Protocol B7981015

A PHASE 2B/3 RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, DOSE-RANGING STUDY TO INVESTIGATE THE EFFICACY AND SAFETY OF PF-06651600 IN ADULT AND ADOLESCENT ALOPECIA AREATA (AA) SUBJECTS WITH 50% OR GREATER SCALP HAIR LOSS

Statistical Analysis Plan
(SAP)

Version: 4

Date: 23 April 2021

TABLE OF CONTENTS

[redacted]

LIST OF TABLES

LIST OF FIGURES

1. VERSION HISTORY

Table 1. Summary of Major Changes in SAP Amendments

Version/Date	Associated Protocol Amendment	Rationale	Specific Changes
1 1 NOV 2018	Original 17 SEP 2018	Not Applicable	Not Applicable
2 11 Aug 2020	Amendment 3 14 July 2020	• To be consistent with the protocol amendments 1-3. • To provide additional details and clarification.	• Changed primary endpoint from SALT ≤10 at Week 24 to SALT ≤20 at Week 24. It was also noted that SALT ≤10 will be analyzed as the primary endpoint in a separate analysis where required (Sections 2.1, 2.2, 3.1, 5.1, 6.1). • Added PGI-C response at Week 24 will be analyzed as a key secondary endpoint where required (Section 2.1. 2.2, 3.2, 5.1, 6.2). • Updated endpoints according to the protocol and provided additional details (Sections 2.1, 3, 6). • Added stratified randomization (Section 2.2). • Updated power estimation (Section 2.2). • Added the study may be unblinded for internal decision-making purposes (Sections 2.2, 7). • Added Per protocol analysis set (Section 4.2). • Added testing procedure for multiple comparisons for SALT ≤10 at Week 24 and SALT ≤20 at earlier timepoints if SALT ≤20 at Week 24 is the primary endpoint. It was also noted that for regions that request a primary endpoint of SALT ≤10 at week 24, a similar approach will be used to control the Type I error for SALT ≤20 at Week 24 and SALT ≤10 at earlier time points (Section 5.1). • Added testing procedure for multiple comparisons incorporating a key secondary endpoint for regions that request PGI-C response at week 24 as a key secondary endpoint where required (Section 5.1.1).

			• Added details on data summary in the extension phase (Sections 5.2, 6). • Updated visit windows (Appendix 1). • Added Appendices 4-6.
Version 3 13 November 2020	Protocol Amendment 4 20 October 2020	• To comply with the VHP feedback that EU specific endpoints and analysis should be explicitly stated (beyond the verbatim that already existed to manage region specific preferences) in a consistent manner and to align with the protocol Amendment 4 which was updated given this VHP feedback	• It has been specifically clarified that for the European Union (EU) (including the Voluntary Harmonisation Procedure [VHP] countries), the primary objective and primary endpoint will utilize an absolute Severity of Alopecia Tool (SALT) Score of ≤10 and not ≤ 20 at Week 24. • Sections 2.1 (study objectives), 2.2 (sample size), 3.1 (primary endpoint), 3.2 (secondary endpoints), 6.1 (analysis of primary endpoint), and 6.2 (analysis of secondary endpoints). • It has been specifically clarified that for the EU (including VHP countries), the evaluation of the effect of PF-06651600 on patient centered outcomes (as measured by Patient's Global Impression of Change [PGI-C] response at Week 24) is a key secondary objective. The PGI-C response at Week 24 will be analyzed as a key secondary endpoint utilizing an appropriate testing procedure to control overall Type I error. • Sections 2.1 (study objectives), 2.2 (sample size), 3.2 (secondary endpoints), 5.1 (hypotheses and decision rules), 5.1.1 (testing procedure for multiple comparisons incorporating a key secondary endpoint) and 6.2.8 (analysis of PGI-C response). • Objectives and Endpoints table (including footnotes) was updated to clarify which secondary endpoints will be controlled for Type I error based on which SALT parameter at Week 24 (ie, ≤20 or ≤10) is utilized as the primary endpoint. • Section 2.1 (study objectives). • Endpoints tables was updated to clarify that for the EU (including

			VHP countries), the endpoints utilizing the HADS will be analyzed as secondary endpoints. • Sections 2.1 (study objectives), and 3.3.1 (exploratory endpoints). • Added text to clarify that more stringent alpha levels have been advised for regulatory submission of B7981015 in the EU and for the FDA. • Sections 2.2 (sample size) and 5.1 (hypothesis and decision rules).
		• General changes, typographical corrections, and clarifications	• Changed the utility weights for the derivation of EQ-5D scores (Section 3.3.1). US utility weights will be used for all participants so that all patients are on the same weighted scale for modeling purposes • Added clarification of the definition of baseline for ECG assessments (Sections 3.4 and 3.5.5). • Correction of grammatical and typographical errors in Sections 2.1, 2.2, 3.3.1, 5.2.1, 5.2.2, 5.3.1, 6.3.8, 6.3.9, 6.3.14, 6.5, 6.6.1, 6.7, 6.7.1, and Appendix 1. • Replaced all text from Sections 5.2.1, 6.1 and 6.2 that mentioned approach for missing data of binary endpoints by a general sentence that refers to Section 5.3.1 • Added Section 6.1.2 to describe the analysis of the primary endpoint in the per protocol analysis set (PPAS). • A standard sentence was used to describe the approach for missing data throughout Section 6.2 to refer back to the missing data approach described in Section 5.3. • HADS cut off values were added to sections 6.3.10 and 6.3.11 for consistency with Section 3.3.1. • Added subset of subjects who received prior pharmacological treatment to AA to Section 6.5 and deleted “requiring discontinuation at any time” in Section 6.7.4.

[redacted]

Version 4 23 April 2021	Protocol Amendment 5 13 April 2021	• To align with protocol Amendment 5, specifically to define the endpoints and alpha levels and hypothesis testing for the overall study and to clarify the endpoints and alpha and hypothesis testing for FDA/PMDA and for EMPA and competent authorities in VHP countries and to incorporate changes to the analysis and approach for handling missing due to COVID-19 after feedback received from regulatory agencies. • The FAS was revised based on regulatory feedback to include all subjects who were randomized regardless of whether they received treatment • To provide further details on the definition of the per protocol population based on regulatory feedback • Also, further details of the E_{max} models for the exposure response, and tipping point analysis were provided • Planned COVID-19 related summaries and listings were added • Unblinding of data for PK/PD modelling was added • Several places in this document were corrected for typographical errors.	• Early unblinding for PK/PD modelling was added to Section 2.2 • Further clarification was added to the sample size section 2.2.1 to account for the alphas for the different regions. • Tables 3 and 4 were added to Section 3 to define the endpoints and alpha for the overall study, and clarify the endpoints and alpha for FDA/PMDA and for EMPA and competent authorities in VHP • FAS and PPAS were updated in Sections 4.1 and 4.2 • Hypothesis testing for the overall study was added to Section 5.1 and was further clarified for FDA/PMDA and for and for EMPA and competent authorities in VHP • Tables 5 was added to Section 5.2.1 to clarify the primary and supportive analyses specifically to address handling of missing data due to COVID-19 for the overall study and for FDA/PMDA and for EMPA and competent authorities in VHP • • Section 5.2.1 was reorganized into subsections 5.2.1.1, 5.2.1.2, 5.2.1.3 and 5.2.1.4 after the introduction of the primary analysis for the EMA and competent authorities in VHP countries, following advice from regulatory agency • • Section 5.2.2. was expanded to provide further details of the E_{max} models for the analysis for exposure response. • Section 5.3.1 was expanded to clarify handling of missing data due to COVID-19 • Section 6.1 was modified to summarize all analyses that are planned for each primary, key secondary and α-controlled endpoints. • Section 6.2 additional detail was provided to clarify handling of

			missing data due to COVID-19 for secondary endpoints • COVID-19 related planned summaries and listings were added to Section 6.7 • In Section 6.7.1, clarification on analyses of AEs of special interest was provided, and the planned summaries of COVID-19 related AEs, and AE summaries by COVID-19 anchor date were added • Analysis of Columbia Suicide Severity Rating Scale analysis was removed from Section 6.7.7 since this scale was only intended for monitoring • Further detail was provided for the interim analysis in Section 7, with regards to the dissemination of results, and to clarify that the interim will have no impact on B7981015 study conduct, analysis, or reporting • Tipping point and E_{max} references were added • Section 9.6 was added to provide further details and specifications of the tipping point analysis. • Section 9.7 was added to provide further specifications for the definition of subgroups for analysis

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study B7981015. This document may modify the plans outlined in the protocol; however, any major modifications of the primary endpoint definition or its analysis will also be reflected in a protocol amendment.

2.1. Study Objectives

The study objectives and corresponding endpoints are listed below in Table 2. There are some region specific differences noted throughout. Table 3 and Table 4 in Section 3.1 lists the primary and secondary endpoints of the study (re)organized by known regional requirements.

Table 2. Study Objectives and Endpoints

Primary Objective(s):	**Primary Endpoint(s):**
To evaluate the efficacy of PF-06651600 compared to placebo in adult and adolescent alopecia areata (AA) subjects with 50% or greater scalp hair loss on regrowth of lost hair (measured by an absolute Severity of Alopecia Tool (SALT) Score ≤20) at Week 24. Note: For the European Medicines Agency (EMA) and competent authorities in the Voluntary Harmonisation Procedure (VHP) countries,[a] the primary objective is to evaluate the efficacy of PF-06651600 compared to placebo in adult and adolescent AA subjects with 50% or greater scalp hair loss on regrowth of lost hair (as measured by an absolute SALT Score ≤10) at Week 24.	• Response based on an absolute Severity of Alopecia Tool (SALT) Score ≤20 at Week 24. Note: For the EMA and competent authorities in the VHP countries response based on an absolute SALT Score ≤10 at Week 24 will be analyzed as the primary endpoint in a separate analysis.
Key Secondary Objective(s)	**Key Secondary Endpoint(s)**
• To evaluate the efficacy of PF-06651600 compared to placebo in adult and adolescent AA subjects with 50% or greater scalp hair loss on regrowth of lost hair (as measured by an absolute SALT Score ≤10) at Week 24. Note: This key secondary objective will be utilized as the primary objective for the EMA and competent authorities in the VHP countries. Additionally, this key secondary objective will not apply for the United States Food and Drug Administration (FDA)/Japan Pharmaceuticals and Medical Devices Agency (PMDA).	• Response based on an absolute SALT Score ≤10 at Week 24. Note: This key secondary endpoint will be utilized as the primary endpoint for the EMA and competent authorities in the VHP countries. Additionally, this key secondary endpoint will not apply for the FDA/PMDA.
• To evaluate the effect of PF-06651600 on patient centered outcomes (as measured by PGI-C response) at Week 24. Note: This key secondary objective is only applicable for the EMA and competent authorities in the VHP countries.	• PGI-C response defined as a score of “moderately improved” or “greatly improved” at Week 24. Note: This key secondary endpoint is only applicable for the EMA and competent authorities in the VHP countries.

Secondary Objective(s):	**Secondary Endpoint(s):**
To characterize the exposure response of PF-06651600 on regrowth of lost hair.	Response based on an absolute SALT Score ≤20 at Week 24 will be used to characterize the exposure response.[b]
To assess the efficacy of PF-06651600 on regrowth of lost hair during the treatment period over time.	Response based on an absolute SALT score ≤20 at Weeks 4, 8, 12, 18, 28, 34, 40, and 48.[c,d] Response based on an absolute SALT score of ≤10 at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.[e] Response based on a 75% improvement in SALT score from baseline (SALT75) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. Change from baseline in SALT scores at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. Response based on at least a 2-grade improvement or a score of 3 in Eyebrow Assessment (EBA) score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. Response based on at least a 2-grade improvement or a score of 3 in Eyelash Assessment (ELA) score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
To evaluate the effect of PF-06651600 on patient-centered outcomes and payer relevant measures to assess treatment benefit from the patient perspective and to demonstrate value. [f]	PGI-C response defined as a PGI-C score of “moderately improved” or “greatly improved” at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.[f] Change from baseline in Alopecia Areata Patient Priority Outcomes (AAPPO) scales at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
Safety Objective(s)	**Safety Endpoint(s)**
To evaluate the safety and tolerability of PF-06651600 in the treatment period over time.	Incidence of treatment-emergent adverse events (AEs). Incidence of serious AEs (SAEs) and AEs leading to discontinuation. The incidence of clinically significant abnormalities in vital signs. The incidence of clinically significant abnormalities in clinical laboratory values.
PK Objective(s)	**PK Endpoint(s)**
To characterize pharmacokinetics of PF-06651600.	Plasma concentrations of PF-06651600 at Weeks 4, and 8 or 12.
Exploratory Objective(s):	**Exploratory Endpoint(s):**
To collect banked biospecimens for exploratory research, unless prohibited by local regulations or ethics committee decision.	Collection of banked biospecimens unless prohibited by local regulations or ethics committee decision. Additional information on collection and potential use is provided in the Banked Biospecimens section in the protocol.
To assess the efficacy of PF-06651600 on regrowth of lost hair during the treatment period over time.	Response based on a 50% improvement in SALT score from baseline (SALT50) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.

