Official Title:	A Phase III, Randomized, Double-Blind, Placebo-Controlled, Multicenter Study to Evaluate the Efficacy and Safety of Etrolizumab as an Induction And Maintenance Treatment For Patients With Moderately to Severely Active Crohn's Disease
NCT Number:	NCT02394028
Document Date:	SAP Version 3: 20-Sep-2021

STATISTICAL ANALYSIS PLAN

TITLE: **A PHASE III, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, MULTICENTER STUDY TO EVALUATE THE EFFICACY AND SAFETY OF ETROLIZUMAB AS AN INDUCTION AND MAINTENANCE TREATMENT FOR PATIENTS WITH MODERATELY TO SEVERELY ACTIVE CROHN'S DISEASE**

PROTOCOL NUMBER: GA29144

STUDY DRUG: Etrolizumab (RO5490261)

VERSION NUMBER: 3

IND NUMBER: 119725

EUDRACT NUMBER: 2014-003824-36

SPONSOR: F. Hoffmann-La Roche Ltd

PLAN PREPARED BY: [redacted] M.Sc.
[redacted] Ph.D.

DATE FINAL: Version 1: 19 December 2019

DATE AMENDED: Version 2: 14 December 2020
Version 3: See electronic date stamp below

STATISTICAL ANALYSIS PLAN AMENDMENT APPROVAL

Date and Time(UTC)	Reason for Signing	Name
20-Sep-2021 16:30:29	Company Signatory	[redacted]

STATISTICAL ANALYSIS PLAN AMENDMENT RATIONALE

- Added Section 2.2 (Trial Objective):
  - Section added to clearly define trial objective for estimands
- Changes to Section 2.3.2 (Secondary Efficacy Outcome Measures):

– Added Crohn's Disease Activity Index (CDAI) remission as secondary endpoint (not included in protocol)

- Changes to Section 2.4.1 (Sample Size):
  - Included statement that Cohort 3 Induction closed early.
- Changes to Section 4.4.3 (Intercurrent Events):
  - Clarified wording for COVID-19 specific Intercurrent Events (ICEs)
  - Added treatment strategy
- Changes to Sections 4.4.4.1 and 4.4.4.2 (Induction Co-Primary Efficacy Endpoints) and Section 4.4.9 (Maintenance Co-Primary Efficacy Endpoints):
  - Clarified wording for coronavirus disease of 2019 (COVID-19) specific ICEs between endoscopy and clinical remission
  - Added treatment strategy
- Changes to Section 4.4.6 (Sensitivity Analyses for Induction Co-Primary Endpoints) and Section 4.4.11 (Sensitivity Analyses for Maintenance Co-Primary Endpoints):
  - Added tipping point analysis and included Multiple Imputation (MI) as sensitivity for clinical remission (further details added in Section 4.10)
- Changes to Section 4.4.8 (Induction Secondary Efficacy Endpoints) and Section 4. 4.13 (Maintenance Secondary Efficacy Endpoints):
  - Added treatment strategy for estimand definitions
  - Clarified handling approach for COVID-19 related ICE for CDAI and clinical remission endpoints
- Changes to Section 4.4.12 (Supplementary Analyses):
  - Added supplementary analysis which changes population attribute in primary estimand i.e., perform analysis among CDAI-100 responders at Week 14
- Removed Section 4.4.14.8 (Histology) from Statistical Analysis Plan (SAP) (and Clinical Study Report [CSR])
- Changes to Section 4.9.3 (Adverse Events):
  - Added long COVID analyses
- Changes to Section 4.10 (Missing Data):
  - Added additional wording and section 4.10.2 on a "well-justified" imputation model and strategy as per FDA feedback. Individual sections have also been added for each endpoint (4.10.2.1, 4.10.2.2 and 4.10.2.3)

- Added multiple imputation as sensitivity analysis (Section 4.10.1) for CDAI and clinical remission endpoints
- Added wording on performing tipping point analysis (Section 4.10.1)

Additional minor changes have been made to improve clarity and consistency.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

GLOSSARY OF ABBREVIATIONS

Abbreviation	Definition
AE	adverse event
AESI	adverse event of special interest
ALT	alanine aminotransferase
AP	abdominal pain
AST	aspartate aminotransferase
CCOD	clinical cutoff date
CD	Crohn's Disease
CDAI	Crohn's Disease Activity Index
CD-PRO/SS	Crohn's Disease- Patient Reported Outcome/ Signs and Symptoms
CMH	Cochran-Mantel-Haenszel
COVID-19	Coronavirus Disease 2019
CRP	C-reactive protein
CS	corticosteroids
CSR	Clinical Study Report
CTCAE	Common Terminology Criteria for Adverse Events
[redacted]	[redacted]
ICE	intercurrent event
iDMC	independent Data Monitoring Committee
IS	immunosuppressant
IxRS	interactive voice/Web based response system
LOCF	last-observation-carried-forward
MAR	missing at random
MedDRA	Medical Dictionary for Regulatory Activities
MI	multiple imputation
MICE	multivariate imputation by chained equations
mITT	modified intent to treat
MMRM	mixed models for repeated measures
NRI	non-responder imputation
PD	pharmacodynamics

Abbreviation	Definition
PK	pharmacokinetic
Q4W	every 4 weeks
SAE	serious adverse event
SAP	statistical analysis plan
SC	Subcutaneous
SES-CD	simplified endoscopic score for Crohn’s disease
SF	stool frequency
SOC	system organ class
TNF	tumor necrosis factor
ULN	upper limit of normal
WOCF	worst-observation-carried-forward

1. BACKGROUND

Study GA29144 (IND119725) is a multicenter, Phase III, double-blind, placebo-controlled study evaluating the efficacy, safety, and tolerability of etrolizumab compared with placebo during Induction and Maintenance treatment phases for patients with moderate to severely active Crohn's disease (CD).

Patients are enrolled sequentially into one of the three Induction cohorts, as described below:

- Cohort 1 (N=300): placebo, etrolizumab 105 mg, etrolizumab 210 mg (1:2:2)
- Cohort 2 (N=350): etrolizumab 105 mg, etrolizumab 210 mg (1:1)
- Cohort 3 (N=496): placebo, etrolizumab 105 mg, etrolizumab 210 mg (2:3:3)

Cohorts 1 and 2 are exploratory Induction cohorts, whilst Cohort 3 serves as the pivotal, confirmatory cohort for the Induction phase analysis.

Patients achieving Crohn's Disease Activity Index−70 (CDAI-70) responses at Week 14, without the use of rescue treatment are eligible to enter the Maintenance Phase. Patients who received etrolizumab (high or low dose) during the Induction Phase will be randomized into the Maintenance Phase in a 1:1 ratio to placebo or etrolizumab 105 mg subcutaneous (SC) every 4 weeks (Q4W). To maintain the blind, patients receiving placebo during the Induction Phase will undergo a sham randomization, and will receive blinded placebo treatment during the Maintenance Phase.

The Maintenance Phase analysis will include patients from all three Induction cohorts.

This Statistical Analysis Plan (SAP) describes the data and analysis methods that will be reported in the GA29144 final Clinical Study Report (CSR). The following data will be evaluated independently:

- Induction Phase – Cohorts 1, 2, and 3
- Maintenance Phase

Data will be reported based on the final protocol and final database, using the analysis methods as described in this SAP.

It is noted that Cohort 1 data were unblinded and analyzed in 2017 to inform the design of the pivotal Induction cohort, prior to its commencement. Topline results from this preliminary analysis will be summarized in the CSR.

2. STUDY DESIGN

Details of the study design can be found in the Protocol Synopsis section of Protocol Version 7 and Study Schema in Figure 1 and Figure 2 of the Protocol Version 7.

2.1 PROTOCOL SYNOPSIS

The Protocol Synopsis is in Protocol Synopsis Section of Protocol Version 7. For additional details, see the Schedule of Assessments in Appendices 1A, 1B, 2 and 3 of the Protocol Version 7.

2.2 TRIAL OBJECTIVE

Primary efficacy objectives for this study will be analyzed separately for the Induction and Maintenance Phases as outlined below:

- Induction Phase: To independently evaluate the efficacy of etrolizumab dose regimens compared with placebo in inducing clinical remission and endoscopic improvement at the end of the Induction Phase (Week 14)
- Maintenance Phase: To independently evaluate the efficacy of etrolizumab compared with placebo in achieving clinical remission and endoscopic improvement at 1 year of maintenance treatment (Week 66), for patients who achieved a Crohn's Disease Activity Index (CDAI)-70 response (defined as a decrease of at least 70 points from baseline CDAI) at Week 14, without the use of rescue therapy.

2.3 OUTCOME MEASURES

Full definitions of all the outcome measures to be analyzed during the Induction and Maintenance Phases, as specified in the protocol, are included in Appendix 1. [redacted] However, as a result of evolving regulatory guidelines, CDAI remission is now considered as a secondary endpoint for the Induction and Maintenance Phases.

2.3.1 Co-Primary Efficacy Outcome Measures

Induction:

- Clinical remission (Stool Frequency [SF] mean daily score≤3 and Abdominal Pain (AP) mean daily score≤1, with no worsening in either sub-score compared to baseline, averaged over the 7 days prior to visit) at Week 14
- Endoscopic improvement (≥50% reduction from baseline in the Simplified Endoscopic Score for Crohn's Disease [SES-CD]) at Week 14

Maintenance:

- Clinical remission at Week 66, for patients achieving a CDAI-70 response (decrease of at least 70 points from baseline CDAI) at Week 14
- Endoscopic improvement at Week 66, for patients achieving a CDAI-70 response at Week 14

2.3.2 Secondary Efficacy Outcome Measures

Induction:

- Clinical remission at Week 6

- CDAI remission at Week 14
- Endoscopic remission (see Appendix 1) at Week 14
- Change from baseline to Week 14 in CD signs and symptoms, as assessed by the Crohn's Disease-Patient Reported Outcome/Signs and Symptoms (CD-PRO/SS) functional domain score
- Change from baseline to Week 14 in CD signs and symptoms, as assessed by the CD-PRO/SS bowel domain score

Maintenance:

As above, all the following are for patients achieving a CDAI-70 response at Week 14 unless otherwise stated:

- Clinical remission at Week 66 among patients achieving clinical remission at Week 14
- CDAI remission at Week 66
- Corticosteroid-free clinical remission at Week 66 among patients who were receiving corticosteroids at baseline
- Endoscopic improvement at Week 66 among patients achieving endoscopic improvement at Week 14
- Endoscopic remission (see Appendix 1) at Week 66
- Durable clinical remission (see Appendix 1)
- Corticosteroid-free clinical remission at Week 66 (off corticosteroid for at least 24 weeks prior to Week 66) in patients who were receiving corticosteroids at baseline
- Change from baseline to Week 66 in CD-PRO/SS bowel domain score
- Change from baseline to Week 66 in CD-PRO/SS functional domain score

2.4 DETERMINATION OF SAMPLE SIZE

2.4.1 Sample Size

The planned sample size and power for each phase specified in the protocol are presented in Table 1, 2 and 3.

Table 1 Planned Sample Sizes for Induction Phase

Cohort	No. of Patients			
	Total	Placebo	etrolizumab 105mg	etrolizumab 210mg
Cohort 1 (Exploratory)	300	60	120	120
Cohort 2 (Exploratory)	350	NA	175	175
Cohort 3 (Pivotal)	496	124	186	186

Table 2 Planned Sample Sizes for Maintenance Phase

	No. of Patients			
	Total	Placebo/ Placebo	etrolizumab[a] / Placebo	etrolizumab[a] / etrolizumab 105 mg
Maintenance Phase	480	~60	~210	~210

[a] Patients randomized to either dose of etrolizumab (i.e., 105 mg or 210 mg) during Induction Phase.

Table 3 Planned Power Estimates for Primary and Key Secondary Efficacy Analyses in Induction Cohort 3 and Maintenance Phase

Study	Endpoint	Power	Assumed Response Rates (placebo vs. active)
Induction (Cohort 3)	Clinical Remission at Week 14	≥85%	15% vs. 30%
	Endoscopic Improvement at Week 14	≥80%	5% vs. 15%
Maintenance	Clinical Remission at Week 66	≥90%	≤ 20% vs. ≤ 35%
	Endoscopic Improvement at Week 66	~90%	≤ 30% vs. = 45%
	Clinical Remission at Week 66 among patients achieving clinical remission at Week 14	80%	30% vs. 50% Based on assumption that 22% patients will achieve clinical remission at Week 14

At the time of study closure a total of 1035 patients were randomized into the Induction Phase across cohorts 1-3, and the sample size of N=480 for Maintenance Phase was met. The protocol specified sample size for Cohort 3 for the pivotal Induction Phase was not met as the study was closed at N =385. Details surrounding the early closure and impact on power will be discussed in the final CSR.

