Statistical Analysis Plan	
Study Code	D3255C00001
Edition Number	4.0
Date	2 September 2022

A Multicenter, Randomized, Double-blind, Parallel-group, Placebo controlled Study to Investigate the Use of Benralizumab as a Treatment Option for Eosinophilic Esophagitis (MESSINA)

Summary of Changes to the Statistical Analysis Plan (SAP)			
Section # and Name	**Description of Change**	**Brief Rationale**	**Substantial /Non-substantial**
Changes since version 3.0			
Section 2.1 Definition of analysis sets	Adding early time point sub-study subjects set	Analyses for early time point sub-study subjects will use this set.	Non-substantial
Section 4.2.5.14 Subgroup analysis	Updated age group subgroup analyses to include <18 vs ≥18 and also ≤21 vs >21 year old analyses	Collapse groups to enable more meaningful analysis of age groups of interest, add supportive young adults (≤21 yr old) subgroup analysis..	Non-substantial
Section 6 CHANGES OF ANALYSIS FROM PROTOCOL	Added PEESS will be descriptively summarized.	Not enough subjects to perform statistical analyses.	Non-substantial
Section 8.1 Accounting for missing data	Clarified the tipping point analyses will only tip subjects with missing data who are not imputed with return-to-baseline MI due to occurrence of treatment failure intercurrent events	Discontinued subjects will be handled by return-to-baseline MI	Non-substantial
Section 3.4.8 Patient Global Impression of Change (PGI-C)	Updated PGI-C score assignment.	Correct an error to ensure consistent with data collection standards.	Non-substantial
Section 3.1.1 Visit window definitions	Added explanation for Week 24 windowing rule	Clarification for how to define Week 24 visit window in relation to first dose of open label benralizumab date.	Non-substantial
3.4.1 Dysphagia Symptom Questionnaire (DSQ) score	Added explanation for 30% improvement threshold in the treatment responder endpoint. Additional threshold pre-defined for supportive responder analyses (-18 points, 50% improvement) after reviewing blinded anchor analyses results.	Pre-define and justify thresholds for clinically meaningful changes from baseline in DSQ scores to use in responder analyses, per health authority feedback	Non-substantial
3.4.1 Dysphagia Symptom Questionnaire (DSQ) score	Added a sensitivity analysis for alternative missing day rules.	Further assess the missing daily score impact on the 14 day overall DSQ score, per health authority feedback	Non-substantial

Summary of Changes to the Statistical Analysis Plan (SAP)			
Section # and Name	**Description of Change**	**Brief Rationale**	**Substantial /Non-substantial**
4.2.4.3 Primary analysis method - change from baseline in DSQ score	Specified and justified target changes / categories for PGI-S and PGI-C to anchor change in DSQ score to	To support anchor analyses to determine clinically meaningful improvement thresholds, per health authority feedback	Non-substantial
8.1.1.4 Tipping point analyses	Updated tipping point analysis methods for the histological response rate endpoint	To align with health authority feedback	Non-substantial
4.2.6.2 Laboratory data	Removed Hy's law section	Not required	Non-substantial
Changes since version 2.0			
Section 3.2 and all WOCF imputed endpoints	Change WOCF to return-to-baseline MI for continuous change from baseline endpoints where treatment failure intercurrent events occur	Updated imputation rule to incorporate uncertainty around imputations using a multiple rather than single imputation approach.	Substantial
Section 8.1 Accounting for missing data	Adding Section 8.1.1.7 for tipping point sensitivity analyses for dual primary endpoints	To explore the plausibility of the missing data assumptions under which the conclusions of the analyses change.	Non-substantial
Section 4.1, 4.2.5.1-4	Adding percent CFB in tissue eosinophils at Week 24, CFB in EoE-HSS grade score at Week 24, CFB in EoE-HSS stage score at Week 24, treatment responder rate at Week 24 as key secondary endpoints.	Updated based on updated CSP.	Substantial
Section 8.3 Partial dates for adverse events and prior/concomitant medications	Updated the imputation rules for partial date AE/CM	Previous rules have some issues which does not work for prior and concomitant meds.	Non-substantial
Section 3.6.1 Adverse Events	Updated the end date rule for TEAEs in the on-treatment period	The visit window should be 3 days, not 7 days.	Non-substantial
Section 4.2.4.3 Primary analysis method –Change from baseline in DSQ score	Added a paragraph for CFB DSQ at Week 24 by PGI-S improvement summaries.	To provide support for suitable MCID scores for the DSQ endpoint	Non-substantial

Summary of Changes to the Statistical Analysis Plan (SAP)			
Section # and Name	**Description of Change**	**Brief Rationale**	**Substantial /Non-substantial**
Section 4.2.5.10 Patient Global Impression of Severity (PGI-S) (PRO)	Added shift table of change from baseline in PGI-S by PGI-S at baseline.	To further support the PGI-S variable.	Non-substantial
Section 3.3.6 Healthcare resource utilization	Added as a secondary endpoint	Missed in pervious SAP version.	Non-substantial
Section 4.2.5.16 Healthcare resource utilization	Added as a secondary endpoint	Missed in pervious SAP version.	Non-substantial
Section 3.4.2 EoE-3D and other symptom questions	Updated the missing rule for 28-day dysphagia episode frequency: updated from 20 out of 28 days to >=8 days in each 14-day period.	To be consistent with DSQ missing rule.	Non-substantial
Section 8.1.1.5 Sensitivity analyses using both MAR and MNAR assumptions	Removed	Since now primary analysis accounts for most sources of missing data, and that along with the tipping point analysis is sufficient to explore the robustness of results. No such sensitivity analyses required anymore.	Non-substantial
Section 8.1.1.6 On treatment Analyses	Removed	Same as above	Non-substantial
Section 8.2 Analysis plan for immunogenicity data	Updated ADA groups	Updated to align with other benralizumab studies	Non-substantial
Overall	Updated ANCOVA model covariate: remove prior response to steroids use for EoE, add baseline steroid use.	To correct inconsistency with the Protocol	Non-substantial
Section3.3.2 EoE histology scoring system (EoE-HSS)	Specify how to derive total score and feature score.	To make the definitions clearer.	Non-substantial

Summary of Changes to the Statistical Analysis Plan (SAP)			
Section # and Name	**Description of Change**	**Brief Rationale**	**Substantial /Non-substantial**
Section 3.4.2 EoE-3D and other symptom questions	Adding the definition of Frequency adjusted severity scores.	For potential analysis perform on Frequency adjusted severity scores	Non-substantial
Section 4.2.5.14 Subgroup analysis	Remove subgroup "Changed EoE medications during first 24 weeks (Yes, No)".	Related analyses will be done in EAP.	Non-substantial
Section 4.2.5.14 Subgroup analysis	Added section for hierarchical Bayesian borrowing methods for age group.	To pre-define analyses that allow data borrowing should the data support it to increase precision around adolescent treatment effects	Non-substantial
Section 4.2.6 Safety analysis	Added subgroup analyses for exposure and AE overview.	To explore exposure and AE in different subgroups.	Non-substantial
Section 4.2.6.2 Laboratory data	Removed liver function related summaries.	Liver function summaries no longer needed.	Non-substantial
Overall	Update to have MI with MAR assumption for missing data as primary analysis.	Accounting in all the missing data at collection.	Substantial
Section 3.4.8 Patient Global Impression of Change (PGI-C)	Update PGI-C score assignment.	To be consistent with data collection standards.	Non-substantial
Section 3.3.3 Centrally-read and Investigator-read EoE EREFS	Added two more categories for EREFS summary.	To better understand EREFS results.	Non-substantial
Overall	Removed +/- 2 weeks window for eosinophil counts.	Include all available data for analyses.	Non-substantial
Changes since version 1.0			
Global	The word 'optional' was removed from the description of the OLE period.	Participation in the OLE period will be offered to all patients who are eligible.	Non-substantial

Summary of Changes to the Statistical Analysis Plan (SAP)			
Section # and Name	**Description of Change**	**Brief Rationale**	**Substantial /Non-substantial**
Section 3.2 Primary outcome variables	Update estimand approach to a composite strategy. Intercurrent events of randomised therapy discontinuation, changes to background medication, addition of a new therapy for EoE or dilation procedures are handled with imputations: non-responders for binary endpoints, WOCF for continuous change from baseline endpoints.	To ensure the analyses better account for these intercurrent events which are considered treatment failures outcomes.	Substantial
Section 3.6.3 Weight	Removed	Combined weight into Vital signs section.	Non-substantial
Section 3.6.4 Vital Signs	Add weight. Change to Vital signs and weight.	Combined weight into Vital signs section.	Non-substantial
Section 4.2.4.X, 4.2.5.X	Clarify the model specification, change MMRM to ANCOVA	As a result of the updated estimand approach to a composite strategy with imputation for treatment failure events.	Substantial
Section 4.2.6.5 Weight	Section deleted.	Combined into Section 4.2.6.6 Vital sign and Weights	Non-substantial
Section 4.2.6.6 Vital sign and Weights	Wording updated Add weight	Clarify parameters to be presented Add weight	Non-substantial
Section 8.1.1.3 Sensitivity analysis under the effectiveness estimand using the Missing at Random (MAR) assumption.	Add paragraph to describe effectiveness MMRM analysis for sensitivity analyses.		Non-substantial

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation or special term	Explanation
ADA	Anti-drug antibodies
AE	Adverse Event
ALT	Alanine aminotransferase
ANCOVA	analysis of covariance
AST	Aspartate aminotransferase
BLQ	Below the lower limit of quantification
BMI	Body mass index
CFB	Change from baseline
CI	Confidence interval
CMH	Cochran-Maentel-Haenszel
COVID-19	Coronavirus disease 2019
CRF	Case Report Form (electronic/paper)
CSP	Clinical Study Protocol
CSR	Clinical Study Report
DAE	Discontinuation of investigational product due to adverse event (AE)
DB	Double-Blinded
DILI	Drug-induced liver injury
DL	Direct Likelihood approach
DRMI	Dropout Reason-based Multiple Imputation
DSQ	Dysphagia Symptom Questionnaire
EAP	Exploratory Analysis Plan
ECG	Electrocardiogram
ECL	electrochemiluminescent
eCRF	Electronic Case Report Form
EndoFLIP	Endolumenal Functional Lumen Imaging Probe
EoE	Eosinophilic esophagitis
EoE-3D	Eosinophilic Esophagitis - Daily Dysphagia Diary
EoE-HSS	Eosinophilic Esophagitis - Histology Scoring System
EoE-QoL-A	Adult Eosinophilic Esophagitis Quality of Life Questionnaire
eos	Eosinophils
EOT	End of Treatment
EREFS	Endoscopic Reference Score
FAS	Full Analysis Set
FU	Follow up
Gamma GT	Gamma-Glutamyl Transferase

Abbreviation or special term	Explanation
GGT	Gamma-Glutamyl Transferase
hpf	high power field
HRQoL	health-related quality of life
HSS	Histology Scoring System
IgE	Immunoglobulin E
IP	Investigational Product
IPD	Investigational Product Discontinuation (visit)
ITT	Intent to treat
LLOQ	Lower limit of quantification
LSMD	Least Square Mean Difference
MAR	Missing At Random
MCAR	Missing Completely At Random
MCID	Minimal Clinically Important Differences
MCS	Mental health Component Summary
MCP	Multiple Comparison Procedure
MedDRA	Medical Dictionary for Regulatory Activities
miRNA	micro RNA
MMRM	Mixed effect Models for Repeated Measures
MNAR	Missing Not At Random
M-N	Miettinen and Nurminen
nAb	neutralizing antibodies
MI	Multiple imputation
NRPS	Numeric Rating Pain Scale
OL	Open-Label
OLE	Open-Label Extension
PCS	Physical health Component Summary
PEESS	Pediatric Eosinophilic Esophagitis Symptom Severity
PGI-C	Patient Global Impression of Change
PGI-S	Patient Global Impression of Severity
PK	Pharmacokinetics
PI	Principal Investigator
PPI	Proton Pump Inhibitor
PRO	Patient-Reported Outcome
PT	Preferred Term
Q4W	Every 4 weeks
Q8W	Every 8 weeks

Abbreviation or special term	Explanation
RNA	Ribonucleic acid
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SF-36 v2	Short Form-36 Version 2.0
SoA	Schedule of Assessments
SOC	System Organ Class
TBL	Total bilirubin
TBNK	T-cell, B-cell, NK-cell
TEAEs	Treatment-Emergent Adverse Events
ULN	Upper Limit of Normal
WHO	World Health Organization
WPAI+CIQ	Work Productivity and Activity Impairment questionnaire plus Classroom Impairment Questions

1 STUDY DETAILS

1.1 Study objectives

This statistical analysis plan (SAP) outlines the analyses to be generated for the global clinical study report (CSR). Additional analyses required for regional submissions will be pre-specified in a separate analysis plan and will be submitted to the appropriate authorities.

The study objectives and endpoints for the 52-week study period (DB+OL treatment periods) and OLE treatment are presented in Section 1.1.1 and Section 1.1.2, respectively.

1.1.1 Study Objectives for the 52-week Study Period (DB+OL Treatment Periods)

The following objectives/endpoints are for the 52-week study period (DB+OL treatment periods):	
Primary Objective:	**Dual-primary Endpoints/Variables:**
To evaluate the effect of benralizumab 30 mg Q4W on histologic signs and symptoms of EoE in patients with symptomatic and histologically active EoE	• Proportion of patients with a histologic response at Week 24, defined as a peak esophageal intraepithelial eosinophil count ≤ 6 eos/hpf • Changes from baseline in DSQ score at Week 24
Secondary Objectives:	**Endpoints/Variables:**
To evaluate the effect of benralizumab 30 mg Q4W on clinical features of EoE and disease activity	• Key secondary endpoint: Percent change from baseline in tissue eosinophils at Week 24 • Key secondary endpoint: Changes from baseline in EoE-HSS grade score at Week 24 • Key secondary endpoint: Changes from baseline in EoE-HSS stage score at Week 24 • Key secondary endpoint: Changes from baseline in centrally-read EoE EREFS at Week 24 • Key secondary endpoint: Treatment responder rate at Week 24, defined as a composite of histologic response (≤6 eos/hpf) and clinically meaningful improvement from baseline in DSQ score (30% improvement). • Centrally-read biopsies for additional histopathology including tissue eosinophil counts at Week 24 • Dysphagia-free days as captured by the DSQ • Frequency of dysphagia episodes as captured by the EoE-3D • Changes from baseline in dysphagia associated pain, discomfort and overall severity as captured by the EoE-3D at Week 24 • Changes from baseline in abdominal pain and nausea as captured by the daily diary at Week 24 • Changes from baseline in PEESS at Week 24
To evaluate the effect of benralizumab 30 mg Q4W on patient reported QOL measures	• Changes from baseline in EoE-QoL-A at Week 24 • SF-36 v2 Health Survey at Week 24

To evaluate the effect of benralizumab 30 mg Q4W on healthcare resource utilization due to EoE	• Percent of patients with relevant concomitant procedures and healthcare resource utilization during the study through Week 24
To evaluate the effect of benralizumab 30 mg Q4W on patient reported measures of disease severity and health status	• PGI-S at Week 24 • PGI-C at Week 24
To assess the PK and immunogenicity of benralizumab 30 mg Q4W in patients with EoE	• Serum benralizumab concentration • ADA and nAb
Other objectives	**Endpoints/Variables:**
To describe the longer-term effect of benralizumab 30 mg Q4W in patients with EoE	• Proportion of patients with a histologic response at Week 52, defined as a peak esophageal intraepithelial eosinophil count ≤ 6 eos/hpf • Changes from baseline in DSQ score at Week 52 • Changes from baseline in centrally-read EoE EREFS at Week 52 • Centrally-read biopsies for histopathology and tissue eosinophil counts at Week 52 • Dysphagia-free days as captured by the DSQ • Frequency of dysphagia episodes as captured by the EoE-3D • Changes from baseline in dysphagia associated pain, and discomfort as captured by the EoE-3D at Week 52 • Changes from baseline in abdominal pain and nausea as captured by the daily diary at Week 52 • Changes from baseline in PEESS at Week 52 • Changes from baseline in EoE-QoL-A at Week 52 • SF-36 v2 Health Survey at Week 52 • Percent of patients with relevant concomitant procedures and healthcare resource utilization during the study through Week 52 • PGI-S at Week 52 • PGI-C at Week 52

Safety Objective:	Endpoints/Variables:
To assess the safety and tolerability of benralizumab 30 mg Q4W in patients with EoE	Safety and tolerability will be evaluated in terms of AEs, Vital signs, and Clinical laboratory values Assessments related to AEs cover • Occurrence/frequency • Relationship to IP as assessed by Investigator • Intensity • Seriousness • Death • AEs leading to discontinuation of IP Vital signs parameters include systolic and diastolic blood pressure, and pulse, as well as respiration rate, body temperature, body weight, and height Assessments related to vital signs cover • Observed value • Absolute and percent change from baseline values over time
Exploratory Objectives:	**Endpoints/Variables:**
[redacted]	[redacted]
To evaluate the effect of benralizumab 30 mg Q4W on clinical features of EoE and disease activity	• Changes from baseline in EndoFLIP (esophageal distensibility) assessment at Week 24 and at Week 52 (sub-study)
To evaluate the effect of benralizumab 30 mg Q4W on early histologic signs, clinical features, symptoms and [redacted] of EoE (sub-study)	• Replicate histology, EREFS and PRO endpoints at weeks 4 and 12 [redacted]
[redacted]	[redacted]
To evaluate the effect of benralizumab 30 mg Q4W on the ability to work, attend classes, and perform regular daily activities	• WPAI+CIQ scores

ADA anti-drug antibody; AEs adverse events; DB Double-blind; DSQ Dysphagia Symptom Questionnaire; EndoFLIP Endolumenal Functional Lumen Imaging Probe; EoE eosinophilic esophagitis; EoE-QoL-A Adult Eosinophilic Esophagitis Quality of Life Questionnaire; eos Eosinophils; EoE HSS Eosinophilic Esophagitis-Histology Severity Score; EREFS Endoscopic Reference Score; hpf high power field; nAb neutralizing antibody; OL Open-label; OLE Open-label Extension; PEESS Pediatric Eosinophilic Esophagitis Symptom Severity Module, Version 2.0, Children and Teens Report; PGI-C Patient Global Impression of Change; PGI-S Patient Global Impression of Severity; PK pharmacokinetics; PRO Patient Reported Outcome; Q4W every 4 weeks; QoL Quality of Life; RNA ribonucleic acid; SF-36v2 Short Form-36 Version 2.0.