	Absolute SALT scores at Baseline, Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
To evaluate the effect of PF-06651600 on patient-centered outcomes and payer relevant measures to assess treatment benefit from the patient perspective and to demonstrate value.	Improvement on PGI-C defined as “slightly improved”, “moderately improved”, or “greatly improved” at Weeks 4, 8, 12, 18, 24, 34, 40, and 48. Improvement on Patient Satisfaction with Hair Growth (P-Sat) items defined as slightly, moderately, or very satisfied at Weeks 4, 8, 12, 18, 24, 34, 40, and 48. Change from baseline in EuroQoL 5 Dimensions (EQ-5D-5L) in adults or EuroQoL 5 Dimensions-Youth (EQ-5D-Y) in adolescents at Weeks 4, 12, 24, and 48. Change from baseline in Alopecia Areata Resource Utilization (AARU) at Weeks 12, 24, 34, and 48. Change from baseline in Work Productivity and Activity Impairment: Alopecia Areata (WPAI:AA) at Weeks 12, 24, 34, and 48. Change from baseline in the depression subscale score of the Hospital Anxiety and Depression Scale (HADS) at Weeks 4, 8, 12, 24, and 48.[g] Change from baseline in the anxiety subscale score of the HADS at Weeks 4, 8, 12, 24, and 48.[g] Improvement on HADS among subjects with a baseline subscale score indicative of depression who achieved a “normal’ subscale score indicative of an absence of depression at Weeks 4, 8, 12, 24, and 48.[g] Improvement on HADS among subjects with a baseline subscale score indicative of anxiety who achieved a “normal’ subscale score indicative of an absence of anxiety at Weeks 4, 8, 12, 24, and 48.[g] Change from baseline in 36-Item Short Form Health Survey version 2 Acute (SF36v2 Acute) at Weeks 4, 8, 12, 24, and 48.
To evaluate the effect of PF-06651600 on the clinician global impression of severity of scalp hair loss.	Change from baseline in Clinician Global Impression – Alopecia Areata (CGI-AA) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
To evaluate efficacy of PF-06651600 in AA nail disease over time.	Change from baseline in fingernails affected by AA at Weeks 12, 24, 34, 40, and 48.
To assess pharmacodynamic and disease-related biomarkers over time.	Change from baseline in interferon gamma-induced protein 10 (IP-10) at Weeks 4, 8, 12, and 24. Change from baseline in percent and absolute lymphocyte subsets (T-cell, B-cell, and natural killer [NK] cells) at Weeks 4, 12, 24, and 48. Change from baseline in immunoglobulins (IgA, IgG, IgM) at Weeks 24 and 48.

a. The VHP countries participating in this study are Czech Republic, Germany, Hungary, Poland, and Spain.

b. For the EMA and competent authorities in the VHP countries, when SALT ≤10 is utilized for the primary endpoint, exposure response will be analyzed utilizing SALT ≤10 in a separate analysis.

c. For the EMA and competent authorities in the VHP countries, when SALT ≤10 is utilized for the primary endpoint, SALT ≤20 at Week 24 will be analyzed as a secondary endpoint.

d. When SALT ≤20 response at Week 24 is utilized as the primary endpoint, responses based on absolute SALT score ≤20 at Weeks 18, 12, 8 and 4 will be analyzed controlling for Type I error utilizing an appropriate testing procedure. When SALT ≤10 response at Week 24 is utilized as the primary endpoint, the response based on absolute SALT score ≤20 at Week 24 will be analyzed controlling for Type I error utilizing an appropriate testing procedure.

e. When SALT ≤20 response at Week 24 is utilized as the primary endpoint, the response based on absolute SALT score ≤10 at Week 24 will be analyzed controlling for Type I error utilizing an appropriate testing procedure. When SALT ≤10 response at Week 24 is utilized as the primary endpoint, the responses based on absolute SALT score ≤10 at Weeks 18, 12, 8 and 4 will be analyzed controlling for Type I error utilizing an appropriate testing procedure.

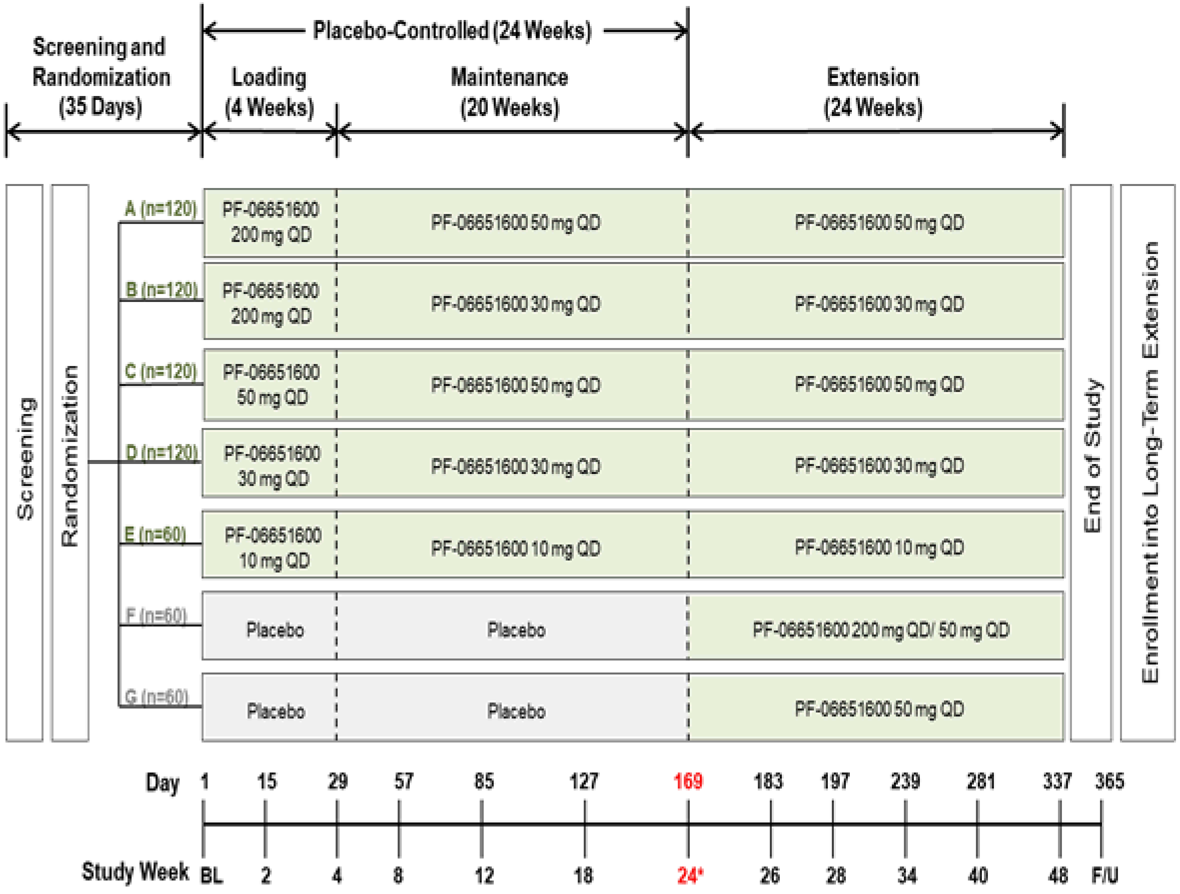
f. For the EMA and competent authorities in the VHP countries, the evaluation of the effect of PF-06651600 on patient centered outcomes (as measured by PGI-C response at Week 24) is a key secondary objective. The PGI-C response at Week 24 will be analyzed as a key secondary endpoint utilizing an appropriate testing procedure to control overall Type I error.

g. For the EMA and competent authorities in the VHP countries, endpoints utilizing the HADS will be analyzed as secondary endpoints.

2.2. Study Design

This is a Phase 2b/3, randomized, double-blind, placebo-controlled, dose-ranging study to investigate PF-06651600 in subjects with AA. The study will have a maximum duration of approximately 57 weeks. This includes an up to 5-week Screening period, a 48-week treatment period, and a 4-week follow-up period as shown below. The treatment period will be comprised of a placebo-controlled period that includes a 4-week loading phase and a 20-week maintenance phase, followed by a 24-week extension phase. The study will enroll a total of approximately 660 subjects. The study will be conducted at approximately 120 sites.

Figure 1. Study Design Schematic



* Primary endpoint is response based on absolute Severity of Alopecia Tool (SALT) Score ≤20 at Week 24.

To be eligible to enroll in this study, subjects must have moderate to severe AA with ≥50% hair loss of the scalp (Severity of Alopecia Tool [SALT] score ≥50) at both Screening and baseline visits, without evidence of terminal hair regrowth within the previous 6 months, and with the current episode of hair loss ≤10 years.

Screening will occur within 35 days prior to the first dose of study drug to confirm that subjects meet selection criteria for the study. Eligible subjects will be randomized as described below.

A stratified randomization will be used for operational purposes in order to achieve a target global composition for AT/AU and adolescent subjects in the enrolled population. The targets for enrollment are approximately 40% AT/AU and approximately 15% adolescents. The randomization will be operationalized as follows. In regions enrolling both adolescents and adults, there will be four strata:

- <18 years of age and AT/AU;
- <18 years of age and not AT/AU;

- ≥ 18 years of age and AT/AU; and
- ≥ 18 years of age and not AT/AU.

Within each of these strata, subjects will be randomized in a 2:2:2:2:1:1:1 manner to blinded PF-06651600 and matching placebo for a total of 7 treatment sequences. In regions enrolling only adults, there will be two strata:

- ≥ 18 years of age and AT/AU; and
- ≥ 18 years of age and not AT/AU.

In these regions, subjects will be randomized using the same ratio as described for regions enrolling both adolescents and adults.

All subjects will begin dosing during the loading phase according to their assigned sequence. Following the 4-week loading phase, subjects will continue dosing according to their assigned sequence in the 20-week maintenance phase. At the end of the maintenance phase, placebo-treated subjects will be advanced in a prespecified, blinded manner to one of 2 active treatment sequences for the remainder of the study (through Week 48). Investigators, subjects, and the sponsor study team will be blinded to treatment throughout the duration of the study. Following the last dose of study drug, both discontinued and completed subjects will enter a 4-week follow-up period for safety monitoring. Subjects who complete treatment may be eligible for enrollment in a long-term study (Phase 3 study B7981032). Subjects who enroll immediately into B7981032 study will not be required to complete the 4-week follow-up period in this study.

The study may be unblinded for internal decision-making purposes when ≥90% of subjects have reached 24 weeks post-randomization (or discontinued prior to the Week 24 visit). The sponsor personnel who are unblinded will be separate from the study team. The sponsor will still maintain the blind for study team members who will be involved in daily study conduct, study management, safety, and data monitoring through the completion of the study. Investigators and subjects will remain blinded.

Early unblinding for PK/PD modeling purposes is planned when ≥90% of subjects have reached 24 weeks post-randomization (or discontinued prior to the Week 24 visit). There will be no impact on B7981015 study conduct, analysis, or reporting based on the early unblinding of the PK/PD data. The results of these analyses will be reported in a separate pharmacometrics analysis report.

2.2.1. Sample Size Determination

The sample size for the study was based on the consideration to have sufficient power to evaluate the primary endpoint. A sample size of 120 subjects per group (for the 200 mg/50 mg once daily (QD), 200 mg/30 mg QD, 50 mg/50 mg QD, or 30 mg/30 mg QD groups) will provide more than 90% power to demonstrate that at least the 200 mg/50 mg QD group is superior to placebo by a difference of 24% in the proportion of subjects achieving the primary endpoint (SALT score ≤20 at Week 24), assuming a placebo response rate of no more than 5%, at α = 0.05 (2-sided significance level). This sample size accounts for multiplicity using a closed testing procedure to ensure strong control of Type I error for all comparisons between active treatment groups and placebo. This sample size additionally provides > 90% power for the SALT ≤10 at Week 24 endpoint assuming that the 200 mg/50 mg QD group is superior to placebo by a difference of 20% in the proportion of subjects achieving SALT ≤10, assuming a placebo response rate of no more than 5%, at α = 0.05 (2-sided significance level). The assumption of the placebo response rate for both SALT ≤20 and SALT ≤10, as well as of the treatment difference, were informed by the Week 24 results from the Phase 2a Study B7931005.

Regulatory requirements for a marketing authorization approval CCI will require significance at an α more stringent than 0.05, although the exact level of stringency required may vary (ie, 0.00125 is required by the FDA/PMDA for their requested primary endpoint of SALT ≤20 while 0.01 is required by the EMA for their requested primary endpoint of SALT ≤10). The sample size of 120 subjects per group provides >90% power for responses based on SALT ≤20 (α = 0.00125) and SALT ≤10 (α = 0.01).

The 10 mg/10 mg QD group is included to address the secondary objective of characterizing the exposure response. The sample size of 60 subjects was chosen to allow for estimation of the exposure response parameters.

For the EMA and competent authorities in the VHP countries that request that the PGI-C response at Week 24 be analyzed as a key secondary endpoint, a sample size of 120 subjects per group also provides more than 90% power for PGI-C response assuming a difference of 35% and a placebo response rate of 20%, at α = 0.05 (2-sided significance level). The assumptions of the treatment difference and placebo response rate for PGI-C response were based on the Concert[10] CTP-543 Phase 2b Study Week 24 data. Under the above assumptions, the power of the study remains >90% at α = 0.01 (EMA required 2-sided significance level).

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary and Secondary Endpoints

Section 2.1 lists the overall study objectives and endpoints. The primary, key secondary and secondary endpoints for the purpose of overall study reporting and publications are also detailed in Table 3. Table 4 lists the primary and secondary endpoints of the study organized by known regional requirements.

Table 3. Primary and Secondary Endpoints for Overall Study Reporting and Publications

Alpha (α)	0.05 (2-sided significance level)
Primary Endpoint	Response based on an absolute SALT Score ≤20 at Week 24.
Key Secondary Endpoint	Response based on an absolute SALT Score ≤10 at Week 24.
Secondary Endpoints	• Response based on an absolute SALT Score ≤20 at Week 24 will be used to characterize the exposure response.

