STATISTICAL ANALYSIS PLAN

CELIM-RCD-002

A PHASE 2A, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, PARALLEL-GROUP STUDY TO EVALUATE THE EFFICACY AND SAFETY OF AMG 714 IN ADULT PATIENTS WITH TYPE II REFRACTORY CELIAC DISEASE, AN IN SITU SMALL BOWEL T CELL LYMPHOMA

Version 2.0

Phase 2a study

Standard: GCP

Date of approval: 16JUN2017

SPONSOR: Celimmune, LLC

Confidential

NCT Number: 02633020
This NCT number has been applied to the document
for purposes of posting on clinicaltrials.gov

TABLE OF CONTENTS

Signature Page: StatFinn

This Statistical Analysis Plan was prepared by:

Name of Author:	PPD
Title:	Statistician
Signature:	PPD
Date:	16JUN2017

SIGNATURE PAGE: STATFINN

This Statistical Analysis Plan was reviewed/approved by:

Name of Reviewer:	PPD [redacted]
Title:	Senior statistician
Signature:	PPD [redacted]
Date:	16JUN2017

SIGNATURE PAGE: CELIMMUNE, LLC

This Statistical Analysis Plan was reviewed/approved by:

Name of Client´s Representative:	PPD [redacted], MD, PhD
Title:	CEO & Chief Medical Officer
Company:	Celimmune, LLC
Signature:	
Date:	16JUN2017

ABBREVIATIONS

18FDG-PET	Fluorodeoxyglucose positron emission tomography
ADA	Anti-drug antibodies
ADR	Adverse drug reaction
AE	Adverse event
AIC	Akaike Information Criterion
ALT	Alanine Aminotransferase
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
AR(1)	First-order auto-regressive
AST	Aspartate Aminotransferase
BMI	Body mass index
BP	Bodily pain
BSA	Body surface area
BSFS	Bristol Stool Form Scale
BUN	Blood Urea Nitrogen
CD3	Cluster of differentiation 3
CD8	Cluster of differentiation 8
CeD PRO	Celiac Disease Patient Reported Outcome
CeD-GSRS	Celiac disease GSRS
C_{max}	Maximum concentration
CRP	C-reactive protein
CS	Compound symmetry
CT	Computer tomography
C_{trough}	Minimum concentration
CV%	Coefficient of variation
$CV\%_{geo}$	Geometric CV%
DSMB	Data Safety Monitoring Board
EATL	Enteropathy-associated T-cell lymphoma
ECG	Electrocardiogram
EQ-5D	European Quality of Life 5 Dimensions questionnaire
FOCBP	Females of child bearing potential
GEE	Generalized estimating equation
GFD	Gluten free diet
GH	General health perceptions
GIP	Gluten immunogenic peptides
GSRS	Gastrointestinal Symptom Rating Scale
GzmB	Granzyme B
HEENT	Head, eyes, ears, nose, throat
Hep B	Hepatitis B
Hep C	Hepatitis C
icCD3+	Intra-cellular CD3-positive
HIV	Human immunodeficiency virus
IEC	Intestinal epithelial cells
IEL	Intraepithelial lymphocyte

IHC	Immunochemistry
IL-15	Interleukin 15
IL-21R	Interleukin 21 receptor
ITT	Intention to treat
LDH	Lactate Dehydrogenase
LLOQ	Lower limit of quantification
LOCF	Last observation carried forward
MAX	Maximum
MCS	Mental component summary
Mean$_{geo}$	Geometric mean
MedDRA	Medical Dictionary for Regulatory Activities
MFI	Mean fluorescence intensity
MH	Mental health
MIN	Minimum
MMRM	Linear mixed effects repeated measures model
MRI	Magnet resonance imaging
NAb	Neutralizing antibodies
NKG2D	Natural killer group 2D (an activating receptor)
NMISS	Number of subjects with missing observations
OR	Odds ratio
PCS	Physical component summary
PD	Pharmacodynamic(s)
PF	Physical functioning
PGA	Physician Global Assessment of Disease
PK	Pharmacokinetic(s)
POC	Proof-of-concept
PP	Per protocol
PRO	Patient reported outcome
PT	Preferred term
PtGA	Patient Global Assessment of Disease
RBC	Red blood cell
RCD-II	Type II Refractory Celiac Disease
RE	Role limitations due to emotional problems
SAE	Serious adverse event
SAP	Statistical analysis plan
sCD3-	Surface CD3-negative
sCD8-	Surface CD8-negative
SD	Standard deviation
SF-12	Short Form 12 questionnaire
SOC	System organ class
TcR	T cell receptor
ULOQ	Upper limit of quantification
UN	Unstructured
VT	Vitality
VH:CD	Villous height to crypt depth ratio
WBC	White blood cell

