Statistical Analysis Plan for Study M21-500

A Phase 3, Multicenter Study to Evaluate the Safety and Efficacy of AGN-151586 for the Treatment of Glabellar Lines

Date: 08 April 2023

Version 5.0

Table of Contents

List of Tables

List of Figures

List of Appendices

1.0 Introduction

This Statistical Analysis Plan (SAP) describes the statistical analyses for AGN-151586 Study M21-500: A Phase 3, Multicenter Study to Evaluate the Safety and Efficacy of AGN-151586 for the Treatment of Glabellar Lines.

Study M21-500 examines the safety and efficacy of AGN-151586 (containing botulinum toxin type E) for treatment in subjects with moderate to severe glabellar lines.

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

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

This SAP includes changes to analyses described in the protocol. Details are outlined in Section 14.0.

2.0 Study Design and Objectives

2.1 Objectives, Hypotheses and Estimands

The study objective is to evaluate the safety and efficacy of AGN-151586 for the treatment of glabellar lines (GL) in subjects with moderate to severe GL.

The clinical hypothesis is that AGN-151586 is more effective than placebo in treating GL, as measured by both investigator and subject assessment of GL severity at maximum frown using the Facial Wrinkle Scale (FWS), and that it has an acceptable safety profile after single and repeat treatments.

Estimands: Primary Endpoints

The attributes of the estimands corresponding to the primary and coprimary efficacy endpoints are summarized in Table 1 (see Section 9.3.2). In addition, the attribute of treatment is a single dose of AGN-151586 or placebo.

For the US FDA, the estimand for the primary composite endpoint is defined as follows:

- Difference in the proportion of subjects who would achieve a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline in composite FWS (as assessed by both investigator and by subject) at Day 7 had they not missed FWS assessments, for the AGN-151586 group in comparison with placebo in the ITT population

For EU Regulatory Agencies, the estimands for the coprimary endpoints are defined as follows:

- Difference in the proportion of subjects who would achieve ≥ 2-grade improvement from baseline in FWS as assessed by investigator at Day 7 had they not missed FWS assessments, for the AGN-151586 group in comparison with placebo in the ITT population, and
- Difference in the proportion of subjects who would achieve ≥ 2-grade improvement from baseline in FWS as assessed by subject at Day 7 had they not missed FWS assessments, for the AGN-151586 group in comparison with placebo in the ITT population

Any missing assessments will be assumed to be missing at random and imputed using multiple imputation. Statistical significance of the difference will be tested using Cochran-Mantel-Haenszel (CMH) test [redacted].

Estimands: Secondary Endpoints

The attributes of the estimands corresponding to the secondary efficacy endpoints are summarized in Table 2 (see Section 9.4.2). Treatment is the same as for the primary efficacy endpoints. For both US FDA and EU variables/endpoints, the population is the ITT population (all randomized). For the EU, subgroup analyses will be performed for the coprimary and key secondary endpoints for the ITT population that also has baseline 11-item Facial Line Outcomes (FLO-11) total scores (transformed) ≤ 50. The

variables/endpoints listed have the same handling of intercurrent events and statistical summary (including population-level summary and analysis methods) within their respective analysis populations for the US FDA and EU, per estimand label.

2.2 Study Design Overview

The schematic of the study is shown in Figure 1 for this study for the double-blind (DB) period and the open-label (OL) period.

[redacted]

Hr = hour

2.3 Treatment Assignment and Blinding

Subjects will be randomized to AGN-151586 [redacted] or placebo in a 3:1 ratio.

[redacted]

2.4 Sample Size Determination

Approximately 600 subjects will be randomized into the study in a 3:1 ratio, yielding approximately 450 subjects in the AGN-151586 group and 150 subjects in the placebo group in the treatment period.

For the US FDA, a sample size of 600 subjects will have > 99% power to detect a difference on Day 7 between the AGN-151586 group and the placebo group using the composite endpoint of ≥ 2-grade improvement, [redacted].

For the EU regulatory agencies, a sample size of 600 subjects for the ITT population will have >99% power (and the subgroup population of an estimated 360 subjects will provide > 99% power) to detect a difference between the AGN-151586 group and the placebo group at Day 7 for each of the coprimary endpoints, [redacted].

Sample size calculations were based on the two group χ^2 two-sided test of equal proportions using nQuery (Version 7.0) at a 0.05 level of significance.

3.0 Endpoints

3.1 Primary Endpoints

For the FDA, the primary composite endpoint is the achievement of a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to both investigator and subject assessments of GL severity using FWS at maximum frown at Day 7.

For EU regulatory agencies, the coprimary endpoints are the achievement of at least a 2-grade improvement from baseline on the FWS according to: (1) subject assessment of GL severity at maximum frown at Day 7, and (2) investigator assessment of GL severity using FWS at maximum frown at Day 7.

3.2 Secondary Endpoints

For US FDA:

- Key secondary endpoint: a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to both investigator and subject assessments (i.e., concurrent composite) of GL severity at maximum frown over time (DB period)*
- Key secondary endpoint: a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to investigator assessment of GL severity at maximum frown over time (DB period)*
- Key secondary endpoint: a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to subject assessment of GL severity at maximum frown over time (DB period)*
- *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (*overall satisfaction*) for GL at Day 7
- *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (*overall satisfaction*) for GL at Hour 24
- *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 4 (*natural look*) for GL at Day 7

For EU regulatory agencies:

- Key secondary endpoint: ≥ 20-point improvement from baseline in FLO-11 total scores for GL at Day 7
- ≥ 2-grade improvement from baseline on the FWS according to subject assessment of GL severity at maximum frown at Hour 24