Table 3. Primary and Secondary Endpoints for Overall Study Reporting and Publications

	- Response based on an absolute SALT score ≤20 at Weeks 4, 8, 12, 18, 28, 34, 40, and 48. NOTE: Response at Weeks 4, 8, 12, and 18 will be analyzed controlling for Type I error. - Response based on an absolute SALT score of ≤10 at Weeks 4, 8, 12, 18, 28, 34, 40, and 48. - Response based on a 75% improvement in SALT score from baseline (SALT75) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. - Change from baseline in SALT scores at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. - Response based on at least a 2-grade improvement or a score of 3 in EBA at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. - Response based on at least a 2-grade improvement or a score of 3 in ELA at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. - PGI-C response defined as a PGI-C score of "moderately improved" or "greatly improved" at Weeks 4, 8, 12, 18, 24, 34, 40, and 48. - Change from baseline in AAPPO scales at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.

Table 4. Primary and Secondary Endpoints Organized by Known Regional Regulatory Requirements

	For the FDA/PMDA	For the EMA and competent authorities in the VHP countries
Alpha (α) CCI [redacted]	0.00125 (2-sided significance level)	0.01 (2-sided significance level)
Primary Endpoint	Response based on an absolute SALT Score ≤20 at Week 24.	Response based on an absolute SALT Score ≤10 at Week 24.
Key Secondary Endpoint	Not applicable	PGI-C response defined as a PGI-C score of "moderately improved" or "greatly improved" at Week 24.
Secondary Endpoints	- Response based on an absolute SALT Score ≤20 at Week 24 will be used to characterize the exposure response. - Response based on an absolute SALT score ≤20 at Weeks 4, 8, 12, 18, 28, 34, 40, and 48. NOTE: Response at Weeks 4, 8, 12, and 18 will be analyzed controlling for Type I error. - Response based on an absolute SALT score of ≤10 at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. NOTE: Response at Week 24 will be analyzed controlling for Type I error. - Response based on a 75% improvement in SALT score from baseline (SALT75) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.	- Response based on an absolute SALT Score ≤10 at Week 24 will be used to characterize the exposure response. - Response based on an absolute SALT score ≤20 at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48. NOTE: Response at Week 24 will be analyzed controlling for Type I error. - Response based on an absolute SALT score of ≤10 at Weeks 4, 8, 12, 18, 28, 34, 40, and 48. NOTE: Response at Weeks 4, 8, 12, and 18 will be analyzed controlling for Type I error. - Response based on a 75% improvement in SALT score from baseline (SALT75) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.

Table 4. Primary and Secondary Endpoints Organized by Known Regional Regulatory Requirements

	For the FDA/PMDA	**For the EMA and competent authorities in the VHP countries**
	• Change from baseline in SALT scores at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.	• Change from baseline in SALT scores at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
	• Response based on at least a 2-grade improvement or a score of 3 in EBA score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.	• Response based on at least a 2-grade improvement or a score of 3 in EBA score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
	• Response based on at least a 2-grade improvement or a score of 3 in ELA score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.	• Response based on at least a 2-grade improvement or a score of 3 in ELA score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
	• PGI-C response defined as a PGI-C score of "moderately improved" or "greatly improved" at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.	• PGI-C response defined as a PGI-C score of "moderately improved" or "greatly improved" at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
	• Change from baseline in AAPPO scales at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.	• Change from baseline in AAPPO scales at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
	-	• Change from baseline in the depression subscale score of the HADS at Weeks 4, 8, 12, 24, and 48.
	-	• Change from baseline in the anxiety subscale score of the HADS at Weeks 4, 8, 12, 24, and 48.
	-	• Improvement on HADS among subjects with a baseline subscale score indicative of depression who achieved a "normal' subscale score indicative of an absence of depression at Weeks 4, 8, 12, 24, and 48.
	-	• Improvement on HADS among subjects with a baseline subscale score indicative of anxiety who achieved a "normal' subscale score indicative of an absence of anxiety at Weeks 4, 8, 12, 24, and 48.

3.2. Other Endpoints

3.2.1. Exploratory Endpoints

- Collection of banked biospecimens unless prohibited by local regulations or ethics committee decision. Additional information on collection and potential use is provided in the Banked Biospecimens section in the protocol.

- Response based on a 50% improvement in SALT score from baseline (SALT50) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.

- Absolute SALT scores at Baseline, Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Improvement on PGI-C defined as "slightly improved", "moderately improved", or "greatly improved" at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
- Improvement on Patient Satisfaction with Hair Growth (P-Sat) items defined as slightly, moderately, or very satisfied at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
  - Will be analyzed for each individual item.
- Change from baseline in EuroQoL 5 Dimensions (EQ-5D-5L) in adults or EuroQoL 5 Dimensions-Youth (EQ-5D-Y) in adolescents at Weeks 4, 12, 24, and 48.
  - Will be analyzed for adults and adolescents separately. US specific utility weights[1] will be applied for the scoring of these endpoints;
- Change from baseline in Alopecia Areata Resource Utilization (AARU) at Weeks 12, 24, 34, and 48.
  - Proportion of subjects with any healthcare professional (HCP) visits.
  - Among those with HCP visits, mean total number of visits for any reason, mean change from baseline in total number of visits for any reason, mean total number of visits related to AA and mean change from baseline in AA-related visits.
- Change from baseline in Work Productivity and Activity Impairment: Alopecia Areata (WPAI:AA) at Weeks 12, 24, 34, and 48.
  - Change from baseline in absenteeism, presenteesism, work productivity loss (overall) and activity impairment in adults.
- Change from baseline in the depression subscale score of the Hospital Anxiety and Depression Scale (HADS) at Weeks 4, 8, 12, 24, and 48.
  - For the EMA and competent authorities in the VHP countries this endpoint will be analyzed as a secondary endpoint.
- Change from baseline in the anxiety subscale score of the Hospital Anxiety and Depression Scale (HADS) at Weeks 4, 8, 12, 24, and 48.
  - For the EMA and competent authorities in the VHP countries this endpoint will be analyzed as a secondary endpoint.
- Improvement on HADS among subjects with a baseline subscale score indicative of depression who achieved a "normal" subscale score indicative of an absence of depression at Weeks 4, 8, 12, 24, and 48.

- For the EMA and competent authorities in the VHP countries this endpoint will be analyzed as a secondary endpoint.
- Among adults, a score of 0-7 is considered normal, a score of >7 is indicative of depression. Among adolescents (at baseline age), a score of 0-6 is considered normal, a score of >6 is indicative of depression;[2,3]

- Improvement on HADS among subjects with a baseline subscale score indicative of anxiety who achieved a "normal" subscale score indicative of an absence of anxiety at Weeks 4, 8, 12, 24, and 48.
  - For the EMA and competent authorities in the VHP countries this endpoint will be analyzed as a secondary endpoint.
  - Among adults, a score of 0-7 is considered normal, a score of >7 is indicative of anxiety. Among adolescents (at baseline age), a score of 0-8 is considered normal, a score of >8 is indicative of anxiety.[2,3]
- Change from baseline in 36-Item Short Form Health Survey version 2 Acute (SF36v2 Acute) at Weeks 4, 8, 12, 24, and 48.
  - 8 domain scales: physical function (PF), bodily pain (BP), role physical (RP), role emotional (RE), Vitality (VT), general health (GH), social function (SF), and mental health (MH).
  - 2 summary scales: Mental Component Summary Score (MCS) and Physical Component Summary Score (PCS).
- Change from baseline in Clinician Global Impression – Alopecia Areata (CGI-AA) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
  - CGI-AA scored as 0='None (no hair loss)', 1='Minimal hair loss', 2-4= 'Moderate-very severe or complete hair loss'.
  - Improvement among participants with a baseline score 2-4 indicating Moderate-very severe or complete hair loss who achieved a score of 0='None (no hair loss)' or 1='Minimal hair loss'.
- Change from baseline in fingernails affected by AA at Weeks 12, 24, 34, 40, and 48.
- Change from baseline in interferon gamma-induced protein 10 (IP-10) at Weeks 4, 8, 12, and 24.
- Change from baseline in percent and absolute lymphocyte subsets (T-cell, B-cell, and natural killer [NK] cells) at Weeks 4, 12, 24, and 48.
- Change from baseline in immunoglobulins (IgA, IgG, IgM) at Weeks 24 and 48.

3.2.2. PK Endpoint

- Plasma concentrations of PF-06651600 at Weeks 4 and 8 (or Week 12).

3.3. Baseline Variables

Baseline is defined as pre-dose on Day 1. Data from the screening period may be used if Day 1 data are missing. If multiple data points are available, we will use the observation closest but no later than Day 1.

For the analysis of ECG data, subject's baseline values will be the ECG assessment at the Screening visit. Data from the Day 1 may be used if screening data are missing.

3.4. Safety Endpoints

The safety endpoints are:

- Incidence of treatment-emergent adverse events (AEs).
- Incidence of serious AEs (SAEs) and AEs leading to discontinuation.
- The incidence of clinically significant abnormalities in vital signs.
- The incidence of clinically significant abnormalities in clinical laboratory values.

3.4.1. Adverse Events

An AE is considered a treatment emergent adverse event (TEAE) if the event starts on or after the first dosing day.

3.4.2. Laboratory Data

See Table 3 in the protocol for the list of clinical laboratory tests to be performed. Laboratory data Criteria for discontinuation and abnormalities are specified in Section 9.3.

3.4.3. Vital Signs, Including Height and Weight

Vital sign measurements are pulse rate, blood pressure.

Height and weight are also collected pre- and post-treatment.

3.4.4. Physical Examinations

Complete physical examinations consist of assessments of general appearance; skin (dermatological full body exam); head, eyes, ears, nose, and throat (HEENT); mouth, heart; lungs; abdomen; extremities; neurologic function, back, and lymph nodes. Targeted physical examinations should include skin, heart, lung, and abdomen, neurologic function, and examination of body systems where there are symptom complaints by the subject.

3.4.5. Electrocardiogram (ECG)

ECG measurements will be collected at Screening, Day 1, Week 4, Week 24, Week 28, and Week 48.

4. ANALYSIS SETS

Data for all subjects will be assessed to determine if subjects meet the criteria for inclusion in each analysis population described below prior to unblinding and releasing the database and classifications will be documented per standard operating procedures.

4.1. Full Analysis Set

The Full Analysis Set (FAS) is defined as all subjects who have been randomized, regardless of whether they received study medication. Subjects will be analyzed in the treatment groups as they are randomized. If a subject was randomized but received the incorrect treatment, then the subject will be reported under their randomized treatment group for all efficacy analyses. If a subject is treated but not randomized, then the subject will be excluded from the FAS.

4.2. Per Protocol Analysis Set

The Per Protocol Analysis Set (PPAS) is defined as all randomized subjects who do not have any major protocol deviation related to inclusion/exclusion criteria, compliance with investigational product, or any other major protocol deviations that might affect the efficacy data through Week 24. The following deviations will be evaluated for each subject prior to unblinding to determine if a subject should be excluded from the PPAS:

1. Subject did not meet the following inclusion criteria for hair loss due to AA at Screening or Day 1:
    a. Clinical diagnosis of AA with no other known etiology of hair loss, including no known androgenetic alopecia;
    b. ≥50% scalp hair loss;
    c. Current episode of hair loss ≤10 years duration at time of Screening;
    d. No evidence of terminal scalp hair regrowth within 6 months at both the screening and baseline visits.
2. Subject met one of the following exclusion criteria regarding medical history diagnoses:
    a. Another scalp disease which could impact hair loss assessments;
    b. Active systemic disease which can cause hair loss.
3. Subjects who were randomized but never received study drug.
4. Prior to the date of the Week 24 visit, subject had a dosing interruption of ≥6 weeks for any reason.
5. Prior to the date of the Week 24 visit, subject used a prior or concomitant prohibited medication that could potentially impact efficacy data at Week 24.

6. Week 24 SALT assessment was impacted in one of the following ways:
    a. Visit/assessment was not performed;
    b. Assessment was not performed properly (including subject had a shaved head).
7. Subjects with other major protocol deviations not listed above that, in the opinion of the sponsor study team, could potentially impact efficacy data at Week 24.

The subjects excluded from this analysis set will be determined and documented before the study is unblinded.

4.3. Safety Analysis Set

The Safety Analysis Set (SAS) will be defined as all subjects who receive at least one dose of study medication. The safety analysis set is the primary population for treatment administration/compliance and safety. A randomized but not treated subject will be excluded from the safety analyses. If a subject was randomized but received partially incorrect treatment, then the subject will be summarized under the treatment received for most of the time during the study, for all safety analyses.

5. GENERAL METHODOLOGY AND CONVENTIONS

The final analysis and reporting of results will be performed after the completion of the study and after the database is locked.

5.1. Hypotheses and Decision Rules

This section describes the testing procedure for multiple comparisons. This CCI will support submission in several regions with different requirements; therefore, this section is organized into 3 sections: Section 5.1.1 describes the testing procedure for the overall study, Section 5.1.2, for the FDA/PMDA, and Section 5.1.3 for the EMA and competent authorities in the VHP countries.

The study is tested at an overall α = 0.05. While the 5% significance level is deemed sufficient for a declaration of effect, more stringent levels have been advised for regulatory submission of B7981015 CCI, ie, 1% for EMA and competent authorities in the VHP countries, and 0.125% for the FDA/PMDA.

The hypotheses to be tested are that each of the active treatment groups (200 mg QD/50 mg QD, 50 mg QD/50 mg QD, 200 mg QD/30 mg QD, and 30 mg QD/30 mg QD) is superior to placebo as measured by the proportion of subjects achieving the primary endpoint at Week 24 and key secondary endpoints where applicable.

The grouping of the dose regimens in the test hierarchy was based on the hierarchy of 2 different dose regimens: in the dosing regimens that have a loading dose, the order is 200 mg QD/50 mg QD → 200 mg QD/30 mg QD; in the regimens without a loading dose, the order is 50 mg QD/50 mg QD → 30 mg QD/30 mg QD. The testing strategy proposed will allow simultaneous testing of the 200 mg QD/30 mg QD and the 50 mg QD/50 mg QD dose regimens against placebo, given that it is not known which dose regimen will have a better response.