2.5 ANALYSIS TIMING

The trial will be formally unblinded, analyzed and reported in the final CSR after the last patient has completed or withdrawn early from the Induction or Maintenance Phase, whichever comes first.

3. STUDY CONDUCT

3.1 RANDOMIZATION

An independent interactive voice/Web-based response system (IxRS) vendor was used to perform all kit assignments, using a stratified permuted block randomization method.

An independent unblinded statistician from an independent data coordinating center (iDCC) performed periodic reviews to confirm that the randomization schema has been implemented correctly for the study.

3.1.1 Induction Phase

Following screening and baseline assessments, patients who fulfil the inclusion/exclusion criteria to enter the study will be enrolled sequentially into one of the three Induction Cohorts. Randomization into each cohort in the Induction Phase is stratified by:

- Concomitant oral corticosteroids (CS) treatment (yes vs. no)

- Concomitant immunosupressants (IS) treatment (yes vs. no)
- Baseline CDAI $\leq$ 330 (yes vs. no)
- Prior anti-tumor necrosis factor (TNF) exposure (yes vs. no)

A stratified permuted block randomization method will be used.

3.1.2 Maintenance Phase

Patients are only eligible to enter the Maintenance Phase if they achieved a CDAI-70 response at Week 14 without the use of rescue therapy. Patients in Cohort 1 or Cohort 3 who were randomized to placebo during the Induction Phase will undergo a sham randomization in order to maintain the blind; these patients will receive placebo during the Maintenance Phase. Patients in Cohort 1-3 who were randomized to etrolizumab (either 105 mg or 210 mg dose) in Induction Phase will be re-randomized into the Maintenance Phase, where randomization is stratified by:

- CDAI Remission at both Week 10 and Week 14 (yes vs. no)
- Induction Dose Regimen (high vs. low)
- Concomitant oral CS (yes vs. no)
- Prior anti-TNF exposure (yes vs. no)

A stratified permuted block randomization method will be used.

3.2 INDEPENDENT REVIEW FACILITY

An independent review facility will provide an independent central review of ileocolonoscopy videos to determine the SES-CD. Details of the review process are described in the central review charter.

3.3 DATA MONITORING

An independent Data Monitoring Committee (iDMC) was incorporated in the study design to monitor safety and study conduct on a twice yearly basis during the study, or more frequently based on the emerging safety data. Members of the iDMC are external to the Sponsor and follow a charter that outlines the iDMC roles and responsibilities.

4. STATISTICAL METHODS

The statistical methods described in this SAP are based on the final protocol and pertain to the following:

- Induction Phase – Cohorts 1, 2, and 3 separately
- Maintenance Phase

Induction data will be presented separately for Cohorts 1–3.

The Maintenance Phase analysis will include all patients re-randomized into the Maintenance Phase; this includes patients across Cohorts 1-3 from the Induction Phase.

4.1 ANALYSIS POPULATIONS

4.1.1 Induction Phase

4.1.1.1 Modified Intent-to-Treat Population

All patients randomized who received at least one dose of study drug in the Induction Phase. Patients will be grouped according to the randomized treatment group at Induction.

4.1.1.2 Safety Population

All patients who received at least one dose of study drug during the Induction Phase. Patients will be grouped by the treatment most frequently received during the Induction Phase.

4.1.1.3 Pharmacokinetic-Evaluable Population

All patients who have received at least one dose of etrolizumab and have had at least one quantifiable concentration measured during the Induction Phase.

4.1.2 Maintenance Phase

4.1.2.1 Modified Intent-to-Treat Population

All patients re-randomized or enrolled into Maintenance who received at least one dose of study drug in the Maintenance Phase and were treated with etrolizumab as induction therapy across Cohorts 1-3.

The placebo/placebo treatment arm will be included in the modified intent to treat (mITT) population but the data will be treated as exploratory.

4.1.2.2 Safety Population

All patients who received at least one dose of study drug during the Maintenance Phase. Patients will be grouped by the treatment most frequently received during the Maintenance Phase.

4.1.2.3 Pharmacokinetic-Evaluable Population

All patients who have received at least one dose of etrolizumab, and have had at least one quantifiable concentration measured during Maintenance Phase.

4.2 ANALYSIS OF STUDY CONDUCT

The following will be summarized by treatment arm(s) for both Induction and Maintenance Phases:

- Analysis populations
- Patient disposition
- Study treatment completion/discontinuation
- Study completion/discontinuation
- Major protocol deviations

- Exposure

4.3 ANALYSIS OF TREATMENT GROUP COMPARABILITY

To review treatment group comparability, a number of variables collected at baseline (see Section 4.6.2) will be summarized and compared across treatment groups.

For continuous variables, descriptive statistics including n, mean, median, standard deviation, minimum, and maximum will be calculated. For categorical variables, number and percentage in each category will be displayed.

Demographics and baseline characteristics for the Maintenance Phase will use the data collected from their Induction baseline visit.

Demography

- Age ($\geq$ 18 to < 40, $\geq$40 to $<$65, $\geq$65)
- Sex
- Race
- Ethnicity
- Region
- Smoking status history
- Body mass index
- Weight

Disease Characteristics

- Duration of disease
- Duration of disease category (<3, $\geq$ 3 to < 8, $\geq$ 8)
- Disease location (Ileum only, Colon only, Ileum + Colon)
- Number of affected ileo-colonic segments
- Ulcerations at baseline
- Mean SF score
- Mean AP score (4-point scale)
- CDAI score
- CDAI Category ($\leq$330,>330)
- SES-CD score
- CD-PRO/SS – Functional score
- CD-PRO/SS – Bowel score
- [redacted]
- [redacted]

- [redacted]
- [redacted]
- [redacted]

Baseline Treatments

- Use of oral CS at Baseline (yes vs. no)
- Use of IS at Baseline (yes vs. no)
- Prior anti-TNF exposure (yes vs. no)

The patients with prior anti-TNF exposure at baseline will be further evaluated to include the number of TNFs previously received and status (e.g., primary non-responder, secondary non-responder, Intolerant, etc).

4.4 EFFICACY ANALYSIS

The Induction and Maintenance Phases will be analyzed and reported as separate studies.

Induction Phase

Cohort 1 statistical analyses will be carried out at the 10% significance level. No control for type I error will be performed; all p-values will be nominal and for exploratory purposes only.

Cohort 2 involves treatment with open label etrolizumab and has no control arm; data will be summarized descriptively only for all efficacy endpoints. Confidence intervals will be reported at the 5% significance level.

Cohort 3 will serve as the pivotal, confirmatory analysis for the Induction phase and will be analyzed using a 5% significance level. The co-primary endpoint and all key secondary endpoints will undergo a multiple testing procedure to provide strong control for type I error, as described in Section 4.4.8.1 and Section 4.4.13.1.

Data for the Induction Phase will be presented separately for Cohorts 1-3 in the study CSR, with Cohort 3 serving as pivotal, confirmatory analysis. Pooling of data across cohorts and pooled treatment comparisons may be undertaken to provide additional supportive information for the Induction Phase; details of such analyses will be documented in a separate SAP (e.g., Summary of Clinical Efficacy SAP).

Maintenance Phase

The Maintenance Phase analysis will include all patients re-randomized into the Maintenance Phase; this includes patients across cohorts 1-3 from the Induction Phase. The pivotal analysis for Maintenance will be performed on patients receiving etrolizumab during the Induction Phase. Patients receiving placebo during the Induction Phase and

who undergo a sham randomization to receive continued treatment with blinded placebo will be reported as exploratory only.

4.4.1 Analysis Methods

Statistical hypotheses for the primary and key secondary endpoints for the Induction Cohort 3 and Maintenance Phases will be evaluated using a multiple testing procedure to ensure an overall type I error rate no greater than 5%. Details of the multiplicity strategies are described separately for the Induction Phase (Section 4.4.8) and Maintenance Phase (Section 4.4.13). All hypothesis testing pertaining to these pivotal analyses will take place using a two-sided 5% significance level. Cohort 1 will be tested using a 10% significance level, no control will be undertaken for multiple testing.

Binary endpoints undergoing formal statistical testing described in this SAP will be analyzed using the Cochran-Mantel-Haenszel (CMH) test as the primary analysis method. The difference between proportions will be presented along with 95% confidence intervals for the pivotal Induction and Maintenance analyses, and 90% for the Cohort 1 Induction analysis, and will be constructed using the Newcombe method. The confidence intervals for the treatment group proportions will be constructed using the Wilson method.

Continuous endpoints undergoing formal statistical testing described in this SAP will be evaluated using a Mixed Model for Repeated Measures (MMRM) as the primary analysis method. The model will include covariates as described in Section 4.4.2, along with terms for treatment, baseline value, visit, and a visit-by-treatment interaction term. Within each participant, the model will incorporate an unstructured variance-covariance matrix for the random error terms. Parameters will be estimated using the restricted maximum likelihood, and the Kenward-Roger method will be used for calculating the denominator degrees of freedom. In the event that the model fails to converge, the spatial variance-covariance matrix will be used.

After accounting for intercurrent events (ICE) according to each endpoint's respective estimand definition (see Section 4.4.2), missing data handling methods for any patient with missing data are detailed in Section 4.10, in addition to Section 4.4.4 and Section 4.4.8 for the co-primary endpoints in Induction and Maintenance.

Approaches for all endpoints (primary, secondary [redacted]) on handling ICE according to each strategy, and the imputation method for true missing data (after accounting for ICE) are summarized in Appendix 2.

4.4.2 Covariate Adjustment

Unless otherwise stated, analyses of efficacy endpoints (co-primary, secondary, [redacted]) will be adjusted for the IxRS stratification factors described below.

Induction

- Concomitant oral CS treatment (yes vs. no)
- Baseline CDAI≤330 (yes vs. no)
- Prior anti-TNF exposure (yes vs. no)

An investigation using Cohort 1 induction data revealed that the analysis was over-stratified using all 4 stratification factors used at randomization, leading to multiple strata with small sample sizes (e.g.,≤5), and some strata being excluded from the CMH analysis. Simulations project that Cohort 3 will also be over-stratified. For this reason, the 4th stratification factor of Immunosuppressant use (yes vs. no) will be dropped from the primary reporting methods for both Cohort 1 and 3.

Maintenance

- Induction Dose Regimen (high vs. low)
- Concomitant oral CS (yes vs. no)
- Prior anti-TNF exposure (yes vs. no)

The fourth stratification factor of "CDAI Remission status at Week 10 and Week 14" used for the Maintenance randomization will be dropped from the primary analysis methods to avoid over-stratification that would otherwise results in too many strata with small (i.e.,<5) sample sizes.

4.4.3 Intercurrent Events

In alignment with the addendum to ICH E9, estimands are described for the co-primary efficacy endpoints in Induction and Maintenance Phase in Section 4.1.1 and Section 4.1.2 respectively. The estimand attributes include the following ICE along with their respective strategy approach:

1. **Patient/Treatment Discontinuation**: The investigator has the right to discontinue a patient from treatment/the study at any time and patients have the right to voluntarily discontinue from treatment/the study for any reason at any time. Patients must discontinue from treatment if any of the following (but not limited to) occur: pregnancy, anaphylaxis or other severe hypersensitivity reaction, development of colonic mucosal dysplasia, malignancy or specific serious infections (more information can be found in the Study GA29144 Protocol). Any patient discontinuing from treatment/study for any reason will not be replaced. If a patient discontinues treatment prematurely, then this patient will be considered as being unsuccessfully treated and will be evaluated under the composite strategy. Therefore, for binary endpoints, these patients will be treated as non-responders. Similarly, for continuous endpoints, as an unfavorable outcome that would reflect a composite

endpoint equivalent to a non-responder, the patient's worst-post-baseline-value for the endpoint will be carried forward, i.e., worst-observation-carried-forward (WOCF) imputation.