1.1.2 Study Objectives for the OLE Period

The following objectives/endpoints are for the OLE period of the study:	
Safety Objective:	**Endpoints/Variables:**
To assess the safety and tolerability of benralizumab 30 mg Q4W in patients with EoE	Safety and tolerability will be evaluated in terms of AEs, Vital signs, Clinical laboratory values Assessments related to AEs cover • Occurrence/frequency • Relationship to IP as assessed by Investigator • Intensity • Seriousness • Death • AEs leading to discontinuation of IP Vital signs parameters include systolic and diastolic blood pressure, and pulse, as well as respiration rate, body temperature, body weight, and height Assessments related to vital signs cover • Observed value • Absolute and percent change from baseline values over time
Other Objectives:	**Endpoints/Variables:**
To describe the longer-term effect of benralizumab 30 mg Q4W in patients with EoE	• Changes from baseline in DSQ; frequency of dysphagia episodes (EoE-3D), dysphagia-free days (DSQ), and associated pain, discomfort, overall severity; abdominal pain and nausea (daily diary); EoE-QoL-A; PEESS • Percent of patients with relevant concomitant procedures and healthcare resource utilization during the study • Centrally-read biopsies for histopathology and tissue eosinophil counts at Week 104 • Changes from baseline in centrally-read EoE EREFS at Week 104 • PGI-S at Week 104
To describe the effect of benralizumab on the use of background EoE medications and related therapies and diet restrictions	• Changes in concomitant medications and diet regimens • Changes in patient experience as reported by PROs

To describe the PK and immunogenicity of benralizumab 30 mg Q4W in patients with EoE	• Serum benralizumab concentration • ADA and nAb
Exploratory Objectives:	**Endpoints/Variables:**
[redacted]	[redacted]
To evaluate the effect of benralizumab 30 mg Q4W on clinical features of EoE and disease activity	• Changes from baseline in EndoFLIP (esophageal distensibility) assessment at Week 104 (sub-study)
[redacted]	[redacted]

ADA anti-drug antibody; AEs adverse events; DSQ Dysphagia Symptom Questionnaire; ECG electrocardiogram; EoE eosinophilic esophagitis; EoE-QoL-A Adult Eosinophilic Esophagitis Quality of Life Questionnaire; EREFS Endoscopic Reference Score; nAb neutralizing antibody; OLE Open-label Extension; PEESS Pediatric Eosinophilic Esophagitis Symptom Severity Module, Version 2.0, Children and Teens Report; PK pharmacokinetics; Q4W every 4 weeks; PRO Patient Reported Outcome

1.2 Study design

This is a randomized, placebo-controlled, double-blind, parallel-group, multicentre, Phase 3 study to compare the efficacy and safety of repeat dosing of benralizumab versus placebo in male and female patients 12 to 65 years of age with symptomatic and histologically active EoE.

The clinical study consists of 4 periods:

- a 2- to 8-week run-in period,
- a 24-week placebo-controlled, double-blind, parallel group treatment period (DB period),
- a 28-week open-label benralizumab treatment period (OL period),
- an additional 52-week open-label extension treatment period (OLE)

All patients will enter a run-in period of 2 to 8 weeks during which inclusion/exclusion criteria are assessed, medical history taken, endoscopy with biopsies performed, and patient reported outcomes (PROs), clinical laboratories, and diet questionnaires administered. Approximately 170 eligible patients (85 per arm) with symptomatic and histologically active EoE before randomization will be randomized in a 1:1 ratio to receive either 30 mg of benralizumab or placebo at 4-week intervals for a 24-week treatment period (DB period). The randomization for adults will be stratified by region (North America vs Rest of World [ROW]) and baseline steroid use. Approximately 20 adolescent patients (12 to 17 years of age) per treatment arm are targeted to be included. Adolescents will be randomized in a separate stratum with no other factors included.

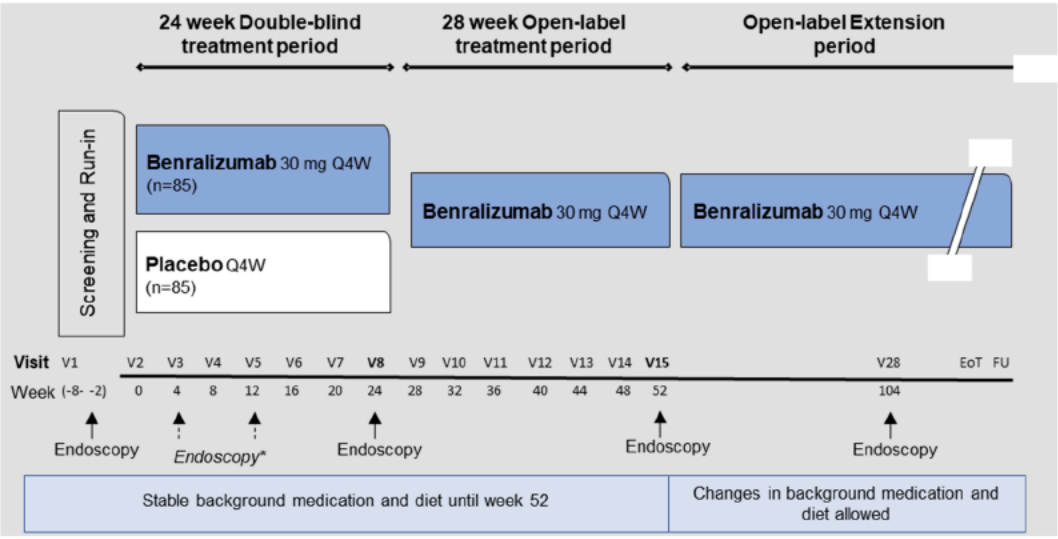
Patients who complete the double-blind (DB), placebo-controlled treatment period on investigational product (IP) and placebo will continue into an open-label (OL) treatment period with benralizumab 30 mg Q4W until Week 52 (OL period).

All patients who complete the 52-week treatment period (the 24-week DB treatment period and the 28-week OL treatment period; DB+OL treatment periods) on IP will be invited to continue into an OLE period on benralizumab 30 mg Q4W (OLE). The OLE period is intended to allow each patient at least 1 year of treatment with OL benralizumab after completion of the 52-week DB+OL treatment periods. The Sponsor may choose to extend the study beyond one year and reserves the right of terminating the OLE early (eg, if development in this indication is terminated or marketing authorisation is obtained). Patients who do not enroll in the OLE period will have a follow-up visit 12 weeks after their last dose of IP.

Please refer to the clinical study protocol (CSP) for more detail.

For an overview of the study design see Figure 1.

Figure 1 Study design flow chart



Q4W every 4 weeks; V Visit; EoT End of Treatment; FU Follow up, *Only for patients in the the Early Time Point Sub-study

1.3 Number of subjects

Approximately 170 patients will be randomised in a 1:1 ratio to benralizumab or matching placebo. This will provide >95% power for the first primary endpoint of proportion of patients achieving histological response to demonstrate an increase from 20% or less on placebo, to 50% on benralizumab at the 2-sided 5% significance level. This is a conservative estimate of the likely placebo histological response rate as lower placebo rates have been reported in previous EoE studies (Hirano et al 2017, Dellon et al 2017). The power calculation for the second primary endpoint of the change from baseline in DSQ score at week 24 is based on detecting similar

effect sizes (mean difference in change from baseline of the PRO / standard deviation) as seen in previous studies (Hirano et al 2017, Dellon et al 2017). Assuming an effect size of 0.6, which equates to a 7.2-point difference in change in the DSQ, 85 patients per arm will allow >95% power for statistical significance at the 5% 2-sided level. The high level of power for the primary endpoints will ensure stronger statistical evidence can be demonstrated in this single phase 3 study.

Success in the adolescent population will be dependent on demonstrating broadly consistent results with the overall population. With 20 adolescents per treatment arm there is a high chance of demonstrating consistent effects, if they truly exist. For the histological response rate at Week 24 endpoint, there is a 99% chance of observing an adolescent treatment difference that is at least half of the overall population effect, assuming the true histological response rates are 65% on benralizumab and 10% on placebo. For the change from baseline in DSQ endpoint at Week 24 there is an 86% chance of observing an adolescent treatment difference that is at least half of the overall population effect, assuming the true treatment effect for the DSQ endpoint is as outlined in the sample size justification above.

The amount of missing data for the histological response rate endpoint is expected to be low, based on rates previously reported; 96% in Hirano et al 2017 had peak eosinophil counts available at Week 12; 94% in Dellon et al 2017 had evaluable post-treatment DSQ and biopsy data at Week 12. There is some uncertainty in the amount of missing data for the DSQ endpoint at Week 24 given the limited data available to date on this tool and differences in length of follow-up between this study and the referenced trials. However, measures are in place in the protocol to limit missing data by excluding non-compliance during the run-in period and monitor overall compliance with the PRO on an ongoing basis. In addition, patients who discontinue randomised therapy are accounted for in the analyses using a composite estimand strategy. Given this estimand strategy which imputes outcomes for the most likely potential source of missing data, it is considered un-likely that there will be much remaining missing data in the analyses, but with missing data rates as high as 25% at Week 24, the study still maintains >90% power for statistical significance under the assumptions highlighted above.

2 ANALYSIS SETS

2.1 Definition of analysis sets

Six patient populations are defined below: All Patients Analysis Set, Full Analysis Set (FAS), Safety Analysis Set, Pharmacokinetic Analysis Set, Open-Label benralizumab Analysis Set and Open-Label extension benralizumab Analysis Set. Patients must have provided their informed consent. If no signed informed consent is collected (important protocol deviation), then the patient will be excluded from all analysis sets defined below.

2.1.1 All subjects analysis set

This analysis set comprises all subjects screened for the study, and will be used for the reporting of disposition and screening failures.

2.1.2 Full analysis set

All randomized subjects who received at least one dose of IP, irrespective of their protocol adherence and continued participation in the study. Subjects will be analysed according to their randomized treatments irrespective of whether or not they have been prematurely discontinued, according to the Intent-to-Treat (ITT) principle. Subjects who withdraw consent, or assent when applicable, to participate in the study will be included up to the date of their study termination.

All efficacy analyses will be performed using an ITT approach based on the full analysis set (FAS). For consistency, demographic and baseline characteristics will be presented using the FAS.

2.1.3 Safety analysis set

All subjects who have received at least 1 dose of IP.

Erroneously treated patients during the DB period (e.g., those randomized to treatment A but actually given treatment B) are accounted for in the treatment group of the treatment they actually received. A subject who has on one or several occasions received active IP is classified as active. Safety summaries and ADA data will be based on this analysis set.

2.1.4 Pharmacokinetic analysis set

All subjects who received benralizumab and from whom PK blood samples are assumed not to be affected by factors such as protocol violations (e.g. received wrong dose) and who had at least 1 quantifiable serum PK observation post first dose. All PK summaries will be based on this analysis set.

2.1.5 Open-label benralizumab analysis set

All subjects who start or carry on receiving at least 1 dose of benralizumab after the end of the Week 24 double blind treatment period.

2.1.6 Open-label extension benralizumab analysis set

All subjects who carry on receiving at least 1 dose of benralizumab after the end of the Week 52 double blind + open label treatment periods.

2.1.7 Early time point sub-study analysis set

All subjects who have Week 4 or Week 12 endoscopy performed.

2.2 Violations and deviations

Patients who do not meet eligibility criteria but are still randomized will be analyzed according to the analysis sets described in Section 2.1. There is no intention to perform a per-protocol (PP) analysis in this study.

2.2.1 Important protocol deviations

The final list of protocol deviations will be finalized and documented before database lock. Only important protocol deviations will be listed and tabulated in the Clinical Study Report (CSR). The important protocol deviations are those that may significantly affect a subject's rights, safety, or wellbeing, as well as those that may affect the true treatment effects with respect to the primary efficacy endpoints, proportion of patients with a histologic response at Week 24, and changes from baseline in DSQ score at Week 24. The important protocol deviations will include but may not be limited to:

- Eligibility criteria not met (patients incorrectly randomized) which are likely to affect the primary endpoints
  - Deviations from key inclusion criteria
  - Deviations from key exclusion criteria
- Deviations from informed consent procedures
- Discontinuation criteria for IP met but patient not withdrawn from study treatment
- Deviations from IP management and administration
- Received prohibited/restricted concomitant medication
- Other important protocol deviations
  - Unblinding of treatment assignment for reasons unrelated to patient safety
  - Other severe non-compliance (such deviations will be clearly described in the CSR)

Only important protocol deviations will be summarized and listed in the CSR. Potential important protocol deviations, either programmable or observable, will be reviewed quarterly at a minimum and at the time of blinded data reviews. Additional details, including key inclusion and exclusion criteria, are provided in the Protocol Deviations Plan.

3 PRIMARY AND SECONDARY VARIABLES

3.1 General definitions

3.1.1 Visit window definitions

The adjusted analysis-defined windows for assessments conducted every 4 weeks are summarized in Table 1.

Table 1 Windows for assessments conducted every 4 weeks

Adjusted defined window visit	Scheduled study day	Maximum windows
Baseline	1	-28≤Study Day ≤ 1
Week 4	29	2 ≤ Study Day ≤ 42
Week 8	57	43 ≤ Study Day ≤ 70
Week 12	85	71 ≤ Study Day ≤ 98
Week 16	113	99 ≤ Study Day ≤ 126
Week 20	141	127 ≤ Study Day ≤ 154
Week 24	169	155 ≤ Study Day ≤ 182 [1]
Week 28	197	183 ≤ Study Day ≤ 210
Week 32	225	211 ≤ Study Day ≤ 238
Week 36	253	239 ≤ Study Day ≤ 266
Week 40	281	267 ≤ Study Day ≤ 294
Week 44	309	295 ≤ Study Day ≤ 322
Week 48	337	323 ≤ Study Day ≤ 350
Week 52	365	351 ≤ Study Day ≤ 378
Week 56	393	379 ≤ Study Day ≤ 406
Week 60	421	407 ≤ Study Day ≤ 434
Week 64	449	435 ≤ Study Day ≤ 462
Week 68	477	463 ≤ Study Day ≤ 490
Week 72	505	491 ≤ Study Day ≤ 518
Week 76	533	519 ≤ Study Day ≤ 546
Week 80	561	547 ≤ Study Day ≤ 574
Week 84	589	575 ≤ Study Day ≤ 602
Week 88	617	603 ≤ Study Day ≤ 630
Week 92	645	631 ≤ Study Day ≤ 658
Week 96	673	659 ≤ Study Day ≤ 686
Week 100	701	687 ≤ Study Day ≤ 714
Week 104	729	715 ≤ Study Day ≤ 742

[1] The windowing will only be performed for assessments within the appropriate periods e.g. double blind versus open label, where the definition of the Double Blind period is all assessments from the date of randomization up to and including the first dose of Open Label benralizumab 30 mg.

For assignment of data to adjusted analysis-defined visit windows, study day will be defined as follows:

Date of assessment – Date of randomization + 1

Study days before randomization will be defined as follows:

Date of assessment – Date of randomization

By this definition, the day of randomization will be study day 1 and the day before the day of randomization will be study day – 1. There is no study day 0. The planned date of Visit 3 (Week 4) will be study day 29 (= 28 + 1), for example.

If multiple assessments are recorded within a single adjusted visit window, please refer to the rules below:

- If there are 2 or more observations within the same visit window, then the non-missing observation closest to the scheduled visit will be used in the analysis.
- If 2 observations are equidistant from the scheduled visit, then the non-missing observation with the earlier collection date will be used in the analysis.
- If 2 observations are collected on the same day, then the non-missing observation with the earlier collection time will be included in the analysis.

If a visit window does not contain any observations, then the data will remain missing.

The daily window for daily assessments starts at 17:00:00 and ends at 4:59:59 the following day. The diary is not available to fill in until 17:00 each day. All responses received after 16:59:59 on day *n* and before 5:00:00 on Day *n+1* will be attributed to Day *n*.

The DSQ score, EoE-3D and other daily symptoms are calculated using data captured over 14-day periods. The scheduled bi-weekly (14-day) study windows are listed in Table 2.

Table 2 Bi-weekly windows for daily diary assessments

Adjusted defined window visit	Scheduled study day	Maximum windows
Baseline	1	The last 14 days from Study Day -14 to Study Day -1
Week 2	15	Study Day 1 to Study Day 14
Week 4	29	Study Day 15 to Study Day 28
Week 6	43	Study Day 29 to Study Day 42
Week 8	57	Study Day 43 to Study Day 56
Week 10	71	Study Day 57 to Study Day 70
Week 12	85	Study Day 71 to Study Day 84
Week 14	99	Study Day 85 to Study Day 98
Week 16	113	Study Day 99 to Study Day 112
Week 18	127	Study Day 113 to Study Day 126

Week 20	141	Study Day 127 to Study Day 140
Week 22	155	Study Day 141 to Study Day 154
Week 24	169	Study Day 155 to Study Day 168 [1]
Week 26	183	Study Day 169 to Study Day 182
Week 28	197	Study Day 183 to Study Day 196
Week 30	211	Study Day 197 to Study Day 210
Week 32	225	Study Day 211 to Study Day 224
Week 34	239	Study Day 225 to Study Day 238
Week 36	253	Study Day 239 to Study Day 252
Week 38	267	Study Day 253 to Study Day 266
Week 40	281	Study Day 267 to Study Day 280
Week 42	295	Study Day 281 to Study Day 294
Week 44	309	Study Day 295 to Study Day 308
Week 46	323	Study Day 309 to Study Day 322
Week 48	337	Study Day 323 to Study Day 336
Week 50	351	Study Day 337 to Study Day 350
Week 52	365	Study Day 351 to Study Day 364
Week 54	379	Study Day 365 to Study Day 378
Week 56	393	Study Day 379 to Study Day 392
Week 58	407	Study Day 393 to Study Day 406
Week 60	421	Study Day 407 to Study Day 420
Week 62	435	Study Day 421 to Study Day 434
Week 64	449	Study Day 435 to Study Day 448
Week 66	463	Study Day 449 to Study Day 462
Week 68	477	Study Day 463 to Study Day 476
Week 70	491	Study Day 477 to Study Day 490
Week 72	505	Study Day 491 to Study Day 504
Week 74	519	Study Day 505 to Study Day 518
Week 76	533	Study Day 519 to Study Day 532

[1] The windowing will only be performed for assessments within the appropriate periods e.g. double blind versus open label, where the definition of the Double Blind period is all assessments from the date of randomization up to the first dose of Open Label benralizumab 30 mg.

For other efficacy by visit summaries that are not based on 4 weeks or 2 weeks windows, as well as all the safety by visit summaries, please use below general windowing rule:

Table 3 General windows for non-4-week assessments

Analysis-defined window visit	Scheduled study day	Maximum windows
Baseline	1	Study Day≤1
Week X*	X*7+1=a	2≤Study Days≤((b-a)/2+a)-1
Week Y*	Y*7+1=b	((b-a)/2+a)≤Study Days≤((c-b)/2+b)-1
Week Z (Follow-up)	Z*7+1=c	((c-b)/2+b)≤Study Days

Study Day is defined as Date of assessment – Date of randomization + 1 for efficacy summaries, Date of assessment – Date of first dose+ 1 for safety summaries.
If (b-a) or (c-b) is odd number, then use (b-a+1) or (c-b+1), respectively.
* For the Week 24 visit, the visit window should end at the first dose date of OL benralizumab, the following visit window should start from the first dose date of OL + 1.

For overall analyses not based on any particular study visit, all data will be listed and/or analysed, including any repeated or unscheduled visits, unless otherwise specified. For safety endpoints, all post-baseline results will be included in the overall analysis up to and including the follow-up visits. For efficacy endpoints, the post-baseline treatment period will be included up to and including the end of treatment (EOT) visit.

3.1.2 Baseline and week 24 definition

In general, the last recorded value on or prior to the date of randomization will serve as the baseline measurement for efficacy endpoints while the last recorded value prior to first dose of study treatment will serve as the baseline measurement for safety endpoints. If there is no value prior to the randomization (or the first dose of study treatment, depending on the endpoint), the baseline value will not be imputed and will be set to missing. No data known to be collected post first dose will be used in determining the baseline value, unless otherwise specified.