5.1.1. Overall Study (Testing Procedure for Multiple Comparisons Incorporating the SALT ≤10 as a Key Secondary Endpoint)

For the overall study the SALT ≤20 response at Week 24 is the primary endpoint and the SALT ≤10 response at Week 24 will be analyzed as a key secondary endpoint. The hypotheses to be tested are that each of the active treatment groups is superior to placebo as measured by the proportion of subjects achieving the primary endpoint and the key secondary endpoint. There are a total of 8 hypotheses to be tested:

- Hypothesis 1 (H1) is to test whether the 200 mg QD/50 mg QD dose regimen is superior to placebo for the primary endpoint;
- Hypothesis 2 (H2) is to test whether the 50 mg QD dose regimen is superior to placebo for the primary endpoint;
- Hypothesis 3 (H3) is to test whether the 200 mg QD/30 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 4 (H4) is to test whether the 30 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 1p (H1p) is to test whether the 200 mg QD/50 mg QD dose regimen is superior to placebo for the key secondary endpoint;

- Hypothesis 2p (H2p) is to test whether the 50 mg QD dose regimen is superior to placebo for the key secondary endpoint;

- Hypothesis 3p (H3p) is to test whether the 200 mg QD/30 mg QD dose regimen is superior to placebo for the key secondary endpoint;

- Hypothesis 4p (H4p) is to test whether the 30 mg QD dose regimen is superior to placebo for the key secondary endpoint.

The family-wise Type I error will be strongly controlled using a gatekeeping approach as described in the following order and shown in Figure 2. The 10 mg group comparison to placebo is not included in the Type I error controlled procedure, as this group is only included to support the estimation of the exposure response.

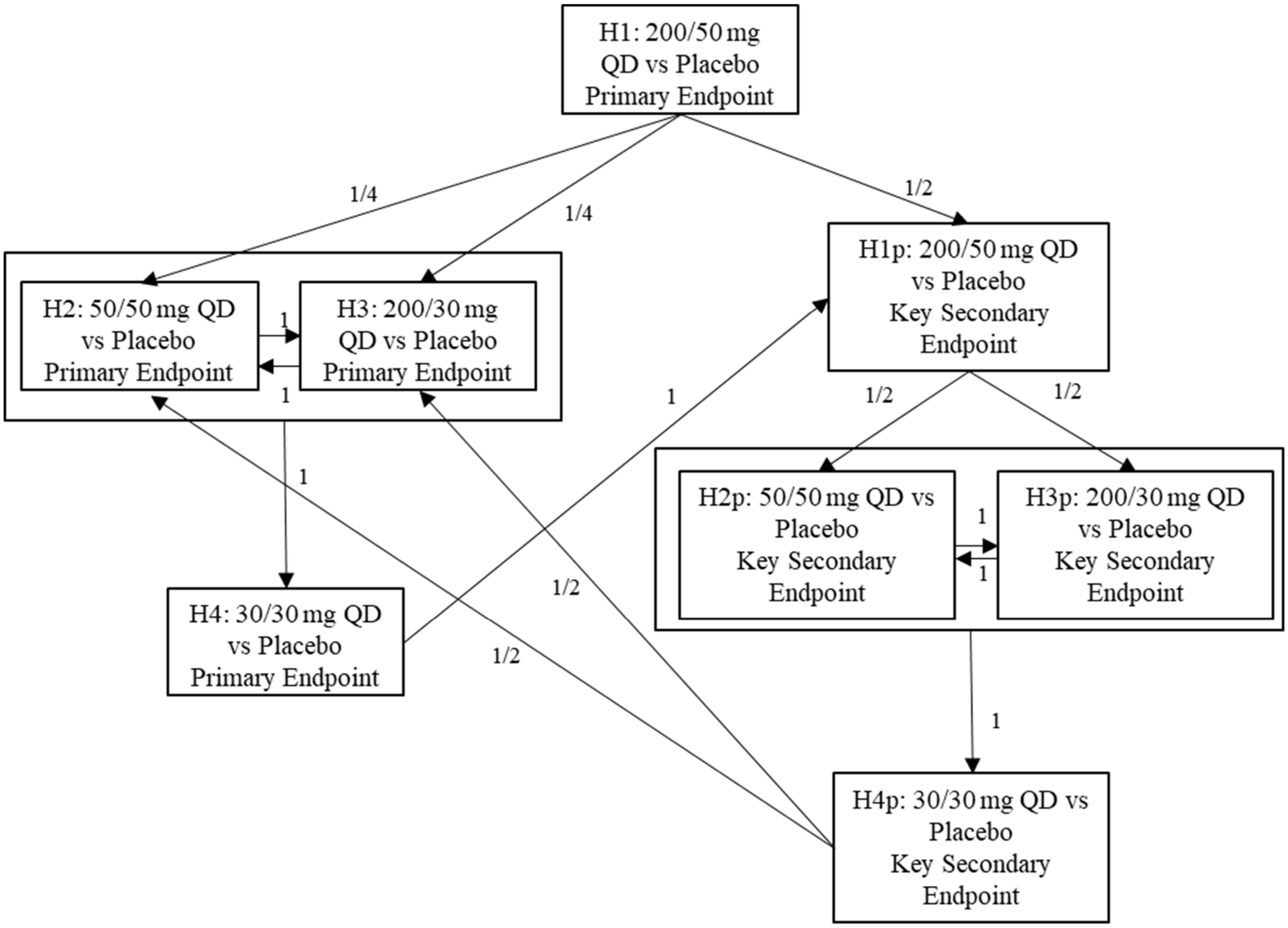
- The primary endpoint for 200 mg QD/50 mg QD dose regimen (H1) will be tested at α significance level first. If significant, then the primary endpoint for 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen (H2 and H3) and the key secondary endpoint for 200 mg QD/50 mg QD dose regimen (H1p) will be tested simultaneously. If both the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen are significant, then the 30 mg QD dose regimen will be tested.

- For the primary endpoint, if both the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen are significant at $\alpha/4$ level, or if one of the dose regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at $\alpha/4$ level and the other is significant at $\alpha/2$ level, both dose regimens will be declared as significant. In this case, the 30 mg QD dose regimen will be tested at $\alpha/2$ level. If only one of the regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at $\alpha/4$ level and the other is not significant at $\alpha/2$ level, only the one significant at $\alpha/4$ level will be declared as statistically significant. In this case, the testing for the 30 mg QD dose regimen will not proceed.

- For the key secondary endpoint, the 200 mg QD/50 mg QD dose regimen will be tested at $\alpha/2$ level first. If significant, then the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen will be tested simultaneously. If both the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen are significant at $\alpha/4$ level, or if one of the dose regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at $\alpha/4$ level and the other is significant at $\alpha/2$ level, both dose regimens will be declared as significant. In this case, the 30 mg QD dose regimen will be tested at $\alpha/2$ level. If only one of the regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at α/4 level and the other is not significant at α/2 level, only the one significant at α/4 level will be declared as statistically significant. In this case, the resting for the 30 mg QD dose regimen will not proceed.

- If the primary endpoint for the 200 mg QD/50 mg QD dose regimen is significant at α level, and all the remaining hypotheses for the primary endpoint and key secondary endpoint are significant at the overall α/2 level, then all 8 hypotheses will be declared as statistically significant.
- If the primary endpoint for all 4 hypotheses or the primary endpoint for H1 is significant and the key secondary endpoint is significant in all 4 hypotheses, then the α/2 level will be passed to the hypothesis testing for the other hypotheses. In this case, the other hypotheses can be tested at the α level instead of the α/2 level following the same procedures for claiming statistical significance outlined in the bullets above with the exception of the increased α.

Figure 2. Schematic and Graphical Presentation for Multiple Testing Procedure Incorporating a Key Secondary Endpoint


H1: 200/50 mg QD vs Placebo Primary Endpoint
1/4
1/4
1/2
H2: 50/50 mg QD vs Placebo Primary Endpoint
1
1
H3: 200/30 mg QD vs Placebo Primary Endpoint
H1p: 200/50 mg QD vs Placebo Key Secondary Endpoint
1
1/2
1/2
1
H2p: 50/50 mg QD vs Placebo Key Secondary Endpoint
1
1
H3p: 200/30 mg QD vs Placebo Key Secondary Endpoint
H4: 30/30 mg QD vs Placebo Primary Endpoint
1/2
1/2
1
H4p: 30/30 mg QD vs Placebo Key Secondary Endpoint

The numbers on each edge indicate the transition rule for the weights on α allocation once a hypothesis is rejected.

The following approach will be used to control the Type I error for SALT ≤20 at earlier time points:

- For a given dose, if the primary endpoint of SALT ≤20 is declared to be statistically significant based on the testing procedure for the primary endpoint, in order to establish the onset of efficacy as measured by SALT ≤20 at the earliest time point, a step-down approach with SALT ≤20 from Week 24 to earlier time points (order of testing: Weeks 24, 18, 12, 8, and 4) will be used for each time point. Although this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the family of SALT ≤20 time points within the same dose group.

5.1.2. For the FDA/PMDA (Testing Procedure for Multiple Comparisons with no Key Secondary Endpoints)

For the FDA/PMDA, the SALT ≤20 response at Week 24 is the primary endpoint and no secondary endpoints will be analyzed as key secondary. The hypotheses to be tested at $\alpha = 0.00125$ are that each of the active treatment groups is superior to placebo as measured by the proportion of subjects achieving the primary endpoint. There are a total of 4 hypotheses to be tested.

- Hypothesis 1 (H1) is to test whether the 200 mg QD/50 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 2 (H2) is to test whether the 50 mg QD/50 mg QD dose regimen is superior to placebo for the primary endpoint;

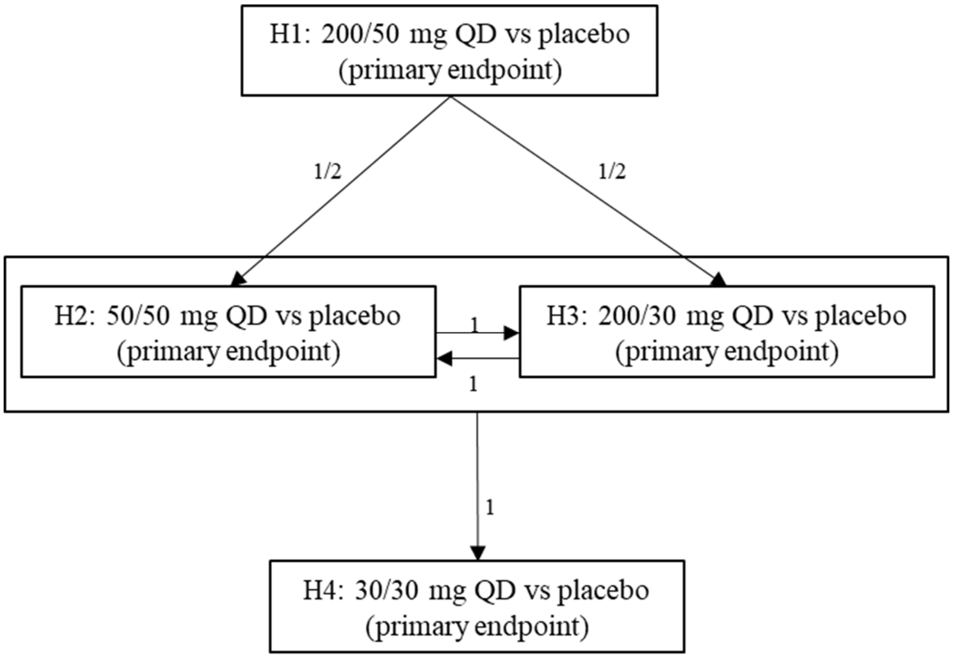
- Hypothesis 3 (H3) is to test whether the 200 mg QD/30 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 4 (H4) is to test whether the 30 mg QD/30 mg QD dose regimen is superior to placebo for the primary endpoint.

The family-wise Type I error will be strongly controlled using a gate-keeping approach to test these hypotheses in 3 groups as shown in Figure 3. The first group includes H1; the second group includes H2 and H3; and the third group includes H4. The statistical significance of the first hypothesis (H1) will allow simultaneous testing of H2 and H3 using Holm’s method. Statistical significance for both H2 and H3 will enable testing the statistical significance for the fourth hypothesis (H4). The 10 mg group comparison to placebo is not included in the Type I error controlled procedure, as this group is only included to support the estimation of the exposure response.

Figure 3 shows a schematic of the procedure for testing each of the active treatment groups versus the placebo group for the primary endpoint.

Figure 3. Schematic and Graphical Presentation for Multiple Testing Procedure


H1: 200/50 mg QD vs placebo
(primary endpoint)
1/2
1/2
H2: 50/50 mg QD vs placebo
(primary endpoint)
1
1
H3: 200/30 mg QD vs placebo
(primary endpoint)
1
H4: 30/30 mg QD vs placebo
(primary endpoint)

The numbers on each arrow indicate the transition rule for the weights on α allocation once a hypothesis is rejected.

In the simultaneous testing of H2 and H3, if both the 50 mg QD/50 mg QD and 200 mg QD/30 mg QD regimens are not significant at α/2 level, the test for statistical significance is stopped. If one of the hypotheses for 50 mg QD/50 mg QD and 200 mg QD/30 mg QD regimens is significant at α/2 level and the other hypothesis is significant at α level, this family hypothesis is significant. If one of the hypotheses for 50 mg QD/50 mg QD and 200 mg QD/30 mg QD is significant at α/2 level and the other hypothesis is not significant at α level, only the hypothesis significant at α/2 level will be declared as statistically significant. In this case, the testing for statistical significance for H4 will not proceed.