WHO DD	World Health Organization Drug Dictionary

1 Introduction

This is a statistical analysis plan (SAP) for study CELIM-RCD-002 which is based on the final study protocol CELIM-RCD-002 Version 3 (dated 11JUL2016). This SAP describes the statistical analyses which will be presented in the clinical study report.

2 Study objectives and endpoints

The study objectives are the following:

The **primary objective** of the study is:

- To assess the efficacy of AMG 714 in treating Type II Refractory Celiac Disease (RCD-II) in adult patients.

The **primary efficacy endpoint** is:

- The Immunological Response 1, the % of aberrant intraepithelial lymphocytes (IELs) vs total IELs as assessed by flow-cytometry.

The primary endpoint of the study will be evaluated by the relative (%) change from baseline to Week 12 in the % of aberrant IELs vs total IELs between the AMG 714 dose arm and the placebo arm.

The **secondary efficacy endpoints** of the study are:

- Immunological Response 2 (i.e. the % of aberrant IELs vs intestinal epithelial cells [IECs]);
- Histological response – small intestinal villous height to crypt depth (VH:CD) ratio;
- Histological response – Marsh score;
- Histological response – total IEL counts;
- Clinical symptoms (i.e. Bristol Stool Form Scale [BSFS], Gastrointestinal Symptom Rating Scale [GSRS] and celiac disease GSRS [CeD-GSRS]).

The secondary endpoints Immunological Response 2, VH:CD ratio, Marsh score and total IEL counts will be evaluated by the change from baseline to Week 12 between the AMG 714 dose arm and the placebo arm. BSFS, GSRS and CeD-GSRS will be evaluated by comparing the change from baseline in weekly scores of AMG 714 dose arm and placebo arm.

For the purpose of this study, and in agreement with leading experts in RCD-II (Malamut *et al*, 2010; Nijeboer *et al,* 2015a, 2015b), aberrant IELs will be defined by flow cytometry as surface CD3-negative, intra-cellular CD3-positive IELs (sCD3-, icCD3+). The cut-off chosen for diagnosis of RCD-II is 20% in accordance with most recent studies (Nijeboer *et al*, 2015b). In IHC, these cells are identified as icCD3+, sCD8- and the cut-off is 50% (Nijeboer *et al*, 2015b).

In addition to this standard definition of aberrant IEL (what can be called “classic RCD-II”), it is not uncommon to observe the diagnosis of RCD-II in subjects with an atypical flow cytometric phenotype of the aberrant IELs, yet otherwise meeting the definition of RCD-II. These “Atypical RCD-II” patients may be enrolled in the study and would be part of the intent-to-treat (ITT) population as well as of the per protocol (PP) population except for the analysis of Immunological Response 1 and 2 (since the aberrant cells object of Immunological Response endpoints are different). In other words, all endpoints except Immunological Response 1 and 2 will be assessed with inclusion of any atypical subject enrolled. However, because it is possible that these patients behave differently, sensitivity analyses may also be conducted excluding the atypical subjects.

The **secondary objective** of the study is:

- To assess the safety and tolerability of AMG 714 when administered to adult patients with RCD-II.

The **safety endpoints** of the study are:

- Adverse events (AEs);
- Clinical laboratory tests;
- Physical examination;
- Vital signs;
- Immunogenicity.

Clinical laboratory tests -including immunogenicity-, physical examinations and vital signs will be tabulated by time point and treatment group and reviewed for potential safety signals. All adverse events (AEs) and serious adverse events (SAEs) will be listed and tabulated by system organ class (SOC), preferred term (PT) and further by severity and relatedness to the study drug.