Additional analyses for the patients who switch from placebo to benralizumab at Week 24 may be performed where the baseline value is set to the last recorded value prior to starting benralizumab at Week 24 (i.e. likely the Week 24 measurement) to obtain an assessment of the changes occurring while actually receiving benralizumab.

For the daily assessment variables including DSQ, EoE-3D and other daily symptoms scores which are calculated during a certain period (e.g., bi-weekly or weekly), the score calculated during the cycle prior to randomization will be the baseline score. The score calculated during the cycle prior to the scheduled Week 24 date (i.e., Day 169) will be used to calculate the Week 24 score (i.e., Days 155 to 168 for bi-weekly scores). However, only days prior to the first day of open label benralizumab will be used for the Week 24 score derivation. If the Week 24 visit and start of open label dosing occurs ahead of schedule (i.e., Day 169) only Day 155 up to the day before the week 24 visit will be used in the week 24 score derivation. The Baseline and Week 24

scores for EoE-3D and other daily symptom items will be defined in the same way as the DSQ score (i.e. 14-days prior to baseline).

For the non-daily dairy endpoints, the Week 24 value will follow the visit windows defined in Table 1 and Table 3 in Section 3.1.1, any record that's collected after the first open label dose date will not be considered as Week 24 value. If no Week 24 record is available or the Week 24 record is collected after the first open label dose date, Week 24 value will be set as missing.

3.1.3 Prior/concomitant medications

A medication will be regarded as prior if it started prior to the date of randomization and was stopped on or before the date of randomization (medication stop date ≤ date of randomization).

A medication will be regarded as concomitant if the start date is on or after the date of randomization, or if it started prior to the date of randomization and was ongoing after the date of randomization. Medications with start date after the on-treatment period will not be considered as concomitant.

If a medication started and stopped on the date of randomization, it will be considered as concomitant.

3.2 Primary outcome variables

The proportion of patients with a histologic response at Week 24 of treatment will be used as a dual-primary efficacy variable.

The analysis of histological response rate at Week 24 will include data collected at the Week 24 visits. Patients with their biopsy data at Week 24 collected after the first dose of OL, or have had more than planned DB dosing before Week 24, or patients with no biopsy data at Week 24 will be considered non-responders.

A histologic response is defined as a peak esophageal intraepithelial eosinophil count ≤ 6 eos/hpf across all available esophageal levels. In the statistical analysis, a binary variable taking on value 1 if a patient has experienced a histologic response during the 24-week double-blind treatment period and 0 otherwise, will be used as the response variable for the primary efficacy analysis. Treatment discontinuation, increases in background medications, or additional new therapies for EoE, or dilation procedures for EoE indicates treatment failure and will be treated as non-responder.

The change from baseline in Dysphagia Symptom Questionnaire (DSQ) score at Week 24 is the second dual-primary efficacy variable and the change from baseline at Week 52 will be used as a supportive variable to the primary outcome. Benralizumab subjects with treatment discontinuation, increases in background medications, or additional new therapies for EoE, or dilation procedure will have their change from baseline value imputed 100 times at each timepoint from the time of the intercurrent events occurring onwards using return-to-baseline

MI (Multiple Imputation) (Seed=779385). Placebo subjects with treatment discontinuation, increases in background medications, or additional new therapies for EoE, or dilation procedure before Week 24 will have their change from baseline value imputed 100 times at each timepoint up to Week 24 from the time of the intercurrent events occurring onwards using return-to-baseine I. Beyond week 24 when these patients switch to benralizumab the intercurrent events re-set and no imputation will be made. For Placebo switched to Benralizumab subjects with treatment discontinuation, increases in background medications, additional new therapies for EoE, or dilation procedure after Week 24 will have their change from baseline value imputed 100 times at each timepoint from the time of the intercurrent events occurring onwards using return-to-baseline MI.

The samples will be drawn from a normal distribution with mean 0 and variance of pooled data. Let X = (Xobs, Xmiss) be the complete data at the timepoint of interest. X is consisted of observed measurements Xobs and the missing observations Xmiss. In return-to-baseline imputation, when X is change from baseline, each missing observation Xmiss is imputed by a random draw from a normal distribution with mean 0 and variance v_{imp}:

$Xmiss \sim N(0, v_{imp})$,

The variance v_{imp} is calculated from the observed changes:

$v_{imp} = (1 + 1/Nc)\, v_c$,

where v_c is the variance of the change among completers across all treatment arms, and Nc is the number of completers.

The missing data at Week 24 which was not due to intercurrent events will be imputed by MI with missing at random (MAR) assumption. The following 4 steps will be used to build the imputation datasets and perform analyses:

1. 100 datasets obtained from return-to-baseline MI will be induced by Markov Chain Monte Carlo (MCMC) method to get the monotone missing pattern (Seed=113165).

2. For each of the imputed datasets obtained in step 1, the remaining missing data will be imputed using the regression method for the monotone pattern with adjustment covariates including treatment groups, region, baseline value of the response variable, baseline steroid use, and presence of strictures at baseline (Seed=352988).

3. Each of the 100 imputed datasets will be analysed using the main statistical model. These 100 datasets will be saved.

4. Apply Rubin's rule (Rubin et al 1986, Rubin 1987) to combine analysis results (point estimates and standard errors) from 100 imputations. Descriptive statistics including

number of patients, mean, standard error, and least squares (LS) means will be provided for each timepoint. In addition, difference in LS means and the corresponding 95% confidence intervals (CI) will be provided along with the p-values for Week 56 and all earlier time points in turn.

3.3 Secondary efficacy outcome variables

3.3.1 Tissue eosinophil counts

The percent change from baseline for the tissue eosinophil counts will be analysed as a key secondary efficacy endpoint.

The analysis of tissue eosinophil counts at Week 24 will include data collected at the Week 24 visits. Patients who experienced intercurrent events before Week 24 will be imputed with return-to-baseline MI. Patients who did not experience intercurrent events before Week 24 and have their biopsy data at Week 24 collected after the first dose of OL, or have had more than planned DB dosing before Week 24, or patients with no biopsy data at Week 24 will be imputed by MI with the missing at random (MAR) assumption. See Section 3.2 for details about imputation process. Supportive analysis at Week 52 will be handled similarly.

3.3.2 EoE histology scoring system (EoE-HSS)

The EoE histology scoring system (EoE-HSS) will be used to derive the change from baseline in EoE-HSS grade and stage scores at Week 24 and Week 52 which will be used as key secondary and supportive efficacy variables respectively. The same estimand rules as mentioned for the primary change from baseline in DSQ score endpoint will be used.

The EoE histology scoring system (EoE-HSS) provides a standardized method to evaluate esophageal biopsies for features in addition to peak eosinophil count. The EoE-HSS scores will be recorded independently in the proximal, mid and distal oesophagus in 8 features:

- Eosinophilic inflammation (EI)
- Basal zone hyperplasia (BZH)
- Eosinophil abscess (EA)
- Eosinophil surface layering (SL)
- Dilated intracellular spaces (DIS)
- Surface epithelial alteration (SEA)
- Dyskeratotic epithelial cells (DEC)
- Lamina propria fibers (LPF)

Each feature will be scored separately for grade (severity) and stage (extent) of abnormality using a 4-point scale (0 = normal; 3 = most severe or extensive).

The maximum total score possible is 24 (maximum grade or stage score of 3×8 features=24), representing the most severe grade or a stage for each esophageal biopsy collected if all 8 features were evaluated.

However, not all eight features are present in all biopsies in the study. For example, LPF is only evaluable in half or fewer biopsies. In order to normalize for missing data, a ratio will be created. The grade score ratios ('grade score') per region is calculated like this: if k ($k \leq 8$) features are evaluated, the sum of the k features will be divided by the maximum possible grade score, 3*k. For example, if only 7 features are evaluated for a patient at certain visit, the score will be the sum of the 7 features divided by 21. The highest possible grade score ratio per region is 1.

The total grade score is the mean of the grade score ratios per region.

Total feature grade score is the raw/observed grade score per feature in each region added together and divided by the number of regions. The max denominator is 3 for each feature; the denominator decreases if there are missing region scores. The highest possible total feature grade score is 3.

Total stage score and total feature stage score will be calculated with the same rules as total grade score and total feature grade score.

The total grade and total stage scores across all regions will be used for the key secondary endpoints. In addition, the region grade/stage scores for all regions will be explored in the analysis to assess how universal were the improvements in the esophagus following therapy.

3.3.3 Centrally-read and Investigator-read EoE EREFS

The change from baseline in worst centrally-read EoE EREFS scores at Week 24 and Week 52 will be used as key secondary and supportive efficacy variables. The same estimand rules as mentioned for the primary change from baseline in DSQ score endpoint will be used. The EREFS will be centrally-read from video recordings and investigator-read during the endoscopies. Centralized imaging data assessments and scoring from expert physician review will be performed for all endoscopies.

The EoE EREFS is a scoring system for assessing the presence and severity of the major endoscopic signs of EoE, including esophageal edema, rings, exudates, furrows, and stricture. The scoring system will be applied separately in the proximal and distal endoscopy findings, and the worst for each individual component from the proximal and distal scores will be used and summed to form the total EREFS score used in the primary analysis of EREFS. The maximum total score is 9. Central reviewers may select not evaluable (NE) if the endoscopic abnormality cannot be graded due to image quality issues. An NE assessment of any category will result in an NE for the sum EREFS (inflammatory, fibrostenotic, total and overall). Secondary analyses will also explore the overall score which is the sum of proximal and distal location (with maximum score of 18). Inflammatory score, fibrostenotic score, total score will

also be summarized separately by proximal and distal location. Table 4 shows the EoE EREFS modified grading.

Table 4 EoE EREFS Modified Grading

Component	Score		Description
Exudates	0	Absent	
	1	Mild	Lesions covering <10% of the esophageal mucosa
	2	Severe	Lesions involving ≥10% of the esophageal mucosa
Rings	0	Absent	
	1	Mild	Subtle circumferential ridges
	2	Moderate	Distinct rings that do not impair passage of a standard adult upper endoscope (outer diameter 8-9.5 mm)
	3	Severe	Distinct rings that do not allow a standard adult upper endoscope to pass
Edema	0	Absent	Distinct vascularity present
	1	present	Loss of vascular marking
Furrows	0	Absent	
	1	Mild	Vertical lines present
	2	Severe	Vertical lines with mucosal depth (indentation)
Strictures	0	Absent	
	1	Present	The inner diameter of the stricture will be estimated by the endoscopist
Sums			
Inflammatory score	0-5		Sum of exudates, edema and furrows
Fibrostenotic score	0-4		Sum of rings and strictures

3.3.4 Proportion of patients with a treatment response at Week 24

The proportion of patients with a treatment response at Week 24 will be used as a key secondary efficacy variable.

Treatment response is defined as composite of histologic response (same histological response criteria defined in Section 3.2) and clinically meaningful improvement (≥30%) from baseline in DSQ score. The same estimand rules as mentioned for the primary histologic response rate endpoint will be used.

3.3.5 Proportion of patients with a histologic response at Week 52 of treatment

The proportion of patients with a histologic response at Week 52 of treatment will be used as a supportive efficacy variable to the primary outcome.

The histologic response is defined in the same way as outlined in Section 3.2. In the statistical analysis, a binary variable taking on value 1 if a patient has experienced a histologic response at week 52 and 0 otherwise, will be used as the response variable for the supportive efficacy analysis.

3.3.6 Healthcare resource utilization

Healthcare resource utilization data, associated with medical encounters related to EoE, will be collected in the eCRF by the Investigator or designee for all patients throughout the study.

The percent of subjects with each healthcare resource utilization during the study will be summarized by visit.

3.4 Patient reported outcome (PRO) variables

Patient reported outcomes (PROs) will be measured using the following questionnaires: Dysphagia Symptom Questionnaire (DSQ), Eosinophilic Esophagitis Daily Dysphagia Diary (EoE-3D) (and other symptom questions), Adult Eosinophilic Esophagitis Quality of Life questionnaire (EoE-QoL-A), SF-36 v2 Health Survey, the Work Productivity and Activity Impairment questionnaire plus Classroom Impairment Questions (WPAI+CIQ), Patient Global Impression of Severity (PGI-S) and Patient Global Impression of Change (PGI-C). These data will be collected when patients enter their own response choices directly into the ePRO handheld device. Daily diary metrics will be recorded each day from Visit 1 to Visit 21. There will be no further collection of PRO information using the ePRO handheld device after Week 76 (Visit 21) when the patients will return the device to the clinic.

The Diet Questionnaire will be administered by the investigator as interviews at specified study visits. Data will be entered by the investigator.

3.4.1 Dysphagia Symptom Questionnaire (DSQ) score

Change from baseline in Dysphagia Symptom Questionnaire (DSQ) score at Week 24 is the second dual-primary efficacy variable. Secondary and other variables supported by the DSQ tool include change from baseline to Week 52, responder analysis, characterization of dysphagia-free days, and pain while swallowing.

The DSQ is a daily PRO developed to capture dysphagia symptoms in EoE patients ≥12 years of age (Dellon et al 2013). Daily DSQ questions, response values (when applicable) are shown in Table 5. If no solid food has been consumed (Q1=No) the patient will be asked to complete two additional, non-scored, items intended to confirm solid food avoidance (Q1a) and characterize the reason for solid food avoidance (Q1b). With confirmation of no solid food intake Questions 2 and 3 will be skipped and the daily DSQ set to missing for the day.

Confirmation and characterization of solid food avoidance items are not included in the DSQ analysis but will be summarized in descriptive tables.

Table 5 Dysphagia Symptom Questionnaire (DSQ) score

Instrument	Item	Content	Response Options and scores	Usage/Endpoint Supported
The Daily Dysphagia Symptom Question-naire (DSQ)	1	Since you woke up this morning, did you eat solid food?	• Yes *{Go to 2}* • No *{Go to 1a}*	• Capture days with solid food avoidance • Characterize proportion of days with avoidance of solid food • **Note**: Q1=No; Items 2 and 3 skipped; Day treated as missing for DSQ scoring
	1a	You have indicated that you did not eat solid food today. Please confirm.	• I did not eat solid food since waking up this morning *{Go to 1b}* • I have eaten solid food *{Go to 1}*	• Operational: Confirm "No" response to Q1 (no solid food) or correct Q1 "No" response as needed
	1b	Select the most important reason for not eating solid food.	• Concerned that solid food would go down slowly or become stuck in my throat or chest *{Response 1}* • Swallowing solid food is too painful *{Response 2}* • Symptoms not related to my eosinophilic esophagitis *{Response 3}* • Reasons other than my symptoms *{Response 4}*	• Supplemental question to characterize the reason for solid food avoidance (Q1=no) and address the possible impact of missing daily DSQ scores specifically due to EOE-symptom related food avoidance.
	2	Since you woke up this morning, has food gone down slowly or been stuck in your throat or chest?	• Yes *{2} {Go to 3}* • No *{0} {Exit}*	• Part of daily DSQ score to capture occurrence of dysphagia • Required for daily DSQ score (Q2+Q3)
	3	For the most difficult time you had swallowing food today (during the past 24 hours), did you have to do	• No, it got better or cleared up on its own *{0}* • Yes, I had to drink liquid to get relief *{1}* • Yes, I had to cough and/or	• Part of daily DSQ score to capture dysphagia severity • Required for daily DSQ score (Q2+Q3)

Instrument	Item	Content	Response Options and scores	Usage/Endpoint Supported
		anything to make the food go down or to get relief?	gag to get relief *{2}* • Yes, I had to vomit to get relief *{3}* • Yes, I had to seek medical attention to get relief *{4}*	
	4	The following question concerns the amount of pain you have experienced when swallowing food. What was the worst pain you had while swallowing food today?	• None, I had no pain *{0}* • Mild *{1}* • Moderate *{2}* • Severe *{3}* • Very Severe *{4}*	• Capture pain associated with swallowing

The DSQ score is comprised of daily values captured over a 14-day period and has a range of 0-84 with higher scores indicating more frequent and/or severe dysphagia. The DSQ Score is calculated as follows:

DSQ score = (Sum of points from questions 2 + 3 in the daily DSQ) × 14 days / (Number of diaries reported with non-missing data).

Daily DSQ scores are calculated as the sum of Q2 and Q3 response values. The daily DSQ score range is 0-6. Higher values indicate worse dysphagia. Daily DSQ score will be set to missing if not recorded or if Q1="No". At least 8 days with an evaluable daily score in a 14-day period are required to calculate a DSQ score; otherwise the DSQ score for the period is set to missing (Hirano et al 2017, Dellon et al 2017). A sensitivity analysis will be performed where at least 4 out of 7 non-missing daily response are required in each of the 2 weeks to compute a 14-day DSQ score. Details on regarding missing data and other sensitivity analyses are included in the Appendix.

The baseline DSQ score will be calculated using data captured in the 14 days prior to randomization. Post-baseline DSQ scores will be calculated every 14 days per the scoring instructions.

A DSQ responder is defined as an improvement from baseline exceeding the minimal clinically important differences (MCID). An MCID of –6.5 points was suggested in Hudgens 2017, however the appropriate threshold for response will be explored using the data generated in this study using anchor-based methods that use the entire distribution of data to establish a clinically meaningful within-patient change threshold range. See section 4.2.4.3 for

details. Anchor-based analyses supplemented with empirical cumulative distribution functions using partially accumulated blinded MESSINA data pooled across treatment arms, using the methods outlined in section 4.2.4.3 were performed and demonstrated that a threshold of approximately 18 points improvement from baseline in DSQ score at week 24 may demonstrate a clinically meaningful improvement to patients, aligning with 1 category improvement in the PGI-S anchor. Therefore, a supportive DSQ responder analysis using the -18 points threshold will be performed at the primary analysis.For the treatment responder key secondary endpoint, the clinically meaningful symptom improvement threshold will use a 30% improvement as a percentage change may be more meaningful for patients and physicians than a number of points improvement. A 30% improvement from baseline in DSQ score is used as the clinically meaningful improvement in symptoms component of the treatment responder endpoint as a threshold considered meaningful to physicians, and aligns with the MCID presented in the currently available literature (Hudgens et al 2017). Limitations with the available literature include differences in trial population and design to the MESSINA study, but the 30% threshold serves as a starting point to assess the potential for clinical relevance. A supportive analysis of the treatment responder endpoint will be performed using 50% improvement from baseline in DSQ score as the clinically meaningful improvement in symptoms component, which aligns to the 18 point improvement demonstrated in blinded anchor-based analyses described above where an 18 point improvement results in an approximately 50% reduction in score for a patient with the average DSQ score of 35 points at baseline.

The monthly number of dysphagia-free days will be reported as the proportion of observed dysphagia-free days (daily DSQ=0) expressed as the number of days in a 28-day period. For the primary assessment of this measure, for any monthly period a minimum of 8 days out of 14 days is required for both 14-day periods in a 28-day month. The number of dysphagia-free days in that period will be scaled up over the 28 days using the proportion of the available days (e.g. 8 dysphagia-free days out of 16 days with data available in a period is 50% dysphagia-free days which will be reported as 14 days dysphagia-free for that month). If the required number of days with data are not available for a monthly period, the monthly number will be missing for that period. If there are less than 8 days (more than 6 days of missing data) out of a 14-day period, this 14-day will be set to missing. In a 28-day month, if one of the 14-day period is missing (the other one is either missing or not), the whole 28-day month will be set to missing. This matches the missing data approach for DSQ score. See section 4.2.4.3 for details.