- ≥ 2-grade improvement from baseline on the FWS according to investigator assessment of GL severity at maximum frown at Hour 24
- ≥ 1-grade improvement from baseline on the FWS according to subject assessment of GL severity at maximum frown at Hour 24
- ≥ 1-grade improvement from baseline on the FWS according to investigator assessment of GL severity at maximum frown at Hour 24
- *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL at Hour 24
- *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 4 (natural look) for GL at Day 7
- Time to the first ≥1-grade improvement from baseline on the FWS according to subject assessment of GL severity at maximum frown over time (DB period)*
- Time to the first ≥1-grade improvement from baseline on the FWS according to investigator assessment of GL severity at maximum frown over time (DB period)*
- ≥ 2-grade improvement from baseline on the FWS according to subject assessment of GL severity at maximum frown over time (DB period)*
- ≥ 2-grade improvement from baseline on the FWS according to investigator assessment of GL severity at maximum frown at over time (DB period)*
- Time to return to baseline FWS criterion[†] according to subject assessment of FWS at maximum frown (DB period)*
- Time to return to baseline FWS criterion[†] according to investigator assessment of FWS at maximum frown (DB period)*
- *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL over time*
- ≥ 4-point improvement from baseline in FLO-11 Item 10 (look angry) for GL at Day 7
- ≥ 4-point improvement from baseline in FLO-11 Item 5 (look less attractive) for GL at Day 7
- Subject-reported global assessment of change in GL based on the Global Assessment of Change in Glabellar Lines (GAC-GL) over time*

* Endpoints will be excluded from hierarchical testing.

[redacted]

3.3 Other Efficacy Endpoints

The primary and secondary efficacy endpoints are listed in Section 3.1 and Section 3.2, respectively. For both the US FDA and EU regulatory agencies, analyses for additional efficacy endpoints will be performed on the ITT population. The additional efficacy endpoints are as follows:

[redacted]

3.4 Safety Endpoints

The safety endpoints are the incidence of AEs for the entire study population, change from baseline in vital signs parameters, change from baseline in electrocardiogram (ECG) parameters, change from baseline in laboratory evaluations (hematology and chemistry), and presence of binding and neutralizing antidrug antibodies. Results of a neurological assessment (which includes assessment of cranial nerves II through VII) will be reviewed as an additional safety measure. After the screening visit, clinically significant (as per investigator) neurological assessment abnormalities (or worsening of a baseline condition) will be captured as AEs.

3.5 Additional Endpoints

Not applicable.

4.0 Analysis Populations

The following population sets will be used for the analyses.

The Intent-to-Treat (ITT) Population includes all randomized subjects. For the DB period analyses, subjects will be included in the analysis according to the treatment groups to which they were randomized. For the OL period analyses, data will be analyzed according to the treatments randomized in the DB period and received in the OL period (i.e., AGN-151586/AGN-151586, AGN-151586/None, Placebo/AGN-151586, Placebo/None).

Baseline analyses and efficacy analyses for US FDA and EU Regulatory agencies will be performed on the ITT population, consisting of all randomized subjects. Subjects will be included in the DB period analysis according to the treatment groups to which they were randomized. For the OL period analyses, data will be analyzed according to the treatments randomized in the DB period and received in the OL period, as above.

Additionally, for EU regulatory agencies only, subgroup analyses will be performed on the ITT population that also has a baseline FLO-11 total score (transformed) ≤ 50. Analyses will consist of the EU coprimary and key secondary endpoints.

The Safety Population consists of all subjects who were treated with at least 1 dose of study drug (i.e., AGN-151586 or placebo). All safety analyses will be performed with subjects analyzed by their actual treatment received in the DB period/OL period (i.e., AGN-151586/AGN-151586, AGN-151586/None, Placebo/AGN-151586, Placebo/None), and will be presented with overall safety data for the study (i.e., either period), for DB period, and for OL period. The Safety Population will be used for all safety analyses.

5.0 Subject Disposition

The total number of subjects who were screened, enrolled (randomized), and treated will be summarized for the ITT population. A summary of subject accountability will be provided where the number of subjects in each of the following categories will be summarized for each treatment group and for the DB period and the OL period:

- Subjects enrolled (randomized) in the study;
- Subjects treated with at least one dose of study drug;
- Subjects who completed the study;
- Subjects who prematurely discontinued study (with reason),

Subjects in each analysis population will also be summarized.

For end of study participation, the number and percentage of subjects who completed the protocol defined follow-up period will be summarized overall and by treatment group.

6.0 Study Drug Duration and Compliance

For the Safety Population, the number of subjects treated will be presented by treatment group for the double-blind period, as well as the open-label period. Duration on treatment is assumed for the duration of the study period or until study exit. If a subject does not receive the full dose, this will be indicated; significant deviations to dosing will be reported.

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

Demographics, baseline or disease characteristics, medical history, prior and concurrent procedures, and prior and concomitant medications will be summarized for the ITT population overall and by treatment group. Categorical variables will be summarized with the number and percentage of subjects; percentages will be calculated based on the number of non-missing observations. Continuous variables will be summarized with descriptive statistics (number of non-missing observations, mean and standard deviation, median, Q1, Q3, minimum and maximum).

7.1 Demographics and Baseline Characteristics

Demographics and baseline characteristics will be summarized descriptively, overall and by treatment group. Unless otherwise specified, baseline is defined as the last non-missing value prior to the first administration of study drug.

Continuous demographic variables include age, weight, height, and body mass index (BMI). Continuous baseline characteristic variables include FLO-11 total scores (transformed) and item scores (for items 1-10), and FLSQ treatment expectations and impact domain scores.