The **exploratory objectives** of the study are:

- To assess the pharmacokinetics (PK), pharmacodynamics (PD), and PK/PD associations of AMG 714.

The **exploratory endpoints** of the study are:

- PK;
- PD;
- Exposure/response (PK/PD).

PK data will be tabulated by timepoint. Exploratory PD endpoints are aberrant and abnormal IELs by flow cytometry, immunochemistry and T cell receptor (TcR) clonality analyses, Physician Global Assessment of Disease (PGA) and Patient Global Assessment of Disease (PtGA), Quality of Life Assessments (i.e. SF-12 v. 2 and EQ-5D), biomarkers of disease activity (i.e., serum IL-15, CD122 and granzyme B) and Celiac Disease Patient Reported Outcome (CeD PRO). The exploratory PD endpoints will be evaluated by investigating the change from baseline and weekly scores, if applicable, of AMG 714 dose arm and placebo arm.

3 Study type and design

CELIM-RCD-002 is designed to be a Phase 2a randomized, double-blind, placebo-controlled, parallel-group study to evaluate the efficacy and safety of AMG 714 for the treatment of adult patients with RDC-II, an in situ small bowel T cell lymphoma.

After signing informed consent, subjects will be screened for the study. All subjects who meet the study entry criteria will be randomized at a 2:1 ratio to receive either 8 mg/kg AMG 714 or placebo for a total of 7 times over 10 weeks, with evaluation of the primary endpoint at Visit 8 (Week 12/Day 84).

AMG 714 (N=16) or placebo (N=8) will be administered at the clinical site in a double-blind fashion via intravenous (IV) infusion of approximately 120 minutes (2 hours) duration.

Subjects will remain confined to the study site for a minimum of 1 hour after the administration of study medication. During this time the investigator and study site staff will assess the subject for adverse events (AEs). As per Appendix 1, PK samples are to be collected prior to the infusion and 1 hour after the infusion, at the end of this observation period. The beginning and end of each infusion, as well as the PK sample collection times, will be recorded.

In addition to receiving study medication (AMG 714 or placebo), concomitant therapy with steroids at a maximum dose of 20 mg of prednisone, prednisolone or equivalent per day and/or oral budesonide at a maximum dose of 9 mg per day will be accepted. Steroid doses must be stable for 4 weeks prior to randomization and remain stable for the duration of the study. Systemic steroids and topical budesonide have been shown to improve symptoms of RCD-II and are considered adequate background therapy on top of which to test experimental medications (Brar *et al*, 2007).

Should AMG 714 show adequate efficacy and safety, as determined elsewhere by the Sponsor, subjects in the study, including those in the placebo arm, may be offered participation in an open label extension study of AMG 714 in due course, but under no circumstances prior to study completion. In the interim, between the end of the study for an individual subject and the start of the possible open label extension, the Sponsor intends to provide a bridging program to allow objective study responders to have access to AMG 714 as determined by their site investigator or physician. The open label extension study and interim bridging program will be described in independent protocols.

Subjects will be expected to maintain total adherence to a gluten free diet (GFD) from 6 months before randomization through the final study visit (Visit 9; Week 16/Day 112). Subject’s adherence to the GFD will be assessed by an expert dietician and monitored via stool sample testing using the iVYLISA gluten immunogenic peptides (GIP) stool gluten test. Subjects with known or suspected GFD transgressions will be counselled and allowed to continue in the study.

A study site staff member will contact each subject by telephone one day after the first study drug administration to assess for AEs. Subjects will return to the clinic for the next administration of study drug after 1 week (Visit 2; Week 1/Day 7). After Visit 2, subjects will return to the clinic for follow-up and study drug administration at Visit 3 (Week 2/Day 14) and every two weeks thereafter as indicated in the study schedule of events (Appendix 1). The final dose of study drug will be administered at Visit 7 (Week 10/Day 70). An end-of-study efficacy visit will be conducted at Visit 8 (Week 12/Day 84). A final follow-up study visit will be conducted 6 weeks after the last dose of study drug at Visit 9 (Week 16/Day 112).