A sensitivity analysis of this measure will be performed using all the number of dysphagia-free days without imputation for the missing data (e.g. 8 dysphagia-free days in a period would be reported as 8 days regardless how many days had data available).

3.4.2 EoE-3D and other symptom questions

The changes from baseline in Eosinophilic Esophagitis Daily Dysphagia Diary symptom metrics (EoE-3D) at Week 24 and Week 52 will be used as secondary and supportive efficacy variables.

The EoE-3D consists of 6 items and focuses on components of an episode of dysphagia. In the EoE-3D, respondents are asked to first report on whether they experienced episodes of difficulty swallowing in the past 24 hours and if so, how many. Patients are asked to report the time at which the episode occurred to facilitate recall and subsequently respond to 3 questions on the pain, discomfort, and overall severity of the event using an 11-point numeric rating scale (0 [no pain] to 10 [worst]). These items are repeated for each episode reported by the respondent. Therefore, the EoE-3D allows for an assessment of the frequency count of episodes over time as well as a characterization of the severity of the episode experience over time.

Patients are also asked several additional questions to characterize their daily experience. These items will be scored separately from EoE-3D. Patients will be asked to report, via two separate questions, the severity of abdominal pain and the severity of nausea at their worst over the past 24 hours using similar 11-point numeric rating scales. Table 6 shows the EoE-3D questionnaire and score for each response option, including additional questions about abdominal pain and nausea.

Table 6 EoE-3D and other symptom questions and scores

Instrument	Item	Content	Response Options	Usage/Endpoint Supported
{Additional daily diary content}				
Abdominal pain (stomach ache) and nausea assessments	A	Please tell us about your abdominal pain (stomach ache) and nausea in the past 24 hours.	• Next	• Operational: Transition from DSQ; Instruction on recall period for abdominal pain and nausea items
	1	In the past 24 hours… Rate your worst abdominal pain (stomach ache).	• 0 (No abdominal pain) • 10 (Abdominal pain as bad as I can imagine) *{NRPS from 0 to 10}*	• Scored as a single item measure of abdominal pain • Other secondary endpoint: Change from baseline in abdominal pain severity
	2	In the past 24 hours… Rate your worst nausea.	• 0 (No nausea) • 10 (Nausea as bad as I can imagine) *{NRPS from 0 to 10}*	• Scored as a single-item measure of nausea • Other secondary endpoint: Change from baseline in nausea severity

Instrument	Item	Content	Response Options	Usage/Endpoint Supported
		{Daily EoE-3D}		
The Eosinophilic Esophagitis Daily Dysphagia Diary (EoE-3D)	1	In the past 24 hours… Did you have any episodes of difficulty swallowing?	• Yes • No	• Capture if the patient experienced any episodes of difficulty swallowing over the past 24 hours
	2	In the past 24 hours… How many episodes of difficulty swallowing did you have?	*{Selection of whole number}* episode(s)	• Capture the number of dysphagia episodes over the past 24 hours • Other secondary endpoint: Change from baseline in mean number of daily episodes of dysphagia
		{Items 3 - 6 will repeat for every episode reported in Item 2. The value for {X} will be based on response to Item 2}		
	3	Please answer this question for episode {X} What time did you have this episode of difficulty swallowing?	*{Selection of hour and minute, am/pm}* Note to translators: Please advise on hour/minute, time format as necessary	• Operational: Patient asked to enter the time of each episode to establish each episode as independent and to reinforce the recall period.
	4	Please answer this question for episode {X} Rate the worst **discomfort** you experienced during this episode.	• 0 (No discomfort) • 10 (Discomfort as bad as I can imagine) *{NRPS from 0 to 10}*	• Capture the level of discomfort associated with each episode of dysphagia • Other secondary endpoint: Change from baseline in discomfort associated with dysphagia
	5	Please answer this question for episode {X} Rate the worst **pain** you experienced during this episode.	• 0 (No pain) • 10 (Pain as bad as I can imagine) *{NRPS from 0 to 10}*	• Capture the severity of pain associated with each episode of dysphagia • Other secondary endpoint: Change from baseline in dysphagia-related pain severity
	6	Please answer this question for episode {X} Rate the **overall**	• 0 (Not severe at all) • 10 (As severe as I can imagine) *{NRPS from 0 to 10}*	• Capture the overall severity of each episode of dysphagia • Other secondary endpoint: Change from baseline in

Instrument	Item	Content	Response Options	Usage/Endpoint Supported
		severity of this episode.		dysphagia episode severity

Dysphagia episode frequency will be summarized as the total number of dysphagia episodes occurring over each 28-day period following randomization. Calculation of the 28-day dysphagia episode frequency will require a minimum of 8 days out of 14 days for both 14-day periods in a 28-day month. If sufficient days are available to calculate a frequency, the frequency of dysphagia episodes will be scaled up over the 28 days using the days with available data in a similar manner to the dysphagia-free days endpoint. If the required number of days with data are not available for a monthly period, the monthly number will be missing for that period. The dysphagia-related pain, discomfort and overall episode severity will be recorded for each episode and summarized individually as average episode severity within a 14-day period. Each episode severity score will be calculated as the sum of daily average NRPS values in the 14-day period divided by the number of days with available episodes of difficulty swallowing during the same 14-day period. Calculation of the 14-day mean scores will require at least 8 days of evaluable data during the period; otherwise the mean scores will be set to missing. Note days with 0 episode of difficulty swallowing still count as evaluable even there is no severity collected. In case all 14 days with 0 episode of difficulty swallowing, the score would be set as missing.

Frequency adjusted severity scores may also be derived for each item individually based on daily severity scores defined as the sum of each episode score divided by the number of daily episodes, and taking the score of zero for days in which the participant reports having no episodes. Frequency adjusted severity scores are then calculated as the sum of the available daily scores in the 14-day period divided by the number of non-missing daily scores in the 14 day period.

Abdominal pain severity and nausea severity will be summarized individually as 14-day means scores. Each 14-day mean score will be calculated as the sum of daily NRPS values divided by the number of non-missing days during the same period. Calculation of the 14-day means will require at least 8/14 days of evaluable data; otherwise the mean score will be set to missing.

3.4.3 EoE-QoL-A score

The change from baseline in EoE-QoL-A domain scores at Week 24 and Week 52 will be used as secondary and supportive efficacy variables.

The Adult Eosinophilic Esophagitis Quality of Life (EoE-QoL-A) questionnaire is a 30-item assessment developed specifically to capture health-related quality of life (HRQoL) in patients with EoE (Taft et al 2011). The assessment is divided into 5 domains: eating/diet impact, social impact, emotional impact, disease anxiety and swallowing anxiety. Response is captured on a 5-point scale: Not at all = 4, Slightly = 3, Moderately = 2, Quite a Bit = 1, Extremely = 0. The total score is calculated as the sum of all responses (total score ranges 0-120; Lower scores indicate a greater degree of impairment). Domain scores are calculated as follows:

- Eating/Diet Impact: sum of Q2, Q9, Q16, Q24, Q25, Q26, Q27, Q28, Q29, Q30
- Social Impact: sum of Q14, Q17, Q19, Q22
- Emotional Impact: sum of Q1, Q5, Q6, Q7, Q11, Q13, Q21, Q23
- Disease Anxiety: sum of Q4, Q10, Q12, Q15, Q18,
- Swallowing Anxiety: sum of Q3, Q8, Q20

3.4.4 Pediatric Eosinophilic Esophagitis Symptom Severity Module, Version 2, Children and Teens Report (PEESS) – sub-study

The change from baseline in PEESS scores at Week 24 and Week 52 will be used as exploratory and supportive efficacy variables. PEESS will only be completed by patients who are age 18 or under at the time of Visit 1.

PEESS is an 18-item assessment of EoE symptom severity and frequency validated for use in patients age 8 to 18. The overall score ranges from 0 to 80, with higher scores representing more severe and frequent EoE symptoms. The first 18 questions alternate between a question about a given symptom's frequency (odd numbered questions) and a question about the symptom's severity (even numbered questions). Q19 asks about frequency of eating less food than others and Q20 asks about the frequency of needing more time to eat than others.

The questions in the assessment are scored on one of two scales:

- Q1, Q3, Q5, Q7, Q9, Q11, Q13, Q15, Q17, Q19, Q20 are scored on a Likert-type scale:
    - Never = 0
    - Almost never = 1
    - Sometimes = 2

- Often = 3
- Almost always = 4

- Q2, Q4, Q6, Q8, Q10, Q12, Q14, Q16, Q18 are scored on a face rating scale with drawings of facial expressions accompanying the following written scale:

  - Not bad at all = 0
  - A little bad = 1
  - Kind of bad = 2
  - Bad = 3
  - Very bad = 4

3.4.5 SF-36 v2 Health Survey score

The SF-36 v2 Health Survey score at Week 24 and Week 52 will be used as secondary and supportive efficacy variables.

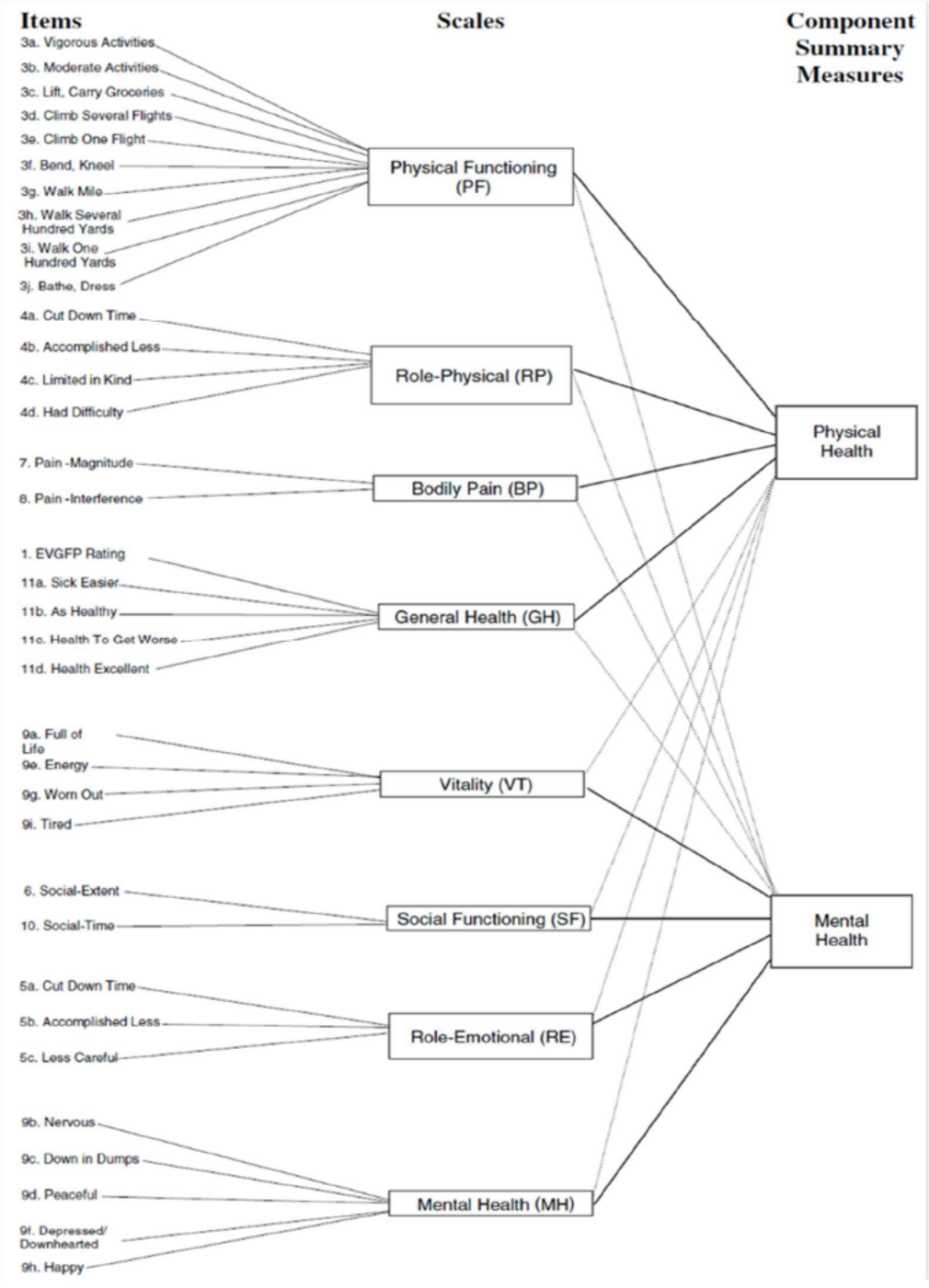
The Short Form 36-item Health survey, Version 2 (acute recall) (SF-36 v2) is a 36-item, self-report survey of functional health and well-being, with a 1-week recall period. Responses to 35 of the 36 items are used to compute an 8-domain profile of functional health and well-being scores. The remaining item, referred to as the 'Health Transition' item, asks patients to rate how their current state of health compared to their state of health 1 year ago and is not used to calculate domain scores.

The 8-domain profile consists of the following subscales: Physical Functioning (PF), Role Limitations due to Physical Health (RP), Bodily Pain (BP), General Health Perceptions (GH), Vitality (VT), Social Functioning (SF), Role Limitations due to Emotional Problems (RE), and Mental Health (MH). Psychometrically-based physical and mental health component summary scores (PCS and MCS, respectively) are computed from subscale scores to give a broader metric of physical and mental HRQoL.

Figure 2 shows the 35 questions used to compute the 8-domain profile of functional health and well-being scores. Question 2 is the remaining item referred to as the 'Health Transition' item not used to calculate domain scores and does not appear in Figure 2.

The threshold values for the SF-36v2 PCS, MCS, and domain scores listed in Table 7 are suitable for interpreting change at the patient level and are referred to as the responder thresholds or responder definitions (QualityMetric 2011). A patient will be classified as a responder if the change from baseline ≥ threshold, or a non-responder if change from baseline < threshold. If data are missing, then the patient will be classified as a non-responder.

Figure 2 The 35 questions for the 8-domain scores and physical & mental health component summary scores (PCS and MCS)



Note. All health domain scales contribute to the scoring of both the Physical and Mental Component Summary measures. Scales contributing most to the scoring of the summary measures are indicated by a connecting solid line (—). Scales contributing to the scoring of the summary measures to a lesser degree are indicated by a dotted line (·····).

Table 7 Responder threshold values for the SF-36 v2 domain and component summary measures

	SF-36v2 score									
Threshold	**PCS**	**MCS**	**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**
Individual change	3.4	4.6	4.3	3.4	6.2	7.2	6.2	6.9	4.5	6.2

BP Bodily Pain; GH General Health Perceptions; MCS mental health component summary; MH Mental Health; PCS physical component summary; PF Physical Functioning; RE Emotional Problems; RP Role Limitations due to Physical Health; SF Social Functioning; VT Vitality.

3.4.6 Work Productivity and Activity Impairment questionnaire plus Classroom Impairment Questions (WPAI+CIQ)

The Work Productivity and Activity Impairment questionnaire plus Classroom Impairment Questions (WPAI+CIQ) at Week 24 and Week 52 will be used as supportive efficacy variables.

The WPAI+CIQ consists of questions about how health and health-related issues impact the ability to work, attend classes, and perform regular daily activities. The questionnaire relates to the previous 7 days. The WPAI+CIQ will be used to measure self-reported productivity loss. The questionnaire will be completed by the patients using the ePRO device at Week 0, 12, 24, 36 and 52.

There are a maximum of 10 questions and a minimum of 3 questions that will be completed by subjects.

1 = currently employed (yes/no)

2 = hours missed work due to health problems

3 = hours missed work due to other reasons

4 = hours actually worked

5 = degree health affected productivity while working (0-10 scale, with 0 meaning no effect)

6 = attends class in an academic setting (yes/no)

7 = hours missed class due to health problems

8 = hours actually attended class

9 = degree health affected productivity while attending class (0-10 scale, with 0 meaning no effect)

10 = degree health affected regular activities (other than work or class) (0-10 scale, with 0 meaning no effect)

If the answer to question 1 is 'No, not currently employed', then the subject should skip to question 6. If the answer to question 6 is 'No, not currently attending class', then the subject should skip to question 10.

The WPAI+CIQ provides 4 types of scores: absenteeism (work or class time missed), presenteeism (impairment at work or class/reduced on-the-job effectiveness), work productivity loss (overall work or class impairment/absenteeism plus presenteeism), and activity impairment. WPAI+CIQ outcomes are expressed as impairment percentages, with higher numbers indicating greater impairment and less productivity.

For the work related questions, the following calculations should be used to create the outcomes of interest:

- Number of work hours missed = Q2
- Absenteeism = Q2/(Q2+Q4)
- Presenteeism = Q5/10
- Work Productivity Loss = Q2/(Q2+Q4) + [(1-Q2/(Q2+Q4))*(Q5/10)]

For the class related questions, the following calculations should be used to create the outcomes of interest:

- Number of class hours missed = Q7
- Absenteeism = Q7/(Q7+Q8)
- Presenteeism = Q9/10
- Class Productivity Loss = Q7/(Q7+Q8) + [(1-Q7/(Q7+Q8))*(Q9/10)]

Additionally, Activity Impairment = Q10/10.

3.4.7 Patient Global Impression of Severity (PGI-S)

The Patient Global Impression of Severity (PGI-S) score at Week 24 and Week 52 will be a secondary and supportive efficacy variable. Patient Global Impression of Severity (PGI-S) is a single item designed to evaluate the patient's perception of overall symptom severity at the time of completion using a 6-point categorical response scale:

Assessment	Score
No symptoms	1
Very mild	2
Mild	3
Moderate	4
Severe	5
Very severe	6

3.4.8 Patient Global Impression of Change (PGI-C)

The Patient Global Impression of Change (PGI-C) score at Week 24 and Week 52 will be a secondary and supportive efficacy variable.

PGI-C instruments are used to evaluate the patient's overall perception of change (change from baseline) in EoE. PGI-C assessments use a single question with 7-point rating scale:

The Patient Global Impression of Change (PGI-C) is a single item assessment to evaluate the patient's perception of change in health status. The patient is asked to report the degree to which they have changed since entering the treatment period using a 7-point scale:

Assessment	Score
Much worse	7
Moderately worse	6
A little worse	5
About the same / No Change	4
A little better	3
Moderately better	2
Much better	1

3.5 Diet Questionnaire

The diet questionnaire is an investigator (or designee) lead interview intended to capture patient-reported diet and eating behaviours related to EoE. The diet questionnaire will be used by investigators to monitor patients during the study and to characterize the patient experience via summary statistics. The patient questionnaire will consist of questions related to initiation or discontinuation of elimination diets and the defining parameters of these diets. Further, the diet questionnaire will ask patients to characterize EoE-related eating behaviours or self-initiated symptom management techniques. Diet questionnaire data will be summarized to characterize patient diet changes and timing of these changes in relation to the study.

The questionnaire is divided into two parts – one to be asked at screening and Week 0 and one to be asked at subsequent visits.