2. **Use of rescue therapy (or prohibited medication)**: Rescue therapy is defined as medication prescribed for new or worsening CD symptoms and includes any new CS or IS therapy for CD and any increase in dose or regimen of baseline Crohn's medications. For patients in Maintenance Phase, increase in the dose of CS therapy above a patient's baseline dose (applicable to patients requiring CS at baseline) is also classed as rescue therapy. Antibiotics, 5-aminosalicylate, anti-diarrheals, probiotics, herbal/ayurvedic, nutritional and homeopathic supplements are not considered as rescue therapies. Patients requiring rescue therapy (or prohibited medication) will be considered as a failure of study drug which is synonymous with the composite strategy. Therefore, for binary endpoints, these patients will be treated as non-responders. Similarly, for continuous endpoints, as an unfavorable outcome that would reflect a composite endpoint equivalent to a non-responder, the patient's worst-post-baseline-value for the endpoint will be carried forward i.e., WOCF imputation.
3. **Deaths**: Any instance of death will be treated as a failure of study treatment. Therefore, for binary endpoints, these patients will be treated as non-responders. Similarly, for continuous endpoints, as an unfavorable outcome that would reflect a composite endpoint equivalent to a non-responder, the patient's worst-post-baseline-value for the endpoint will be carried forward i.e., WOCF imputation.
4. **Coronavirus disease -19-related:** Corona virus disease 2019 (COVID-19) related events are defined as any systematic disruptions such as:

- Visit attended but key assessments not performed at Week 14 or 66 due to logistical constraints, or restrictions put in place by the government, hospital or investigator (e.g., ileocolonoscopy cancelled due to not qualifying as an 'urgent' medical procedure)
- Treatment discontinuation with COVID-19 cited as a reason
- Visit not done, leading to key endpoint evaluations not performed (e.g., ileocolonoscopy at Weeks 14 and 66, and CDAI assessments at Weeks 14 and 66). Reasons for missing the visit include logistical constraints at site, restrictions that impede a patient/investigator from carrying out the visit, or due to patient's fear of the pandemic.

A hypothetical strategy will be employed in order to estimate the treatment effect in the hypothetical world in the absence of COVID-19. It is determined that COVID-19 is only likely to impact collection of data from site-based assessments and endoscopy procedures. Therefore, COVID-19 is not considered to have a major impact on a patient's ability to complete eDiary data. Any missing eDiary data will be considered as "true missing" and handled according to the missing data handling method as detailed in Section 4.4.5. For site based assessments, data from the previous site visit will be used (i.e., last-observation-carried-forward [LOCF] imputation). For endoscopy data, i.e.,

SES-CD score, where intermediate data collection isn't available, a multiple imputation (MI) approach will be used. For continuous endpoints where MMRM is used for analysis, MMRM will also be used to impute data missing due to COVID-19-related ICE under the hypothetical strategy.

4.4.4 Induction Co-Primary Efficacy Endpoints

The Induction Phase of Study GA29144 has two co-primary efficacy endpoints: Clinical remission and endoscopic improvement, both measured at Week 14.

No type I error control will be undertaken for treatment comparisons in Cohort 1. Cohort 2 involves treatment with open label etrolizumab and has no control arm, data will be summarized descriptively.

For Cohort 3, to control for multiple testing the primary efficacy treatment comparison will take place for placebo vs. etrolizumab 210 mg, where both endpoints need to be statistically significant at the 5% significance level. The testing for the etrolizumab 105 mg dose group will follow thereafter, and is described further in Section 4.4.5.

The estimands for the pivotal Cohort 3 primary analysis is described below.

4.4.4.1 Definition of Co-Primary Estimand: Clinical Remission

The estimand corresponding to clinical remission at Week 14 is described by the following attributes:

a) **Population**: Adult patients with moderate to severely active CD who have experienced intolerance, refractory disease or no response to one of CS therapy, IS therapy or anti-TNF therapy. See Protocol Synopsis section of Protocol Version 7 for the full list of inclusion/exclusion criteria.

b) **Variable**: Clinical remission at Week 14, measured using stool frequency and abdominal pain scores averaged over 7 days.

c) **Intercurrent Events Strategy:** Patients with an ICE due to treatment discontinuation, use of rescue therapy, or death as highlighted in Section 4.4.3, will be considered as treatment failures and therefore, non-responders under the composite strategy. COVID-19 is not considered an ICE for this endpoint as it is not likely to impact a patient's ability to enter eDiary data.

d) **Treatment:** The treatment of interest is etrolizumab 210mg, which will be compared to a placebo treatment group.

e) **Population-Level Summary**: Difference in the proportion of patients in clinical remission at Week 14 between placebo and etrolizumab 210 mg in Cohort 3.

Definition:

Clinical remission is defined as a SF mean daily score$\leq$3 and AP mean daily score$\leq$1, with no worsening in either sub-score compared to baseline, averaged over 7 days. For details of calculations, see Section 4.6.

Analysis Methods:

The following hypothesis will be tested:

Null hypothesis H0: ρ etrolizumab 210 mg $-$ ρ Placebo $= 0$; the proportion of patients achieving clinical remission at Week 14 in the placebo arm is the same as the proportion of patients achieving clinical remission in the etrolizumab 210 mg arm.

Alternative hypothesis H1: ρ etrolizumab 210 mg $-$ ρ Placebo $\neq 0$; the proportion of patients achieving clinical remission at Week 14 in the placebo arm is not the same as the proportion of patients achieving clinical remission in the etrolizumab 210 mg arm.

Clinical remission at Week 14 will be analyzed as described in Section 4.4.1.

4.4.4.2 Definition of Co-Primary Estimand: Endoscopic Improvement

The estimand corresponding to endoscopic improvement at Week 14 is described by the following attributes:

a) **Population**: Adult patients with moderate to severely active CD who have experienced intolerance, refractory disease or no response to one of CS therapy, IS therapy or anti-TNF therapy. See Protocol Synopsis section of Protocol Version 7 for the full list of inclusion/exclusion criteria.

b) **Variable**: Endoscopic improvement at Week 14, as measured by the SES-CD score.

c) **Intercurrent Events Strategy:** ICE are outlined in Section 4.4.3 and the strategy described for binary endpoints will be employed i.e., a composite strategy. In other words, patients with an ICE due to treatment discontinuation, use of rescue therapy, or death, will be considered as treatment failures and therefore, non-responders. For COVID-19 related ICE, since no additional data are collected at time-points other than Weeks 14 and 66, a multiple imputation approach will be used as detailed in Section 4.10.

d) **Treatment:** The treatment of interest is etrolizumab 210 mg, which will be compared to a placebo treatment group.

e) **Population-Level Summary**: Difference in proportion of patients achieving endoscopic improvement at Week 14 between placebo and etrolizumab 210 mg in Cohort 3.

Definition:

Endoscopic improvement is defined as a $\geq$50% reduction in the baseline SES-CD score. For details of the scoring methodology and handling of partial scores, see Section 4.6.

Analysis Methods:

The following hypothesis will be tested:

Null hypothesis H0: ρ etrolizumab 210 mg $-$ ρ Placebo $= 0$; the proportion of patients achieving endoscopic response at Week 14 in the placebo arm is the same as the proportion of patients achieving endoscopic response in the etrolizumab 210 mg arm.

Alternative hypothesis H1: ρ etrolizumab 210 mg $-$ ρ Placebo $\neq 0$; the proportion of patients achieving endoscopic response at Week 14 in the placebo arm is not the same as the proportion of patients achieving endoscopic response in the etrolizumab 210 mg arm.

Endoscopic response at Week 14 will be analyzed as described in Section 4.4.1.

4.4.5 Missing Data Handling Methods for Induction Co-Primary Endpoints

The handling of ICE explained in Section 4.4.2 will be applied before using the missing data handling approaches in this Section. The handling of ICE will be applied in the order that the ICE occurs within the study.

Patients with insufficient daily diary entries (fewer than 4 out of the 7 days) for calculation of the weekly average will be flagged. For these patients, the 7-day window will be extended to up to 3 additional days to a maximum of a 10-day window until there are at least 4 diary entries available. For example, if diary entries are only available for 3 out of the 7 days prior to the visit, the 4th entry will be used from either the 8th, 9th, or 10th day, depending on when the 4th entry is first observed. Clinical remission will then be determined as defined in Section 4.4.4.1. If at least 4 diary entries are not available using the extended 10 day window, patients will be set to non-responders. Patients will be flagged if their clinical remission is calculated using the extended 10 window, and separate sensitivity analyses will be performed to assess robustness under alternative data handling methods, see Section 4.4.6.

Missing data on endoscopic improvement at Week 14 will be assumed to be missing at random (MAR) and imputed using Multivariate Imputation by Chained Equations (MICE) as described by White et. al. (2011). This is a popular Multiple Imputation (MI) approach and is detailed in Section 4.10. Details for the handling of partially available data are described in Section 4.6, here the method of adjusting for segments available at baseline will be used as the main analysis method.

4.4.6 Sensitivity Analyses for Induction Co-Primary Endpoints

Sensitivity analyses will be carried out on both co-primary endpoints using the following methods:

- A logistic regression model using the same stratification factors used in the CMH test for the primary analysis as covariates. The adjusted rates and differences will be reported.
- CMH test using stratification factors obtained from the electronic Case Report Form instead of IxRS, if data is discrepant between the two data sources (e.g., > 10% discrepancy for a stratification factor)
- A Fisher's Exact test will also be performed on both co-primary endpoints. For endoscopic improvement, this will be performed on non-responder imputed data only and for clinical remission it will be performed on the data imputed using the extended 10 day window.

Additionally, to assess the robustness of analyses and sensitivity to departures from the missing data assumptions, the following analyses will be produced.

- Non-responder imputation (replacing widening the window to 10 days for clinical remission, and replacing MI for endoscopic improvement).
- Multiple imputation for imputing missing data in clinical remission without widening the window.
- A tipping point analysis will be conducted to explore the impact of independently varying the MAR assumptions about missing outcomes on the two treatments arms of interest i.e., etrolizumab 210mg and placebo. This will be performed as a sensitivity to the missing data handling methods for both co-primary endpoints. The handling of ICEs corresponding to the composite strategy (i.e., withdrawal, rescue therapy, death) will remain unchanged in the tipping point analysis and patients will be set to non-responders. For endpoints with COVID-19 as an ICE, these will be included in the tipping point analysis along with the patients with otherwise missing data to test the MAR assumptions. The tipping point analysis is described further in Section 4.10.1

4.4.7 Supplementary Analyses

The following analysis will be carried out for the endoscopic endpoints:

- Endoscopic 50% improvement/remission using all available data (i.e., no adjustment for segments available at Baseline

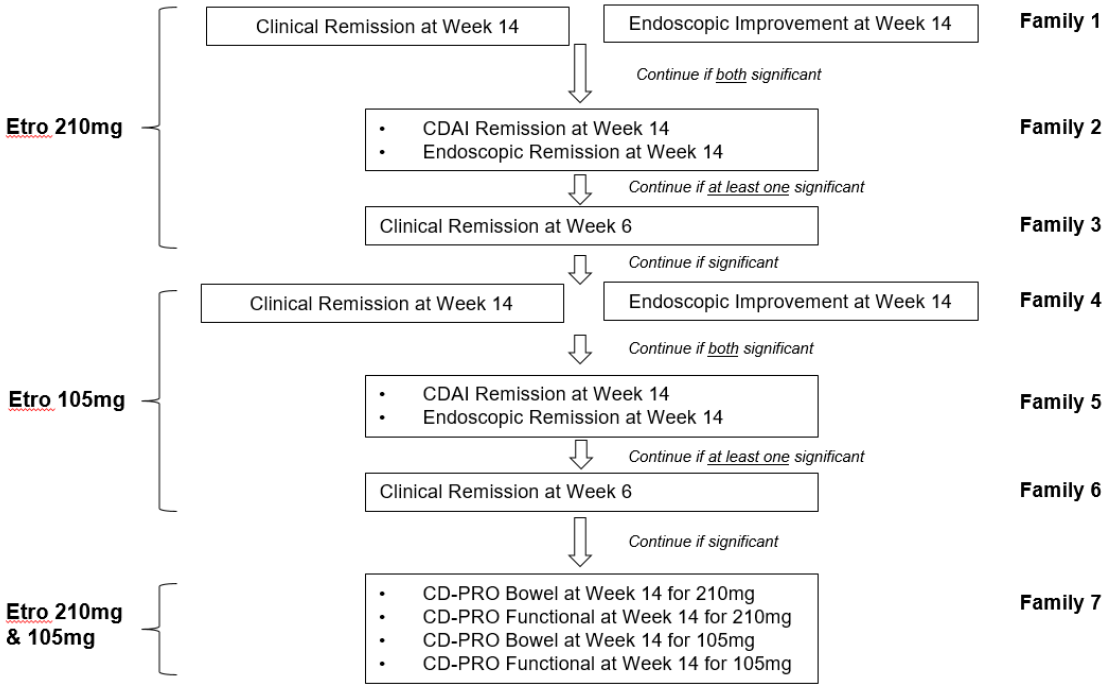
4.4.8 Induction Secondary Efficacy Endpoints

4.4.8.1 Control of Type I Error

The testing of the co-primary and key secondary endpoints will undergo a multiple testing procedure to ensure that the family-wise control for type I error does not exceed 5%. For the context of multiple testing, the etrolizumab 210 mg dose will be considered as the co-primary endpoint, where testing of the etrolizumab 105 mg dose and all other secondary endpoints will be gated on success of the etrolizumab 210 mg co-primary.