Categorical demographic and baseline characteristic variables include:

- Sex
- Ethnicity (Hispanic or Latino, Not Hispanic or Latino)
- Race
- Age (< 65, ≥ 65 years)
- Region (North America, EU)
- BMI (< 25, ≥ 25 kg/m^2).
- Nicotine user (current, former, never, unknown)
- Alcohol user (current, former, never, unknown)
- Prior aesthetic toxin use (prior use, no prior use)
- FWS at maximum frown
  - As assessed by investigator
  - As assessed by subject
- FWS at rest
  - As assessed by investigator
  - As assessed by subject
- FLO-11 total score (transformed) (≤50, >50)
- FLSQ item scores
- Fitzpatrick Skin Type

7.2 Medical History and Prior and Concurrent Procedures

Medical history data will be reported and summarized separately from prior procedures data for the ITT population. Medical history data and prior procedure data will be coded using the Medical Dictionary for Regulatory Activities (MedDRA). The actual version of the MedDRA coding dictionary will be noted in the statistical tables and clinical study report. The number and percentage of subjects in each medical history category (by MedDRA system organ class and preferred term) will be summarized overall and by treatment group. Similarly, the number and percentage of subjects in each prior procedures category will be summarized overall and by treatment group. The system organ class (SOC) will be presented in alphabetical order, and the preferred terms will be

presented in alphabetical order within each SOC. Subjects reporting more than one condition/diagnosis will be counted only once in each row (SOC or preferred term).

In addition, any concurrent procedures, defined as any procedure performed on or after the date of first treatment, will be summarized by term in each treatment group for the ITT population.

7.3 Prior and Concomitant Medications

Prior and concomitant medications will be summarized by generic name for the ITT population. A prior medication is defined as any medication taken prior to the date of the first dose of study drug. A concomitant medication is defined as any medication that started prior to the date of the first dose of study drug and continued to be taken after the first dose of study drug or any medication that started on or after the date of the dose of study drug, but not after the date of the last visit assessment. The number and percentage of subjects taking medications will be summarized by generic drug name based on the World Health Organization (WHO) Drug Dictionary for both prior and concomitant medications.

8.0 Handling of Potential Intercurrent Events for the Primary and Key Secondary Endpoints

The primary composite and coprimary endpoints for the US FDA and for the EU regulatory agencies (defined in Section 3.1), respectively, will be analyzed based on the ITT population, and the following methods will be used to address potential intercurrent events:

- Subjects who did not receive any dose of study drug but are randomized will still be included in the ITT population. For the EU, those who also had baseline FLO-11 total score (transformed) of ≤ 50 will be included in subgroup analyses of the coprimary and key secondary endpoints.
- Subjects who are randomized but prematurely discontinued the study before assessment of the primary composite or coprimary endpoints will be considered as part of the ITT population.

- Subjects who die before assessment of the primary composite or coprimary endpoints will count as though they hypothetically continued in the study.
- Subjects who are lost to follow-up and are missing data for the primary composite or coprimary endpoints will count as though they hypothetically continued in the study.
- Subjects who are missing assessments or data due to the COVID-19 pandemic and are missing data for the primary composite or coprimary endpoints will count as though they hypothetically continued in the study.
- Subjects who are missing data for any other reason for the primary composite or coprimary endpoints will count as though they hypothetically continued in the study.

The secondary efficacy analysis of secondary endpoints (defined in Section 3.2) will be analyzed based on the same populations as above with similar methods for addressing potential intercurrent events.

9.0 Efficacy Analyses

9.1 General Considerations

All efficacy analyses will be conducted in the ITT Population for the US and EU regulatory agencies.

Analyses for the DB period will be conducted using a gated hierarchical testing procedure to preserve a familywise Type I error rate of $\alpha = 0.05$. DB treatment groups will be AGN-151586 and placebo. Analyses for the OL period will be performed outside of the testing hierarchy. Treatment groups reported for the OL period will include both the DB and OL period treatment group combinations (i.e., AGN-151586/AGN-151586, AGN-151586/None, Placebo/AGN-151586, Placebo/None).

For the DB period, to control the family-wise Type I error rate at 0.05 for multiplicity across the primary and secondary analyses, a hierarchical testing strategy[1] will be used. The hierarchical testing strategy starts with the primary efficacy endpoint followed by

sequential testing of the secondary endpoints, in the order listed for the US and EU regulatory agencies, respectively (see Section 13.0). Statistical significance will be evaluated in order. If statistical significance at the 0.05 alpha level (2-sided testing) is shown for the first variable in the ranking order, then the next variable in the immediate subsequent order will be evaluated for statistical significance. If statistical significance is shown for the second variable in the ranking order, then the analysis will continue in the same manner for subsequent variables. If no statistical significance is shown at the 0.05 alpha level for any variable in the ranking order, then that variable and all subsequent variables will not be considered statistically significant, regardless of their p-values. Nominal p-values will then be provided.

The Primary Analysis will be performed after all ongoing subjects have completed both the DB period and the OL period and the database has been locked. This will be the only and final analysis for the composite, coprimary and secondary efficacy endpoints, as well as all other efficacy endpoints in the DB period. Descriptive analyses will be provided for the OL period measures over time.

Unless otherwise specified, categorical variables will be analyzed using Cochran-Mantel-Haenszel (CMH) test, [redacted]. Continuous variables will be analyzed using analysis of covariance (ANCOVA) [redacted].

A subject who is randomized based on a wrong stratum will be analyzed according to the subject's actual stratum. However, if a subject has a baseline FWS lower than moderate severity at maximum frown, then the subject will be included in stratified analyses within the moderate GL severity at maximum frown stratum. In analyses not stratified by baseline FWS, the observed baseline FWS will be used.