Diet Questionnaire **At Screening and Week 0**			
Item	**Content**	**Response options**	**Usage/Endpoint Supported** (Do Not Translate)
1	Do you avoid any food because it previously has caused an acute allergy reaction (e.g. rash, hives, throat itching/swelling/closing, anaphylaxis)?	Yes / No If Yes, to what food did you have a reaction a. Milk/dairy b. Egg c. Soy d. Wheat e. Gluten f. Fish g. Shellfish h. Peanuts i. Tree nuts j. Other	
2	Do you have any experience of making dietary changes to help address your EoE symptoms?	Yes / No (If yes - answer all following questions) (If no – answer Q7 and Q8)	
3	Are you on a modified diet <u>now</u> to help address your EoE symptoms?	Yes / No	
4	If yes, what kind of dietary changes have you made? Please select all that apply.	a. Single food elimination diet b. Multiple food elimination diet c. Elemental (remove all protein, consume an amino acid formula) d. Allergy test-based modification e. Self-directed elimination (food trial) f. Reintroduction of previously eliminated foods g. Other	
5	If yes on Q2, what foods do you eliminate? Please select all that apply.	a. Milk/dairy b. Egg c. Soy d. Wheat e. Gluten f. Fish g. Shellfish h. Peanuts i. Tree nuts j. Other than these 9 foods, if yes how many?	
6	Have you changed the diet described above <u>in the past 6 weeks</u>? Please select all that apply.	a. No, I have not changed my diet b. Yes, I reintroduced Milk/dairy c. Yes, I reintroduced Egg d. Yes, I reintroduced Soy e. Yes, I reintroduced Wheat f. Yes, I reintroduced Gluten g. Yes, I reintroduced Fish	

		h. Yes, I reintroduced Shellfish i. Yes, I reintroduced Peanuts j. Yes, I reintroduced Tree nuts k. Yes, I eliminated 1 or more foods l. Other	
7	Are you using any behavioral approach to manage your EoE? Please select all that apply.	a. Chewing food very thoroughly to ensure food goes down b. Drinking plenty of water to ensure food goes down c. Repeated swallows to facilitate food goes down d. Chew food into a mush e. Blend food into a smoothie-like consistency f. Lubricating food (dip in to sauce/oil) g. Eating slowly h. Cutting food into very small pieces i. Avoidance of troublesome foods j. Crushing or avoiding pills k. Other	
8	Have environmental factors affected your EoE symptoms? Please select all that apply.	a. Seasonal changes b. Climate or weather conditions c. Other environmental factors (animal dander, dust mites, pollutants in air, etc.)	

Diet Questionnaire **From Visit 2 and onwards**			
Item	**Content**	**Response options**	**Usage/Endpoint Supported** (Do Not Translate)
1	Compared to baseline, have you reintroduced any food during the past month that you previously had an acute allergic reaction to? (e.g. rash, hives, throat itching/swelling/closing, anaphylaxis)?	Yes / No If Yes, what have you reintroduced? _____	
2	Are you on a modified diet now to help address your EoE symptoms?	Yes / No (if no - go to Q4)	
3	Compared to baseline, have you made any changes to your diet to help address your EoE symptoms during the past month? Please select all that apply.	a. No, I have not changed my diet b. Yes, I reintroduced Milk/dairy c. Yes, I reintroduced Egg d. Yes, I reintroduced Soy e. Yes, I reintroduced Wheat f. Yes, I reintroduced Gluten g. Yes, I reintroduced Fish h. Yes, I reintroduced Shellfish i. Yes, I reintroduced Peanuts j. Yes, I reintroduced Tree nuts k. Yes, I eliminated 1 or more foods l. Other	

4	Compared to baseline, have you changed any behavioral approach to help address your EoE during the past month?	Yes/No If yes, please select all that apply; a. Chewing food very thoroughly to ensure food goes down ____initiated ____increased ____continued same ____decreased ____stopped b. Drinking plenty of water to ensure food goes down ____initiated ____increased ____continued same ____decreased ____stopped c. Repeated swallows to facilitate food bolus passage ____initiated ____increased ____continued same ____decreased ____stopped d. Chew food into a mush ____initiated ____increased ____continued same ____decreased ____stopped e. Blend food into a smoothie-like consistency ____initiated ____increased ____continued same ____decreased ____stopped	

		f. Lubricating food (dip in to sauce/oil) ____initiated ____increased ____continued same ____decreased ____stopped g. Eating slowly ____initiated ____increased ____continued same ____decreased ____stopped h. Cutting food into very small pieces ____initiated ____increased ____continued same ____decreased ____stopped i. Avoidance of troublesome foods ____initiated ____increased ____continued same ____decreased ____stopped j. Crushing or avoiding pills ____initiated ____increased ____continued same ____decreased ____stopped	

3.6 Safety outcome variables

Safety and tolerability will be evaluated in terms of: reported AEs (including SAEs), vital signs, and clinical laboratory assessments related to AEs.

All safety measurements will use all available data for analyses, including data from unscheduled visits and repeated measurements.

Change from baseline to each post-treatment time point where scheduled assessments were made will be calculated for relevant measurements. AEs will be summarised by means of descriptive statistics and qualitative summaries.

No safety data will be imputed. The handling of partial/missing dates for AEs and prior/concomitant medications is detailed in Appendix 8.3. Duration of AEs and prior/concomitant medications will not be calculated using imputed dates and will instead be set to missing.

3.6.1 Adverse Events

Adverse events experienced by the patients will be collected throughout the entire study and will be coded using the latest version of the Medical Dictionary for Regulatory Activities (MedDRA) per the Data Management Plan.

The following events are considered treatment emergent:

- Adverse events with an onset date on or after first dose of IP
- Worsening of pre-existing events on or after first dose of IP.

Adverse event data will be categorized according to their onset date into the following study periods:

- AEs in the on-treatment period are defined as those with onset day between day of first dose of study treatment and scheduled end of treatment (EOT) visit for patients who complete study treatment or investigational product discontinuation visit (IPD) for patients who prematurely discontinue study treatment, inclusive. In the event that the EOT or IPD visit is beyond the protocol-defined visit window, AEs with onset after the last dose of study treatment date +28 days +3 days (visit window) will be excluded from the on-treatment period and instead assigned to the post-treatment period.

- AEs in the on-study period are defined as those with onset between day of first dose of study treatment and the day of the scheduled follow-up visit, inclusive.

- AEs in the post-treatment period are defined as those with onset after the on-treatment period defined above.

- On-study AEs in the DB period will be defined as those with onset date between day of first dose of study DB treatment and the day prior to the first dose of OL period (up to and including the day of the scheduled follow-up visit for patients who do not roll over to OL).

- On-study AEs in the OL period are defined as those with onset date on or after the day of the first dose of OL treatment and the day prior to the first dose of OLE period (up to and including the day of the scheduled follow-up visit for patients who do not roll over to OLE).

- On-study AEs in the OLE period are defined as those with onset date on or after the day of the first dose of OLE treatment and up to EOT.

For instances where a patient attends the safety follow-up visit only, but does not attend an earlier IPD visit or EOT visit, adverse events occurring on or after the day of first dose of study treatment and on or before the last dose of study medication + 31 days will be assigned to the on-treatment period, while AEs with onset date after this time will be assigned to the post-treatment period.

If an AE has a missing onset date it will be considered an on-treatment event unless the stop date of the AE indicates otherwise. Similarly, if an AE has a partial onset date it will be considered an on-treatment AE unless the partial onset date or the stop date indicates otherwise.

Adverse events that have missing causality (after data querying) will be assumed to be related to study drug.

3.6.2 Clinical laboratory variables

Blood and urine samples for determination of clinical chemistry, haematology and urinalysis parameters will be taken at the times detailed in the CSP, and will be assessed in a central laboratory. The parameters outlined in Section 8.2.1, Table 11 of the CSP will be collected.

In summaries, listings and figures, lab results and normal ranges will be presented in the International System (SI) unit. Eosinophils data will be presented in both SI and conventional units (eos/HPF) in summaries.

Changes in haematology and clinical chemistry variables between baseline and each postbaseline assessment will be calculated. Baseline is defined as the last available value measured prior to the first dose of randomized treatment. The change from baseline is defined as the post-baseline visit value minus the baseline visit value. There will be no imputation for missing values. For values recorded with a leading greater than or less than ('>', '<') symbol, the reported numeric value will be used for analysis and the value with the symbol will be included in the listings, unless otherwise specified. For example, a value of <0.01 will be analyzed as 0.01 and listed as <0.01.

Absolute values will be compared to the relevant reference range and classified as low (below range), normal (within range or on limits) or high (above range). The central laboratory reference ranges will be used for laboratory variables. All absolute values falling outside the reference ranges will be flagged.

Urinalysis data will be categorized as negative (0), positive (+), or strongly positive (++, +++, or > +++) at each timepoint.

For the purposes of haematology, clinical chemistry and urinalysis shift tables, baseline will be defined as the last available non-missing value prior to first dose of randomized treatment,

and maximum or minimum value post-baseline will be calculated over the entire post-baseline period, including the post-treatment period.

3.6.3 Vital signs and weight

Pre-dose vital signs and weight (pulse, systolic blood pressure, diastolic blood pressure, respiration rate, and body temperature) will be obtained in accordance with the schedule provided in the protocol. Weight will be recorded in kilograms (kg).

Changes in vital signs and weight variables between baseline and each subsequent scheduled assessment will be calculated. Baseline is defined as the last value prior to the first dose of randomized treatment. The change from baseline is defined as the post-baseline visit value minus the baseline visit value. There will be no imputation for missing values.

Body mass index (BMI) will be calculated from the height and weight as follows:

$$\text{BMI } (kg/m^2) = \text{Weight } (kg) / (\text{Height } (m))^2$$

3.6.4 ECGs

The outcome of the overall evaluation is to be recorded as normal/abnormal in the eCRF, with any abnormalities being recorded as not clinically significant or clinically significant.

3.6.5 Physical examination

Complete and brief physical examinations will be performed at time points specified in Table 1 and Table 2 of the CSP. What is included in the assessment will be dependent on whether the examination is complete or brief, as described in Section 8.2.3 of the CSP. For the brief physical examination only information on whether the assessment was performed or not is to be recorded. Any new finding(s) or aggravated existing finding(s), judged as clinically significant by the investigator or designee, will be reported as an AE.

3.7 Exploratory outcome variables

Details of exploratory outcome variables and their analysis methods will be defined in a separate exploratory analysis plan (EAP), and will be reported in a separate report to the CSR.

3.8 Pharmacokinetic variables

Blood samples (processed to serum) for pharmacokinetic assessments will be collected from all subjects at baseline prior to first benralizumab administration at Week 0 Day 1, at Weeks 8, 16, 24, 36, 52, 76 and 104 before benralizumab administrations during the treatment period, and at the 12 weeks after last IP dose for follow-up visit, or 4 weeks after last IP dose if IPD/EOT. Serum concentrations of benralizumab will be determined using a validated

electrochemiluminescent (ECL) immunoassay. Results below the lower limit of quantification (BLQ) will be set to LLOQ/2 for analysis and will be listed as <LLOQ.

3.9 Immunogenicity variables

Anti-drug antibodies (ADA) variables, such as ADA and neutralizing antibodies (nAb) responses, will be generated and analysed as per the details in Section 8.2 (Appendix).

3.10 EndoFLIP (esophageal distensibility) – sub-study

Details of endoFLIP outcome variables and their analysis methods will be defined in a separate exploratory analysis plan (EAP), and will be reported in a separate report to the CSR.

4 ANALYSIS METHODS

All subjects involved with the analysis of the study will remain blinded until primary database lock and Clinical Study Protocol deviations identified.

4.1 General principles

The primary efficacy analyses will be based on the double-blind placebo-controlled first 24 weeks of the study (DB period). In this part of the study all efficacy analyses will use the full analysis set (FAS) as defined in Section 2.1.2, and patients will be analysed according to their randomized treatment, following the Intention-to-Treat (ITT) principle.

A composite estimand strategy will be used for the primary analyses of endpoints at week 24 whereby any patient with intercurrent events of either randomized therapy discontinuation, an increase of background medications, addition of a new therapy for EoE, or having a dilation procedure will be considered as treatment failures at week 24. A review of all concomitant medications and procedures during the study will be performed prior to database lock to identify events to be considered as treatment failure in these analyses, whereby only medications / procedures considered likely to have a meaningful impact on EoE outcomes would be considered intercurrent events. For the histologic response rate endpoint, patients with these intercurrent events prior to week 24 will be considered non-responders at week 24; for the change in DSQ at week 24 endpoint, and for other change from baseline continuous endpoints, any patients experiencing the described intercurrent events will have their change from baseline value at Week 24 imputed with return-to-baseline MI (see Section 3.2 for details).

All patients who prematurely discontinue from IP or have any changes to background therapies for EoE as described above are asked to come in for all visits and study assessments up to week 52. Therefore, sensitivity analyses can be performed to assess the robustness of the

efficacy results to these estimand approaches and missing data assumptions as described in individual endpoint analysis methods and Section 8.1.

The statistical analyses of the DB period, is designed to compare both efficacy and safety of benralizumab to placebo while the OL period is designed to evaluate the long-term safety and tolerability and persistence of effect of benralizumab in this patient population. The OLE will provide an opportunity to assess long term safety and tolerability. Details regarding primary and key secondary estimands are provided in Table 8, with additional details including sensitivity analyses provided in Section 8.1.

All analyses of Week 52 endpoints will be descriptive as no placebo control is available at that timepoint and so no hypothesis testing will be performed. Week 52 analyses will primarily be presented on the FAS, but a repeat of key analyses may also be produced on the open-label benralizumab analysis set to ensure only patients who switched to receive benralizumab after 24 weeks are included in the denominator for that group and to ensure a meaningful interpretation of the placebo-to-benralizumab patients.

Demography and baseline characteristics will be summarized by treatment group for the FAS. In the event that there are major differences between the FAS and safety analysis set, these summaries may also be repeated for the safety analysis set.

Summary data will be presented in tabular format by treatment group. Categorical data will be summarized by the number and percentage of subjects in each category. Continuous variables will be summarized by descriptive statistics including N, mean, SD, median, and range. Data listings will be sorted by treatment group and patient number.

All hypothesis testing will be reported using 2-sided tests. Any *p*-values presented for endpoints other than those included in the hierarchical testing strategy (or those in the testing strategy but after a failed endpoint) will be nominal (i.e., not multiplicity adjusted). All *p*-values will be rounded to 4 decimal places.

The data analyses will be conducted using the SAS® System version 9.4 or above (SAS Institute Inc., Cary, NC). All SAS® programs used to generate analytical results will be developed and validated according to AstraZeneca SAS® programming standards and validation procedures. Pharmacokinetic analyses will be performed using NONMEM or other appropriate software, and will be reported in a separate report to the CSR.

Table 8 Primary, key secondary efficacy and main safety estimands

Statistical Category	Estimand[1]			
	Treatment Condition[1]	**Endpoint (Population[1])**	**Intercurrent Event Strategy[1]**	**Population Level Summary[1] (Analysis)**
Primary Objective: To evaluate the effect of benralizumab 30 mg Q4W on signs and symptoms of EoE in patients with symptomatic and histologically active EoE				
Primary/MCP Section 4.2.4.2	Treatment with benralizumab 30 mg versus placebo, where treatment discontinuation or increases in background medications or dilation procedures for EOE indicate treatment failure.	• Proportion of patients with a histologic response at Week 24 (FAS)	• Treatment discontinuation - composite (treated as non-responder) • Missing data – composite (treated as non-responder) • Increases in background therapy or addition of new therapies for EoE, or dilation procedures – composite (treated as non-responder)	• CMH test. Week 24 is the primary timepoint.
Primary/MCP Section 4.2.4.3		• CFB in DSQ score at Week 24 (FAS)	• Treatment discontinuation – composite (Return-to-baseline MI) • Increases in background therapy or addition of new therapies for EoE, or dilation procedures – composite (Return-to-baseline MI) • Missing data not due to intercurrent events – composite (MI MAR)	• Mean difference between interventions (LSMD from CFB ANCOVA). Week 24 is the primary timepoint.
Key Secondary Objective: To evaluate the effect of benralizumab 30 mg Q4W on clinical features of EoE and disease activity				
Secondary/MCP Section 4.2.5.1-4	Treatment with benralizumab 30 mg versus placebo, where treatment discontinuation or increases in background medications or dilation procedures for EOE indicate treatment failure.	• Percent CFB in tissue eosinophils at Week 24 (FAS) • CFB in EoE HSS grade score at Week 24 (FAS) • CFB in EoE HSS stage score at Week 24 (FAS) • CFB in centrally-read EoE EREFS at Week 24 (FAS)	• Treatment discontinuation – composite (Return-to-baseline MI) • Increases in background therapy or addition of new therapies for EoE, or dilation procedures – composite estimand (Return-to-baseline MI) • Missing data not due to intercurrent events – composite (MI MAR)	• Mean difference between interventions (LSMD from CFB ANCOVA). Week 24 is the primary timepoint.

Table 8 Primary, key secondary efficacy and main safety estimands

Statistical Category	Estimand[1]			
	Treatment Condition[1]	**Endpoint (Population[1])**	**Intercurrent Event Strategy[1]**	**Population Level Summary[1] (Analysis)**
Secondary/MCP Section 4.2.5.5	Treatment with benralizumab 30 mg versus placebo, where treatment discontinuation or increases in background medications or dilation procedures for EOE indicate treatment failure.	• Proportion of patients with a treatment response at Week 24, defined as composite of histologic response (as per primary endpoint) and clinically meaningful improvement (30%) from baseline in DSQ score. (FAS)	• Treatment discontinuation - composite (treated as non-responder) • Missing data – composite (treated as non-responder) • Increases in background therapy or addition of new therapies for EoE, or dilation procedures – composite (treated as non-responder)	• CMH test. Week 24 is the primary timepoint.
Safety Objective: To assess the safety and tolerability of benralizumab 30 mg Q4W in patients with EoE				
Safety Section 4.2.9.1	Treatment with benralizumab 30 mg, versus placebo, regardless of compliance with background medications.	• Presence of AEs DB+OL+OLE (Safety) • Presence of SAEs DB+OL+OLE (Safety) • Vital Signs values DB+OL (Safety) • CFB and percent CFB of Vital Signs DB+OL (Safety)	• Remained adherent to intervention (on-treatment)	Categorical descriptive

MCP = Multiple comparisons procedure; EOT = End of treatment; LSMD = Least squares mean difference; CFB = Change from baseline; MMRM = Mixed model for repeated measures; CMH test = Cochran-Maentel-Haenszel test; ANCOVA = Analysis of covariance; AE = Adverse event; MI = Multiple imputation; SAE = Serious adverse event.

[1] All estimand attributes explicitly identified for primary and key secondary estimands only.

4.1.1 Testing strategy to account for multiplicity considerations

To account for multiplicity testing for the dual primary endpoints (histological response rate and changes from baseline in DSQ at Week 24) and the key secondary endpoints (percent change from baseline in tissue eosinophil counts at Week 24, change from baseline in EoE-HSS total grade score at Week 24, change from baseline in EoE-HSS total stage score at Week 24, change from baseline in centrally-read EREFS at Week 24, Proportion of patients with a treatment response at Week 24), a hierarchical testing strategy will be used to strongly control the overall type 1 error rate at the 0.05 level.

If the null hypothesis for the first primary endpoint of histological response rate at 24 weeks is not rejected (i.e. p-value > 0.05, or worsening compared to placebo), no null hypotheses will be rejected for any other endpoint in the study. If the null hypothesis is rejected for the first dual-primary endpoint, then hierarchical fixed-sequence testing will continue at the $\alpha = 0.05$ level moving to the second dual-primary endpoint of change from baseline in DSQ at Week 24 and subsequently to the key secondary endpoints in the order listed above. At any time that a null hypothesis cannot be rejected (i.e. p-value > 0.05, or worsening compared to placebo), further testing will stop and no subsequent null hypothesis in the testing hierarchy will be rejected.