A hierarchical gatekeeping method will be employed, involving a mixture of serial and parallel testing. A serial test will be applied to the endpoints of clinical remission and endoscopic improvement (Families 1 and 4 in Figure 1) and will therefore require statistical significance of both endpoints in order to proceed to the next family. [Technical note: to achieve this, the two endpoints within the family will be treated as two sub-families, which will be ordered by the p-values where the sub-family with the lowest p-value will be considered first. The two sub-families will be tested using a truncated Hommel procedure with a truncation parameter of 1. The single endpoint of clinical remission at Week 6 (Families 3 and 6) will require statistical significance in order to proceed to the next family as only hypothesis is tested. For all other endpoints (Families 2,5,7), testing of endpoints within a family will occur in parallel using a truncated Hommel procedure with a truncation value of 0.8, except for the last family (Family 7) which will use a regular Hommel test (truncation parameter of 1). At each level of testing, the truncation parameter of 0.8 prioritizes power for the endpoints within a family, at the cost of a substantial drop in power to subsequent families in the event that at least one non-significant result is obtained within the family. Where testing occurs in parallel, the order of the endpoints for testing will be based on their raw p-values, and continuation to the next family may only take place if at least one endpoint is statistically significant, otherwise testing will stop. After the hierarchy has stopped, any remaining secondary endpoints [redacted] will be considered as supportive information, and no claims of statistical significance will be made. An 'overview of efficacy results' table will be generated that will show the adjusted and raw p-values to help guide interpretation.

Figure 1 Multiple Testing Procedures for Induction Endpoints



CDAI=Crohn's disease activity index; CD-PRO=Crohn's disease patient reported outcome; Etro=etrolizumab.

4.4.8.2 Estimand Definitions for Induction Secondary Efficacy Endpoints

For the secondary efficacy endpoints described in the proceeding sections, we define the following:

- **Population:** Adult patients with moderate to severely active CD who have experienced intolerance, refractory disease or no response to one of CS therapy, IS therapy or anti-TNF therapy. See Protocol Synopsis Section of Protocol Version 7 for the full list of inclusion/exclusion criteria.
- **Intercurrent Events strategy:** Patients with an ICE due to treatment discontinuation, use of rescue therapy, or death, as highlighted in Section 4.4.3, will be considered as treatment failure under the composite strategy and treated with an unfavorable outcome, i.e., non-responders for binary endpoints and WOCF for continuous endpoints. Any COVID-related ICE (where relevant) will be handled according to the hypothetical strategy. The imputation approach will be dependent on the nature and frequency of collected data as detailed in Section 4.4.2.
- **Treatment:** The treatments of interest are the etrolizumab 105 mg and 210 mg treatment arms. These arms will be compared to the placebo treatment arm separately i.e., etrolizumab 105 mg vs placebo and etrolizumab 210 mg vs placebo

The remainder of attributes, namely, the variable and population-level summary, are defined for each secondary efficacy endpoint in the sections below.

4.4.8.3 Crohn's Disease Activity Index Remission at Week 14

The proportion of patients achieving CDAI remission at Week 14, as defined in Appendix 1. CDAI remission is measured using 8 weighted sub-scores. Details of the scoring method and handling of missing sub-scores can be found in Section 4.6 and Section 4.10.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the placebo treatment arm will be compared with the etrolizumab 105 mg and 210 mg treatment arms separately using the difference in proportions of patients in CDAI remission at Week 14 in Cohort 3.

COVID-19-related events are not to be considered as an ICE for the eDiary component of the CDAI score.

4.4.8.4 Endoscopic Remission at Week 14

The proportion of patients achieving endoscopic remission at Week 14, as defined in Appendix 1.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the placebo treatment arm will be compared with the etrolizumab 105 mg and 210 mg treatment arms separately using the difference in proportions of patients in endoscopic remission at Week 14 in Cohort 3. Patients with SES-CD scores that are not available from any reader for Week 14 will be treated as missing and handled according to the approach outlined in Section 4.10. Details for the handling of partially available data are described in Section 4.6, here the method of adjusting for segments available at Baseline will be used as the main analysis method.

4.4.8.5 Clinical Remission at Week 6

The proportion of patients achieving clinical remission at Week 6, as defined in Appendix 1. Clinical remission at Week 6 is measured using SF and AP averaged over at least 4 out of the 7 days prior to the visit.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the placebo treatment arm will be compared with the etrolizumab 105 mg and 210 mg treatment arms separately using the difference in proportions of patients in clinical remission at Week 6 in Cohort 3. If less than 4 days diary entries are available out of the 7 days prior to the visit, then the data will be flagged and handled according to the approach outlined in Section 4.10.

COVID-19-related events are not to be considered as an ICE for the eDiary data.

4.4.8.6 Crohn's Disease-Patient Reported Outcome/Signs and Symptoms Score at Week 14-Functional Domain

The change from baseline at Week 14 in CD-PRO/SS score for the functional domain.

This will be analyzed as described in Section 4.4.1 for continuous endpoints. The mean change from baseline to Week 14 CD-PRO/SS score for the functional domain will be estimated in Cohort 3 for each of the etrolizumab 105 mg and 210 mg treatment arms and compared to the placebo treatment arm separately using difference of means. Missing CD-PRO/SS score for the functional domain at Week 14 will be imputed using the MMRM as described in Section 4.10.

4.4.8.7 Crohn's Disease-Patient Reported Outcome/Signs and Symptoms Score at Week 14-Bowel Domain

The change from baseline at Week 14 in CD-PRO/SS score for the bowel domain.

This will be analyzed as described in Section 4.4.1 for continuous endpoints. The mean change from baseline to Week 14 CD-PRO/SS score for the bowel domain will be estimated in Cohort 3 for each of the etrolizumab 105 mg and 210 mg treatment arms and compared to the placebo treatment arm using difference of means. Missing CD-PRO/SS score for the bowel domain at Week 14 will be imputed using the MMRM as described in Section 4.10.

4.4.9 Maintenance Co-Primary Efficacy Endpoints

The Maintenance Phase of Study GA29144 has two co-primary efficacy endpoints: clinical remission and endoscopic improvement, both measured at Week 66. To control for multiple testing, both endpoints will need to be statistically significant for the co-primary endpoint to be deemed as significant.

4.4.9.1 Definition of Co-primary Estimand: Clinical Remission

The estimand corresponding to clinical remission at Week 66 is described by the following attributes:

a) **Population:** Adult patients that achieved a CDAI-70 response at Week 14 without the use of rescue therapy, and who were randomized to etrolizumab 105 mg or 210 mg in Cohorts 1–3 as Induction therapy. See Protocol Synopsis Section of Protocol Version 7 for the full list of inclusion/exclusion criteria.

b) **Variable**: Clinical remission at Week 66, measured using stool frequency and abdominal pain averaged over 7 days.

c) **Intercurrent Events Strategy:** ICE are outlined in Section 4.4.3 and the strategy described for binary endpoints will be employed i.e., a composite strategy. In other words, patients with an ICE due to treatment discontinuation, use of rescue therapy, or death, will be considered as treatment failures and therefore, non-responders. COVID-19 is not considered to be an ICE for this endpoint as it is not likely to impact a patient's ability to enter eDiary data.

d) **Treatment:** The treatment of interest is etrolizumab/etrolizumab 105 mg, which will be compared to the etrolizumab/placebo treatment group.

e) **Population-Level Summary:** Difference in proportion of patients in clinical remission at Week 66 between etrolizumab/etrolizumab 105 mg and etrolizumab/placebo in the Maintenance Phase.

Definition:

Clinical remission is defined as a SF mean daily score≤3 and AP mean daily score≤1, with no worsening in either sub-score compared to baseline, averaged over the 7 days prior to visit. For details of calculations, see Section 4.6.

Analysis Methods:

The following hypothesis will be tested:

Null hypothesis H0: ρ etrolizumab/etrolizumab 105 mg $-$ ρ etrolizumab/Placebo $=0$; the proportion of patients achieving clinical remission at Week 66 in the etrolizumab/placebo arm is the same as the proportion of patients achieving clinical remission in the etrolizumab/etrolizumab 105 mg arm.

Alternative hypothesis H1: ρ etrolizumab/etrolizumab 105 mg $-$ ρ etrolizumab/Placebo $\neq 0$; the proportion of patients achieving Clinical Remission at Week 66 in the etrolizumab/placebo arm is not the same as the proportion of patients achieving Clinical Remission in the etrolizumab/etrolizumab 105 mg arm.

Clinical remission at Week 66 will be analyzed as described in Section 4.4.1.

4.4.9.2 Definition of Co-primary Estimand: Endoscopic Improvement

The estimand corresponding to endoscopic improvement at Week 66 is described by the following attributes:

a) **Population**: Adult patients that achieved a CDAI-70 response at Week 14 without the use of rescue therapy, and who were randomized to etrolizumab 105 mg or 210 mg in Cohorts 1-3 as Induction therapy. See protocol synopsis from Protocol Version 7 for the full list of inclusion/exclusion criteria.

b) **Variable**: Endoscopic improvement at Week 66, as measured by the SES-CD score.

c) **Intercurrent Event Strategy:** Patients with an ICE due to treatment discontinuation, use of rescue therapy, or death, as highlighted in Section 4.4.3, will be considered as treatment failures and therefore, non-responders under the composite strategy. For COVID-19 related ICE, under the hypothetical strategy, since no additional data are collected at time-points other than Weeks 14 and 66, a MI approach will be used as detailed in Section 4.10.

d) **Treatment:** The treatment of interest is etrolizumab/etrolizumab 105mg, which will be compared to the etrolizumab/placebo treatment group.

e) **Population-Level Summary:** Difference in proportion of patients achieving endoscopic improvement at Week 66 between etrolizumab/etrolizumab 105mg and etrolizumab/placebo in the Maintenance Phase.

Definition

Endoscopic improvement is defined as a ≥50% reduction in the baseline SES-CD score. For details of calculations, see Section 4.6.

Analysis Methods

The following hypothesis will be tested:

Null hypothesis H0: ρ etrolizumab/etrolizumab 105 mg $-$ ρ etrolizumab/Placebo $= 0$; the proportion of patients achieving endoscopic response at Week 66 in the etrolizumab/placebo arm is the same as the proportion of patients achieving endoscopic response in the etrolizumab/etrolizumab 105 mg arm.

Alternative hypothesis H1: ρ etrolizumab/etrolizumab 105 mg $-$ ρ etrolizumab/Placebo $\neq 0$; the proportion of patients achieving endoscopic response at Week 66 in the etrolizumab/placebo arm is not the same as the proportion of patients achieving endoscopic response in the etrolizumab/etrolizumab 105 mg arm.

Endoscopic response at Week 66 will be analyzed as described in Section 4.4.1.

4.4.10 Missing Data Handling Methods for Maintenance Co-Primary Endpoints

The handling of ICE explained in Section 4.4.3 will be applied before using the missing data handling approaches in this section. The handling of ICE will be applied in the order that the ICE occurs within the study.