For efficacy analyses of FWS assessments, "baseline" refers to the last non-missing observation before either the first administration of study drug or randomization if no study drug is given or the administration of study drug during retreatment. The change

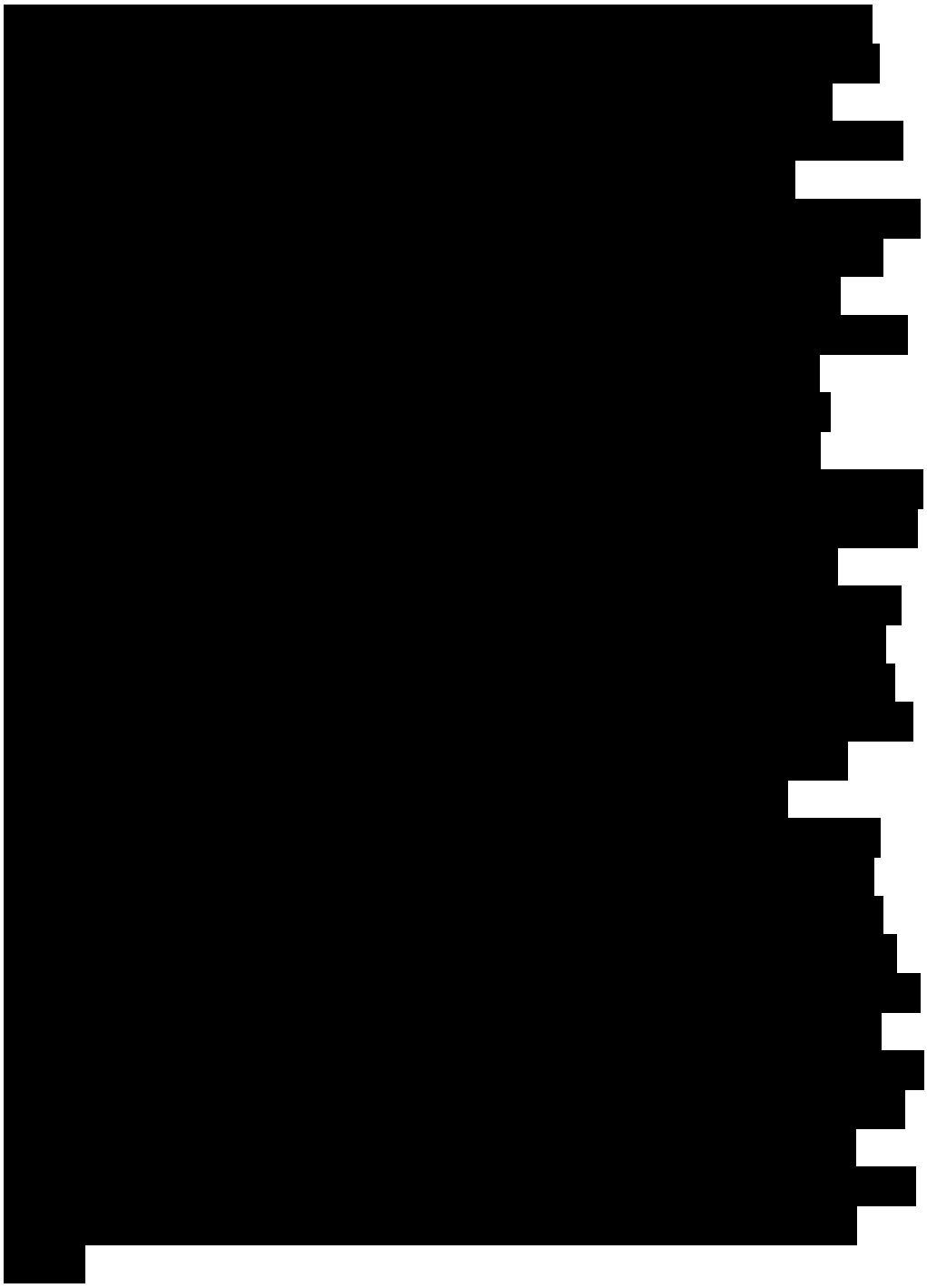
from baseline values will be computed as the post-baseline value minus the observed baseline value.

[redacted]

9.2 Handling of Missing Data

Missing data will be imputed for the DB Period only using the following methods for the efficacy analyses:

- Multiple Imputation: The Multiple Imputation approach will be used as a primary analysis for the primary endpoint, as well as for all secondary endpoints. [redacted]



- Non-responder Imputation (NRI): The NRI analysis will categorize any subject who does not have evaluation during a specific visit window as a non-responder for that visit. The NRI will be one of the sensitivity analysis approaches in the analyses of categorical variables for the primary efficacy endpoints.
- As Observed (AO): The AO analysis will not impute values for missing evaluations, and thus a subject who does not have an evaluation on a scheduled visit will be excluded from the AO analysis for that visit. AO will include all values collected in the study and will be used for sensitivity analyses for the primary efficacy endpoints.

9.3 Primary Efficacy Endpoints and Analyses

9.3.1 Primary Efficacy Endpoints

The assessments of the severity of GL at maximum frown using the validated FWS is based on the following scale (for both investigator and subject):

- 0 = None
- 1 = Mild
- 2 = Moderate
- 3 = Severe

The primary composite endpoint for FDA is the following:

- A Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to both investigator and subject assessments of GL severity at maximum frown at Day 7

The coprimary endpoints for EU regulatory agencies are the following:

- ≥ 2-grade improvement from baseline on the FWS according to subject assessment of GL severity at maximum frown at Day 7, and

- ≥ 2-grade improvement from baseline on the FWS according to investigator assessment of GL severity at maximum frown at Day 7.

9.3.2 Main Analysis of Primary Efficacy Endpoints

The attributes of the estimands corresponding to the primary efficacy endpoints for the US FDA and EU regulatory agencies are summarized in Table 1. In addition, the attribute of treatment is a single dose of AGN-151586 or placebo.

US FDA: The primary composite responder endpoint for a Grade 0 or 1 (none or mild) and at least a 2-grade improvement will be analyzed using the CMH method [redacted]. The primary composite endpoint must meet $p \leq 0.05$ to be considered successful.

EU regulatory agencies: The coprimary responder endpoints for ≥ 2-grade improvement from baseline in the FWS will be analyzed separately using the CMH method [redacted]. Both coprimary endpoints must meet $p \leq 0.05$ to be considered successful.

The evaluation of the equality of the proportions of responders for the primary composite or coprimary endpoints at Day 7 will be based on the CMH test [redacted]. The 95% confidence interval for the treatment responder rate differences will be constructed using the normal approximation to the binomial distribution.