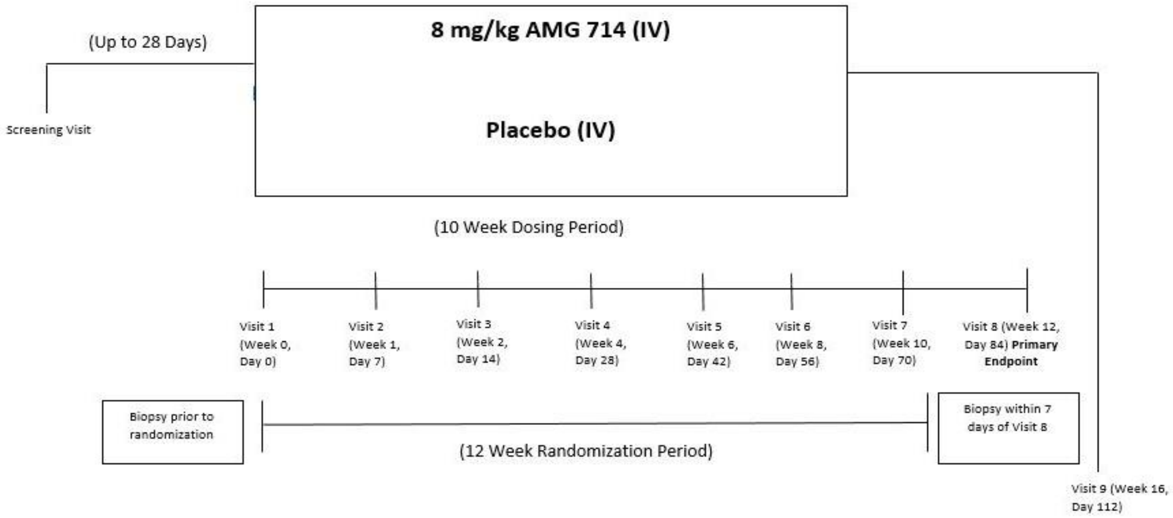
Subjects who meet all study entry criteria will undergo upper gastrointestinal endoscopy with biopsy collection prior to baseline (i.e., prior to Visit 1, Week 0/Day 0) and within 7 days of Visit 8 (Week 12/Day 84) in order to assess changes from baseline to end of treatment (defined as Week 12 for the purpose of efficacy testing) in aberrant and abnormal intraepithelial lymphocytes (IELs), villous height to crypt depth ratio (VH:CD ratio), T cell receptor (TcR) clonality and Marsh score.

Subjects enrolled in the study will complete the Bristol Stool Form Scale (BSFS) at the time of each bowel movement from baseline (Visit 1; Week 0/Day 0) up to the final study visit, Visit 9 (Week 16/Day 112). Subjects will complete the Celiac Disease Patient Reported Outcome (CeD PRO) daily from baseline up to the final study visit. Subjects will also complete the Gastrointestinal Symptom Rating Scale (GSRS) beginning at Visit 1 (Week 0/Day 0) and, thereafter, weekly from the time of randomization through the final study visit. The BSFS, GSRS and the daily CeD PRO will be completed using a handheld electronic diary. In addition, subjects will complete paper quality of life diaries (Short Form 12 questionnaire (SF-12) v. 2, European Quality of Life 5 Dimensions questionnaire (EQ-5D) and Patient Global Assessment of Disease (PtGA)). Diaries will be completed at the times specified in the Schedule of Events (Appendix 1).

Safety will be monitored on an ongoing basis and subjects may undergo unscheduled visits if needed for safety reasons. Safety will be assessed throughout the study by clinical laboratory tests, physical examination, vital signs and AE monitoring. Immunogenicity will also be monitored.

Figure 1 represents a schematic drawing of the study periods and visits.

Figure 1 CELIM-RCD-002 study schematic



4 Interim analysis

An interim analysis for safety and PK data will be conducted when the tenth randomized subject reaches Visit 4 (Week 4/Day 28; subjects not replaced if dropped out). The interim analysis, specified in more detail in the Data Safety Monitoring Board (DSMB) charter, will be unblinded and performed by an independent DSMB which will also monitor unblinded safety data throughout the study. In addition to safety and PK, other information available at the time of the analysis may be considered and a recommendation to stop, continue or modify the study will be made by the DSMB to the Sponsor. Should the exposure of AMG 714 be below the anticipated range, a root cause analysis will be done, which could result in a protocol amendment. The recommendation and decision will be shared with the investigational sites and ethics committees. No adjustments in significance levels will be applied to the final analyses as this interim assessment is DSMB-driven for the purpose of safety assessments and to assure that minimal steady-state exposures (10 µg/mL) are achieved. No efficacy data will be assessed during this interim analysis.