For the purpose of US marketing approval both dual-primary endpoints (histological response and change from baseline in DSQ score) would need to be statistically significant.

4.2 Analysis methods

4.2.1 Patient disposition

Patient disposition will be summarized using the all patients analysis set. The total number of patients will be summarized for the following groups: those who enrolled, those who entered run-in, and those who were not randomized (and reason). The number and percentage of patients within each treatment group will be presented by the following categories: randomized, received treatment with study drug, did not receive treatment with study drug (and reason), completed treatment with study drug in DB treatment period, discontinued treatment with study drug in DB treatment period (and reason), discontinued treatment with study drug in DB treatment period but completed study follow-up, completed DB treatment period study, and withdrawn from study in DB treatment period (and reason).

The number and percentage of patients within each treatment group will be presented by the following categories: enrolled in OL treatment period, did not enrol in OL treatment (and reason), completed OL treatment, discontinued OL treatment (and reason), discontinued OL treatment but completed study follow-up, completed OL treatment, and withdrawn from OL treatment (and reason).

For patients who completed the OL treatment period, the number and percentage of patients within each treatment group will be presented by the following categories: enrolled in the OLE treatment period, did not enrol in the OLE treatment (and reason), completed OLE treatment, discontinued OLE treatment (and reason), completed OLE treatment, and withdrawn from OLE treatment (and reason).

Screen failure information will be listed for the all patients analysis set.

The number of patients remaining on treatment, patients discontinued IP but still in study follow-up, and patients who withdraw from the study will be summarized by treatment group and scheduled visit, separately for patients in the full analysis set.

The number of patients randomized by country and centre will also be summarized by treatment group in the FAS.

4.2.2 Demography data and patient characteristics

Demography and baseline characteristics will be summarized by treatment group and for 'total' in the FAS, using frequency and percentages (for categorical variables) and descriptive statistics of *n*, mean, standard deviation, minimum, Q1, median, Q3, and maximum (for continuous variables). If there are major differences between the FAS and safety analysis set, these summaries may also be repeated for the safety analysis set.

Age will be derived from the date of informed consent-date to birth, rounded down to the nearest integer. For patients in countries where date of birth is not recorded, the age as recorded in the electronic case report form (eCRF) will be used.

Various baseline characteristics will also be summarized, including patient characteristics (weight, height, BMI, baseline eosinophil count, historical eosinophil count, etc). Medical history will be summarized separately for past and current conditions. Specific medical and surgical histories will be summarized separately.

4.2.3 Prior/concomitant medications

The number and percentage of patients who take prior medications, those who take allowed concomitant medications and those who take disallowed concomitant medications during the study, will be presented by treatment group. Concomitant medications will be classified according to the WHO-Drug. The summary tables will present data by generic term within ATC code.

4.2.4 Primary efficacy outcome variable

4.2.4.1 Statistical hypotheses

For the first primary endpoint of proportion of patients achieving a histological response at Week 24, the null hypothesis is that the odds of responding on benralizumab 30mg Q4W is equal to the odds of responding on placebo (this can be interpreted as the proportion of responders on benralizumab 30mg Q4W is equal to the proportion of responders on placebo). The alternative hypothesis is that the odds of responding on benralizumab 30mg Q4W is not equal to the odds of responding on placebo, i.e.:

H_0: *Odds ratio* (*benralizumab 30mg / Placebo*) = *1*

H_a: *Odds ratio* (*benralizumab 30mg / Placebo*) ≠ *1*

For the second primary endpoint of change from baseline in DSQ score at Week 24, the null hypothesis is that the change in DSQ score for patients on benralizumab 30mg Q4W is equal to the change in DSQ score for patients on placebo. The alternative hypothesis is that the change in DSQ score for patients on benralizumab 30mg Q4W is not equal to the change in DSQ score for patients on placebo, i.e.:

H_0: *Difference in change from baseline in DSQ score at week 24* (*benralizumab 30mg – Placebo*) = *0*

H_a: *Difference in change from baseline in DSQ score at week 24* (*benralizumab 30mg – Placebo*) ≠ *0*

Hypothesis testing for the primary analyses will be performed at the 2-sided 5% significance level. If the *p*-value is less than 0.05, reject H_0 and accept H_a. The multiple testing procedure requires the effect to favour Benralizumab to continue testing the dual-primary endpoint.

The estimated treatment effects (the proportion of patients achieving a histological response/ treatment response, the difference in DSQ, tissue eosinophil counts, EoE-HSS, EREFS, EoE-QoL-A, SF-36 v2 PCS and MCS score changes, and episode pain, discomfort, severity as well as abdominal pain, nausea scores) from baseline of benralizumab versus placebo, corresponding 95% confidence intervals (CI), and two-sided p-values for the differences of score changes will be presented.

A multiple testing procedure will be applied to the primary endpoints and key secondary endpoints, details are provided in Section 4.1.1.

4.2.4.2 Primary analysis method – Histological response rate

The first of the dual-primary endpoints, the proportion of patients achieving a histological response at Week 24, will be compared between benralizumab and placebo using a Cochran-

Maentel-Haenszel (CMH) test stratified by region, baseline steroid use, and presence of strictures at baseline.

The results of the analyses will be presented using an odds ratio, together with its associated 95% CIs and 2-sided *p*-value. Results will be transformed into a difference in proportions for ease of interpretation. The number and percentage of histological responders will also be summarized by randomized treatment with confidence intervals around the proportions.

Patients with their biopsy data at Week 24 collected after the first dose of OL, or have had more than planned DB dosing before Week 24, or patients with no biopsy data at Week 24 will be considered non-responders. Any patients with an intercurrent event of randomised therapy discontinuation, an increase in background medications or additional new therapies for EoE, or having a dilation procedure at or before Week 24 will also be considered non-responders at week 24.

Sensitivity analyses will be performed including all post baseline biopsy data to assess the impact of any additional data collected after intercurrent events. If the amount of missing data warrants further investigation, sensitivity analyses to alternative missing data assumptions described in Section 8.1 (Appendix) may also be explored for the histological response rate endpoint.

4.2.4.3 Primary analysis method –Change from baseline in DSQ score

The second of the dual-primary endpoints, the change from baseline in DSQ score at Week 24, will be compared between the benralizumab and placebo treatment groups using a composite estimand strategy as described above. Return-to-baseline MI imputations will be made 100 times for patients with the intercurrent events, missing data not due to intercurrent events will be imputed 100 times using multiple imputation with missing at random assumption. The change from baseline in DSQ score at Week 24 will then be analysed using an analysis of covariance (ANCOVA) model for each imputation. Apply Rubin's rule (Rubin et al 1986, Rubin 1987) to combine analysis results from 100 imputations for the final analysis result (see Section 3.2 for details).

The model will include change from baseline in DSQ score at Week 24 as the dependent variable, baseline DSQ score as a continuous covariate, and region, baseline steroid use, and presence of strictures at baseline as categorical covariates. The model will be used to estimate the mean change from baseline at Week 24 for each treatment group and the difference versus placebo, with corresponding 95% confidence limits. A p-value, corresponding to a 2-sided test, will be presented to compare the benralizumab and placebo treatment groups.

The exclusion criteria exclude patients with strictures severe enough to prevent easy passage of a standard endoscopy or any critical esophageal stricture that requires dilation. It is expected that there may be some patients with milder strictures at baseline. The intent in the

analysis is to include the presence of strictures at baseline in the ANCOVA models, to ensure the imbalances in all key factors are adjusted for in the treatment effect estimated. However, in the instance of non-convergence, presence of strictures at baseline may be the first covariate to be excluded from the ANCOVA models.

The missing at random assumption (MAR) for the analysis is considered appropriate as once the composite estimand approach is applied it is not considered likely that there will be much remaining missing data and any remaining sources are appropriate to consider missing at random. Sensitivity analyses to alternative missing data assumptions may be performed as described in Section 8.1 (Appendix).

Sensitivity analysis will also be performed by including all available data regardless of intercurrent events (i.e., treatment policy strategy).

Descriptive statistics to summarize reasons for non-evaluable daily DSQ score data due to the patient reporting no solid food consumption will be provided.

In addition, the associations between the change from baseline in DSQ score at Week 24 and baseline tissue/blood eosinophil counts will be evaluated by Loess plots.

Other supportive analyses based on the DSQ tool

Descriptive summaries will be provided for the number of dysphagia-free days per patient derived from the DSQ tool. Dysphagia-free days will be summarized monthly (28-days).

The primary summaries of the number of dysphagia free days will be based on the endpoint derived with missing data rules described in section 3 where as long as sufficient days (8 in each 14 day period) are available a number of dysphagia free days will be estimated for the whole period scaling up the missing days based on the data available. A supportive summary of dysphagia free days will be performed where only the days actually dysphagia free will be included and no scaling up for missing days will be performed. Refer to section 3 for the derivation of the 2 version of this endpoint.

A DSQ responder analysis will be performed. A DSQ responder is defined as decrease in DSQ score of exceeding the minimal clinically important differences (MCID) from baseline to Week 24. The DSQ responder analysis will be conducted using a logistic regression model adjusted for treatment group, baseline DSQ score, and region. From this model, the absolute difference in response rates (benralizumab-placebo) will be estimated with the associated 2-sided 95% CI. Results will be presented in terms of adjusted response rates and difference in response rates with 95% confidence intervals and p-values. The responder analysis will be supported with a cumulative distribution function plot of change from baseline at Week 24 and descriptive summaries of the proportion of responders by treatment group and visit.

Marginal standardization methods (Bartlett 2018) will be used for the model estimates for all rate analyses, including logistic regression, unless otherwise specified.

To further explore the threshold of clinically meaningful within-patients DSQ changes noted in Hudgens 2017 where an MCID of –6.5 points was suggested, anchor-based methodology will be implemented.

Anchor-based approaches estimate a threshold by 'anchoring' the results on a separate variable, often a patient-reported outcome (PRO). In this study, the anchor-based analyses will employ PGI-C and/or change in PGI-S as anchors. Patients will be grouped by PGI-C and/or change in PGI-S scores. The PGI-C survey is selected to determine the anchor-based estimates for the MCID because of the strong positive correlation between PGI-C and DSQ scores (correlation coefficient >0.3, Hudgens 2017). Spearman correlation coefficients between PGI-S and DSQ scores will be assessed. The larger the correlation coefficient between an anchor and the endpoint, the greater the confidence in the classifications. An anchor is considered adequate if it has a Spearman correlation coefficient of 0.3 or greater (Hudgens 2017, Coon 2018). If there is also a strong correlation between changes in PGI-S and changes in DSQ scores, an anchor-based analysis categorized by changes in PGI-S will be performed.

The PGI-C and PGI-S anchors will be categorized to provide a clearer difference between patients who have and have not experienced a meaningful change according to the anchors. The ordinal responses to PGI-C and PGI-S at Week 24 will be assigned the numeric values listed in Section 3.4.8 and Section 3.4.7. One category improvement on the PGI-S or the response category of "a little better" on the PGI-C will be considered as the primary target response categories to anchor the change in DSQ score against. While there are limited successful biologic clinical trials in Eosinophilic Esophagitis to model from, and given that the MESSINA trial recruited patients who are histologically active and symptomatic at baseline despite availability of standard of care approaches, **any** improvement on these anchor scales are considered likely clinically meaningful to patients.

Hudgens 2017 focused only on the means of the anchor categories. However, in this analysis the entire distribution of the anchor categories will be used to ensure that there is adequate separation between different anchor categories. Empirical cumulative distribution functions (eCDF) and probability density function (ePDF) curves using data that are pooled across both treatment arms (but grouped by anchor categories) will be provided to establish a clinical meaningful within-patient change threshold range.

Descriptive statistics (mean, SD, median, Q1, Q3, minimum and maximum), eCDF and ePDF curves will be presented for each combination of anchor, category and endpoint. The eCDF curves display a continuous plot of the change from baseline on the horizontal axis, and the cumulative percent of patients experiencing changes from baseline up to that level on the vertical axis. The ePDF curves are useful in aiding the interpretation of eCDF curves. Compared

to eCDF curves, ePDF curves provide an easier overview of the shape, dispersion, and skewness of the distribution of the change from baseline in DSQ score across various anchor categories. Examples of eCDF and ePDF curves are shown in Figure 3 and Figure 4 (FDA 2018). Un-collapsed change categories of PGI-S and categories of PGI-C will be used for the summaries. If the sample size within a response category is too small, grouping of the response categories will be conducted in another analysis.

Figure 3 Example of Empirical Cumulative Distribution Function (eCDF) Curves of Change in DSQ Score from Baseline to Week 24 grouped by Patient Global Impression of Change (PGI-C) score.

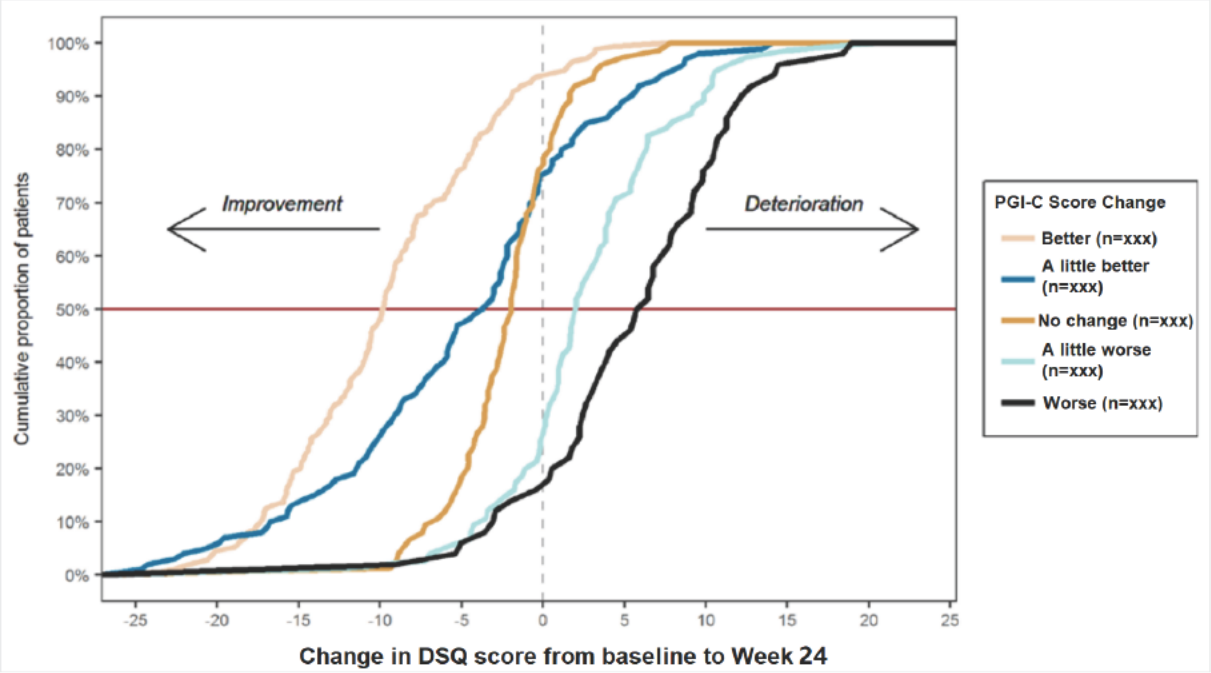
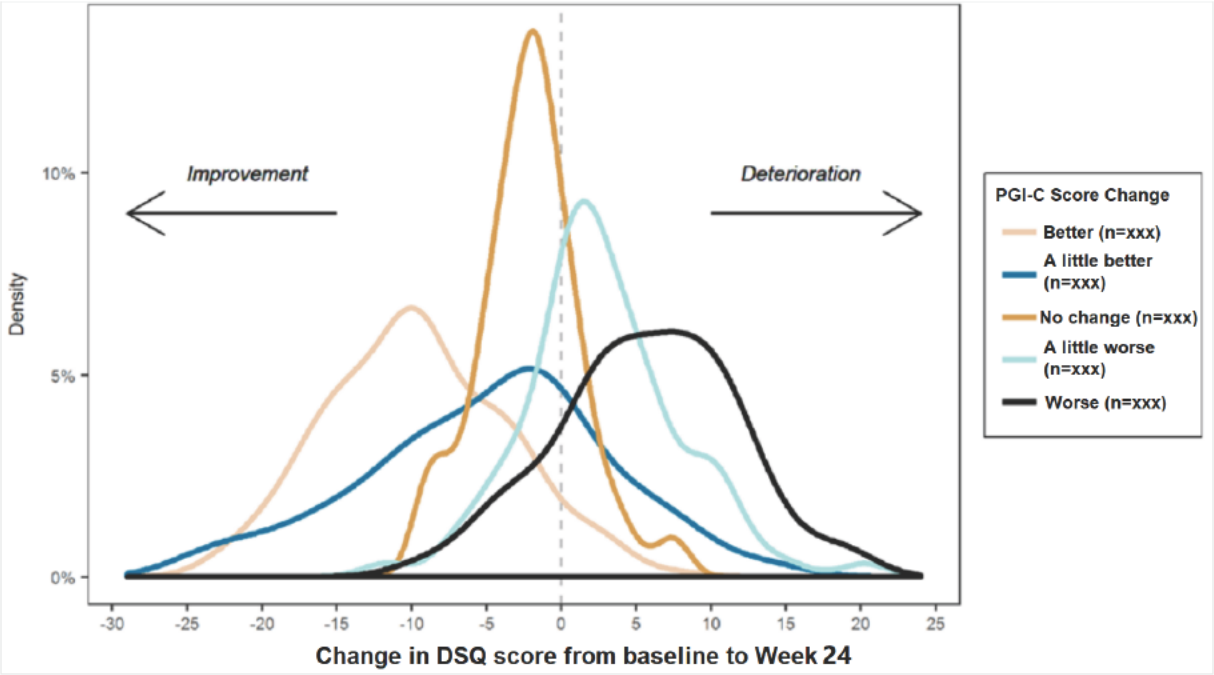


Figure 4 Example of Empirical Density Function (ePDF) Curves of Change in DSQ Score from Baseline to Week 24 grouped by Patient Global Impression of Change (PGI-C) score.



In addition, the change from baseline in DSQ score at Week 24 by baseline PGI-S will be summarized for the subjects who achieved 1-category, 2-category and 3-category PGI-S improvement, respectively.

Correlations between the two primary endpoints – the proportion of patients achieving a histological response at Week 24, and the change from baseline in DSQ score at Week 24, will be explored using box plots and tables of change in DSQ at Week 24 by histological response or not, and/or box plots and tables of change in DSQ over time split by histological response or not. Additional statistical analyses may be performed if appropriate to explore further. An ANCOVA model may be used. The dependent variable will be the change from baseline in DSQ score up to Week 24, histological response and treatment will be included as covariates along with region, prior response to steroids for EoE and presence of strictures at baseline. Presence of strictures at baseline will be the first covariate to be excluded from the ANCOVA models if in the instance of convergence issues or not enough data.

4.2.5 Secondary efficacy outcome variable

4.2.5.1 Centrally read biopsies for tissue eosinophil counts and additional histopathology

Key secondary endpoint of percent change from baseline in tissue eosinophil counts at Week 24 will be compared between the benralizumab and placebo treatment groups using an analysis of covariance (ANCOVA) model. A composite strategy estimand will be used

whereby the occurrence of randomised treatment discontinuation, increases in background therapies, addition of a new therapy for EOE, or dilation procedures prior to week 24 will result in the week 24 change from baseline value to be imputed using return-to-baseline MI.