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

Estimand Label	Attributes of the Estimand			
	Variable (Endpoint)	Population	Handling of Intercurrent Events	Statistical Summary
Hypothetical estimand for primary composite endpoint (US FDA)	Achievement of a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to both the investigator and subject assessments (composite) of GL severity at maximum frown on Day 7	Intent-to-treat (ITT) (all randomized)	Subjects who discontinue study prior to Day 7 assessments, or who do not have FWS assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FWS assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; Cochran-Mantel- Haenszel (CMH) test [redacted]
Hypothetical estimand for coprimary endpoints (EU)	Achievement of at least a 2-grade improvement from baseline on the FWS according to: • Subject's assessment of GL severity at maximum frown at Day 7, and • Investigator's assessment of GL severity at maximum frown at Day 7	ITT	Subjects who discontinue study prior to Day 7 assessments, or who do not have FWS assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FWS assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted]

CMH = Cochran-Mantel-Haenszel; EU = European Union; FDA = Food and Drug Administration; FWS = Facial Wrinkle Scale; GL = Glabellar lines; ITT = intent-to-treat; US = United States

9.3.3 Sensitivity and Supplementary Analyses of the Primary Efficacy Endpoints

Sensitivity analyses of the primary efficacy variables will be performed to establish their consistency and robustness as well as to further characterize the extent of subjects'

responses. As-observed data analysis, as well as missing as non-responder (using NRI) analysis, [redacted] for sensitivity analyses.

[redacted]

[redacted]

9.4 Secondary Efficacy Endpoints and Analyses

9.4.1 Secondary Efficacy Endpoints

Analysis of the secondary endpoints will be conducted on the ITT population based on treatment as randomized for the US FDA and EU regulatory agencies.

Secondary efficacy endpoints for the US FDA and EU regulatory agencies are listed in Section 3.2, and include other response definitions for the FWS by investigator and subject, as well as for the FLO-11, FLSQ, and GAC-GL.

The FLO-11, FLSQ, and GAC-GL instruments are described below.

FLO-11: The FLO-11 is a validated 11-item COA instrument to assess emotional and appearance-related impacts of upper facial lines from a subject's perspective. Each item is on an 11-point numeric rating scale that ranges from 0 (not at all) to 10 (very much) with a middle label of somewhat. For scoring, all items are transformed to a 0 to 100 scale, with higher scores indicating better outcomes. Total score is calculated as the sum of the reverse score (i.e., 10 minus the original score) of Items 1 to 10 plus the raw score of Item 11 (Item 11 does not need to be a reverse score). Total scores are transformed to a 0 (worst) to 100 (best) point scale by setting the lowest and highest possible scores at 0 and 100, respectively; these are the transformed scale scores. The 50% rule is applied for the total score, such that scales with at least 50% non-missing item responses will be scored as the average of the non-missing responses. If the total scale is missing more than 50% of the item responses, then the total scale score will be set as missing.

FLSQ: The FLSQ consists of 2 versions (baseline and follow-up versions) to assess study intervention (i.e., treatment) expectations, treatment satisfaction, and impact of hyper functional upper facial lines (LCL, forehead lines, and/or GL). The baseline version consists of 11 items (2 multi-item domains: treatment expectations and impact), while the follow-up version consists of 13 items (2 multi-item domains: treatment satisfaction, and impact; 3 single item domains: met treatment expectations, continue treatment, and recommend to others). Additionally, FLSQ follow-up items 2, 3, 4, 5, 9, and 11 are

validated for use as stand-alone items. [redacted]

[redacted]

GAC-GL: The GAC-GL is a 2-item measure that assesses the appearance of the subject's GL "now" in comparison with their perspective before treatment at rest and at maximum frown. The response options are scored on a 7-point numerical rating scale ranging from *Very much improved* to *Very much worse*. Scores are unique for at rest and maximum contraction, and they are simply the responses to the items.

[redacted]

[redacted]

9.4.2 Main Analyses of Secondary Efficacy Endpoints

The attributes of the estimands corresponding to the secondary efficacy endpoints are summarized in Table 2. Analysis of the secondary endpoints will be conducted on the ITT population based on treatment as randomized for the US FDA and EU regulatory agencies.

For the secondary endpoints, the proportion of responders will be analyzed using the CMH test [redacted]. The continuous variables will be analyzed using ANCOVA [redacted].

[redacted]

Table 2. Summary of the Estimand Attributes of the Secondary Efficacy Endpoints

	Attributes of the Estimand		
Estimand Label	**Variables (Endpoints)**	**Handling of Intercurrent Events**	**Statistical Summary**
Hypothetical estimand for key secondary categorical endpoints (US FDA)	Key (US FDA): Achievement of a Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline on the FWS according to: • both the investigator and subject assessments (i.e., concurrent composite) of GL severity at maximum frown over time (DB period)*, • investigator assessment of GL severity at maximum frown over time (DB period)*, • subject assessment of GL severity at maximum frown over time (DB period)*	Subjects who discontinue study in the DB period, or who do not have FWS assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FWS assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups at each time point; CMH test [redacted] Endpoints will be evaluated outside of the gated hierarchical testing.
Hypothetical estimand for secondary categorical endpoints (EU)	Achievement of at least a 2-grade improvement from baseline on the FWS according to: • investigator assessment of GL severity at maximum frown over time (DB period)*, • subject assessment of GL severity at maximum frown over time (DB period)*	Subjects who discontinue study in the DB period, or who do not have FWS assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FWS assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups at each time point; CMH test [redacted] Endpoints will be evaluated outside of the gated hierarchical testing.

