927	0.948	0.672	0.095	0.105	0.108	0.07
In-trial PBO rate = Historical rate + 10%	0.811	0.904	0.935	0.549	0.133	0.168	0.189	0.069

* To determine the number of matched subjects, additional scenario of historical rate + 6% was also evaluated and the power and false positive rate are given in the table.

The simulation results show that power with matched placebo subjects from historical data is higher than that with in-trial placebo subjects only in all of the scenarios. Power increases with the increase in the number of matched placebo subjects. In the meantime, when the in-trial placebo rate is lower than or similar to the historical placebo rate, the false positive rates are controlled at approximately 0.05. When the in-trial placebo rate is higher than historical placebo rate, the false positive rate increases with the increase in the number of matched subjects. The inflation in the false positive rate is larger when higher in-trial placebo rate is observed. The number of matched subjects is therefore determined by both achieving a high power and maintaining a slightly inflated false positive rate.

To determine the number of matched placebo subjects in the analysis, the following two criteria should be satisfied:

- One sided false positive rate is controlled at 10%;
- Not to exceed 68 (approximately 3 times the number of in-trial placebo subjects).

The numbers of matched placebo subjects are given in the following table:

In-Trial PBO Rate	Number of Matched Subjects
≤ 33%	68
>33% and ≤ 34%	23
>34%	0

3.2 NRS30

With historical placebo rate at 26%, the following scenarios of the proportion of subjects achieving NRS30 at Week 12 were evaluated to assess the power and false positive rates with different number of matched subjects:

Scenario	Placebo Rate	Upadacitinib Rate*
In-trial PBO rate = Historical rate – 10%	0.16	0.46
In-trial PBO rate = Historical rate – 5%	0.21	0.46
In-trial PBO rate = Historical rate	0.26	0.46
In-trial PBO rate = Historical rate + 5%	0.31	0.46
In-trial PBO rate = Historical rate + 10%	0.36	0.46

* Assume that upadacitinib performs at least as good as min TPP.

One-sided power and false positive rates for each scenario and different numbers of matching subjects were calculated, and the results from 10000 simulations (random seed: 12345) are summarized in the following table:

	Power				False Positive Rate			
	Number of Matched Subjects			In-Trial Only	Number of Matched Subjects			In-Trial Only
Scenario	23	46	68	UPA vs PBO: 45:23	23	46	68	UPA vs PBO: 45:23
Historical rate - 10%	0.813	0.788	0.775	0.899	0.006	0.001	0	0.094
Historical rate - 5%	0.721	0.723	0.724	0.769	0.014	0.006	0.003	0.084
Historical rate (Base case)	0.625	0.65	0.668	0.603	0.03	0.018	0.012	0.077
Historical rate + 5%	0.521	0.573	0.607	0.434	0.057	0.045	0.038	0.077
Historical rate + 10%	0.41	0.492	0.542	0.279	0.089	0.098	0.096	0.073
Historical rate + 11%*	0.391	0.476	0.528	0.249	0.099	0.112	0.115	0.073

* To determine the cutoff number in the matched subjects, additional scenario of historical rate + 11% is also conducted and the power and false positive rate are given in the above table.

The same conclusion can be reached with the results provided from the simulations.

The numbers of matched placebo subjects are given in the following table using the same criteria as for HiSCR:

In-Trial PBO Rate	Number of Matched Subjects
≤ 36%	68
>36% and ≤ 37%	23
>37%	0

Appendix G. List of Matching Subjects

Table A. List of matching subjects for HiSCR:

Subject ID	Distance	Rank
[redacted]	2.57E-06	1
	8.96E-05	2
	0.000158	3
	0.000268	4
	0.000327	5
	0.000411	6
	0.000466	7
	0.000475	8
	0.000638	9
	0.000688	10
	0.000781	11
	0.000937	12
	0.000982	13
	0.000998	14
	0.001223	15
	0.001303	16
	0.001496	17
	0.001548	18
	0.001571	19
	0.001755	20
	0.001853	21
	0.001972	22
	0.00217	23
	0.002261	24
	0.002472	25
	0.002498	26
	0.002688	27
	0.002832	28
	0.002928	29

Subject ID	Distance	Rank
[redacted]	0.003055	30
	0.003187	31
	0.004881	32
	0.005842	33
	0.006369	34
	0.006551	35
	0.007607	36
	0.008428	37
	0.009143	38
	0.011198	39
	0.012126	40
	0.012508	41
	0.013673	42
	0.014632	43
	0.015039	44
	0.015057	45
	0.015973	46
	0.016307	47
	0.0166	48
	0.016938	49
	0.018109	50
	0.018142	51
	0.018631	52
	0.025821	53
	0.026235	54
	0.027435	55
	0.030238	56
	0.035562	57
	0.036589	58
	0.04232	59
	0.045276	60
	0.049032	61

Subject ID	Distance	Rank
[redacted]	0.058936	62
	0.072734	63
	0.142095	64
	0.199571	65
	0.203572	66
	0.212218	67
	0.253149	68

Note: the subject's rank is based on the distance, a subject with the closer distance has a higher rank.

Table B. List of matching subjects for NRS30:

Subject ID	Distance	Rank
[redacted]	0.000269	1
	0.000516	2
	0.000929	3
	0.001143	4
	0.001297	5
	0.001443	6
	0.001773	7
	0.002621	8
	0.002963	9
	0.003434	10
	0.003834	11
	0.004059	12
	0.004439	13
	0.004735	14
	0.004945	15
	0.005183	16
	0.006073	17
	0.00641	18
	0.006571	19
	0.00688	20
	0.008785	21

Subject ID	Distance	Rank
[redacted]	0.011996	22
	0.012155	23
	0.012481	24
	0.012774	25
	0.013329	26
	0.014019	27
	0.018572	28
	0.019276	29
	0.019987	30
	0.020912	31
	0.022264	32
	0.022325	33
	0.027872	34
	0.029373	35
	0.030355	36
	0.034626	37
	0.035711	38
	0.037138	39
	0.040258	40
	0.041608	41
	0.046707	42
	0.048336	43
	0.049609	44
	0.050289	45
	0.050512	46
	0.051712	47
	0.051713	48
	0.05409	49
	0.055272	50
	0.055707	51
	0.056069	52
	0.057397	53

Subject ID	Distance	Rank
[redacted]	0.058898	54
	0.059654	55
	0.06074	56
	0.061507	57
	0.064318	58
	0.068315	59
	0.074845	60
	0.09129	61
	0.09648	62
	0.099541	63
	0.100568	64
	0.100701	65
	0.102387	66
	0.106828	67
	0.112646	68

Note: the subject's rank is based on the distance, a subject with the closer distance has a higher rank.