The following approach will be used to control the Type I error for testing SALT ≤10 at Week 24:

- For a given dose, if SALT ≤20 at Week 24 is declared to be statistically significant based on the testing procedure for the primary endpoint(s), then SALT ≤10 at Week 24 will be tested at the same significance level as the SALT ≤20 at Week 24 was tested. Though this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the endpoints within the same dose group.

The following approach will be used to control the Type I error for testing SALT ≤20 at earlier time points:

[redacted]

- For a given dose, if the primary endpoint of SALT ≤20 at Week 24 is declared to be statistically significant based on the testing procedure for the primary endpoint, in order to establish the onset of efficacy as measured by SALT ≤20 at the earliest time point, a step-down approach with SALT ≤20 from Week 24 to earlier time points (order of testing: Weeks 24, 18, 12, 8, and 4) will be used for each time point. Though this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the family of SALT ≤20 time points within the same dose group.

5.1.3. For the EMA and Competent Authorities in the VHP Countries (Testing Procedure for Multiple Comparisons Incorporating the PGI-C Response as a Key Secondary Endpoint)

For the EMA and competent authorities in the VHP countries, the SALT ≤10 response at Week 24 will be the primary endpoint and the PGI-C response at Week 24 will be analyzed as a key secondary endpoint. The hypotheses to be tested at $\alpha = 0.01$ are that each of the active treatment groups is superior to placebo as measured by the proportion of subjects achieving the primary endpoint and the key secondary endpoint. There are a total of 8 hypotheses to be tested:

- Hypothesis 1 (H1) is to test whether the 200 mg QD/50 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 2 (H2) is to test whether the 50 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 3 (H3) is to test whether the 200 mg QD/30 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 4 (H4) is to test whether the 30 mg QD dose regimen is superior to placebo for the primary endpoint;

- Hypothesis 1p (H1p) is to test whether the 200 mg QD/50 mg QD dose regimen is superior to placebo for the key secondary endpoint;

- Hypothesis 2p (H2p) is to test whether the 50 mg QD dose regimen is superior to placebo for the key secondary endpoint;

- Hypothesis 3p (H3p) is to test whether the 200 mg QD/30 mg QD dose regimen is superior to placebo for the key secondary endpoint;

- Hypothesis 4p (H4p) is to test whether the 30 mg QD dose regimen is superior to placebo for the key secondary endpoint.

The family-wise Type I error will be strongly controlled using a gate-keeping approach as described in the following order and shown in Figure 2. The 10 mg group comparison to placebo is not included in the Type I error controlled procedure, as this group was only included in the study design to support the estimation of the exposure response.

- The primary endpoint for 200 mg QD/50 mg QD dose regimen (H1) will be tested at α significance level first. If significant, then the primary endpoint for 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen (H2 and H3) and the key secondary endpoint for 200 mg QD/50 mg QD dose regimen (H1p) will be tested simultaneously. If both the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen are significant, then the 30 mg QD dose regimen will be tested.

- For the primary endpoint, if both the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen are significant at α/4 level, or if one of the dose regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at α/4 level and the other is significant at α/2 level, both dose regimens will be declared as significant. In this case, the 30 mg QD dose regimen will be tested at α/2 level. If only one of the regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at α/4 level and the other is not significant at α/2 level, only the one significant at α/4 level will be declared as statistically significant. In this case, the testing for the 30 mg QD dose regimen will not proceed.

- For the key secondary endpoint, the 200 mg QD/50 mg QD dose regimen will be tested at α/2 level first. If significant, then the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen will be tested simultaneously. If both the 50 mg QD dose regimen and the 200 mg QD/30 mg QD dose regimen are significant at α/4 level, or if one of the dose regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at α/4 level and the other is significant at α/2 level, both dose regimens will be declared as significant. In this case, the 30 mg QD dose regimen will be tested at α/2 level. If only one of the regimens (50 mg QD dose regimen or 200 mg QD/30 mg QD dose regimen) is significant at α/4 level and the other is not significant at α/2 level, only the one significant at α/4 level will be declared as statistically significant. In this case, the testing for the 30 mg QD dose regimen will not proceed.

- If the primary endpoint for the 200 mg QD/50 mg QD dose regimen is significant at α level, and all the remaining hypotheses for the primary endpoint and key secondary endpoint are significant at the overall α/2 level, then all 8 hypotheses will be declared as statistically significant.

- If the primary endpoint for all 4 hypotheses or the primary endpoint for H1 is significant and the key secondary endpoint is significant in all 4 hypothesis, then the α/2 level will be passed to the hypothesis testing for the other hypotheses. In this case, the other hypotheses can be tested at α level instead of the α/2 level following the same procedures for claiming statistical significance outlined in the bullets above with the exception of the increased α.

For the EMA and competent authorities in the VHP countries, the following approach will be used to control the Type I error for SALT ≤20 at Week 24:

- For a given dose, if SALT ≤10 at Week 24 is declared to be statistically significant based on the testing procedure for the primary endpoint incorporating the key secondary endpoint, then SALT ≤20 at Week 24 will be tested at the same significance level as SALT ≤10 at Week 24 was tested. Although this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the endpoints within the same dose group.

The following approach will be used to control the Type I error for SALT ≤10 at earlier timepoints:

- For a given dose, if the primary endpoint of SALT ≤10 is declared to be statistically significant based on the testing procedure for the primary endpoint incorporating the key secondary endpoint, in order to establish the onset of efficacy as measured by SALT ≤10 at the earliest time point, a step-down approach with SALT ≤10 from Week 24 to earlier time points (order of testing: Weeks 24, 18, 12, 8, and 4) will be used for each time point. Although this testing scheme does not protect the Type I error for the family of all possible comparisons, it will provide Type I error protection for testing the family of SALT ≤10 time points within the same dose group.

5.2. General Methods

In general, number and percent will be presented for binary variables. Number, mean, standard deviation (or standard error of the mean), median, minimum, and maximum will be presented for continuous variables.

For analyses up to Week 24, in the comparison to placebo, the data from both placebo groups (F and G) will be pooled to form a combined placebo group. For analyses up to Week 48, data will be summarized by 7 treatment sequences (Figure 1) without combining two placebo groups.

5.2.1. Binary Endpoints

Planned analysis for the overall study, for the FDA/PMDA and for the EMA and competent authorities in the VHP countries are summarized in Table 5.

Table 5. Statistical Methods of Planned Analysis for Primary and Key Secondary Endpoints for the Study and for each Regulatory Region

Study or Region	Endpoint/ designation	Analysis #	Analysis Designation	Statistical Method	Missing due to COVID-19	Missing due to other reasons
Overall Study	SALT ≤20 at Week 24/ primary	Analysis 1	Primary	MN	Exclude	non-responders
		Analysis 2	Supplementary	GLMM	MAR	MAR
		Analysis 2a	Supplementary	GLMM/TP	MNAR	MNAR
		Analysis 3	Supplementary	MN	non-responders	non-responders
	SALT ≤10 at Week 24/ key secondary	Analysis 1	Key Secondary	MN	exclude	non-responders
		Analysis 3	Supplementary	MN	non-responders	non-responders
FDA/ PMDA	SALT ≤20 at Week 24/ primary	Analysis 1	Primary	MN	exclude	non-responders
		Analysis 2	Supplementary	GLMM	MAR	MAR
		Analysis 2a	Supplementary	GLMM/TP	MNAR	MNAR
		Analysis 3	Supplementary	MN	non-responders	non-responders
EMA and competent authorities in VHP countries	SALT ≤10 at Week 24/ primary	Analysis 4	Primary	GLMM	MAR	non-responders
		Analysis 4a	Supplementary	GLMM/TP	MNAR	non-responders
		Analysis 1	Supplementary	MN	exclude	non-responders
		Analysis 3	Supplementary	MN	non-responders	non-responders
	PGI-C Response at Week 24/ key secondary	Analysis 4	Key Secondary	GLMM	MAR	non-responders

Abbreviations: GLMM = Generalized linear mixed model; MN = Miettinen and Nurminen; MAR = Missing at random; MNAR = Missing not at random; TP = Tipping point

5.2.1.1. Miettinen and Nurminen – Analysis 1 and Analysis 3

The Miettinen and Nurminen (MN) method will be used to calculate 95% confidence intervals (CI) and p-values for the difference in the proportion of responders between each active treatment group and placebo. Subjects with missing SALT score at Week 24 due to COVID-19 related reasons will be excluded from the analysis at that time point, whereas subjects with missing SALT scores due to other reasons will be counted as non-responders at that time point (Analysis 1).

As a supplementary analysis to assess the impact of COVID-19-related missing data on the results of the primary analysis, a similar analysis will be conducted with all missing data considered as non-responders regardless of the reason for missingness (Analysis 3).

5.2.1.2. Generalized Linear Mixed Model and Tipping Point – Analyses 2 and 2a

As a supplementary analysis, the primary endpoint for the overall study and for the FDA/PMDA will also be analyzed utilizing a generalized linear mixed model (GLMM) for longitudinal binary data of response based on SALT ≤ 20 over time up to Week 24. This binary outcome will be modeled using a logistic-normal distribution. Fixed factors in this model are treatment (6 levels), visit (5 levels) and treatment-by-visit interaction. Visit will be modeled as a categorical covariate. A subject-specific random intercept will be used to model the correlation within a subject over time. Due to the expected negligible placebo response, a Bayesian estimation will be carried out in which diffusive priors will be used and the placebo response probability will be bounded away from 0 to avoid the potential convergency issues. From this model, one can derive the estimates of the marginal probability of response for each treatment group at each visit as well as inferential comparisons (the proportional difference between active and placebo groups, CIs and p-values) between treatment groups after the random effect is integrated out from the posterior samples. Given that this random intercept model specifies the full joint distribution for the outcome, the inference for the marginal proportions will be valid in the presence of missing data when the missing mechanism is missing at random (MAR). This inference will be with respect to the difference in treatments assuming all subjects maintained their randomized therapy. This analysis will be referred to as Analysis 2.

A tipping point analysis will be conducted under the assumption of missing not at random (MNAR). The imputations will be implemented under the framework of the same logit normal GLMM as above. Imputation of missing SALT response is performed only at Week 24 based on the predictive distribution of the GLMM. Under MNAR assumption, a fixed MNAR quantity (favorable or unfavorable), delta, will be applied to the probability of response at Week 24 for subjects with missing SALT response in the PF 06651600 dose groups and placebo group independently to assess when the conclusion might change (ie, tipping). Imputations will be repeated to obtain completed datasets. The Rubin's rules[4,6] will be used to combine the results of multiple imputed samples for inference. This analysis will be referred to as Analysis 2a.

A scenario included in the tipping point analysis framework is an analysis under MAR, if there is no fixed quantity applied to the probability of response for each subject in the PF-06651600 dose groups or the placebo group, ie, when delta is zero.

Detailed descriptions of the tipping point analyses are provided in Section 9.6.

5.2.1.3. Generalized Linear Mixed Model and Tipping Point – Analyses 4 and 4a

The primary analysis of the primary endpoint for the EMA and competent authorities in the VHP countries will be based on a MAR assumption and use multiple imputation methods similar to those described in Section 5.2.1.2. In this analysis, while subjects with missing SALT scores due to COVID-19 will be assumed to be MAR, subjects with missing SALT scores due to other reasons will be treated as non-responders. This analysis will be referred to as Analysis 4.

A tipping point analysis will be conducted similar to that described in Section 5.2.1.2. While the analysis for the study and for the FDA/PMDA will treat missing as MNAR regardless of the reason for missingness, the analysis for the EMA will treat missing due to COVID-19 as MNAR but missing due to other reasons will be considered as non-responders. This analysis will be referred to as Analysis 4a.

The key secondary endpoint, PGI-C response at Week 24, will also be analyzed using this approach.

5.2.1.4. Descriptive Analysis of Binary Endpoints

All secondary binary endpoints in the placebo-controlled phase will be analyzed in the same way as Analysis 1 in Section 5.2.1.1.

Exploratory binary endpoints will be summarized using descriptive statistics unless specified otherwise.

Summaries of data from baseline to Week 48 (over the entire study) will include number of responders, percent, standard error, 95% CI based on normal approximation. These summaries will be based on observed cases, and data will be presented by 7 treatment sequences (ie, without combining the placebo groups F and G, Figure 1).

5.2.2. Exposure response analysis

A Bayesian three-parameter maximum effect attributable to the drug (E_{max}) exposure-response model will be used as the primary analysis approach to characterize the dose-response relationship. The response function will be the log odds (logit) of the proportion of subjects with response based on SALT ≤20 at Week 24 (or SALT ≤10 at Week 24, as applicable). In modeling the exposure response, the effect of loading dose will be included as a fixed factor in the model. Model-based estimation of the treatment effect for each dose compared to placebo will be presented.

5.2.2.1. Logistic regression analysis for the evaluation of effect of loading dose

The evaluation of the effect of loading versus no loading dose on the response at Week 24 will be assessed by logistic regression applied to the data from the 30 mg and 50 mg maintenance doses, with and without the loading dose. Data from placebo and the 10 mg dose will not be included in this analysis. There will be a single indicator variable for loading (treatment during the first 24 weeks) versus no-loading dose, and a single indicator variable for the effect of the 30 mg versus 50 mg maintenance doses. If there is evidence of a loading dose effect, an E_{max} model as specified in Model 1 will be used. In this case, an additional term for the effect of loading dose is added to the E_{max} model.

Although an interaction effect large enough to produce a significant interaction would be surprising, a check for interaction will be conducted using a similar logistic regression model, which includes interaction terms between the loading and maintenance doses. If the interaction effect is significant, exploratory analyses and follow-up work will be needed, and E_{max} modeling will not be performed. No models will be pre-specified for this condition.