Estimand Label	Attributes of the Estimand		
	Variables (Endpoints)	**Handling of Intercurrent Events**	**Statistical Summary**
Hypothetical estimand for secondary categorical endpoints (EU only)	Achievement of at least a 2-grade improvement from baseline in GL severity at maximum frown based on: • Subject assessment using FWS at Hour 24 • Investigator assessment using FWS at Hour 24	Subjects who discontinue study prior to Hour 24 assessments, or who do not have FWS assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FWS assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted] Gated hierarchical testing will be conducted.
Hypothetical estimand for secondary categorical endpoints (EU only)	Achievement of at least a 1-grade improvement from baseline in GL severity at maximum frown based on: • Subject assessment using FWS at Hour 24 • Investigator assessment using FWS at Hour 24	Subjects who discontinue study prior to Hour 24 assessments, or who do not have FWS assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FWS assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted] Gated hierarchical testing will be conducted.
Hypothetical estimand for secondary categorical endpoints (US FDA)	*Mostly satisfied* or *Very satisfied* on the Facial Lines Satisfaction Questionnaire (FLSQ) follow-up version Item 5 (overall satisfaction) for GL at Day 7	Subjects who discontinue study prior to Day 7 assessments for the FLSQ follow-up assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FLSQ assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted] Gated hierarchical testing will be conducted.

Estimand Label	Attributes of the Estimand		
	Variables (Endpoints)	**Handling of Intercurrent Events**	**Statistical Summary**
Hypothetical estimand for secondary categorical endpoints (US FDA; EU)	• *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL at Hour 24 • *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 4 (natural look) for GL at Day 7	Subjects who discontinue study prior to Hour 24 or Day 7 assessments for the FLSQ follow-up assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FLSQ assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted] Gated hierarchical testing will be conducted.
Hypothetical estimand for secondary categorical endpoints (EU)	*Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL over time*	Subjects who discontinue study or who do not have FLSQ follow-up version Item 5 will be included in the analysis as a hypothetical scenario in which they had not missed the FLSQ assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted] Endpoint will be evaluated outside of the gated hierarchical testing.
Hypothetical estimand for secondary categorical endpoints (EU only)	• Key: Achievement of at least a 20-point improvement from baseline in FLO-11 total scores for GL at Day 7 • Achievement of at least a 4-point improvement from baseline in FLO-11 Item 10 (look angry) for GL at Day 7 • Achievement of at least a 4-point improvement from baseline in FLO-11 Item 5 (look less attractive) for GL at Day 7	Subjects who discontinue study prior to Day 7 assessments or who do not have FLO-11 Day 7 assessments will be included in the analysis as a hypothetical scenario in which they had not missed the FLO-11 assessments	Response rates and rate differences between AGN-151586 and placebo treatment groups; CMH test [redacted] Gated hierarchical testing will be conducted.

Estimand Label	Attributes of the Estimand		
	Variables (Endpoints)	Handling of Intercurrent Events	Statistical Summary
Hypothetical estimand for secondary categorical endpoints (EU only)	• Subject-reported global assessment of change in GL based on the GAC-GL over time*	Subjects who discontinue study or who do not have the GAC-GL assessments will be included in the analysis as a hypothetical scenario in which they had not missed the GAC-GL assessments	Mean and mean differences between AGN-151586 and placebo treatment groups; ANCOVA [redacted] Endpoints will be evaluated outside of the gated hierarchical testing.
Hypothetical estimand for secondary continuous endpoints (EU only)	• Time to the first ≥1-grade improvement from baseline on the FWS according to subject assessment (DB period)* • Time to the first ≥1-grade improvement from baseline on the FWS according to investigator assessment (DB period)* • Time to return to baseline FWS criterion† according to subject assessment of FWS (DB period)* • Time to return to baseline FWS criterion† according to investigator assessment of FWS (DB period)*	Subjects who discontinue study or do not have FWS assessments will be included in the analysis as a hypothetical scenario in which the assessments were conducted as if they had not missed the FWS assessments	[redacted] Endpoints will be evaluated outside of the gated hierarchical testing.

ANCOVA = analysis of covariance; CMH = Cochran-Mantel-Haenszel; EU = European Union; FDA = Food and Drug Administration; FLO-11 = 11-item Facial Line Outcomes; FLSQ = Facial Line Satisfaction Questionnaire; FWS = Facial Wrinkle Scale; GAC-GL = Global Assessment of Change in Glabellar Lines; GL = Glabellar lines; US = United States

* Endpoints will be excluded from hierarchical testing.

[redacted]

9.4.3 Sensitivity and Supplementary Analyses for Secondary Efficacy Endpoints

Additional sensitivity analyses may be performed as needed to assess impact of missing data for secondary endpoints due to a pandemic or natural disaster.

9.4.4 Supportive Secondary Efficacy Endpoints and Analyses

Not applicable.

9.5 Additional Efficacy Analyses

For both the US FDA and EU regulatory agencies, analyses for additional efficacy endpoints will be performed on the ITT population using data as observed. The additional efficacy endpoints are listed in Section 3.3.

9.6 Efficacy Subgroup Analyses

For the primary and key secondary endpoints, analyses will be conducted using the following subgroups and descriptive summary statistics provided:

Analyses will be conducted using the following subgroups:

- Sex (Female, Male)
- Age group (<65, ≥65 years)
- Race (White, non-White)
- Ethnicity (Hispanic or Latino, not Hispanic or Latino)
- Region (North America, EU)
- [redacted]
- Baseline FWS at maximum frown (moderate, severe)
- Baseline FLO-11 total score (transformed) ≤ 50
- Fitzpatrick skin type (Types I & II, Types III & IV, Types V & VI)

10.0 Safety Analyses

10.1 General Considerations

Safety data will be summarized for the Safety Population. Safety summaries will be presented by treatment group, by overall study, by study period, and by treatments received in the DB Period and in the OL Period (i.e., AGN-151586/AGN-151586, AGN-151586/None, Placebo/AGN-151586, Placebo/None).

For the safety analysis, subjects are assigned to a treatment group based on the treatment actually received, regardless of the treatment randomized.

The safety parameters will include AEs, vital signs, 12-lead ECGs, laboratory evaluations (hematology and chemistry), neurological assessment (including examination of cranial nerves II through VII), and presence of antidrug antibodies. For change from baseline safety parameters, "baseline" refers to the last non-missing observation before the first study drug administration in the study.