5 Randomisation

Randomisation, i.e. the random allocation of treatments to subject numbers, will be performed according to the design of the study. A detailed description of the randomisation method, including the size(s) of randomly permuted blocks used to balance the randomisation, will be stored at the restricted area of StatFinn server, where only the randomization expert and the unblinded statistician will have access to. Once the data base is locked and the treatment code open, the randomization documents will be moved to the non-restricted area.

Subjects will be randomized at a 2:1 allocation ratio to receive 8 mg/kg AMG 714 or placebo for a total of 7 administrations over 10 weeks. The randomization will not be stratified since this is the first-ever placebo controlled randomized clinical trial in RCD-II and there are no known confounding factors for response to treatment.

6 Statistical hypotheses

The primary endpoint of the study is the Immunological Response 1: relative (%) change at Week 12 in the % of aberrant IELs vs total IELs as assessed by flow-cytometry between the AMG 714 dose arm and the placebo arm.

The primary endpoint of Immunological Response 1 will be tested as follows:

$$H_O\colon \mu_{AMG\ 714} = \mu_{placebo}$$

against the alternative

$$H_1\colon \mu_{AMG\ 714} \neq \mu_{placebo}$$

where $\mu_{AMG\ 714}$ and $\mu_{placebo}$ denote the mean baseline to Week 12 relative (%) change in % aberrant IELs vs total IELs as assessed by flow-cytometry in the AMG 714 and placebo arm, respectively. The hypotheses will be tested using a two-sided, 0.10 level of significance.

7 Estimation of sample size

The CELIM-RCD-002 study is an exploratory proof-of-concept (POC) study of an experimental medication, AMG 714, in RCD-II patients. While it will be the first such study sponsored by a pharmaceutical company, the straightforward design is based on several academic studies (Goerres *et al*, 2003; Brar *et al*, 2007; Tack *et al*, 2011a, 2011b) as well as on the nonclinical proof-of-principle signal obtained with AMG 714 in intestinal explants from patients with RCD-II (Malamut *et al*, 2010).

The sample size is based on the size of previous academic studies. Published prospective academic studies have ranged from 13 to 18 subjects total (Tack *et al*, 2011a, 2011b) and CELIM-RCD-002 will be the largest prospective trial ever conducted in RCD-II, a very rare malignant disease.

The sample size of 24 subjects (16 subjects in AMG 714 arm and 8 subjects in placebo arm) has been calculated to achieve at least ~ 80% power to detect a 20 percentage point difference in the primary endpoint, the difference in the baseline-to-Week 12 reduction of % aberrant IELs vs total IELs between the AMG 714 arm and placebo arm.

This sample size calculation was based on the following assumptions:

- Two-sided type one error rate $\alpha = 0.1$.
- Power $1\text{-}\beta = 0.8$.
- Analysis method: one-way analysis of variance (ANOVA) by SAS® (proc power).
- 2:1 allocation ratio between the AMG 714 and placebo arms.
- Common SD = 17.4 for the baseline to Week 12 reduction in % aberrant IELs as assessed by flow-cytometry.

- Mean change of 20 percentage points and 0 percentage points in the baseline to Week 12 reduction in % aberrant IELs in the AMG 714 and placebo arms, respectively.

The standard deviation used for the sample size calculation is computed from the % aberrant IELs (obtained by flow-cytometry) data from 13 subjects treated with cladribine (subset of the subjects reported in Tack et al, 2011).

8 Statistical methods

8.1 Data sets to be analysed

The populations for analysis will be the intention to treat (ITT, safety population, at least one dose of the investigational product received) and the per protocol population (PP, efficacy, i.e., available and evaluable pre- and post-biopsy information [Week 6]).