Patients with insufficient daily diary entries (less than 4 out of the 7 days) for calculation of the weekly average will be flagged. For these patients, the 7-day window will be extended to up to 3 additional days to a maximum of a 10-day window until there are at least 4 diary entries available. For example, if diary entries are only available for 3 out of the 7 days prior to the visit, the 4th entry will be used from either the 8th, 9th, or 10th day, depending on when the 4th entry is first observed. Clinical remission will then be determined as defined in Section 4.4.4.1. If at least 4 diary entries are not available using the extended 10 day window, patients will be set to non-responders. Patients will be flagged if their clinical remission is calculated using the extended 10 window, and separate sensitivity analyses will be performed to assess robustness under alternative data handling methods, see Section 4.4.11. Missing data on endoscopic improvement at Week 66 will be assumed to be MAR and imputed using MICE as described by White et. al. (2011). This approach is detailed in Section 4.10. Details for the handling of partially available data are described in Section 4.6, here the method of adjusting for segments available at baseline will be used as the main analysis method.

4.4.11 Sensitivity Analyses for Maintenance Co-Primary Endpoints

Sensitivity analyses will be carried out on both co-primary endpoints using:

- A logistic regression model using the same stratification factors used in the CMH test for the primary analysis as covariates. The adjusted rates and differences will be reported.
- A Fisher's Exact test will also be performed on both co-primary endpoints. For endoscopic improvement, this will be performed on non-responder imputed data only and for clinical remission it will be performed on the data imputed using the widened windows.

Additionally, to assess the robustness of analyses and sensitivity to departures from the missing data assumptions, the following analyses will be produced:

- Non-responder imputation (replacing widening the window for clinical remission, and MI for endoscopic improvement).
- Multiple imputation for imputing missing data in clinical remission without widening the window.
- A tipping point analysis will be conducted to explore the impact of independently varying the assumptions about missing outcomes on the two treatments arms of interest i.e., etrolizumab/etrolizumab 105 mg and etrolizumab/placebo. This will be performed as a sensitivity to the missing data handling methods for both co-primary endpoints (handling of ICE will remain unchanged). The tipping point analysis is described further in Section 4.10.1.

4.4.12 Supplementary Analyses

The following analysis will be carried out for the endoscopic endpoints:

- Endoscopic 50% improvement/remission using all available data (i.e., no adjustment for segments available at baseline

In addition, a supplementary analysis will be done on each Maintenance co-primary endpoint with an alternative estimand with the following change to the population attribute:

- **Population:** Adult patients that achieved a CDAI-100 response at Week 14 without the use of rescue therapy, and who were randomized to etrolizumab 105 mg or 210 mg in Cohorts 1–3 as Induction therapy. See Protocol Synopsis Section of Protocol Version 7 for the full list of inclusion/exclusion criteria.

All other attributes will remain the same.

4.4.13 Maintenance Secondary Efficacy Endpoints

4.4.13.1 Control of Type 1 Error

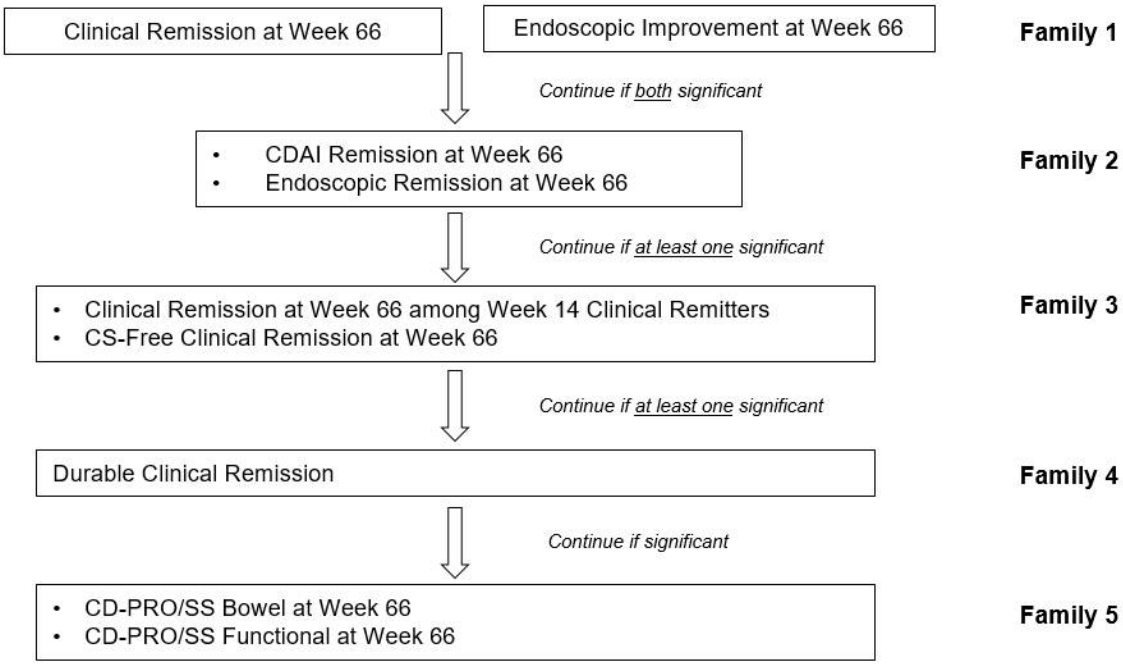
The testing of the co-primary and key secondary endpoints will undergo a multiple testing procedure to ensure that the family-wise control for type I error does not exceed 5%.

The study for the Maintenance Phase will be considered positive if both clinical remission and endoscopic improvement endpoints at Week 66 demonstrate statistical significance at the 5% significance level.

A hierarchical gatekeeping method will be employed for the key secondary endpoints, involving a mixture of serial and parallel testing. A serial test will be applied to the endpoints of clinical remission at Week 66 and endoscopic improvement at Week 66 (Family 1 in Figure 2) and will therefore require statistical significance of both endpoints in order to proceed to the next family. To achieve this, Family 1 will be treated as two sub-families, which will be ordered by the p-values where the sub-family with the lowest p-value will be considered first. The two sub-families will be tested using a truncated Hommel procedure with a truncation parameter of 1. The endpoint of durable remission (Family 4) will require statistical significance in order to proceed to the next family since there is only one hypothesis being tested. For all other endpoints (Families 2,3,5), testing of endpoints within a family will occur in parallel using a truncated Hommel procedure; Families 2 and 3 will involve a truncated test using a truncation value of 0.5, and Family 5 will use a regular Hommel test (with a truncation parameter of 1). For Families 2, 3, and 5, the order of the endpoints for testing will be based on their raw p-values, and continuation to the next family (if applicable) may only take place if at least one endpoint is significant, otherwise testing will stop. A truncation parameter of 0.5 preserves power for endpoints in families lower down the hierarchy, with the trade-off for a reduction in power for endpoints within families higher up the hierarchy. After the hierarchy has stopped, any remaining secondary endpoints [redacted] will be considered as supportive information, and no claims of statistical significance will be made.

The secondary endpoints of corticosteroid-free (24 weeks) clinical remission and endoscopic improvement among Week 14 endoscopic improvers will not be formally controlled for type I error; nominal p-values will be presented.

Figure 2 Multiple Testing Procedure for Maintenance Endpoints



CDAI=Crohn's Disease Activity Index; CD-PRO/SS=Crohn's Disease-Patient Reported Outcome/Signs and Symptoms.

4.4.13.2 Estimand Definitions for Maintenance Secondary Efficacy Endpoints

For the secondary efficacy endpoints described in the proceeding sections, we define the following:

- **Population:** Adult patients that achieved a CDAI-70 response at Week 14 without the use of rescue therapy, and who were randomized to etrolizumab 105 mg or 210 mg in Cohorts 1–3 as Induction therapy. See Protocol Synopsis of Protocol Version 7 for the full list of inclusion/exclusion criteria.
- **Intercurrent Event strategy:** Patients with an ICE due to treatment discontinuation, use of rescue therapy, or death, as highlighted in Section 4.4.3, will be considered as treatment failure under the composite strategy and treated with an unfavourable outcome i.e., non-responders for binary endpoints and WOCF for continuous endpoints. Any COVID-related ICE (where relevant) will be handled according to the hypothetical strategy. The imputation approach will be dependent on the frequency of collected data as detailed in Section 4.4.2.
- **Population-Level Summary:** Difference between etrolizumab/etrolizumab 105mg and etrolizumab/placebo in the Maintenance Phase.
- **Treatment:** The treatment of interest is etrolizumab/etrolizumab 105mg, which will be compared to a etrolizumab/placebo treatment group.

The remainder of the attributes, namely, the variable and population-level summary are defined for each secondary efficacy endpoint in the sections below, including any other more refined definition of the patient population.

4.4.13.3 Crohn's Disease Activity Index Remission at Week 66

The proportion of patients in CDAI remission at Week 66, as defined in Appendix 1. CDAI remission is measured using 8 weighted sub-scores. Details of the scoring method and handling of missing sub-scores can be found in Section 4.6 and Section 4.10.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions.

COVID-19-related events are not to be considered as an ICE for the eDiary component of the CDAI score.

4.4.13.4 Clinical Remission at Week 66 among Patients in Clinical Remission in Week 14

The proportion of patients in clinical remission (as defined in Appendix 1) at Week 66, among the subset of Maintenance mITT population achieving clinical remission at Week 14. Clinical remission is measured using SF and AP averaged over at least 4 out of the 7 days prior to the visit.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions. If less than 4 days diary entries are available out of the 7 days prior to the visit, then the data will be flagged and handled according to the approach outlined in Section 4.10.

COVID-19-related events are not to be considered as an ICE for eDiary data.

4.4.13.5 Corticosteroid-free Clinical Remission at Week 66 among Patients Receiving Corticosteroids at Baseline

The proportion of patients achieving clinical remission and are not taking oral corticosteroids at Week 66, among the subset of Maintenance mITT population receiving oral corticosteroids at baseline.

This will be analyzed as described in Section 4.4.1 for binary endpoints, except the oral corticosteroid stratification factor will be dropped. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions. If less than 4 days diary entries are available out of the 7 days prior to the visit, then the data will be flagged and handled according to the approach outlined in Section 4.10.

4.4.13.6 Endoscopic Remission at Week 66

The proportion of patients achieving endoscopic remission at Week 66, as defined in Appendix 1.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions. Patients with SES-CD scores that are not available from any reader for Week 66 will be treated as missing and handled according to the approach outlined in Section 4.10. Details for the handling of partially available data are described in Section 4.6, here the method of adjusting for segments available at Baseline will be used as the main analysis method.

4.4.13.7 Endoscopic Improvement at Week 66 among Patients Achieving Endoscopic Improvement at Week 14

The proportion of patients achieving endoscopic improvement at Week 66, among the subset of Maintenance mITT population achieving endoscopic improvement at Week 14. Patients with no available SES-CD score at Week 14 are excluded and not considered.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions. Patients with SES-CD scores that are not available from any reader will be treated as missing and handled according to the approach outlined in Section 4.10. Details for the handling of partially available data are described in Section 4.6, here the method of adjusting for segments available at baseline will be used as the main analysis method.

4.4.13.8 Durable Clinical Remission

The proportion of patients achieving durable clinical remission during the Maintenance Phase, defined as achieving clinical remission in at least 4 of the 6 in-clinical assessment visits (Weeks 24, 28, 32, 44, 56, and 66), which must include the Week 66 visit.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions. If less than 4 days diary entries are available out of the 7 days prior to the visit at Weeks 24, 28, 32, 44, and 56, then the data will be considered missing and handled according to the approach outlined in Section 4.10.

4.4.13.9 Corticosteroid-Free Clinical Remission for 24 Weeks at Week 66 among Patients Receiving Corticosteroids at Baseline

The proportion of patients achieving clinical remission at Week 66 and are not taking oral CS for at least 24 weeks prior to Week 66. This analysis will be performed among the subset of Maintenance mITT population who were receiving oral corticosteroids at baseline.

This will be analyzed as described in Section 4.4.1 for binary endpoints. Patients in the etrolizumab/placebo treatment arm will be compared with the etrolizumab/etrolizumab 105 mg treatment arm using the difference in proportions. If less than 4 days diary entries are available out of the 7 days prior to the visit, then the data will be considered missing and handled according to the approach outlined in Section 4.10.

4.4.13.10 Crohn's Disease-Patient Reported Outcome/Signs and Symptoms Score at Week 66-Functional Domain

The change from baseline at Week 66 in CD-PRO/SS score for the functional domain.