10.2 Adverse Events

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

10.2.1 Treatment-Emergent Adverse Events

Treatment-emergent AEs (TEAEs) are defined as any AE with the onset that is after the first dose of study drug. Events where the onset date is the same as the study drug start date are assumed to be treatment-emergent, unless known to have started prior to study

treatment administration. All TEAEs will be summarized overall, as well as by primary MedDRA SOC and PT. The SOCs will be presented in alphabetical order, and the PTs will be presented in alphabetical order within each SOC. The number and percentage of subjects experiencing treatment-emergent AEs will be summarized.

In addition, for both possible distant spread of toxin (PDSOT) TEAEs and treatment-related TEAEs, onset after study drug administration and duration will be evaluated. TEAE onset will be reported by number of days subsequent to the most recent study drug administration and will be summarized with descriptive statistics. TEAE duration will be defined as ((end date minus start date) + 1 day) and will be summarized with descriptive statistics.

10.2.2 Adverse Event Overview

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

- Any TEAE
- Any TEAE related to study treatment according to the investigator
  - Any TEAE related to study drug according to the investigator
  - Any TEAE related to study procedure according to the investigator
- Any mild TEAE
  - Any mild TEAE related to study treatment according to the investigator
- Any moderate TEAE
  - Any moderate TEAE related to study treatment according to the investigator
- Any severe TEAE
  - Any severe TEAE related to study treatment according to the investigator
- Any serious TEAE
  - Any serious TEAE related to study treatment according to the investigator
- Any TEAE leading to discontinuation of study drug
- Any TEAE leading to death

- Any PDSOT TEAE
- Any neurological assessment associated TEAE
- All deaths

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

TEAEs will be summarized by SOC and PT; TEAEs considered related to study treatment as assessed by the investigator (i.e., reasonable possibility) by SOC and PT; by maximum severity by SOC and PT; and listed by subject number by SOC and PT. Specific adverse events will be counted once for each subject for calculating percentages, unless stated otherwise. In addition, if the same adverse event occurs multiple times within a subject, the adverse event will be counted only once and the highest severity and level of relationship to study drug will be reported.

In addition, TEAEs will be summarized by PT and sorted by decreasing frequency for the total active group after first dose of AGN-151586. Frequent TEAE incidence will be reported for PTs with ≥5% subject incidence, listed by SOC and PT.

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

Not applicable.

10.2.5 SAEs (Including Deaths) and Adverse Events Leading to Study Drug Discontinuation

Serious AEs (including deaths) and AEs leading to study drug discontinuation will be summarized by SOC and PT and in listing format.

10.2.6 Possible Distant Spread of Toxin Adverse Events

PDSOT TEAEs will be summarized by PT. The PTs [redacted] will be based on the current MedDRA version used for the study prior to database lock.

Tabular listings of PDSOT TEAEs will be provided.

10.2.7 Neurological Assessment Adverse Events

The neurological assessment, which is comprised of a Focused Symptoms Questionnaire and a Focused Neurological Examination, will be conducted at selected study visits per protocol. Neurological assessment-associated TEAEs will be summarized by PT. The PTs will be based on the current MedDRA version prior to database lock.

Tables and listings of AEs related to the neurological assessment will be provided.

10.3 Analysis of Laboratory Data

Laboratory evaluations consisting of chemistry and hematology, and changes from baseline at each assessment timepoint will be summarized by treatment group and study period. In addition, potentially clinically important (PCI) laboratory values will be summarized.

Each laboratory value will be summarized for all time points (starting with Baseline) with the number of non-missing observations, mean and standard deviation, median, minimum, and maximum. Mean change from baseline to each applicable post-baseline visit will be summarized for each laboratory parameter, with the number of observations, baseline mean, and visit mean. The change from baseline mean, standard error, and 95% confidence interval will be presented for the mean change from baseline within each treatment group.

Changes in laboratory values will be tabulated using shift tables categorized as low, normal, or high based on the normal ranges of the laboratory used for each sample [redacted].

Laboratory values will be evaluated [redacted]. For each laboratory PCI criterion, the number and percentage of subjects who have a post-baseline laboratory value meeting the criteria will be summarized. [redacted]

[redacted].

Listings will be provided to summarize subject-level laboratory data for subjects meeting PCI criteria.

10.4 Analysis of Vital Signs

Vital sign measurements of systolic and diastolic blood pressure, pulse rate, and temperature, and changes from baseline at each assessment timepoint will be summarized by treatment group and study period. In addition, potentially clinically important vital signs values will be summarized.

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

Vital sign variables will be evaluated based on PCI criteria [redacted]

[redacted]

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

10.5 Safety Subgroup Analyses

Not applicable.

10.6 Other Safety Analyses

Electrocardiogram (ECG): Descriptive statistics for quantitative ECG parameters (i.e., heart rate, QRS duration, QTcB interval, QTcF interval, RR interval, and PR interval) at baseline, post-baseline, and changes from baseline values at each post-baseline timepoint will be presented.

ECG parameters are considered PCI if ECG values meet either the actual value or change from baseline PCI high criteria [redacted]. The number and percentage of subjects with PCI post-baseline values will be tabulated. [redacted]

A supportive listing of subjects with PCI post-baseline values will be provided.

Immunogenicity Assessment: Immunogenicity results, manifested as the presence of binding antibodies (positive) and neutralizing antibodies (positive) to AGN-151586, will be summarized in a table for baseline and post-baseline data. Immunogenicity findings (positive) will be tabulated with the number and percentage of subjects at each visit separately for each treatment group. Percentages will be based on the number of treated subjects with interpretable antibody assays in each treatment group at the specified visit.

11.0 Other Analyses

Not applicable.

12.0 Interim Analyses

No interim analysis is planned for this study.

12.1 Data Monitoring Committee

No data monitoring committee (DMC) will be used for this study.