The dependent variable will be the percent change from baseline in tissue eosinophil counts at Week 24, baseline tissue eosinophil counts, and treatment will be included as covariates along with region, baseline steroid use and presence of strictures at baseline. Presence of strictures at baseline may be the first covariate to be excluded if not enough data.

Sensitivity analyses may be performed by including all post baseline data reported.

In addition, summaries of number of patients achieving histological response to certain levels will be produced (<1, 1 to ≤6, 7 to <15, 15 to ≤60, >60 eos/hpf, etc.).

The associations between the percent change from baseline in tissue eosinophil counts at Week 24 and baseline tissue/blood eosinophil counts will be evaluated by Loess plots.

4.2.5.2 EoE HSS

Key secondary endpoints of change from baseline in EoE HSS total grade score and change from baseline in EoE HSS total stage score at Week 24 will be compared between the benralizumab and placebo treatment groups using an analysis of covariance (ANCOVA) model. Please follow the ANCOVA model and composite strategy as described for the analysis of tissue eosinophil counts endpoint (see Section 4.2.5.1 for details).

Sensitivity analyses may be performed by including all post baseline data reported.

4.2.5.3 Centrally-read EoE EREFS

Another key secondary endpoint of change from baseline in centrally-read EREFS at Week 24 will be compared between the benralizumab and placebo treatment groups using an analysis of covariance (ANCOVA) model. A composite strategy estimand will be used whereby the occurrence of randomised treatment discontinuation, increases in background therapies, addition of a new therapy for EOE, or dilation procedures prior to week 24 will result in the week 24 change from baseline value to be imputed using return-to-baseline MI.

The dependent variable will be the change from baseline in centrally-read EREFS score (total score, taking the worst of each of the proximal and distal results for each individual component, with the maximum result of 9) at Week 24, baseline centrally-read EREFS score and treatment will be included as covariates along with region, baseline steroid use and presence of strictures at baseline. Presence of strictures at baseline may be the first covariate to be excluded if not enough data.

In addition, the change from baseline in total EREFS score (taking the sum of each of the proximal and distal results for each individual component, with the maximum result of 18) at Week 24 will be analysed using the same ANCOVA model and composite estimand strategy as a supplementary analysis.

Sensitivity analyses may be performed to explore any potential missing data including imputation rules for patients who discontinue prior to Week 24, and also including all post baseline data reported. In addition, exploration into the concordance between site recorded and centrally-read data may be performed.

Descriptive analyses of the EREFS score by location (proximal vs distal) and the worst scores across the components of each location before and after therapy will be produced.

4.2.5.4 Treatment responder at Week 24

The last key secondary endpoint is the proportion of patients with a treatment response at Week 24. Treatment response is defined as composite of histologic response (same histological response criteria defined in Section 3.2) and clinically meaningful improvement (30%) from baseline in DSQ score.

The proportion of patients who achieve treatment response at Week 24 will be compared between benralizumab and placebo using CMH test controlling for region, baseline steroid use, and presence of strictures at baseline.

Patients with not enough data at Week 24 to be determined as treatment responder will be considered non-responders. Any patients with an intercurrent event of randomised therapy discontinuation, an increase in background medications or additional new therapies for EoE, or having a dilation procedure at or before Week 24 will also be considered non-responders at week 24.

A supportive analysis of the same endpoint but using a 50% improvement in DSQ score threshold rather than the 30% improvement threshold will be performed.

The same sensitivity analyses mentioned in histological response rate (Section 4.2.4.2) may be performed if the amount of missing data warrants further exploration.

4.2.5.5 EoE-3D score (PRO)

Change from baseline in EoE-3D items and symptoms scores at Week 24, will be compared between the benralizumab and placebo treatment groups using ANCOVA models. Treatment group will be included as an explanatory variable along with the baseline item scores. Other explanatory variables considered in the analysis include region, baseline steroid use, and presence of strictures at baseline. Presence of strictures at baseline may be the first covariate to be excluded from the ANCOVA models if in the instance of convergence issues or not enough data. Return-to-baseline MI imputations will be made 100 times for patients with the

intercurrent events, missing data not due to intercurrent events will be imputed 100 times using multiple imputation with missing at random assumption. The change from baseline in EoE-3D scores will then be analysed using an analysis of covariance (ANCOVA) model for each imputation. Apply Rubin's rule (Rubin et al 1986, Rubin 1987) to combine analysis results from 100 imputations for the final analysis result.

Dysphagia episode frequency will be summarized as monthly (28-day) counts. Descriptive statistics will be provided by period.

The pain, discomfort, and overall severity of the event (Questions 4, 5, 6) and severity of abdominal pain/nausea (additional questions 2 and 3) will be reported using 11-point Numeric Rating Pain Scale (NRPS, 0-10) in which 0 indicates no pain and 10 indicates the worst pain imaginable. The pain, discomfort and overall severity scores will be summarized as 14-day means and change from baseline to Week 24 will be analysed using three separate ANCOVA models in a similar way to that described for the change in DSQ primary endpoint. For the three ANCOVA models the dependent variables will be the changes from baseline of the respective scores at Week 24, and each will include the relevant baseline score as a covariate. Items for abdominal pain and nausea will be evaluated as separate items following the same approach used for EoE-3D content.

Summary statistics for episode frequency and severity and change from baseline in item scores will be produced by treatment group and visit.

4.2.5.6 EoE-QoL-A score (PRO)

Changes from baseline in Adult Eosinophilic Esophagitis Quality of Life (EoE-QoL-A) summary score and domain scores will be analysed using ANCOVA models in a similar way to that described for the change in DSQ primary endpoint. The dependent variable will be the change from baseline in the EoE-QoL-A summary or domain score at Week 24 and each will include the relevant baseline score as a covariate. Return-to-baseline MI imputations will be made 100 times for patients with the intercurrent events, missing data not due to intercurrent events will be imputed 100 times using multiple imputation with missing at random assumption. The change from baseline in EoE-QoL-A scores will then be analysed using an analysis of covariance (ANCOVA) model for each imputation. Apply Rubin's rule (Rubin et al 1986, Rubin 1987) to combine analysis results from 100 imputations for the final analysis result.

Descriptive summary statistics for change from baseline in EoE-Qol-A summary score and domain scores will be produced by treatment group and visit.

4.2.5.7 Pediatric Eosinophilic Esophagitis Symptom Severity Module, Version 2, Children and Teens Report (PEESS) – sub-study (PRO)

Descriptive summary statistics for change from baseline in PEESS overall score will be produced by treatment group and visit.

4.2.5.8 SF-36 v2 Health Survey score (PRO)

Changes from baseline in SF-36 v2 health survey scores will be analysed using ANCOVA models in a similar way to that described for the change in DSQ primary endpoint. Ten separate ANCOVA models will be fitted, for each of the 8 domain scores (PF, RP, BP, GH, VT, SF, RE, MH) and 2 physical and mental health component summary scores (PCS and MCS). Return-to-baseline MI imputations will be made 100 times for patients with the intercurrent events, missing data not due to intercurrent events will be imputed 100 times using multiple imputation with missing at random assumption. The change from baseline in SF-36 v2 health survey scores will then be analysed using an analysis of covariance (ANCOVA) model for each imputation. Apply Rubin's rule (Rubin et al 1986, Rubin 1987) to combine analysis results from 100 imputations for the final analysis result.

The dependent variable will be the change from baseline in the relevant domain scores at Week 24 and each will include the relevant baseline domain score as a covariate.

Summary statistics for change from baseline in SF-36v2 PCS, MCS and Domain scores will be produced by treatment group and visit.

SF-36 responder analysis where a responder from baseline to Week 24 will be conducted using logistic regression adjusted for treatment group, baseline DSQ score, and region. The responder analysis will be supported with a cumulative distribution function plot of change from baseline at Week 24 and descriptive summary tables. Marginal standardization methods (Bartlett 2018) will be used for the model estimates for all rate analyses, including logistic regression, unless otherwise specified.

4.2.5.9 Work Productivity and Activity Impairment questionnaire plus Classroom Impairment Questions (WPAI+CIQ) (PRO)

The WPAI+CIQ data will be summarized by treatment as described in Section 3.4.6.

The number and percentage of patients with health specific resource utilization (defined in Section 3.3.6) will be presented by treatment group.

4.2.5.10 Patient Global Impression of Severity (PGI-S) (PRO)

The number and percentage of patients in each PGI-S response category will be summarized by treatment group and visit.

A shift table will be generated to present changes from baseline to Week 24 with the change from baseline in PGI-S category.

4.2.5.11 Patient Global Impression of Change (PGI-C) (PRO)

The number and percentage of patients in each PGI-C response category will be summarized by treatment group and visit.

4.2.5.12 Diet Questionnaire

Descriptive summaries of the diet questionnaire responses will be produced by treatment group and visit.

4.2.5.13 Sensitivity analyses

If the occurrence of the intercurrent events of randomised treatment discontinuation, increases of background medications, addition of a new therapy for EOE or dilation procedures is high enough to warrant further exploration, sensitivity analyses to explore the impact of alternative estimand strategies dealing with these intercurrent events as described in the individual endpoint sections above will be performed. These may be performed on secondary endpoints if needed. The different approaches that may be considered are:

- Repeated measures endpoints: to explore the primary composite estimand with ANCOVA analyses proposed, an effectiveness strategy will be considered as a sensitivity analysis whereby all data will be included up to the point of the described intercurrent events prior to week 24 and mixed effect models for repeated measures (MMRM) analyses will be performed.

- Appropriate for all endpoints: treatment policy strategy analyses including all data regardless of the occurrence of intercurrent events.

In addition, if the amount of missing data for reasons other than randomised treatment discontinuation and background therapy intercurrent events is high enough to warrant further exploration, sensitivity analyses for the primary and key secondary endpoints based on different missing data mechanism assumption will be used to explore the robustness of any treatment effect including multiple imputation approaches. See Section 8.1 (Appendix) for details.

Sensitivity analyses will also be considered to explore the effect of extreme outliers on individual endpoints, such as rank based methods.

Tipping point analyses will be performed for the dual primary endpoints if they reach statistical significance level (*p*-value less than 0.05).

4.2.5.14 Subgroup analysis

To explore the uniformity of the detected overall treatment effect on the primary efficacy variables, subgroup analyses and statistical modelling including testing for interaction between treatment and covariates will be performed for each of the dual-primary endpoints for the subgroup factors listed below. Analyses will only be performed if sufficient patients in each level of the subgroup are available, condensing of groupings may be considered if necessary.

- Age group (years) (age<18 vs age≥18, age≤21 vs age>21)
- Geographic region (North America, Rest of World)
- Sex (Male, Female)
- Race (White, Asian, Other)
- Baseline steroid use (Yes, No)
- Refractory to steroid (Yes, No)
- Prior PPI use (Yes, No)
- Prior PPI response (Yes, No)
- Baseline steroid or PPI use (Yes, No)
- Duration of EoE symptoms (years) (<5, 5-10, >10)
- Presence of strictures at baseline (Yes, No)
- History of stricture dilations (Yes, No)
- Baseline DSQ score (<median, ≥median)
- Baseline blood eosinophils (cell/µL) (≥150 vs <150, ≥300 vs <300, ≥400 vs <400)
- Baseline tissue eosinophils (<median, ≥median)

For subgroup analyses of the change in DSQ at week 24 endpoint, for each of the subgroup factors in turn, a separate ANCOVA model will be fitted using the same model terms as used for the primary analysis (defined in Section 4.2.4.3), with additional terms for the subgroup main effect and the treatment×subgroup interaction.

Subgroup analyses will also be performed for the proportion of patients achieving a histological response at Week 24, comparing benralizumab and placebo using logistic regression models. The dependent variable will be achieving a histological response at Week 24 (Yes, No) and the independent variables will include the same covariates as in the primary analysis along with additional terms for the subgroup main effect and the treatment×subgroup interaction. Marginal standardization methods (Bartlett 2018) will be used for the model estimates for all rate analyses, including logistic regression, unless otherwise specified.

It is noted that if there are low counts in some of the treatment by subgroup response groups, the logistic regression models may not be reliably estimable, in which case data will be presented descriptively without formal analysis.

For the age<18 subgroup, hierarchical Bayesian borrowing methods may be used to achieve better precision. This analysis would assess the possibility of borrowing data from the 18-21 population as well as the >21 population. Weighting would be assigned to the data to determine what level of data can be called similar to the <18 subgroup and that data then used in the analysis to provide a higher precision.

Subgroup analyses results will be shown in forest plots.

Additional exposure-efficacy subgroup analyses will also be performed separately for each of the co-primary endpoints. These analyses will split the benralizumab treatment group into patients above and below median observed trough PK concentrations and will compare these patients to the full group of placebo patients, for each efficacy endpoint. The difference in efficacy endpoint between treatment groups for each level of the subgroup (> median benralizumab concentrations vs placebo, and <= median benralizumab concentrations vs placebo) and their 95% confidence intervals will be presented in forest plots. These analyses may be repeated using predicted AUC or C_{ave} values from population PK modelling if any differences are observed there that warrant further exploration.

It is important to note that the study has not been designed or powered to assess efficacy within any of these pre-defined subgroups, and as such these analyses are considered as exploratory.

4.2.5.15 Impact on analyses due to COVID-19 pandemic

Given the uncertainty surrounding the future impact of the COVID-19 worldwide pandemic on clinical trials, operational procedures are being implemented in this study to maintain the integrity of collected data. Efforts may be made to collect data via alternative means where possible, when on-site visits cannot be performed.

If there is a sufficient number of protocol deviations or study disruptions as a result of COVID-19, then sensitivity analyses may be conducted to evaluate their impact on the interpretation of results. Protocol deviations, including doses or visits missed due to COVID-19 related protocol deviations will be described separately in the CSR. Confirmed or suspected cases of COVID-19 will be listed and included as AEs as appropriate.

4.2.5.16 Healthcare resource utilization

Proportion of patients with relevant HRU and number of events by HRU type (including but not limited to hospitalizations, length of hospital stay, office visits, emergency room visits, tests and procedures) will be summarized by randomised treatment and visit.

4.2.5.17 Early time point sub study

Participants recruited to the early time point sub-study will have additional endoscopies and biopsies at weeks 4 and 12, along with the other timepoints assessed by all study participants (including baseline, week 24 and week 52). Analysis of tissue eosinophils will be performed in the subset of participants recruited to the early time point sub study. This will include analysis of percent change from baseline in tissue eosinophils which will be analysed at weeks 4, 12, 24, and 52 with separate ANCOVA models at each timepoint, using similar methods and intercurrent event handling to that outlined for the percent change in tissue eosinophils at week 24 endpoint outlined in section 4.2.5.1, removing covariates if needed due to the smaller n. Supportive summary statistics of absolute levels and changes from baseline in tissue eosinophils by timepoint will also be produced. Plots of absolute values and percent change from baseline in tissue eosinophils at each time point up to week 24 will also be produced.

Analyses of other exploratory endpoints including HSS scores, EREFS and exploratory biomarkers will be described in the EAP and reported outside the CSR.

4.2.6 Safety analysis

All safety variables will be summarized using the safety analysis set and data presented according to actual treatment received.

The first analysis of safety data will include only data from the double-blind, placebo-controlled first 24 weeks of the study (DB period). Patients will be analyzed according to the treatment they actually received (benralizumab or placebo). A second analysis of safety data will include all data reported in the study for patients receiving benralizumab from the start of treatment. Safety data from patients' entire duration on benralizumab during the DB period, along with the benralizumab OL period and OLE period will be summarized. Additional safety data presentations based on the open-label benralizumab set will be included to summarize safety data from patients who switched from placebo to receive benralizumab after 24 weeks, with only their safety data while receiving benralizumab included. If there is considerable drop out between the first 52 weeks of the study and the OLE, additional safety summaries may be produced on the OLE benralizumab analysis set to avoid any concern around rollover bias between parts of the study.

Plot of frequencies and risk differences (forest plots) between treatment arms will be presented for the most common adverse events and other specific events of interest. Estimates and confidence intervals based on the Miettinen Nurminen (M-N) method will also be presented for the most common adverse events and any other specific events of interest included in the structured assessment of benefit risk.

Summaries of exposure and overall adverse events by category will be produced in the following subgroups.

- Age group (age<18 vs age≥18, age≤21 vs age>21)
- Sex (Male vs Female)
- Race (White, Asian, Other)
- BMI (≤30 vs >30 kg/m2)
- Geographic region (North America vs Rest of World)

Forest plots showing the differences in the proportion of patients (benralizumab – placebo) reporting at least 1 AE, at least 1 SAE, and at least 1 AE leading to discontinuation by the above subgroups, with associated 95% confidence intervals using the Miettinen Nurminen (M-N) method will be constructed to illustrate consistency across subgroups.

4.2.6.1 Adverse events (AEs)

Adverse events (AEs) will be summarized separately for the on-treatment and on-study periods, as defined in Section 3.6.1. Additionally, only serious adverse events (SAEs) in the pre-treatment period (with start date prior to the first dose of IP) will be listed. All AEs will be listed for each subject. All summaries will be presented by treatment group and will be exposure-adjusted to account for the variability in follow-up periods beyond 24 or 52 weeks.

The rate of AEs per person-years at risk will be calculated as (number of patients reporting the AE)/(total IP exposure with patients at risk of AE) for on-treatment and on-study periods. The post-treatment AEs will be listed in listings. The total period at risk for each patient will be the duration of the on-treatment, post-treatment and on-study periods as defined in Section 3.6.1. Rates will be expressed in terms of events per 100 patient-years.

An overall summary table will be produced showing the number, percentage, and exposure-adjusted rate of patients with at least 1 AE in any of the following categories; AEs, serious adverse events (SAEs), AEs with outcome of death, and AEs leading to discontinuation of investigational product (DAEs).

AEs, AEs with outcome of death, SAEs and DAEs will be summarised by System Organ Class (SOC) and Preferred Term (PT) assigned to the event by MedDRA. For each PT, the number, percentage and exposure-adjusted rate of patients reporting at least one occurrence will be presented (ie, multiple occurrences of an AE for a patient will only be counted once).

A summary of the most common (frequency of >3%) AEs will be presented by PT. Additionally, a summary of non-serious AEs occurring in >5% of patients in any treatment group will be presented by PT. AEs causing discontinuation of the study treatment or from the study will also be summarised.

AEs and SAEs will be summarised by preferred term and investigator's causality assessment (related vs. not related) and maximum NCI CTCAE grade intensity. If a patient reports multiple occurrences of the same AE within the same study period, the maximum intensity will be taken as the highest recorded maximum intensity (the order being mild, moderate, and severe).

Other significant adverse events will include but may not be limited to injection site reactions and hypersensitivity events. Adverse events of injection site reactions (high level term of administration and injection site) and hypersensitivity [standardized MedDRA query of hypersensitivity (narrow)] will be summarised by preferred term. The summary of injection site reactions will be summarised by injection site location and number of IP administrations. The summary of AEs of hypersensitivity will be presented overall and repeated for events causally related to IP as assessed by the investigator.