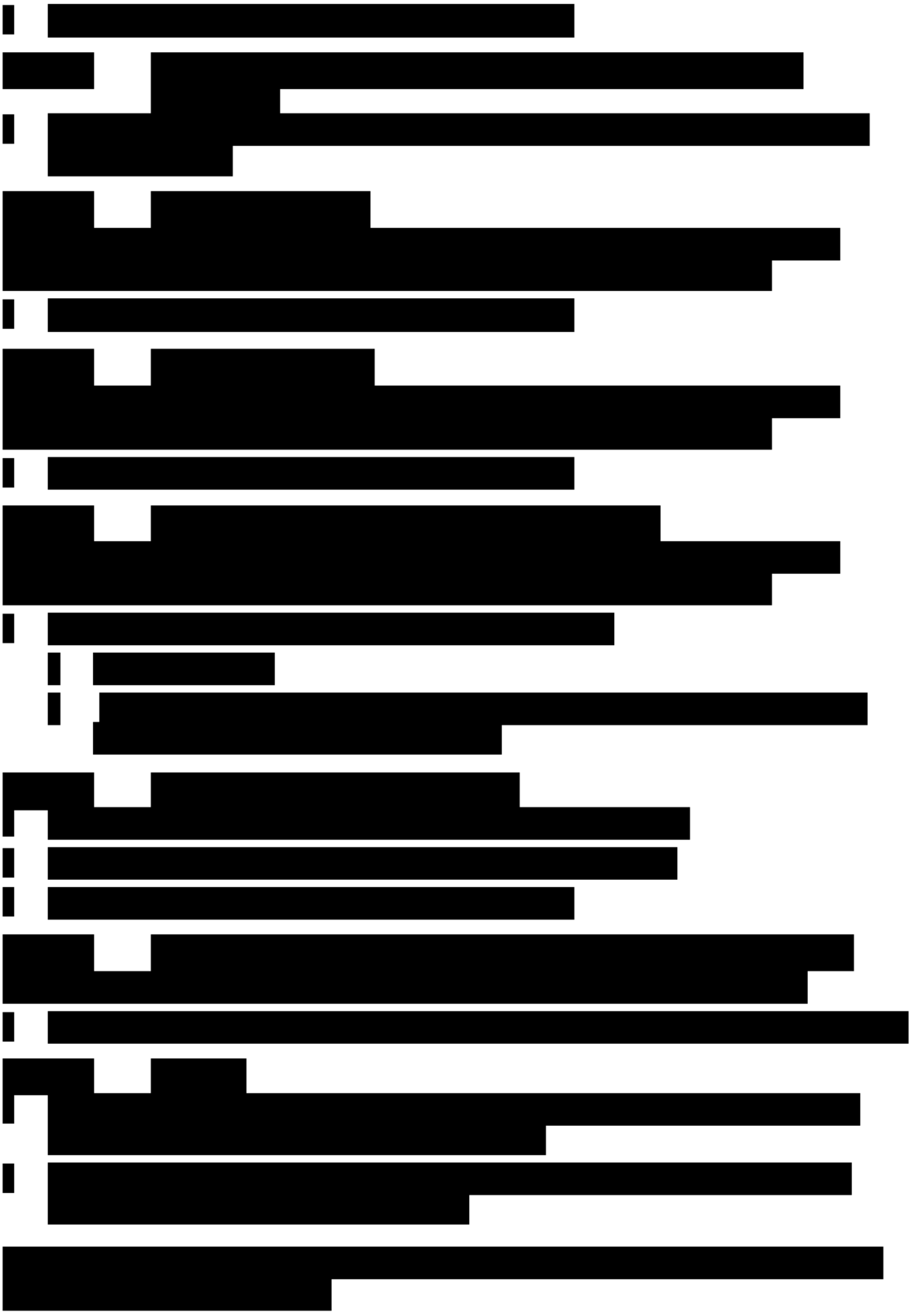
This will be analyzed as described in Section 4.4.1 for continuous endpoints. The mean change from baseline to Week 66 CD-PRO/SS score for the functional domain will be estimated for the etrolizumab/etrolizumab 105 mg treatment arm and compared to the etrolizumab/placebo treatment arm using difference of means. Missing CD-PRO/SS score for the functional domain at Week 66 will be imputed using the MMRM as described in Section 4.10.

4.4.13.11 Crohn's Disease-Patient Reported Outcome/Signs and Symptoms Score at Week 66 – Bowel Domain

The change from baseline at Week 66 in CD-PRO/SS score for the bowel domain.

This will be analyzed as described in Section 4.4.1 for continuous endpoints. The mean change from baseline to Week 66 CD-PRO/SS score for the bowel domain will be estimated for the etrolizumab/etrolizumab 105 mg treatment arm and compared to the etrolizumab/placebo treatment arm using difference of means. Missing CD-PRO/SS score for the bowel domain at Week 66 will be imputed using the MMRM as described in Section 4.10.

[redacted]



4.4.15 Subgroup Analyses

The following subgroup analyses will be conducted on the co-primary endpoints for both the Induction and Maintenance Phase.

- TNF status
- CS at baseline
- IS at baseline
- CDAI score at baseline
- Age
- Gender
- Region
- ADA status
- Disease location
- Disease duration
- Induction dose – Maintenance only

Subgroup categories may be combined if numbers are too low (e.g., $n < 10$) to ensure a meaningful assessment.

4.5 INTERIM ANALYSIS

The GA29144 study contains distinct patient cohorts for the Induction phase, where the cohorts are independent of each other. Cohorts 1 and 2 are exploratory Induction cohorts that have been unblinded and analyzed previously by the Sponsor to support decision making for Cohort 3. In particular, Cohort 1 was considered to be a Phase II study within the GA29144 Phase III trial, and Cohort 2 is open-label etrolizumab without a placebo control arm.

After all patients in Cohort 1 completed the Week 14 visit (or withdrew early), a preliminary, exploratory analysis was performed that led to the Protocol being amended in global version 6 for the following:

- Primary endpoint for Induction and Maintenance
- Inclusion criteria
- Reduced sample size for Cohort 3 due to changes in type I error control strategy for testing of the two etrolizumab doses

The protocol amendment took place prior to enrollment of any patients in Cohort 3.

A summary of the preliminary results for key endpoints will be included in the CSR for background and context. The CSR main results section however will report data from Cohorts 1 and 2 using the final locked database. It is acknowledged that data may

change slightly between the preliminary and final analysis, although these are expected to be minimal. Further, it is acknowledged that the analysis methods may differ.

No interim analysis is planned for the pivotal Induction Phase (i.e., Cohort 3) or for the pivotal Maintenance Phase analysis.

No type I error adjustment for evaluating multiple Induction cohorts will be undertaken since the pivotal analysis for Induction will take place on an independent set of patients (Cohort 3). Furthermore, no analyses have been conducted on the Maintenance Phase data, and treatment assignment for the pivotal treatment arms in Maintenance Phase (i.e., etrolizumab/placebo and etrolizumab/etrolizumab 105 mg) remains blinded to the Sponsor until formal database lock.

4.6 EFFICACY DERIVATIONS

4.6.1 Study Day 1 Definition

Study Day 1 is defined as the day of first receipt of study drug.

4.6.2 Baseline Definition

Baseline is defined as the last available assessment prior to first receipt of study drug.

4.6.3 Abdominal Pain and Liquid/Soft Stool Frequency

Patients are to report their symptoms daily throughout the duration of study using the e-diary devices provided by the Sponsor. The AP and SF scores will be determined weekly, as follows:

The score will be taken as the average of 4-7 daily scores within a 7-day window. Since the endoscopy procedure and associate bowel preparations will affect outcomes, the selected scores used in the calculation will exclude the days surrounding such procedures. Therefore, the 7 days will be taken prior to the earliest of the following dates:

- Visit date or projected date (if visit date unavailable/not applicable)
- Endoscopy date - 1
- Bowel preparation date

At least 4 (not necessarily consecutive) non-missing scores will need to be available to calculate the average; otherwise the patient's score will be flagged as missing.

The abdominal pain is reported on a 0-3 ordinal scale (0=none, 1=mild, 2=moderate, 3=severe), stool frequency denotes the number of loose or liquid stools passed by the patient. A higher score indicates worse disease activity for both abdominal pain and stool frequency.

4.6.4 Crohn's Disease Activity Index

The CDAI score is a composite endpoint of 8 sub-scores that are weighted and summed together, as shown in Appendix 9 of the Protocol, Version 7. Sub-scores 1-3 consist of patient reported outcomes, as reported daily using e-diaries provided by the Sponsor. Sub-scores 4-8 are assessed by the investigator at site visits. The CDAI score ranges from 0-600, where higher scores indicate worse disease activity.

CDAI Remission is defined as an absolute CDAI score<150.

CDAI-70 response is defined as a decrease from CDAI baseline score of at least 70 points.

CDAI-100 response is defined as a decrease from CDAI baseline score of at least 100 points.

4.6.5 Simplified Endoscopic Score for Crohn's Disease

The SES-CD assessment consists of 20 individual component scores, where 4 characteristics are measured across 5 segments of the intestine. Each score can take a value between 0-3, giving rise to an overall SES-CD total score of 0-60, with higher scores indicating worse disease.

The endoscopic assessment will undergo an independent central reading process, where up to 3 readers (local reader, central reader 1, and central reader 2) will determine a SES-CD score via an adjudication model. It should be noted that during Cohort 1 Induction phase, all three readers were always required, and a fixed 2 point scale adjudication model was used to determine subject eligibility. This later changed to a sliding scale adjudication model, and the 2nd central reader was only requested if required (i.e., non-agreement of the first two readers). For analysis purposes to ensure consistent handling across cohorts, the sliding scale adjudication model will be used for all endoscopic outcomes described in this SAP. Details of the adjudication process can be found in the Bioclinica Independent Review Charter, latest Version this includes details for both the 2 point scale and sliding scale models.

Partial scores may arise when less than 20 individual scores are available for evaluation. In particular, scores for an entire segment may be missing due narrowing, video quality (applicable to central reads only), or procedural reasons at site. Three data handling methods will be used and are described below; this will take place at a reader level when determining the total SES-CD score for a given reader, unless specified otherwise.

SES-CD using segments available at Baseline

To enable a like-for-like assessment between baseline and the post-baseline visit, the segments evaluable at baseline will first be assessed and the post-baseline SES-CD score handled as follows:

- Segment missing at baseline and post-baseline: the segment will not be used when calculating the SES-CD total score at both visits.
- Segment missing at Baseline: the post-baseline SES-CD score will only consider the segments that were available at baseline when determining the total SES-CD score. Any segments that were not evaluable at baseline that are later evaluable at the post-baseline visit will be ignored.
- Segment missing post-baseline: If the segment is missing for the central reader(s) only, the central reader(s) score will not be used for the adjudication process when deriving the post-baseline total SES-CD score. If the segment is missing for all three readers then the patient will be set to non-responder for the response/remission binary endpoints.

SES-CD using all available scores

The SES-CD total score at a visit will be calculated using all available scores. For post-baseline visits, all available segments will be included in the total score calculation irrespective of whether the segment was available at baseline or not.

SES-CD using Baseline Observation Carried Forward (BOCF)

Segments that are available at Baseline but missing at a post-baseline visit will have the segment score imputed using the Baseline segment score.

See Appendix 10 of the Protocol Version 7 for SES-CD scoring form.

[redacted]

[redacted]

4.7 PHARMACOKINETIC AND PHARMACODYNAMIC ANALYSES

Patients in all cohorts will be required to provide blood samples for population pharmacokinetic (PK) analysis and pharmacodynamics (PD) characterization.

Patients had the option to consent and participate in a PK/PD sub study to collect samples at additional time points.

4.7.1 Pharmacokinetic Analyses

Serum concentrations at various times during Induction and Maintenance Phases will be listed and summarized by descriptive summary statistics including means, geometric means, ranges, standard deviation and coefficients of variation. Analyses will be split by treatment group and cohort as appropriate.

Individual and mean concentration versus time data will be tabulated and plotted if more than two post dose time points are available.

The PK data from each Cohort will be combined to form a population PK analysis, details of which will be documented in a separate PK analysis plan. Population typical value of PK parameters will be estimated for the entire study population, along with estimates of intra/interpatient variance and estimation of random error. Individual patient parameter estimates will be computed using the post hoc analysis procedure. Impacts of covariates on relevant PK parameters will also be evaluated.