PP population: the PP population will exclude non-evaluable subjects and subjects with major protocol deviations thought to impact the ability to assess the effect of treatment. The PP population will also exclude atypical RCD-II patients (patients with a different phenotype of the aberrant IELs) for the purpose of analysis of Immunological Response 1 and 2; however, they will be included in all other efficacy assessment where a sensitivity analysis maybe performed excluding these subjects if these subjects are determined to be influential. Exclusion of subjects from the PP set will be reviewed, documented and approved before the study is unblinded to the study Sponsor.

Non-evaluable subjects will include subjects missing one of the two biopsies (or with biopsies of insufficient quality to generate valid data as determined by the Sponsor prior to database lock) and subjects dropping out of the study before Week 6. If a discontinuation occurs on or after Week 6 and the second biopsy is collected, then the subject is considered evaluable.

Atypical patients may also be analysed as a separate population if the results and sample size suggest such analysis is valuable.

The following major protocol deviations, as determined by the Sponsor prior to database lock, will lead to exclusion of subject from the PP population and will be thoroughly documented in the clinical study report:

- Intake of any forbidden concomitant medication or participation in other medical procedures, in case there is reason to believe, by the judgement of the Sponsor or investigator, that such concomitant medication/procedures would have a significant effect on the efficacy data obtained;
- Significant deviations from the inclusion/exclusion criteria of the study not previously approved by the Sponsor and determined prior to database lock to have impact on interpretation of results.

ITT (safety) population: this population consists of all randomized subjects who have received at least one dose of the study drug. The safety population is by definition the same as the ITT population. If a subject had a misallocated treatment on a specific visit, observed upon unblinding, a secondary sensitivity analysis will be performed on as treated basis.

Main efficacy measures which require pre- and post-treatment biopsies will be analysed using the PP population (at least 2 biopsies, the second on Week 6 or later). Continuous efficacy variables will be analysed based on the ITT population. Demographic and baseline variables will be assessed using both ITT and PP populations. Safety parameters will be analysed using the ITT population. Compliance data will be reviewed to assure subjects were treated as randomized.

A detailed description of the study populations and subjects included will be given in the Subject Classification Document, which will be finalized before the database lock.

Table 1 Data sets to be used in the analysis

Primary variable[2]	Secondary variables[1,2]	Exploratory/PD variables[1]	Safety variables	Demographic and baseline variables
PP	ITT PP	ITT PP	ITT	ITT PP

[1]PP set used for biopsy-dependant variables and ITT for all other variables.
[2]PP population will exclude atypical RCD-II patients (patients with a different phenotype of the aberrant IELs) for the purpose of analysis of Immunological Response 1 and 2.

8.2 General statistical considerations

Descriptive statistics (e.g. mean, median, standard deviation (SD), minimum (MIN), maximum (MAX), and number of subjects with an observation (N) or missing observation (NMISS)) are used for summarizing continuous variables. Additional statistics will be provided for PK-related data, including the geometric mean, SD of log-transformed data, geometric CV% and geometric N (i.e. number of subjects with an observation that are included in the natural logarithmic transformation).

Frequencies and percentages are used for summarizing categorical variables.

All summary statistics will be presented by treatment arm and if repeated measures, then by visit/collection time point.

All listings will be created using ITT population. The PP and ITT flags for each subject will be included in all listings.

All tests will be two-sided, if not stated differently. P-values smaller than 0.1 (for primary hypothesis) or 0.05 (all other hypotheses) will be considered statistically significant. In addition to the inferential statistics, 90% or 95% confidence intervals will be constructed. For the primary endpoint a 90% confidence interval of the treatment difference will be constructed. Ninety-five percent (95%) confidence intervals will be constructed for the individual treatment point estimates.

8.2.1 Definition of derived variables

Geometric mean will be calculated as:

$$\mathrm{mean_{geo}} = \exp\{\tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ln x_i\}.$$

Values that are missing or not available will be ignored in calculation of geometric mean.

$CV\%_{geo}$ will be calculated using the following formula:

$$CV\%_{geo} = 100\% \cdot \sqrt{\exp V_{ln} - 1}.$$

Values that are missing or not available will be ignored in calculation of $CV\%_{geo}$.