4.2.6.2 Laboratory data

All continuous laboratory parameters will be summarized by absolute value at each visit by treatment group, together with the corresponding changes from baseline. The summary statistics presented will be the minimum, 1st quartile, median, 3rd quartile, maximum, mean and SD. Mean changes from baseline over time will also be plotted by treatment group.

AstraZeneca defined extended reference ranges will be used for the identification of individual clinically important abnormalities, and a shift table will be produced for each laboratory parameter to display low, normal, high, and missing values. The shift tables will present baseline and maximum/minimum on-treatment value, as applicable for each parameter.

Shift plots showing each individual patient's laboratory value at baseline and at maximum/minimum will be produced for each continuous laboratory variable. If any laboratory variables show any unusual features (high or low values or a general shift in the data points) at other time points then shift plots of these data may be produced. A diagonal line indicating no change, and horizontal and vertical reference lines indicating the limits of the AstraZeneca defined reference ranges will also be displayed on the shift plots.

Data for patients who have treatment-emergent changes outside the predefined criteria will be presented. This data presentation will include all visits for this subset of patients.

The frequency of changes with respect to normal ranges between baseline and each post-treatment time point will be tabulated. Frequencies of clinically noteworthy values (using AstraZeneca defined reference ranges) occurring during the clinical study will also be given.

For urinalysis data, a shift table will be generated to present changes from baseline to EOT. The number of patients with treatment-emergent changes will also be summarized. Here,

treatment-emergent changes are defined as 1) None/Trace at baseline to +, ++, +++, ++++ at any visit after baseline or 2) Increase of at least ++.

Any data outside the AstraZeneca normal and extended reference ranges will be explicitly noted on the listings that are produced.

4.2.6.3 ECGs

The Investigator's assessment of the 12-lead ECG (normal or abnormal) will be listed for all patients, along with detailing whether any abnormalities were clinically significant or not.

The number and percentage of patients with clinically significant abnormal ECGs will be summarized by treatment group. Only ECG at baseline will be included.

4.2.6.4 Physical Examination

No summary of physical examinations will be presented.

4.2.6.5 Vital sign and Weight

Vital sign parameters will be presented for each treatment group. Summary statistics for continuous variables cover n, mean, SD, Minimum, Q1, median, Q3, and Maximum. Frequency tables cover number and percentage of patients in the respective category.

For each scheduled post-baseline visit, descriptive statistics for all vital sign parameters (systolic and diastolic blood pressure, pulse, respiration rate, oral temperature) and body weight will be presented for observed values and change from baseline.

Changes in vital signs and weight will be examined according to Table 1. Frequencies of clinically noteworthy values occurring during the clinical study will be presented using AstraZeneca defined reference ranges, and clinically important change criteria.

All recorded vital signs data will be listed.

5 OLE TREATMENT PERIOD

For patients entering the OLE, at the OLE analysis, summaries from the OLE will be presented for the overall population, and by prior randomized treatment (benralizumab or placebo).

In addition, selected efficacy and safety data may be integrated for those patients randomized to benralizumab, to describe efficacy and safety data over the entire study follow-up period. The only OLE data that will be presented at the primary analysis (when the double blind period has completed) is a top level overview of exposure and AEs, integrated with the double-blind and open-label periods data to give a view of the safety profile over the longest follow-up accrued in the study at that point.

6 CHANGES OF ANALYSIS FROM PROTOCOL

The protocol allowed enrolment to the Early Time Point Sub-study to continue once the required number for the primary analysis population had been recruited in the event that the required sample size for the Early Time Point Sub-study had not yet been reached. If this extension for Sub-study recruitment occurred, the aim was that the additional Sub-study patients would be analysed at a later point and not included in the primary analysis. This note is to clarify that this extension to enrolment for the Early Time Point Sub-study was not needed and so the primary analysis population consists of the complete population of all patients randomised in the trial, including the complete sub-study population.

The final full analysis set population is larger than the originally planned 170 patients (approximately 200 patients actually randomised) due to a large number of patients being in screening at the time recruitment completed who then became eligible for randomisation.

PEESS was descriptively summarized only since there are not enough paediatric subjects available.

7 REFERENCES

Bartlett 2018
Bartlett J W, Covariate adjustment and estimation of mean response in randomised trials. Pharm Stat. 2018 Sep; 17(5):648-666.

Coon et al 2018
Coon C D, Cook K F. Moving from significance to real-world meaning: methods for interpreting change in clinical outcome assessment scores. Qual Life Res 2018; Jan 27; (1):33-40.

Dellon et al 2013
Dellon, E. S., Irani, A. M., Hill, M. R., & Hirano, I. Development and field testing of a novel patient-reported outcome measure of dysphagia in patients with eosinophilic esophagitis. Alimentary Pharmacology & Therapeutics, 38, 634–42.

Dellon et al 2017
Dellon ES, Katzka DA, Collins MH, Hamdani M, Gupta SK, Hirano I, on behalf of the MP 10 06 Investigators. Budesonide Oral Suspension Improves Symptomatic, Endoscopic, and Histologic Parameters Compared with Placebo in Patients with Eosinophilic Esophagitis. Gastroenterology. 2017;152:776–786.

FDA 2018
Food and Drug Administration. Methods to Identify What is Important to Patients Select,

Develop or Modify Fit-for-Purpose Clinical Outcomes Assessments, PATIENT-FOCUSED DRUG DEVELOPMENT GUIDANCE PUBLIC WORKSHOP. October 15-16, 2018.

FDA 2019
Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Draft guidance for industry. Eosinophilic Esophagitis: Developing Drugs for Treatment. February 2019.

FDA meeting minutes 2019
Food and Drug Administration, Center for Drug Evaluation and Research (CDER). Meeting minutes. March 7, 2019.

Hirano et al 2017
Hirano I, Dellon ES, Hamilton JD, Collins MH, Peterson K, Chehade M et al. Dupilumab Efficacy and Safety in Adult Patients With Active Eosinophilic Esophagitis: a Randomized Double-Blind Placebo Controlled Phase 2 Trial. Presented at World Congress of Gastroenterology ACG, Orlando, Florida, USA, 13-18 October 2017.

Hudgens et al 2017
Hudgens S, Evans C, Phillips E, Hill M. Psychometric validation of the Dysphagia Symptom Questionnaire in patients with eosinophilic esophagitis treated with budesonide oral suspension. J Patient Rep Outcomes. 2017; 1(1):3

QualityMetric 2011
QualityMetric, I. User's manual for the SF-36v2 Health Survey (3 ed.). 2011. Lincoln, RI

Rubin et al 1986
Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81: 366-374. 1986.

Rubin 1987
Rubin DB. Multiple Imputation for Nonresponse in Surveys. 1987. New York: Wiley.

Taft et al 2011
Taft T H, Kern E, Kwiatek M A, Hirano I, Gonsavles N, Keefer L. The adult eosinophilic oesophagitis quality of life questionnaire: a new measure of health-related quality of life. Aliment Pharmacol Ther. 2011 Oct; 34(7):790-8.

8 APPENDIX

8.1 Accounting for missing data

8.1.1 Accounting for missing data for change in DSQ at Week 24

In this study some patients dropping out of the study potentially leads to unobserved events. The amount of missing data is minimized in this study as all patients switch to receive benralizumab after the first 24 weeks and are encouraged to complete visits until they withdraw from the study even if they discontinue treatment. In addition, in the primary analyses most sources of missing data are accounted for with the composite estimand strategy that imputes outcomes for patients who discontinue randomised treatment.

This section summarizes how we will describe the pattern of and reasons for missing data from the study. It will also describe how we plan to account for missing data, including both the primary and sensitivity analyses to assess the robustness of the treatment effect under different underlying assumptions to account for missing data.

The methodology is outlined below for the change in DSQ primary endpoint, but similar techniques may also be used for other endpoints if appropriate.

8.1.1.1 Primary analysis under the Composite Strategy Estimand

The primary analysis for the primary endpoint of DSQ at Week 24 allows for differences in outcomes over the study treatment period up to 24 weeks. In this analysis, all the data up to Week 24 visit will be included with imputation of return-to-baseline MI for patients experiencing intercurrent events of randomized treatment discontinuation, increase in background medications or additional new therapies for EOE, or a dilation procedure. The missing data at Week 24 which was not due to intercurrent events will be imputed by MI with missing at random (MAR) assumption. The primary analysis uses the ANCOVA method, treatment group will be included as an explanatory variable along with the baseline DSQ score, region, baseline steroid use, and presence of strictures at baseline as explanatory variables, and assumes that missing data is missing at random (MAR) and is a direct likelihood approach (DL).

8.1.1.2 Sensitivity analysis under the effectiveness estimand using the Missing at Random (MAR) assumption.

An effectiveness estimand sensitivity analysis will be explored where instead of return-to-baseline MI for intercurrent events, data after the intercurrent event until week 24 will instead be treated as missing and a mixed effect model for repeated measures (MMRM) analysis will be used for the remaining data. The dependent variable in the MMRM model will be the change from baseline in the continuous outcome at Week 24 visit. Treatment group, baseline values, region, baseline steroid use, and presence of strictures at baseline, visit, and treatment group × visit will be the covariates. The variance-covariance matrix will be assumed to be

unstructured (UN). If the procedure does not converge, then the Toeplitz, first-order autoregressive (AR(1)), compound symmetric (CS), and variance components (VC) variance-covariance matrices will be tried in that order. The estimate of the treatment effect will be based on a contrast from this MMRM model.

It is noted that if the primary analysis is statistically significant, it is not necessarily expected that all sensitivity analyses will also give statistically significant results. If the results of the sensitivity analyses provide reasonably similar estimates of the treatment effect to the primary analysis, this will be interpreted as providing assurance that neither the lost information nor the mechanisms which cause the data to be missing have an important effect on the primary analysis conclusions. Based on these outputs and the drug's mechanism of action, the plausibility of the assumptions we make about missing data in the different analyses will be considered and described in the clinical study report.

8.1.1.3 Tipping point Analyses

To examine the impact of missing data for the dual-primary endpoints, tipping point analyses may be performed. These analyses will systematically vary the assumptions about outcomes among the subsets of participants on the treatment arms who have missing data at Week 24 for any reason other than the occurrence of the specified treatment failure intercurrent events which are handled with non-response imputation. Tipping point analyses are intended to identify the point at which the results would tip from statistically significant to not statistically significant. Thus, the tipping point analyses will only be performed if an endpoint achieves a nominally statistically significant result (ie, nominal p-value < 0.05).

The analyses will be performed following below steps for histologic response:

- For the primary analysis, participants who have missing data at Week 24 for any reason other than the treatment failure intercurrent events are by definition imputed as non-responders. For this sensitivity analysis, first all the non-responders on placebo arm will be imputed as responders and check if the result can be tipped.
- If the result tips then subjects with missing data will be imputed using multiple imputation with missing at random (MAR) assumption. Placebo subjects will have their first imputed value improved by δP in log odds. This results in a one-time shift towards a better value in the outcomes of placebo subjects. Benralizumab subjects will have their first imputed value worsened by δT in log odds. This results in a one-time shift towards a worse value in the outcomes of Benralizumab subjects. Tipping points are defined as the range of smallest values (δP, δT) which would result in a change of conclusion.

For the tipping point analysis of change from baseline in DSQ score at Week 24, only the subjects with missing DSQ score at Week 24 (after intercurrent event imputations are performed) will be shifted. Patients with intercurrent events will be handled with return to baseline multiple imputation as per the primary analysis. In this analysis, various degrees of

improvement in the placebo group δP, and various degrees of worsening in the benralizumab group δT, will be simultaneously explored.

Placebo subjects who have missing data at Week 24, without having previously had one of the other intercurrent events causing return to baseline multiple imputation, will have their first imputed value improved by δP. This results in a one-time shift towards a better value in the outcomes of placebo subjects. Benralizumab subjects who discontinued early or have missing data at Week 24 will have their first imputed value worsened by δT. This results in a one-time shift towards a worse value in the outcomes of Benralizumab subjects. The maximum shift factor is 84 which is the worst possible score for 14-day DSQ.

Tipping points are defined as the range of smallest values (δP, δT) which would result in a change of conclusion.

8.2 Analysis plan for immunogenicity data

Serum samples for ADA assessments will be conducted utilizing a tiered approach (screen, confirm, titre) and ADA data will be collected at scheduled visits shown in the CSP. ADA result from each sample will be reported as either positive or negative. If the sample is positive, the ADA titre will be reported as well. In addition, the presence of neutralizing antibodies (nAb) will be tested in all ADA-positive samples using a ligand binding assay. The nAb results will be reported as positive or negative.

In general, patients with a missing baseline ADA assessment will be assumed to be ADA negative at baseline as a conservative approach to ensure that all subjects are included in all analyses. If a positive ADA titre result is reported as ≤50, then the titre will be imputed as 50 for titre summaries. ADA results from samples collected post-dose instead of pre-dose on an IP administration day are considered unreliable and should be excluded from all derivations.

For each subject, the following ADA and nAb responses will be evaluated over the double blind as well as double blind combined with open label period:

- Subjects who are ADA positive at any time during the study, including baseline and/or post-baseline (also generally referred to as ADA positive). The proportion of ADA-positive subjects in a population is known as ADA prevalence.

- Subjects who are ADA negative at all assessments, including baseline and post-baseline (also generally referred to as ADA negative).

- Treatment-emergent ADA positive (referred to as ADA incidence). A positive post-baseline result and either of the following statements holds:

  - Baseline is ADA negative and at least one post-baseline assessment is ADA positive. This is called treatment-induced ADA positive.

  - Baseline is ADA positive, and the baseline titre is boosted by greater than the variability of the assay (i.e. > 4-fold increase) at ≥1 post-baseline timepoint. This is called treatment-boosted ADA positive.

- Subjects who are persistently ADA positive, which is defined as ADA negative at baseline and having at least 2 post-baseline ADA positive measurements with ≥ 16 weeks between first and last positive, or an ADA positive result at the last available post baseline assessment.

- Subjects who are ADA positive with maximum titre > median of maximum titres. The median of maximum titres will be calculated based on the maximum titre of each ADA positive subject within each treatment group (including both baseline and post-baseline measurements).

- nAb positive. Defined as nAb positive at any visit including baseline and/or post-baseline (also referred to as nAb prevalence)

- Subjects who are persistently ADA positive and nAb positive.

The responses above will be summarized as counts and percentages by treatment group. The maximum ADA titre over the on-study period will also be summarized for patients in each of the ADA positive response categories listed above. The maximum titre will be derived based on all available ADA titres reported for each subject, including any unscheduled assessments.

ADA response (positive or negative) and titre will be summarized at baseline and at all scheduled post-baseline visits by treatment group using derived visit windows (refer to Section 3.1.1 for detailed definition of visit windows). In the event a patient has more than one result within a given visit window, the maximum ADA titre will be used in the by-visit summary. In addition, the ADA response will be presented cumulatively. The cumulative ADA response is positive for a specific visit if a positive ADA result is detected at any time point up to and including the specific visit. If all ADA result are negative up to the specific visit, then the cumulative ADA response is negative for that visit. A summary of the number and percentage of patients who are ADA positive at a post-baseline assessment for the first time by visit will also be presented. A line plot of the proportion of subjects who are ADA positive at each visit will be provided.

The proportion of patients with positive nAb response will be summarized by visit. The summary will be repeated for ADA persistently positive patients.

Key patient information will be listed for patients with positive ADA results, including ADA status, nAb status, titer, benralizumab serum concentration, and eosinophil level.

All analyses will be conducted on the safety analysis set by treatment group unless otherwise specified. All ADA results will be listed.

ADA and eosinophil levels

Blood and tissue eosinophil levels will be summarised by visit for the following ADA response categories of patients: ADA positive, ADA negative, treatment-emergent ADA positive, ADA persistently positive, ADA positive with titer > median of maximum, nAb-positive, both ADA persistently positive and nAb positive. A line plot of eosinophil levels by visit and ADA status will also be presented.

ADA and efficacy

No statistical comparisons of benralizumab versus placebo by ADA status (positive/negative) are planned. The effects of ADA on the primary endpoints will be evaluated through summary statistics by ADA status (ADA positive, ADA negative, treatment-emergent ADA positive, ADA persistently positive, ADA positive with titer > median of maximum, nAb-positive, both ADA persistently positive and nAb positive).

ADA and safety

Adverse events during the study (separately for on-treatment and on-study periods) will be summarized by ADA status (ADA positive, ADA negative, treatment-emergent ADA positive, ADA persistently positive, ADA positive with titer > median of maximum). The on-treatment and on-study periods are as defined in Section 3.6.1.

ADA and PK

Benralizumab serum concentrations will be summarised by visit and ADA status (ADA positive, ADA negative, treatment-emergent ADA positive, ADA persistently positive, ADA positive with titer > median of maximum, nAb-positive, both ADA persistently positive and nAb positive) for patients in the PK analysis set.

8.3 Partial dates for adverse events and prior/concomitant medications

Dates missing the day or both the day and month of the year will adhere to the following conventions in order to classify treatment-emergent AEs and to classify prior/concomitant medications:

Adverse Events

- The missing day of onset of an AE will be set to:
  - First day of the month that the event occurred, if the onset YYYY-MM is after the YYYY-MM of first study treatment
  - The day of the first study treatment, if the onset YYYY-MM is the same as YYYY-MM of the first study treatment
  - The date of informed consent, if the onset YYYY-MM is before the YYYY-MM of the first treatment.
- The missing day of resolution of an AE will be set to:
  - The last day of the month of the occurrence. If the patient died in the same month, then set the imputed date as the death date.
- If the onset date of an AE is missing both the day and month, the onset date will be set to:
  - January 1 of the year of onset, if the onset year is after the year of the first study treatment
  - The date of the first treatment, if the onset year is the same as the year of the first study treatment
  - The date of informed consent, if the onset year is before the year of the first treatment
- If the resolution date of an AE or end date of an IP is missing both the day and month, the date will be set to:
  - December 31 of the year of occurrence. If the patient died in the same year, then set the imputed date as the death date.

Prior/concomitant medication

- The missing day of start date of a therapy will be set to the first day of the month that the event occurred.
- The missing day of end date of a therapy will be set to the last day of the month of the occurrence.
- If the start date of a therapy is missing both the day and month, the onset date will be set to January 1 of the year of onset.
- If the end date of a therapy is missing both the day and month, the date will be set to December 31 of the year of occurrence.

- If the start date of a therapy is null and the end date is not a complete date then the start date will be set to the earliest of the imputed partial end date and the date of the first study visit.
- If the start date of a therapy is null and the end date is a complete date
    - and the end date is after the date of the first study visit then the start date will be set to the date of the first study visit.
    - otherwise the start date will be set to the end date of the therapy.
- If the end date of a therapy is null and the start date is not a complete date then the end date will be set to the study end date.
- If the end date of a therapy is null and the start date is a complete date
    - and the start date is prior to the study end date then the end date will be set to the study end date.
    - otherwise, the end date will be set to the start date of the therapy.

SIGNATURE PAGE

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature

Document Name: d3255c00001-sap-ed-4		
Document Title:	Statistical Analysis Plan Edition 4	
Document ID:	Doc ID-004177148	
Version Label:	4.0 CURRENT LATEST APPROVED	
Server Date **(dd-MMM-yyyy HH:mm 'UTC'Z)**	**Signed by**	**Meaning of Signature**
06-Sep-2022 13:04 UTC	[redacted]	Content Approval
06-Sep-2022 07:44 UTC	[redacted]	Content Approval

Notes: (1) Document details as stored in ANGEL, an AstraZeneca document management system.