4.7.2 Pharmacodynamic Analyses

PD biomarkers to be analyzed are as follows:

- Serum concentrations of sMAdCAM-1 (in all study enrolled patients)
- $\beta 7$ receptor occupancy on peripheral blood "intestinal" homing beta7^{high} T and B lymphocyte subsets as measured by flow cytometry (data from patients enrolled in the PKPD sub-study only)

Data will be evaluated based on availability, and expressed as percentage of pre-dose baseline value over time. Absolute value and/or percentage change from baseline will be listed and summarized by descriptive summary statistics at each study time point, including but not limited to means, standard deviation, medians, minimum and maximum. Analyses will be split by treatment group and cohort as appropriate.

Additional [redacted] analyses or modelling may be conducted as appropriate.

4.8 BIOMARKER ANALYSES

Additional biomarker strategies and analyses will be detailed in the Biomarker Analysis Plan (BAP).

4.9 SAFETY ANALYSES

Safety data will be assessed through descriptive summaries using the Induction and Maintenance Safety populations respectively. All data available up until the clinical cut-off date (CCOD) will be reported. No formal statistical comparisons between treatment groups will be conducted.

Safety evaluations for the Induction Safety population will include data from first dose of treatment in the Induction Phase until the earliest of the following: withdrawal during the Induction treatment period (including withdrawal at Week 14 for patients not eligible for Maintenance), first Maintenance dose, withdrawal or completion of Safety Follow-Up.

Safety evaluations for the Maintenance Safety Population will include data from first dose of treatment in the Induction Phase (i.e., data from the Induction Phase will be included) until the latest of the following: withdrawal from the Maintenance treatment period, completion of Maintenance treatment period, withdrawal or completion of Safety Follow-Up.

4.9.1 Exposure to Study Drug

Exposure to study treatment will be summarized by:

- treatment duration (weeks)
- number of patients receiving a dose at each visit
- number of doses received/missed for a patient

A listing of any dosing errors (i.e., incorrect treatment received) will be produced, if applicable.

In the event that a treatment group cannot be attributed to a Kit ID, it will be assumed that the patient received the planned (i.e., randomized) treatment.

4.9.2 Adverse Events

Adverse events (AEs) will include all terms recorded on the AE Case Report Form (CRF) pages (except pregnancies). For each recorded AE, the term entered by the investigator describing the event (the “reported term”) will be assigned a standardized term (the “preferred term”) and assigned to a superclass term on the basis of the Medical Dictionary for Regulatory Activities (MedDRA) World Health Organization (WHO) dictionary of terms. All analyses of AE data will be performed using the preferred terms unless otherwise specified.

For the etrolizumab program, the adverse events of special interest (AESIs) are the following:

- Systemic hypersensitivity reactions and anaphylactic and anaphylactoid reaction which will be reported using the MedDRA anaphylactic reaction Standard MedDRA Query (SMQ) algorithmic and Hypersensitivity SMQ narrow.
- Neurological signs, symptoms, and AEs that may suggest possible progressive multifocal leukoencephalopathy (see Appendices 5 and 6 of Protocol)
- Suspected transmission of an infectious agent by the study drug

- Cases of potential drug-induced liver injury that include an elevated alanine aminotransferase (ALT) or aspartate aminotransferase (AST) in combination with either an elevated bilirubin or clinical jaundice, as defined by Hy's Law

Specific AEs listed below will also be reported:

- Serious infections
- Gastrointestinal Infections
- Opportunistic infections
- Malignancies
- Injection site reactions

A listing and if applicable, summary table (e.g., $n > 10$) of all AEs suspected, confirmed or associated with COVID-19 will be generated. Associated AEs are defined as all AEs reported within a time window of 7 days prior and 30 days after the confirmed COVID-19 start date (dates inclusive). Additionally, a listing of patients experiencing 'long COVID-19' will be generated. This listing will include all AEs with a duration >30 days occurring after a confirmed COVID-19 infection or a positive PCR test.

All AESIs and specific AEs will be determined using a selection of eCRF tick box or MedDRA SMQs, Adverse Events Group Terms (AEGTs), High Level Term (HLTs), High Level Group Terms (HLGTs), System Organ Class (SOC), as appropriate.

Summary tables will be generated for AEs, serious adverse events (SAEs), deaths, AEs leading to discontinuation of study drug, and AESIs. A listing will be generated for AEs, deaths, selected AESIs/specific AEs, and COVID-19 AEs. Pregnancies will be reported as narratives in each individual study CSR.

Summary tables of AEs will summarize the incidence of treatment-emergent AEs only. Treatment-emergent events are defined as any new AE reported or any worsening of an existing condition on or after the first dose of study drug. If the onset date of the AE is prior to the day of first dose, the AE will be considered treatment emergent only if the most extreme intensity is greater than the initial intensity (i.e., the intensity for a given AE increases and its end date is on or after the date of the first dose). For all summary tables, the AEs will be sorted by SOC (in decreasing order of overall incidence) and then by preferred term (PT) (in decreasing order of overall incidence).

For each treatment group, the incidence count for each AE PT will be defined as the number of patients reporting at least one treatment-emergent occurrence of the event (multiple occurrences of the same AE in 1 patient will be counted only once). The proportion of patients with an AE will be calculated as the incidence count divided by the total number of patients in the population. Each table will also present the total number of AEs reported where multiple occurrences of the same AE in an individual are counted separately.

To account for potential exposure differences between arms during the Maintenance phase (e.g., due to differences in drop-out rate), rate tables will be generated for selected AEs of interest. The rate per 100 patient years and 95% CIs will be summarized by treatment group and will be calculated by:

$$AE\ Rate\ (per\ 100\ patient\ years) = \frac{Total\ number\ of\ AEs}{Total\ number\ of\ patient\ years\ at\ risk}\ x\ 100$$

where the total patient-years at risk is the sum over all patients of the time intervals (in years) from the first dose of study treatment in the Induction phase up until the patient completes/withdraws from the study.

All summary tables and listings will report AEs using the safety population split by treatment arm. For the Induction phase, cohorts will be summarized separately.

In addition to evaluations on the Maintenance Safety Population which include AEs during the Induction phase, evaluations will also be made on the Maintenance Safety Population that only include AEs that occur during the Maintenance Phase (i.e., AEs occurring during Induction will be excluded).

4.9.3 Laboratory Data

Laboratory abnormalities and the patient's worst Common Terminology Criteria for Adverse Events (CTCAE) grade during study will be summarized for hematology and serum chemistry parameters, in addition to change from Baseline summaries.

Elevated liver enzyme tests will be summarized by the following upper limit of normal (ULN) categories as these are indicators of severe liver injury:

- ALT or AST>3ULN and total bilirubin>2 ULN as defined by Hy's law
- ALT or AST>3 ULN

4.9.4 Immunogenicity

The number and percentage of patients with positive serum antibodies to etrolizumab at baseline and during the study will be summarized by treatment arm.

4.9.5 Vital Signs

Vital signs will be summarized using summary statistics and change from baseline. The proportion of patients experiencing clinically significant changes relative to baseline will be reported if appropriate.

4.9.6 Electrocardiogram

A shift table for the qualitative ECG assessments will produced, summarizing the Baseline and worst post-baseline results.

4.9.7 Medical History

Medical history data collected in the electronic-CRF (eCRF) will be summarized using summary statistics, reporting the proportion of patients with at least one medical condition and the total number of medical conditions. The medical conditions will then be split out by type.

4.9.8 Concomitant Medications

Concomitant medications include any medication being used at any time from first dose of study drug through to 7 days after last dose of study drug, or medication being used at any time up to the start of study treatment. The data will be summarized, and report the total number of patients taking at least one medication, and total number of medications. Summaries will also be split by medication class and preferred medication. Medication terms will be mapped and reported using the WHO drug dictionary.

4.10 MISSING DATA

Missing data for longitudinal continuous endpoints (e.g., CD-PRO/SS) will be imputed as part of the MMRM analysis framework, as described in Section 4.4.1.

For e-diary data excluding CD-PRO/SS, patients with insufficient daily diary entries (less than 4 out of the 7 days) for calculation of the weekly average will be flagged. For these patients, the 7-day window will be extended up to 3 additional days to a maximum of a 10-day window until there are 4 diary entries available. For example, if diary entries are only available for 3 out of the 7 days prior to the visit, the 4^{h} entry will be used from either the 8th, 9th, or 10th day, depending on when the 4^{h} entry is first observed. These 4 entries will then be used to define clinical remission, and to calculate the 7-day total needed for the CDAI sub-component. If 4 diary entries are not available within the extended 10-day window, patients will be set to non-responders in the clinical remission and CDAI endpoints respectively. Patients will be flagged if their clinical remission is calculated using the extended 10 day window, and separate sensitivity analyses will be performed to assess robustness under alternative missing data handling methods (see Sections 4.4.6 and 4.4.11).

All other data that are exploratory or descriptive in nature will handle missing data as detailed in Appendix 2

Missing data for endoscopy endpoints at Weeks 14 and 66 will be imputed using Multiple Imputation as described in Section 4.10.2.1.

4.10.1 Sensitivity Analysis

Sensitivity and robustness to LOCF, using a wider 10 day window for e-diary data, and departures from missing data assumptions in MICE, will be investigated using non-responder imputation (NRI) for the co-primary endpoints.

In addition, to assess robustness of the approach for imputing missing values, multiple imputation will be performed for the clinical remission (see Section 4.10.2.2) and CDAI remission (see Section 4.10.2.3) endpoints as sensitivity analyses.

A tipping point analysis will be conducted to explore the impact of varying the assumptions about missing outcomes (MAR) on the treatments arms of interest e.g., etrolizumab 210 mg and placebo or etrolizumab/etrolizumab 105 mg and etrolizumab/placebo. This will be performed as a sensitivity to the missing data handling methods for co-primary endpoints for induction and maintenance. The tipping point is defined as the difference in the number of missing data on clinical remission/endoscopic improvement between the treatment groups that result in a change in the primary outcome conclusions (Yan et al. 2009). A two-dimensional plot will be produced for each primary comparison of the etrolizumab arm vs. the comparator to evaluate where the tipping point lies. The clinical plausibility of the combinations on the boundary will be discussed in the CSR to evaluate robustness of study conclusions to missing data. Within the tipping point analysis, the intercurrent events of withdrawal, rescue therapy and death will continue to follow the composite strategy and will be set to non-responders. COVID-19 is an intercurrent event for the endoscopy co-primary endpoint only and yields non-observable data due to the pandemic impact making the site based assessments unfeasible. The patients with non-observable data due to COVID-19 intercurrent event will also be included in the tipping point analysis to test the MAR assumption that was implied in the multiple imputation model used in the primary analysis method.

4.10.2 Multiple Imputation Strategy

The endpoints below will be imputed using a Fully Conditional Specification (FCS) logistic regression model. This is also known as multiple imputation by chained equations (MICE) and will follow the method described in White et. al (2011). The model for imputation was developed based on Cohort 1 unblinded data for Induction and blinded data for all Cohorts in Maintenance. Development of the imputation model for each of the endpoints is described below and follows guidance by White et. al (2011) and van Buuren et. al (1999). Separate imputation models for each Induction and Maintenance phase were chosen. Data will be imputed after implementing the ICE strategy for events that relate to treatment withdrawals, rescue medications and death which will be handled by non-responder imputation under the composite strategy. COVID-19 related events for CDAI and Endoscopy endpoints and any other remaining missing data (i.e., true missing data), will be multiply imputed.

A candidate list of variables was put forward based on clinical knowledge and significant predictors for inclusion in the final imputation model were chosen using a stepwise model selection procedure using the corrected AIC method. The imputation model will also include treatment and all IxRS stratification factors as listed in Section 4.4.2 for both Induction and Maintenance as forced variables for inclusion in the analysis model.

Correlations were also investigated between the auxiliary variables and the endpoint of interest, where point-biserial correlation was investigated between the continuous variables and the binary endpoint of interest. If correlation exceeded |0.2|, these would be included in the final imputation model regardless of the results from the model selection procedure. The selection of auxiliary variables was performed individually for each endpoint in Induction and Maintenance. However, for consistency, the union of significant auxiliary variables for each endpoint was chosen as a single final imputation model.