5.2.2.2. E_{max} models for exposure response analysis

a. If the loading dose is significant in the logistic regression model in Section 5.2.2.1, then an indicator variable will be added representing the loading dose effect to the E_{max} model.

$$\log\left(\frac{\pi}{1-\pi}\right) = E_0 + \frac{E_{max}*Dose}{ED_{50}+Dose} + \beta I_{(loading)} \text{ (Model 1)}$$

b. If the effect of loading dose is not significant in the logistic regression model in Section 5.2.2.1, the data with/without loading doses will be pooled for each maintenance dose and analyzed using the E_{max} model.[5]

$$\boldsymbol{\log\left(\frac{\pi}{1-\pi}\right) = E_0 + \frac{E_{max}*Dose}{ED_{50}+Dose}} \text{ (Model 2)}$$

In either case, the results on the logistic scale of the E_{max} model will be back-transformed to the probability scale to improve interpretability and the posterior distribution for the difference on the proportion of SALT responders between the treatment groups determined by each dose and placebo will also be displayed.

5.2.2.3. Bayesian Estimation of Dose Response Models

The Bayesian estimation of the E_{max} model uses prior distributions on the placebo response (E_0), the dose that produces half the maximal drug effect (ED_{50}), E_{max} parameters, and the β loading dose parameter, if it is included.

The specification of the E_{max} model and the prior distributions for some of its parameters are based on three meta-analyses of clinical dose response from more than 100 compounds.[7,8,9] The E_{max} model specification is also supported by data from another compound with similar target biology, for which dose finding studies were performed in three indications.

We will use a *t* prior distribution for the logit of the placebo response centered at logit(0.05). A prior scale parameter (analogous to the standard deviation) of 4.0 (logistic scale) will be used, which yields a diffuse prior distribution for the placebo response. A *t* prior distribution for the E_{max} parameter is used, which is also specified on the logistic scale. It is centered at 0. This prior distribution will also be diffuse on the logistic scale with a prior scale parameter of 4.0. If a loading dose effect is included, its *t* prior distribution will also be centered at 0 with a prior scale parameter of 4.0. Our current projected value for the ED_{50} is P_{50}=30 mg. The prior distribution for the ED_{50} is derived from a t prior distribution for the normalized parameter, log(ED_{50}/ P_{50}). The t distribution for log(ED_{50}/ P_{50}) is centered at 0 with scale parameter, 1.73. This is the recommended distribution derived from the meta-analyses, and implemented in the R package, clinDR. It implies the ED_{50} is likely to be within 30-fold of its initial projected value (P_{50}). The degrees of freedom (df) for each prior t-distribution is 5, also obtained from the meta-analyses of dose response data.

5.2.3. Continuous Endpoints

For continuous secondary endpoints up to Week 24, a mixed-effect model with repeated measures (MMRM) will be used. This model will include the factors (fixed effects) for treatment group, visit, treatment-by-visit interaction, and relevant baseline value when modeling the change from baseline. Within the framework of MMRM, the treatment difference will be tested at each time point. Visit will be modeled as a categorical covariate. Unstructured covariance matrix will be assumed for the model errors.

When modeling the change from baseline values, the variable for visit will start with the first post-baseline visit, and the actual baseline value will be included as a covariate. At each visit, estimates of least square mean (LSM) values and the LSM differences between the PF-06651600-treated groups and the placebo group will be derived from the model. The corresponding p-values, standard errors and 95% CIs will also be derived from the model. Estimates of the difference in LSMs along with the two-sided 95% CI will also be provided for the treatment groups versus placebo group for the initial 24 weeks.

Exploratory endpoints will be summarized using descriptive statistics unless specified otherwise.

The MMRM method described above will also be applied to the continuous data up to Week 48 (over the entire study) for initial active treatment groups. LSM, standard error and 95% CI will be presented for each treatment group without treatment comparison. Descriptive statistics based on the observed case data will also be provided from baseline to Week 48 by 7 treatment sequences (Figure 1), without combining two placebo groups, as supportive.

5.3. Methods to Manage Missing Data

In general, for analyses using descriptive statistics, missing values will not be imputed. Missing values for exploratory endpoints will not be imputed unless specified otherwise. In addition, for safety endpoints, missing values will not be imputed. Other methods for handling missing values are discussed below.

5.3.1. Binary Endpoints

The same rule will be used for any missing data whether it is a missed visit, or it is a missing value due to early discontinuation. Depending on the analysis and the endpoint being analyzed, the following approaches will be used for managing missing data in binary endpoints:

a. Subjects with missing SALT scores due to COVID-19 in a given visit will be excluded from the analysis at that timepoint, whereas subjects with missing SALT scores due to other reasons will be treated as non-responders at that timepoint (Analysis 1, Section 5.2.1.1).

b. Subjects with missing SALT scores due to any reason in a given visit will be treated as non-responders at that timepoint (Analysis 3, Section 5.2.1.1).

c. Imputation of missing SALT response at Week 24 is performed based on the predictive distribution of the GLMM under MAR (Analysis 2, Section 5.2.1.2).

d. Imputation of missing SALT response at Week 24 is performed based on the predictive distribution of the GLMM under MNAR using tipping point analysis (Analysis 2a, Section 5.2.1.2).

e. Imputation of missing SALT response is performed based on the predictive distribution of the GLMM under MAR for subjects with missing data due to COVID-19, whereas subjects with missing SALT scores due to other reasons will be treated as non-responders at that timepoint (Analysis 4, Section 5.2.1.3).

f. Imputation of missing SALT response is performed based on the predictive distribution of the GLMM under MNAR using tipping point analysis for patients with missing data due to COVID-19, whereas subjects with missing SALT scores due to other reasons will be treated as non-responders at that timepoint (Analysis 4a, Section 5.2.1.3).

5.3.2. Continuous Endpoints

For continuous endpoints such as absolute SALT score or change from baseline in SALT score measured longitudinally, missing values post-baseline will not be imputed. For longitudinal continuous endpoints, assuming that the missing data mechanism is MAR, the data will be analyzed based on the full likelihood using a linear mixed-effect model with repeated measures for these continuous variables (see Section 5.2.3). This model will yield unbiased estimates and valid inferences in the presence of a missing data mechanism that is MAR.

For the continuous PRO variables such as AAPPO, HADS, EQ-5D-5L/EQ-5D-Y, SF-36v2, and WPAI:AA, rules suggested by Pfizer (for AAPPO) or the developers of these instruments will be followed in calculating the missing item-level and scale values. If these rules are not enough for imputing a value, then the missing values will be handled in the same way as efficacy variables.

6. ANALYSES AND SUMMARIES

Analysis will be done by testing the difference between each of the PF-06651600 QD-treated groups and placebo for the first 24 weeks. For the 25-48 week period, point estimates of the PF-06651600 QD-treated groups and the associated 95% CIs will be presented without testing any hypotheses.

Visit windows to be used for all efficacy analyses and some relevant safety analyses are detailed in Section 9.1.

The analyses of the primary endpoint(s) are described separately for the overall study and for specific regulatory regions in order to address differences in advice received from some regulatory agencies.

6.1. Primary Endpoints, Key Secondary and α-controlled Endpoints

Since this CCI [redacted] supports submission in several regions with different regulatory requirements the analyses of some endpoints may overlap to support multiple regions with different designations. This section describes the analyses planned for endpoints that are either primary, key secondary or α-controlled endpoints for the overall study reporting or for at least one of the known regional regulatory requirement.

6.1.1. Response Based on SALT Score ≤20 at Week 24

- Response based on SALT Score ≤20 at Week 24
- Population: FAS
- Statistical methods:
    - Analysis 1, Section 5.2.1.1 – this is the primary analysis and primary endpoint for the study and for the FDA/PMDA. This analysis will be repeated in the PPAS. This method will be also used for the analysis of earlier timepoints.
    - Analyses 2 and 2a, Section 5.2.1.2 – this is a supplementary analysis for the study and for the FDA/PMDA
    - Analysis 3, Section 5.2.1.1 – this is a supplementary analysis for the study, for the FDA/PMDA, and for the EMA and competent authorities in the VHP countries.
    - Analysis 4, Section 5.2.1.3 – this is an α-controlled analysis for the EMA and competent authorities in the VHP countries.

6.1.2. Response Based on SALT Score ≤10 at Week 24

- Response based on SALT Score ≤10 at Week 24
- Population: FAS
- Statistical methods:
    - Analysis 4, Section 5.2.1.3– this is the primary analysis and primary endpoint for the EMA and competent authorities in the VHP countries. This analysis will be repeated in the PPAS. This method will be also used for the analysis of earlier timepoints. This analysis could also be considered as supplementary for the overall study.
    - Analysis 1, Section 5.2.1.1 – this is the key secondary endpoint for the study, and considered a supplementary analysis for the EMA and competent authorities in the VHP countries. For the FDA/PMDA this is an α-controlled analysis. This analysis will be repeated in the PPAS.
    - Analysis 3, Section 5.2.1.1 – this is considered a supplementary analysis for the study and for the EMA and competent authorities in the VHP countries.

    - Analysis 4a, Section 5.2.1.3 – this analysis is supplementary for the EMA and competent authorities in the VHP countries.

6.1.3. PGI-C Response at Week 24

- PGI-C response at Week 24
- PGI-C response is defined as PGI-C score of "moderately improved" or "greatly improved" at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
- Population: FAS
- Statistical methods:
    - Analysis 4, Section 5.2.1.3 – this is the key secondary endpoint for the EMA and competent authorities in the VHP countries. This analysis will be repeated in the PPAS.
    - Section 6.2.9 describes other analysis that will be conducted for the PGI-C as a secondary endpoint the overall study and for the FDA/PMDA.

6.2. Secondary Endpoints

The analysis of binary secondary endpoints in the placebo-controlled phase will use the same approach in Analysis 1 (Section 5.2.1.1).

6.2.1. Exposure Response for Response Based on SALT Score ≤20 at Week 24

- Response based on SALT Score ≤20 at Week 24.
- Population: FAS.
- Statistical Method: Bayesian three-parameter E_{max} model with the log odds (logit) of the probability of response based on SALT score ≤20 at Week 24 (Section 5.2.2).
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.2. Exposure Response for Response Based on SALT Score ≤10 at Week 24

- Response based on SALT Score ≤10 at Week 24.
- Population: FAS.
- Statistical Method: Bayesian three-parameter E_{max} model with the log odds (logit) of the probability of response based on SALT score ≤10 at Week 24 (Section 5.2.2).
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.3. Response Based on SALT Score ≤20 at all Visits

- Response based on SALT Score ≤20 at all visits except Week 24.
- Population: FAS.
- Statistical Method: MN up to week 24. Descriptive method for entire duration.
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.4. Response Based on SALT Score ≤10 at all Visits

- Response based on an absolute SALT Score ≤10 at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Population: FAS.
- Statistical Method: MN up to week 24. Descriptive method for entire duration.
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.5. 75% Improvement in SALT Score (SALT75)

- Response based on a 75% improvement in SALT score from baseline (SALT75) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Population: FAS.
- Statistical Method: MN up to week 24. Descriptive method for entire duration
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.6. Change From Baseline in SALT Score

- Change from baseline in SALT scores at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Population: FAS.
- Statistical Method: MMRM up to Week 24. MMRM for initial active treatment groups for entire duration. Descriptive statistics for all treatment groups over the entire duration.
- Missing Data: Handled in MMRM model. Observed Cases will be used for the descriptive analysis.

6.2.7. Eyebrow Assessment Score

- Response based on at least a 2-grade improvement or a normal EBA score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Population: FAS on the subset of subjects who do not have normal (score of 3 on the EBA) eyebrow at baseline.
- Statistical Method: MN up to week 24. Descriptive method for entire duration.
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.8. Eyelash Assessment Score

- Response based on at least a 2-grade improvement or a normal ELA score at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Population: FAS on the subset of subjects who do not have normal (score of 3 on the ELA) eyelash at baseline.
- Statistical Method: MN up to week 24. Descriptive method for entire duration.
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.9. PGI-C Response

The analysis of PGI-C as a key secondary endpoint for the EMA and competent authorities in the VHP countries was described in Section 6.1.3. This section describes the analysis of the PGI-C response at all timepoints as a secondary endpoint for the overall study and for the FDA/PMDA.

- PGI-C response defined as PGI-C score of “moderately improved” or “greatly improved” at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
- Population: FAS.
- Statistical Methods: MN up to week 24. A descriptive summary of the distribution of subjects according to each category of PGI-C will be provided.
- Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders.

6.2.10. Alopecia Areata Patient Priority Outcomes (AAPPO) Scale

AAPPO was measured at baseline, Weeks 4, 8, 12, 18, 24, 34, 40, and 48. Endpoints include:

  - Current hair loss on scalp, eyebrows, eyelash, and body hair (Items 1-4) for each individual item scored as 0=‘no hair loss’, 1=‘little hair loss’ and 2-4=‘moderate-complete hair loss’.

  - Improvement on AAPPO items 1-4 among participants with a baseline score 2-4 indicating moderate-complete hair loss who achieved a score of 0=‘no hair loss’ or 1=’little hair loss’.

  - Change from baseline in AAPPO Emotional Symptoms domain score, where Emotional Symptoms domain score is defined as mean of items 5-8.

  - Change from baseline in AAPPO Activity Limitations subscore, where Activity Limitations subscore is defined as mean of items 9-11.

- Population: FAS.

- Statistical Methods:

  - MN for Proportion of patients with Improvement on Items 1-4. A descriptive summary of the distribution of subjects according to each category of AAPPO Items 1-4 will be provided.

  - MMRM for change from baseline in AAPPO Emotional Symptoms and Activity Limitations up to week 24. MMRM for initial active treatment groups for entire duration. Descriptive statistics for all treatment groups over the entire duration.

  - Missing Data: Exclude missing due to COVID-19, consider missing due to other as non-responders. For change from baseline, missing values are handled in the MMRM model. Observed Cases will be used for the descriptive analysis.

6.3. Exploratory Endpoints

Exploratory binary endpoints will be summarized using descriptive statistics unless specified otherwise.