13.0 Overall Type-I Error Control

For the DB period, to control the family-wise Type I error rate at 0.05 for multiplicity across the primary and secondary analyses, a hierarchical testing strategy[1] will be used. The hierarchical testing strategy starts with the primary endpoint followed by sequential testing of the secondary endpoints. The statistical significance will be evaluated in the following order: if a statistical significance at the 0.05 level (2-sided) is shown for the first endpoint in the ranking order, then the next endpoint in the immediate subsequent order will be evaluated; evaluation of subsequent endpoint will continue in the same manner. If no statistical significance is shown at $\alpha = 0.05$ at any endpoint, then the endpoint and all subsequent endpoints will not be considered statistically significant, regardless of their nominal p-values.

The hierarchical testing order follows for US FDA:

1. Primary: A Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline based on the following:
    - **Investigator and subject assessments** (composite) of GL severity at maximum frown at Day 7
1. Secondary: *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL at Day 7

2. Secondary: *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL at Hour 24
3. Secondary: *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 4 (natural look) for GL at Day 7

For the US FDA, the following secondary endpoints will be evaluated outside of the gated hierarchical testing sequence:

1. Key Secondary: A Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline in GL severity at maximum frown according to the **investigator and subject assessments** (i.e., concurrent composite) using the FWS over time (DB period)
2. Key Secondary: A Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline in GL severity at maximum frown according to the **investigator assessment** using the FWS over time (DB period)
3. Key Secondary: A Grade 0 or 1 (none or mild) and at least a 2-grade improvement from baseline in GL severity at maximum frown according to the **subject assessment** using the FWS over time (DB period).

The hierarchical testing order follows for EU regulatory agencies:

1. Coprimary: At least a 2-grade improvement based on the following:
   - **Subject assessment** using FWS at maximum frown at Day 7, and
   - **Investigator assessment** using FWS at maximum frown at Day 7.
2. Key Secondary: At least 20-point improvement from baseline in FLO-11 total scores for GL at Day 7
3. Secondary: At least 2-grade improvement from baseline in GL severity at maximum frown according to the **subject assessment** using the FWS at Hour 24

4. Secondary: At least 2-grade improvement from baseline in GL severity at maximum frown according to the **investigator assessment** using the FWS at Hour 24
5. Secondary: At least 1-grade improvement from baseline in GL severity at maximum frown according to the **subject assessment** using the FWS at Hour 24
6. Secondary: At least 1-grade improvement from baseline in GL severity at maximum frown according to the **investigator assessment** using the FWS at Hour 24
7. Secondary: *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) for GL at Hour 24
8. Secondary: *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 4 (natural look) for GL at Day 7
9. Secondary: At least 4-point improvement from baseline in FLO-11 Item 10 (look angry) for GL at Day 7
10. Secondary: At least 4-point improvement from baseline in FLO-11 Item 5 (look less attractive) for GL at Day 7

For the EU, the following secondary endpoints will be evaluated outside of the gated hierarchical testing sequence:

1. Secondary: Time to the first ≥ 1-grade improvement from baseline on the FWS according to subject assessments of GL severity at maximum frown (DB period)
2. Secondary: Time to the first ≥ 1-grade improvement from baseline on the FWS according to investigator assessments of GL severity at maximum frown (DB period).
3. Secondary: At least 2-grade improvement from baseline on the FWS according to subject assessments of GL severity at maximum frown over time (DB period)

4. Secondary: At least 2-grade improvement from baseline on the FWS according to investigator assessments of GL severity at maximum frown over time (DB period)
5. Secondary: Time to return to baseline FWS criterion† according to subject assessments of FWS at maximum frown (DB period)
6. Secondary: Time to return to baseline FWS criterion† according to investigator assessments of FWS at maximum frown (DB period)
7. Secondary: *Mostly satisfied* or *Very satisfied* on the FLSQ follow-up version Item 5 (overall satisfaction) over time
8. Secondary: Subject-reported global assessment of change in GL based on the GAC-GL over time

[redacted]

14.0 Version History

Table 3. SAP Version History Summary

Version	Date	Summary
1.0	20 May 2021	Original version
2.0	15 February 2022	Updated study design, primary and secondary efficacy endpoints and estimand attributes for US and EU regulatory agencies, removed mITT population Updated hierarchical testing sequences with included/excluded endpoints [redacted] Added safety assessments for neurological assessment, ECG assessments, and laboratory assessments Updated sample sizes Added PCI criteria for laboratory, vital signs, and ECG
3.0	15 November 2022	Updated key secondary, secondary, and other efficacy endpoints and estimand attributes for US and EU regulatory agencies Updated specifications for the multiple imputation analysis method Updated efficacy analyses to include GAC-GL and responder definition [redacted] Clarified baseline definition and retreatment baseline definition for FWS-related efficacy analyses and safety analyses Clarified that FLO-11 total scores refer to the transformed scores Updated safety analyses to include TEAE onset and duration analyses, and frequent TEAE analyses Updated PDSOT terms and PCI criteria in appendices
4.0	23 November 2022	Updated primary and key secondary endpoints and estimands for US: Grade 0 or 1 (none or mild) and at least a 2-grade improvement

Version	Date	Summary
5.0	08 April 2023	Removed gatekeeping strategy from sample size section [redacted] Clarified that the 95% confidence interval would be based on the responder rate differences [redacted] Removed FWS at baseline from the multiple imputation model Added other efficacy analyses to be performed for other efficacy endpoints Updated baseline characteristics to include FWS at maximum frown as assessed by investigator and by subject Removed vital signs shift tables Updated MedDRA versioning to be the version used for the study at the time of database lock [redacted]

15.0 References

1. Dmitrienko A, Molenberghs G, Chuang-Stein C, et al. Analysis of Clinical Trials Using SAS: A Practical Guide. SAS Institute, Cary, North Carolina. 2005;104-8.

Appendix A. Protocol Deviations

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

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