The number of burn-in iterations of 100 and a seed of 983164 will be used with 15 imputed datasets. A Wilson-Hilferty transformation is also applied to the CMH test statistic for normalization (Ratitch et al. 2013). Newcombe confidence intervals were obtained for the difference in proportions estimated from the CMH test. The standard errors associated with the Newcombe confidence intervals for the difference in proportions from the CMH test are also required in order to pool analyses using Rubin's rules. However, this is not obtained as part of the CMH output in SAS. Therefore, these were calculated using the following:

$$Standard\ Error = \frac{(Newcombe\ Upper\ Limit - Newcombe\ Lower\ Limit)}{Z}$$

Where $Z = 3.29$ for 90% confidence intervals, or $Z = 3.92$ for 95% confidence intervals (Higgins et al. 2021).

4.10.2.1 Imputation Model for Endoscopy Endpoints

Patients with partially available SES-CD scores are handled as part of the efficacy derivations, see Section 4.6.5. Providing that an ICE has not occurred (as described in Section 4.4.3), patients where SES-CD scores that are not available from any reader will be treated as missing and assumed to be MAR. This will lead to missing response for endoscopic improvement and endoscopic remitters for the following endpoints:

- Induction:
  - Endoscopic Improvement at Week 14
  - Endoscopic Remission at Week 14
  - SES-CD=0 at Week 14
- Maintenance:
  - Endoscopic Improvement at Week 66
  - Endoscopic Remission at Week 66
  - Endoscopic Improvement at Week 66 Among Patients Achieving Endoscopic Improvement at Week 14
  - SES-CD=0 at Week 66

The SES-CD=0 endpoint is highly difficult to achieve, with few patients available to provide a reliable imputation model. Patients with missing data will be considered as non-responders for this endpoint.

The final imputation model for the Induction endpoints included the following:

- Treatment group (etrolizumab 105 mg, etrolizumab 210 mg, or Placebo)
- Concomitant oral CS treatment at baseline (yes or no)
- Prior anti-TNF exposure at baseline (yes or no)
- Baseline CDAI≤330 (yes or no)
- Sex
- Age at Baseline
- Albumin Level at Baseline
- [redacted]
- [redacted]
- Average AP score at Baseline
- Average AP score at Week 10
- Average AP score at Week 14
- Average SF score at Baseline
- Average SF score at Week 10
- Average SF score at Week 14
- SES-CD Score at Baseline
- [redacted]
- Disease location

The final imputation model for the Maintenance endpoints included the following:

- Treatment group (etrolizumab/etrolizumab or etrolizumab/Placebo)
- Induction dose regimen (high or low etrolizumab)
- Concomitant oral CS treatment at baseline (yes or no)
- Prior anti-TNF exposure at baseline (yes or no)
- Immunosuppressant use (yes or no)
- Sex
- CDAI score at Baseline
- [redacted]
- [redacted]
- Disease Duration
- SES-CD score at Baseline

- SES-CD score at Week 14
- Average AP score at Baseline
- Average AP score at Week 32
- Average SF score at Baseline
- Average SF score at Week 14
- Average SF score at Week 32
- Average SF score at Week 56
- Average SF score at Week 66
- [redacted]
- [redacted]
- Albumin Level at Baseline
- Albumin Level at Week 14
- Albumin Level at Week 66
- Baseline Age

The endoscopic improvement at Week 66 among patients achieving endoscopic improvement at Week 14 will only be conducted in the subset of patients with non-missing data at Week 14.

4.10.2.2 Imputation Model for Clinical Remission

Patient symptom data was extensively collected during the study using daily diaries, this data provides a weekly assessment of the patient's disease severity. The patient's abdominal pain and stool frequency scores at key timepoints during the study were selected preferentially to predict the missing Week 14/66 timepoint.

The variables to be included in the final multiple imputation model for clinical remission at Week 14 are:

- Induction treatment (etrolizumab 105 mg, etrolizumab 210 mg or Placebo)
- Prior aTNF experience (Yes or No)
- Corticosteroid use at baseline (Yes or No)
- Baseline CDAI score ($\leq$330 or$>$330)
- Average AP and average SF at Baseline
- Average AP and average SF at Week 6
- Average AP and average SF at Week 13
- Average AP and average SF at Week 14

The variables to be included in the final multiple imputation model for clinical remission at Week 66 are:

- Maintenance treatment (etrolizumab 105 mg or Placebo)
- Induction treatment (etrolizumab 105 mg or etrolizumab 210 mg)
- Prior aTNF experience (Yes or No)
- Corticosteroid use at baseline (Yes or No)
- Average AP and average SF at Week 32
- Average AP and average SF at Week 56
- Average AP and average SF at Week 65
- Average AP and average SF at Week 66

4.10.2.3 Imputation Model for CDAI Remission

In addition to applying ICE strategies before imputation, any patients with partial CDAI scores available that sum to more than 150 will be imputed to non-remitters. Since the CDAI score is formed from 8 sub-scores, the CDAI sub-scores were included in the model preferentially over baseline disease characteristics and laboratory data (except haematocrit). Sub-scores that are ordinal will be treated as factors in the model to ensure only valid numeric values are imputed by the model (e.g., abdominal mass can only take values of 0, 20 or 50).

The variables to be included in the final multiple imputation model for CDAI remission at Week 14 are:

- Induction treatment (etrolizumab 105 mg, etrolizumab 210 mg or Placebo)
- Prior aTNF experience at baseline (Yes or No)
- Corticosteroid use at baseline (Yes or No)
- Baseline CDAI score ($\leq$330 or$>$330)
- All 8 weighted CDAI subscores at Baseline
- All 8 weighted CDAI subscores at Week 10
- All 8 weighted CDAI subscores at Week 14

The variables to be included in the final imputation model for CDAI remission at Week 66 are:

- Maintenance treatment (etrolizumab 105 mg or Placebo)
- Induction treatment (etrolizumab 105 mg or etrolizumab 210 mg)
- Prior aTNF experience at baseline (Yes or No)
- Corticosteroid use at baseline (Yes or No)
- All 8 weighted CDAI subscores at Week 32
- All 8 weighted CDAI subscores at Week 56

- All 8 weighted CDAI subscores at Week 66

5. REFERENCES

Higgins JPT, Li and Deeks JJ. Chapter 6: Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J and Chandler J et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.2 (updated February 2021). Cochrane, 2021. Available from www.training.cochrane.org/handbook.

Ratitch B, Lipkovich I and O'kelly M. Combining Analysis Results from Multiply Imputed Categorical Data. PharmaSUG 2013 - Paper SP03.

van Buuren S, Boshuizen HC and Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18(6):681-94. doi: 10.1002/(sici)1097-0258(19990330)18:6<681::aid-sim71>3.0.co;2-r. PMID: 10204197.

White, I.R., Royston, P. and Wood, A.M. Multiple imputation using chained equations: Issues and guidance for practice. Statist. Med. 2011;30:377-399. https://doi.org/10.1002/sim.4067

Yan X, Lee S, Li N. Missing data handling methods in medical device clinical trials. J Biopharm Stat. 2009;19(6):1085-98.

Appendix 1 Outcome Measures Definitions

Clinical Remission	SF mean daily score ≤ 3 and AP mean daily score ≤ 1, with no worsening in either sub-score compared to baseline, averaged over the 7 days prior to visit.
Mean Stool Frequency Score	Daily number of liquid or very soft stools (unweighted mean over the 7 days immediately prior to the study visit) reported in the e-diary according to the Bristol Stool Form Scale, which will be provided. This score should not use e-diary entries that correspond to day(s) of bowel preparation, endoscopy or the day after endoscopy.
Mean Abdominal Pain Score	Daily abdominal pain score on a 3-point scale: 0=none, 1=mild, 3=severe (unweighted mean over the 7 days immediately prior to the study visit) reported in the e-diary. This score should not use e-diary entries that correspond to day(s) of bowel preparation, endoscopy or the day after endoscopy.
Endoscopic Improvement	A ≥ 50% reduction in the baseline SES-CD score
Endoscopic Remission	SES-CD ≤ 4 (≤ 2 for ileal patients), with no segment having a subcategory score > 1
SES-CD	A composite of four assessments, each rated from 0 to 3: size of ulcers, proportion of the surface covered by ulcers, proportion of the surface with any other lesions, and presence of narrowings (stenosis)
CDAI	A composite of eight assessments: number of liquid or soft stools (liquid/soft), abdominal pain, general well-being, presence of complications, taking Lomotil® (diphenoxylate/atropine) or other opiates for diarrhea, presence of an abdominal mass, hematocrit, and percentage deviation from standard weight.
CDAI-70 Response	A decrease from CDAI baseline score of at least 70 points
CDAI-100 Response	A decrease from CDAI baseline score of at least 100 points
CDAI Remission	CDAI score < 150
[redacted]	[redacted]
[redacted]	[redacted]
Durable Clinical Remission	Patients who achieve clinical remission at ≥ 4 of the 6 in-clinic assessment visits that are conducted during the 1 year of Maintenance phase
[redacted]	[redacted]

Appendix 2 Summary of Imputation Methods for Intercurrent Events (ICE) and Missing Data

INDUCTION PHASE

Endpoint	Intercurrent Event (Strategy)				Missing Data
	Treatment Discontinuation (Composite)	Rescue Medication (Composite)	Death (Composite)	COVID-19 (Hypothetical)	
Primary Efficacy Endpoints					
Clinical Remission at Week 14	NRI	NRI	NRI	n/a	10 day window/NRI*
Endoscopic Improvement at Week 14	NRI	NRI	NRI	MI	MI
Secondary Efficacy Endpoints					
Clinical Remission at Week 6	NRI	NRI	NRI	n/a	10 day window/NRI*
Endoscopic Remission at Week 14	NRI	NRI	NRI	MI	MI
CD-PRO/SS Score at Week 14 (Functional & Bowel Domain)	WOCF	WOCF	WOCF	MMRM	MMRM
CDAI Remission at Week 14	NRI	NRI	NRI	LOCF for non-diary sub-components; retain minimum 4/7 day window for diary data	LOCF for non-diary sub-components; 10 day window for diary components/ NRI*

Appendix 2 Summary of Imputation Methods for Intercurrent Events (ICE) and Missing Data (cont.)

MAINTENANCE PHASE

Endpoint	Intercurrent Event (Strategy)				Missing Data
	Treatment Discontinuation (Composite)	Rescue Medication (Composite)	Death (Composite)	COVID-19 (Hypothetical)	
Primary Efficacy Endpoints					
Clinical Remission at Week 66	NRI	NRI	NRI	n/a	10 day window/NRI*
Endoscopic Improvement at Week 66	NRI	NRI	NRI	MI	MI
Secondary Efficacy Endpoints					
Clinical Remission at Week 66 Among Patients in Clinical Remission at Week 14	NRI	NRI	NRI	n/a	10 day window/NRI*
Endoscopic Improvement at Week 66 Among Patients Achieving Endoscopic Improvement at Week 14	NRI	NRI	NRI	MI	MI
Endoscopic Remission at Week 66	NRI	NRI	NRI	MI	MI
CDAI Remission at Week 66	NRI	NRI	NRI	LOCF for non-diary sub-components; retain minimum 4/7 day window for diary data	LOCF for non-diary sub-components; 10 day window for diary components/ NRI*
CD-PRO/SS Score at Week 66 (Functional & Bowel Domain)	WOCF	WOCF	WOCF	MMRM	MMRM
Durable Clinical Remission	WOCF	WOCF	WOCF	n/a	10 day window/NRI*
Corticosteroid-Free Clinical Remission for 24 Weeks at Week 66 Among Patients Receiving Corticosteroids at Baseline	NRI	NRI	NRI	n/a	10 day window/NRI*

Appendix 2 Summary of Imputation Methods for Intercurrent Events (ICE) and Missing Data (cont.)

[redacted]

Appendix 2 Summary of Imputation Methods for Intercurrent Events (ICE) and Missing Data (cont.)

CDAI = Crohn's Disease Activity Index; CD = Crohn's Disease; CD-PRO/SS = Crohn's Disease- Patient Reported Outcome/Signs and Symptoms (functional, bowel, [redacted] domains); COVID-19 = Coronavirus Disease-19; LOCF = last observation carried forward; MI = multiple imputation; MMRM = mixed models for repeated measures; NRI= non-responder imputation; SES-CD = simplified endoscopic index for Crohn's disease; WOCF =worst-observation-carried-forward.

* The 10 day extended window will first be applied, followed by non-responder imputation if 4 days ediary data aren't available within a maximum of 10 days.

[redacted]