6.3.1. 50% Improvement in SALT Score (SALT50)

- Response based on a 50% improvement in SALT score from baseline (SALT50) at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.

- Population: FAS.

- Statistical Method: Descriptive statistics.

- Missing Data: Observed Case.

6.3.2. Absolute SALT Score

- Absolute SALT scores at Baseline, Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.

[redacted]

- Population: FAS.
- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.3. Improvement on PGI-C

- Improvement on PGI-C defined as “slightly improved”, “moderately improved”, or “greatly improved” at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
- Population: FAS.
- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.4. Improvement on P-Sat

- Improvement on each item of the P-Sat (ie, amount of hair, quality of new hair, and overall) defined as slightly, moderately, or very satisfied for each item at Weeks 4, 8, 12, 18, 24, 34, 40, and 48.
- Population: FAS.
- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.5. Change From Baseline in the EQ-5D-5L/EQ-5D-Y

- Change from baseline in EuroQoL 5 Dimensions (EQ-5D-5L) in adults or EuroQoL 5 Dimensions-Youth (EQ-5D-Y) in adolescents at Weeks 4, 12, 24, and 48.
- Population: FAS.
- Statistical Method: Descriptive statistics for EQ-5D VAS score and Index Value for adults and adolescents separately. Counts and percent of patients in each category of for each of the 5 Dimensions, separately for adults and adolescents.
- Missing Data: Observed Case.

6.3.6. Change From Baseline in the AARU Questions

- Proportion of subjects with any HCP visits at Weeks 12, 24, 34, and 48.
- Among those with HCP visits, mean total number of visits for any reason, mean change from baseline in total number of visits for any reason, mean total number of visits related to AA and mean change from baseline in AA-related visits at Weeks 12, 24, 34, and 48.

- Population: FAS.
- Statistical Method: Descriptive Statistics.
- Missing Data: Observed Case.

6.3.7. Change From Baseline in the WPAI:AA Items

- Change from baseline in absenteeism (% work time missed due to AA), presenteeism (% impairment while working due to AA), work productivity loss (overall) due to AA and activity impairment due to AA at Weeks 12, 24, 34, and 48.
- Population: Adults in FAS.
- Statistical Method: Descriptive Statistics.
- Missing Data: Observed Case.

6.3.8. Change From Baseline in the Depression Subscale Score of HADS

- Change from baseline in the depression subscale score of the HADS at Weeks 4, 8, 12, 24, and 48.
- Population: FAS.
- Statistical Method: MMRM up to week 24. MMRM for initial active treatment groups for entire duration. Descriptive statistics for all treatment groups over the entire duration.
- Missing Data: handled in MMRM model. Observed Cases will be used for the descriptive analysis.

6.3.9. Change From Baseline in the Anxiety Subscale Score of HADS

- Change from baseline in the anxiety subscale score of the HADS at Weeks 4, 8, 12, 24, and 48.
- Population: FAS.
- Statistical Method: MMRM up to week 24. MMRM for initial active treatment groups for entire duration. Descriptive statistics for all treatment groups over the entire duration.
- Missing Data: handled in MMRM model. Observed Cases will be used for the descriptive analysis.

6.3.10. Improvement on Depression Subscale of HADS

- Improvement on HADS among subjects with a baseline subscale score indicative of depression who achieved a “normal” subscale score indicative of an absence of depression at Weeks 4, 8, 12, 24, and 48.
- Population: Subjects from the FAS with a baseline subscale score indicative of depression (ie, depression subscale score >7 for adults, depression subscale score >6 for adolescents).
- Statistical Method: MN up to week 24. Descriptive method for entire duration.
- Missing Data: Observed Case.

6.3.11. Improvement on Anxiety Subscale of HADS

- Improvement on HADS among subjects with a baseline subscale score indicative of anxiety who achieved a “normal” subscale score indicative of an absence of anxiety at Weeks 4, 8, 12, 24, and 48.
- Population: Subjects from the FAS with a baseline subscale score indicative of anxiety (ie, anxiety subscale score >7 for adults, anxiety subscale score >8 for adolescents).
- Statistical Method: MN up to week 24. Descriptive method for entire duration.
- Missing Data: Observed Case.

6.3.12. Change From Baseline in SF36v2 Acute

- Change from baseline in the physical and mental component scores, and 8 domain scores of the SF36v2 Acute at Weeks 4, 8, 12, 24, and 48.
- Population: FAS.
- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.13. Change From Baseline in CGI-AA

- CGI-AA scored as 0=‘None (no hair loss)’, 1=‘Minimal hair loss’, 2-4=‘Moderate-very severe or complete hair loss’ at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Improvement among participants with a baseline score 2-4 indicating moderate-very severe or complete hair loss, who achieved a score of 0=‘None (no hair loss)’ or 1=‘Minimal hair loss’ at Weeks 4, 8, 12, 18, 24, 28, 34, 40, and 48.
- Population: FAS.

[redacted]

- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.14. Change From Baseline in Number of Fingernails Affected by AA

- Change from baseline in the number of fingernails affected by AA at Weeks 12, 24, 34, 40, and 48.
- Population: FAS: on the subset of subjects who have at least one affected fingernail at baseline.
- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.15. Change From Baseline in IP-10

- Change from baseline in IP-10 at Weeks 4, 8, 12, and 24.
- Population: FAS.
- Statistical Method: Descriptive statistics.
- Missing Data: Observed Case.

6.3.16. Change From Baseline in T-cell, B-cell, and NK Cells

- Change from baseline in percent and absolute lymphocyte subsets (T-cell, B-cell, and natural killer [NK] cells) at Weeks 4, 12, 24, and 48.
- Population: FAS.
- Statistical Methods: Descriptive statistics for absolute or percent change from baseline. Counts and percent of patients in each category of CTCAE grade.
- Missing Data: Observed Case.

6.3.17. Change From Baseline in IgA, IgG, IgM

- Change from baseline in immunoglobulins (IgA, IgG, IgM) at Weeks 24 and 48.
- Population: FAS.
- Statistical Method: Descriptive statistics for change from baseline. Counts and percent of patients categories determined by clinical team.
- Missing Data: Observed Case.

6.4. PK Endpoints

- Plasma concentrations of PF-06651600 at Weeks 4, and 8 or 12.
- Population: PK analysis set.
- Statistical Method: Descriptive statistics of plasma concentrations of PF-06651600 at Weeks 4, and 8 or 12. Population PK model detailed in a separate PK analysis plan.

6.5. Subset Analyses

Summary statistics for SALT ≤20 at week 24, SALT ≤10 at week 24 and PGI-C response at Week 24 will be presented by subgroups as below:

- Age (years) at baseline (12-17, ≥18);
- Age (years) at baseline (12-17, 18-44, 45-64, ≥65);
- Body Mass Index (BMI) group at baseline (<25, 25-30, ≥30);
- Weight at baseline (<median, ≥median);
- Gender (Male, Female);
- Race (White, Black, Asian, Other);
- Region of enrollment (North America, Europe, Asia, Rest of World) as specified in Section 9.4;
- Baseline disease severity (AT/AU, non AT/AU);
- AA duration since first diagnosis (years) (<median, ≥median);
    - The median of the entire sample will be used as the cut-off value.
- Duration of onset of current AA episode (years) (<median, ≥median);
    - The median of the entire sample will be used as the cut-off value.
- Prior pharmacological treatment of AA (yes, no); Section 9.7.

Estimates of the response rates for each dose compared to the placebo group and its 95% CI based on MN method, will be presented for each defined category of each subgroup. No p-values will be presented.

The primary purpose of the subgroup analyses is to check for consistency of results across subgroups. There is no intention to have any specific inference within subgroups. The analysis described in this SAP may be repeated on the specific country and/or region subpopulation sets in order to meet country-specific regulatory requirements.

6.6. Baseline and Other Summaries and Analyses

6.6.1. Baseline Summaries

Medical history will be summarized by treatment group using frequency and percentage of subjects for each medical condition according to Pfizer Standards. Demographics will be summarized by treatment group. Continuous demographics characteristics data, such as age will be summarized using descriptive statistics. Categorical data such as gender will be summarized using frequency and percentages.

Targeted medical history using the list in Section 9.5 will also be summarized.

6.6.2. Study Conduct and Subject Disposition

Subjects evaluation, disposition, discontinuation will be summarized according to Pfizer standards.

6.6.3. Study Treatment Exposure

A summary of compliance and the number of doses received as well as the median total dose by treatment group will be provided.

The exposure to study drug will be summarized by the total number of days of dosing.

6.6.4. Concomitant Medications and Non-Drug Treatments

Prior drug and non-drug treatment, concomitant drug and non-drug treatment will be summarized according to Pfizer standards.

6.7. Safety Summaries and Analyses

Safety analysis will be based on the safety analysis set (SAS).

Safety data will be presented in tabular and/or graphical format and summarized descriptively, where appropriate. All safety endpoints will be listed and summarized in accordance with Pfizer Standards. Categorical outcomes (eg, AEs) will be summarized by subject counts and percentage. Continuous outcome will be summarized using N, mean, median, standard deviation, etc. Selected subject listings will be produced for these safety endpoints accordingly.

In order to report the impact of COVID-19 on clinical trial populations and study data, the following additional listings and summaries will be produced:

- Listing of subjects affected by COVID-19 related study disruption;
- Protocol deviations related to COVID-19;
- Summary of drug interruption before and after the COVID-19 pandemic anchor date (09 January 2020 for Taiwan/China sites, and 11 March 2020 for the remaining countries);

- Discontinuations from study drug due to COVID-19 related AEs.

6.7.1. Adverse Events

The safety data will be summarized in accordance with Pfizer Standards. Adverse events of special interest will be summarized based on a list of preferred terms that will be provided by safety risk lead to the programming team prior to database lock. All safety data will be summarized descriptively through appropriate data tabulations, descriptive statistics, categorical summaries, and graphical presentations. Safety endpoints for the study include:

- Treatment-emergent (including treatment-related) AEs and SAEs;
- AEs leading to discontinuation;
- AEs will be displayed for the following periods: Baseline to Week 24, and Baseline to Week 48;
- Treatment-emergent COVID-19 related AEs by SOC and PT;
- Summary of TEAE before and after the COVID-19 pandemic anchor date.

6.7.2. Laboratory Data

Laboratory data will be listed and summarized in accordance with the Pfizer reporting standards. Summaries of incidence of clinically significant abnormalities will include frequency and percentages. Laboratory data criteria for discontinuation and abnormalities can be found in Section 9.3.

6.7.3. Vital Signs, including Height and Weight

Vital signs will be summarized at baseline and all post-baseline visits. The data will be listed and summarized in accordance with the Pfizer reporting standards. Summaries of incidence of clinically significant abnormalities will include frequency and percentages.

Height and weight will be summarized at enrollment/baseline, Week 24, and Week 48.

6.7.4. Electrocardiogram

Abnormal ECG findings will be summarized.

6.7.5. Physical Examination

Physical examinations will be summarized at baseline and all-available post-baseline visits.

6.7.6. Hearing Examination

Audiometry will be performed at screening, Week 24, and Week 48 visits. Summaries of incidence of clinically significant worsening in hearing from screening, as assessed by the investigator following review of the audiometry reports, will be presented.

7. INTERIM ANALYSES

The study may be unblinded for Sponsor internal decision-making purposes when ≥90% of subjects have reached 24 weeks post-randomization (or discontinued prior to the Week 24 visit), and dissemination of these results will be limited to the unblinded reporting team and Sponsor management. The sponsor personnel who are unblinded will be separate from the study team. The sponsor will still maintain the blind for study team members who will be involved in daily study conduct, study management, safety, and data monitoring through the completion of the study. Investigators and subjects will remain blinded. There will be no impact on B7981015 study conduct, analysis, or reporting.

This study will use an external data monitoring committee (E-DMC).

The E-DMC will be responsible for ongoing monitoring of the safety of subjects in the study according to the charter.

8. REFERENCES

1. EQ5D country specific weights. https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/valuation-standard-value-sets/crosswalk-index-value-calculator/.

2. Snaith RP, Zigmond AS. The Hospital Anxiety and Depression Scale with the Irritability-Depression-Anxiety Scale and the Leeds Situational Anxiety Scale Manual. GL Assessment, 1994.

3. White et al. Validation of the Hospital Anxiety and Depression Scale for use with adolescents. Br J Psy, 1999; 175:452-54.

4. Rubin, D. B. (1987) “Multiple Imputation for Nonresponse in Surveys.” (1st ed.) New York: Wiley.

5. Kallen A. (2007). PK/PD modelling (Chapter 6). Computational Pharmacokinetics. CRC. Boca Raton, Florida

6. Rubin, D. B. (2004) “Multiple Imputation for Nonresponse in Surveys.” (2nd ed.) New York: Wiley Interscience.

7. Thomas N, Sweeney K, Somayaji V. (2014) Meta-analysis of clinical dose response in a large drug development portfolio. Statistics in Biopharmaceutical Research. 6(4):302-17.

8. Thomas N, Roy D (2017) Analysis of clinical dose-response in small-molecule drug development: 2009-2014. Statistics in Biopharmaceutical Research. 9(2):137-46.

9. Wu J, Banerjee A, Jin B et al.(2018) Clinical dose-response for a broad set of biological products: A model-based meta-analysis. Stat Methods in Med Res. 27(9):2694 721.

10. Cassella J, Hamilton C, von Hehn J, Braman V. CTP-543, an Oral JAK Inhibitor, Achieves Primary Endpoint in Phase 2 Randomized, Placebo-Controlled Dose-Ranging Trial in Patients with Moderate-to-Severe Alopecia Areata [Conference presentation]. In: 28th European Academy of Dermatology and Venereology (EADV) Annual Congress; 2019 Oct 9-13, Madrid, Spain. https://www.concertpharma.com/wp-content/uploads/2019/10/EADV-Late-Breaker-CTP543-Presentation-FINAL-12OCT2019.pdf

9. APPENDICES

9.1. Definition and Use of Visit Windows in Reporting

Visit windows will be used for efficacy and PRO variables, and for any safety data that display or summarize by study visit.

Table 6. Visit Windows

Visit Label	Target Day	Definition
Baseline	1	<=Day 1
Week 2	15	Days 2 to 21
Week 4	29	Days 22 to 42
Week 8	57	Days 43 to 70
Week 12	85	Days 71 to 105
Week 18	127	Days 106 to 147
Week 24	169	Day 148 to 179 and <=EDay[a] 1
Week 26	183	Days 180 to 189 and >EDay 1
Week 28	197	Days 190 to 217 and >EDay 1
Week 34	239	Days 218 to 259 and >EDay 1
Week 40	281	Days 260 to 308 and >EDay 1
Week 48	337	≥Day 309 and >EDay 1

[a] EDay =relative day of the first dose date in the extension phase based on dosing log data after week 24 visit. If Week 24 dose log is missing, earliest dosing date from Week 26 and onward will be considered start of Extension.

The Baseline value is defined in Section 3.3.

Some endpoints were not scheduled to be assessed at each visit. In this case, the unscheduled 'windowed' visit will be moved to the next scheduled windowed visit for analysis and reporting. For example, HADS were not to be assessed at Week 18, Week 28, Week 34 and Week 40, if data falls under Week 18 it will be treated as Week 24; and data that falls under Weeks 28, 34, 40 will be treated as Week 48.

If more than one observation falls within a visit window, the observation that is the closest to the targeted day will be used for the summary tables. If two visits are equally distant from the target day, the data from the later visit will be used.

9.2. Details on Selected PRO Endpoints

AAPPO

The AAPPO is a 11-item scale. Items 1-4 are an assessment of current hair loss, eyebrow loss, eyelash loss and body hair loss and will be as such analyzed separately on a scale of 0-4, with 0 =‘no hair loss’ and 4=‘complete hair loss’. Items 5-8 are an assessment of emotional symptoms. Response choices on these items are scored from 0 =‘never’ to 4=‘always’. Items 9-11 are an assessment of activity limitations. Response choices on these items are scored from 0=‘not at all’ to 4= ‘completely’.

- AAPPO Items 1 through 4 are scored separately as 0=‘no hair loss’, 1=‘little hair loss’ and 2-4=‘moderate-complete hair loss’.
- Improvement on AAPPO items (1-4) among participants with a baseline score 2-4, indicating moderate-complete hair loss who achieved a score of 0=‘no hair loss’ or 1=‘little hair loss’.
- Emotional Symptoms: Mean of Items 5, 6, 7, 8 (missing rule: requires at least 2 non-missing responses; otherwise missing).
- Activity Limitations: Mean of Items 9, 10, 11 (missing rule: requires at least 2 non-missing responses; otherwise missing).

AARU:

- Assessment of total in-person and remote healthcare visits.
- Assessment of AA-related in-person and remote healthcare visits.
- Assessment of utilization of hair prostheses or camouflaging agents.
- Assessment of # of days hair prostheses or camouflaging agents were utilized.
- Percentage of patients not currently employed.
- Percentage of patients who have not sought employment due to alopecia areata.
- Number of opportunities for employment not sought due to alopecia areata.

WPAI –AA:

WPAI-AA is an adaptation of the WPAI:SHP (Specific Health Problem) version. Outcomes from this scale are expressed as impairment percentages, with higher numbers indicating greater impairment and less productivity. The questionnaire is scored as follows:

Questions:

- 1 = currently employed.

[redacted]

- 2 = hours missed due to specified problem.
- 3 = hours missed other reasons.
- 4 = hours actually worked.
- 5 = degree problem affected productivity while working.
- 6 = degree problem affected regular activities.

Scoring.

- Multiply scores by 100 to express in percentages.
- Percent work time missed due to problem (Absenteeism): Q2/(Q2+Q4).
- Percent impairment while working due to problem (Presenteeism): Q5/10.
- Percent overall work impairment due to problem (Overall Work Impairment): Q2/(Q2+Q4)+[(1-(Q2/(Q2+Q4)))x(Q5/10)].
- Percent activity impairment due to problem (Activity Impairment): Q6/10.

9.3. Laboratory Data Criteria for Discontinuation and Abnormalities

Labs Meeting Discontinuation Criteria, as per Appendix 3.2 of the protocol

Table 7. Laboratory Values Meeting Discontinuation Criteria

		Criterion #1	**Criterion #2**
Hematology	Hemoglobin	<90 g/L or a decrease of >30% from baseline	n/a
	Platelets	<75 × 10^9/L	n/a
	Lymphocytes (ABSOLUTE)	<0.5 × 10^9/L	n/a
	Neutrophils (ABSOLUTE)	<0.75 × 10^9/L	n/a
Chemistry	Aspartate Aminotransferase	>3× ULN W/Bilirubin >2× ULN	Two Sequential Elevations >5× ULN
	Alanine Aminotransferase	>3× ULN W/Bilirubin >2× ULN	Two Sequential Elevations >5× ULN
	Creatine Kinase	>10× ULN	

Lab Test Abnormalities

Table 8. Laboratory Test Abnormalities

		Lower Limit	**Upper Limit**
Hematology	Hemoglobin	<0.8× LLN	n/a
	Hematocrit	<0.8× LLN	n/a
	Red Blood Cell Count	<0.8× LLN	n/a
	MCV	<0.9× LLN	>1.1× ULN
	MCH	<0.9× LLN	>1.1× ULN
	MCHC	<0.9× LLN	>1.1× ULN
	Platelets	<0.5× LLN	>1.75× ULN
	White Blood Cell Count	<0.6× LLN	>1.5× ULN
	Reticulocytes	<0.5× LLN	>1.5× ULN
	Reticulocytes/Erythrocytes	<0.5× LLN	>1.5× ULN
	Leukocytes	<0.6× LLN	>1.5× ULN
	Lymphocytes (ABSOLUTE)	<0.8× LLN	>1.2× ULN
	Lymphocytes/Leukocytes	<0.8× LLN	>1.2× ULN
	Neutrophils (ABSOLUTE)	<0.8× LLN	>1.2× ULN
	Neutrophils/Leukocytes	<0.8× LLN	>1.2× ULN
	Basophils/Leukocytes	n/a	>1.2× ULN
	Eosinophils	n/a	>1.2× ULN
	Eosinophils/Leukocytes	n/a	>1.2× ULN
	Monocytes	n/a	>1.2× ULN
	Monocytes/Leukocytes	n/a	>1.2× ULN
	Prothrombin Time	n/a	>1.1× ULN
	Basophils	n/a	>1.2× ULN

Table 8. Laboratory Test Abnormalities

		Lower Limit	**Upper Limit**
	Prothrombin Intl. Normalized Ratio	n/a	>1.1× ULN
Chemistry	Blood Urea Nitrogen	n/a	>1.3× ULN
	Urea	n/a	>1.3× ULN
	Creatinine	n/a	>1.3× ULN
	Glucose	<0.6× LLN	>1.5× ULN
	Calcium	<0.9× LLN	>1.1× ULN
	Sodium	<0.95× LLN	>1.05× ULN
	Potassium	<0.9× LLN	>1.1× ULN
	Chloride	<0.9× LLN	>1.1× ULN
	Bicarbonate	<0.9× LLN	>1.1× ULN
	Aspartate Aminotransferase	n/a	>3.0× ULN
	Alanine Aminotransferase	n/a	>3.0× ULN
	Bilirubin	n/a	>1.5× ULN
	Direct Bilirubin	n/a	>1.5× ULN
	Indirect Bilirubin	n/a	>1.5× ULN
	Alkaline Phosphatase	n/a	>3.0× ULN
	Uric Acid	n/a	>1.2× ULN
	Albumin	<0.8× LLN	>1.2× ULN
	Creatine Kinase	n/a	>2.0× ULN
	Cholesterol	n/a	>1.3× ULN
	HDL Cholesterol	<0.8× LLN	n/a
	LDL Cholesterol	n/a	>1.2× ULN
	Triglycerides	n/a	>1.3× ULN
	Protein	<0.8× LLN	>1.2× ULN
	Gamma Glutamyl Transferase	n/a	>3.0× ULN
Urinalysis	pH (Scalar)	<4.5	>8
	Glucose (Scalar)	n/a	≥1
	Ketones (Scalar)	n/a	≥1
	Protein (Scalar)	n/a	≥1
	Hemoglobin (No Unit)	n/a	≥1
	Nitrite (No Unit)	n/a	≥1
	Leukocyte Esterase (No Unit)	n/a	≥
	Urobilinogen	n/a	≥1
	Erythrocytes (Scalar)	n/a	≥20
	Leukocytes (Scalar)	n/a	≥20
	Epithelial Cells (Scalar)	n/a	≥6
	Granular Casts (Scalar)	n/a	>1
	Hyaline Casts (Scalar)	n/a	>1
	Bacteria (No Unit)	n/a	>20

9.4. List of Countries by Region

Table 9. Countries by Region

Region	Countries
North America	Canada, United States
Europe	Czech Republic, Germany, Hungary, Poland, Russia, Spain, United Kingdom
Asia	China, Japan, South Korea, Taiwan
Rest of World	Australia, Argentina, Chile, Colombia, Mexico

9.5. List of Targeted Medical History Terms

Table 10. Targeted Medical History Terms

Term	PT
Atopic dermatitis	DERMATITIS ATOPIC
Diabetes mellitus Type I	TYPE 1 DIABETES MELLITUS
Crohn’s disease	CROHN'S DISEASE
Ulcerative colitis	COLITIS ULCERATIVE
Systemic lupus erythematosus	SYSTEMIC LUPUS ERYTHEMATOSUS
Sjogren's syndrome	SJOGREN'S SYNDROME
Rheumatoid arthritis	RHEUMATOID ARTHRITIS
Psoriatic arthritis	PSORIATIC ARTHROPATHY
Psoriasis	PSORIASIS
Vitiligo	VITILIGO
Hashimoto’s thyroiditis	AUTOIMMUNE THYROIDITIS
Graves' disease	BASEDOW'S DISEASE

9.6. Tipping Point Analysis for Binary Data

The primary endpoint for the study and for the FDA/PMDA is response based on SALT ≤20 (a binary endpoint) at Week 24; for the EMA and competent authorities in the VHP countries the primary endpoint is response based on SALT ≤10 at Week 24. This tipping point analysis will be conducted for both endpoints, with a subtle difference: while the analysis for the study and for the FDA/PMDA will treat missing as MNAR regardless of the reason for missingness, the analysis for the EMA will treat missing due to COVID-19 as MNAR but missing due to other reasons will be considered as non-responders.

Treatment comparisons are performed for each PF-06651600 dose group versus placebo. These binary endpoints are also assessed at post-baseline visits of Weeks 4, 8, 12, and 18. And are coded as a numeric variable with values of 1 if SALT score is ≤20 or 0 if SALT score is >20. When there is insufficient information to evaluate the endpoint at a visit (eg, due to discontinuation of study participation or other reasons), they will be coded as missing at that visit. Tipping point analysis is used to evaluate the impact of these missing values on the trial conclusion based on the primary endpoints. Since the focus is on the Week 24 visit, this analysis will use longitudinal data of these endpoints from Week 4 up to Week 24 only (5 post-baseline visits) without imputation of the missing responses. Data of visits after Week 24 will not be included in these analyses.

A SAS macro “alopbinarytippingV2.sas” (Version 2) is developed to fit the purpose of implementing the tipping point analysis for this study. Longitudinal data of the endpoints are imported into the macro.

The prior density specified for the baseline logit is Normal(-1.7, sd=3) and those for all other logit effect parameters are Normal(0, sd=1). This gives the prior median and mean of 15% and 32% respectively for the probability of response (SALT ≤20) for the placebo group at the post-baseline visits. The prior density for variance of random effects is Inverse-Gamma(shape=1, scale=1). In this prior, the prior 90th percentile for σ^2 is approximately 9.

The following values will be specified for the implementation of the analyses for each of the active treatment and placebo groups.

- Active treatment group: deltat = -0.5, -0.3, -0.1, 0, 0.1, 0.3, 0.5 (7 of them)
- Placebo group: deltac = -0.9, -0.7, -0.5, -0.3, -0.1, 0 (6 of them)

Since the delta value is subtracted from the MAR probability, a negative delta value means favoring the probability; a positive delta value means penalizing the probability; and a 0 value means the same MAR probability. The values of these sensitivity parameters (deltat, deltac) may be adjusted if needed.

The number of imputations/imputed datasets per combination of sensitivity parameters for the tipping point analysis will be at least 100.

The seeds used for tipping point analysis are defined below:

Endpoint	**Seed for MCMC (seed4mcmc)**	**Seed for imputation (seed4imp)**
Response based on SALT ≤20	3355	6688
Response based on SALT ≤10	1286	9276
PGI-C Response	4562	1520

9.7. Prior Pharmacological Treatments for AA – Definition of Subgroups

The following categories of medications for AA were identified by the Sponsor as relevant for the determination of the subgroup 'Prior Pharmacological Treatment for AA' in Section 6.5. A list of medications in each of these categories will be reviewed by the clinical team before the study is unblinded.

- Biologics
- Intralesional corticosteroid injection
- Intralesional minoxidil
- Oral anti-inflammatory
- Oral immunosuppressant
- Oral vasodilator
- Oral/IV/IM Steroids
- Other (non-oral) systemic immunosuppressant
- Other immunotherapy
- Other topical anti-inflammatory
- Subcutaneous immunotherapy
- Topical anthralin/dithranol
- Topical corticosteroid
- Topical immunotherapy
- Topical JAK
- Topical vasodilators
- Unknown steroid
- Unknown methotrexate
- Unknown minoxidil