Absolute change from baseline for a given treatment needs to be calculated, the following formula will be used:

Absolute change from baseline = Post-baseline value – Pre-dose value.

For weekly scores, Week 0 score is considered to be the baseline value. If Week 0 score is not available, the first non-missing result within 2 weeks of treatment will be used as baseline. For daily scores (e.g. CeD PRO), Day 0 score is considered to be the baseline; in case Day 0 score is not available, the first non-missing result within 2 weeks of treatment will be used as baseline. Otherwise, the given subject will be considered non-evaluable for the specific test. In case pre-dose value doesn't exist (i.e. for biopsy endpoints), screening value will be used as baseline.

In the case where relative (%) change from baseline needs to be calculated, the following formula will be used:

Relative (%) change from baseline = (Post-baseline value – Pre-dose value) / (Pre-dose value) · 100%.

For weekly scores, Week 0 score is considered to be the baseline value. If Week 0 score is not available, the first non-missing result within 2 weeks of treatment will be used as baseline. For daily scores (e.g. CeD PRO), Day 0 score is considered to be the baseline; in case Day 0 score is not available, the first non-missing result within 2 weeks of treatment will be used as baseline. Otherwise, the given subject will be considered non-evaluable for the specific test. In case pre-dose value doesn't exist (i.e. for biopsy endpoints), screening value will be used as baseline.

Body mass index (BMI) is calculated using the formula:

$$\mathrm{BMI}\ (\mathrm{kg/m^2}) = \mathrm{weight\ (kg)} / [\mathrm{height\ (m)} * \mathrm{height\ (m)}].$$

Body surface area (BSA) is calculated using Dubois' formula:

$$\mathrm{BSA}\ (\mathrm{m^2}) = 0.007184 * \mathrm{height\ (cm)}^{0.725} * \mathrm{weight\ (kg)}^{0.425}.$$

Table 2 Decimal places for summary statistics of continuous and categorical variables

Statistic	Number of digits
Minimum, maximum	Same as in original data
Mean, median, $mean_{geo}$	1 more than in original data
SD	2 more than in original data
Frequencies (%)	1 decimal place
$CV\%_{geo}$	1 decimal place

8.2.2 Missing values

If a discontinuation occurs on or after Week 6 and the second biopsy is collected then the subject is considered evaluable, the results of the second biopsy (pre-maturely done before Week 12) will be used to investigate the change from baseline to Week 12, i.e. the values of parameters obtained from this biopsy will be carried forward for Week 12 efficacy assessments (LOCF). This LOCF will only be done for PP population and only for variables depending on the biopsy.

AMG 714 concentrations below the lower limit of quantification (LLOQ) will be assigned a value of 0.5 x LLOQ in mean calculations for the summary of AMG 714 concentrations. A similar rule will be used for any other assay results below the LLOQ. All assay results over the upper limit of quantification (ULOQ) will be assigned a value of ULOQ.

8.2.3 Handling of data from discontinued subjects

Subjects who are randomized but discontinue before receiving study treatment will not be included in any efficacy or safety analysis, but will be included in the disposition of subjects table. Subjects who receive at least one dose of study treatment will be included in ITT analysis population.

Subjects discontinuing from study drug administration before Week 6 will be excluded from the PP analysis population and the second biopsy will not be collected. Subjects discontinuing on or after Week 6 will be included in the PP analysis population sets if the second biopsy can be collected (in case a subject is lost to follow-up on or after Week 6 and a biopsy cannot be collected, the subject is considered non-evaluable and excluded from the PP population).

Subjects will be considered study completers at Visit 8 (Week 12/Day 84), regardless of whether or not the Final Study Visit (Visit 9; Week 16/Day 112) is attended.

8.3 Disposition of subjects

The number of subjects screened, randomized, completed, or discontinued from the study and the reason for study discontinuation will be tabulated by treatment group and site as appropriate. Subject count by analysis population will also be tabulated. Major protocol deviations will be summarized by treatment group.

Disposition of subjects, informed consent signing information, and inclusion/exclusion criteria will also be listed by subject.

8.4 Demographic and baseline characteristics

Demographic and baseline characteristics collected and presented for this study, include: age, sex, race, ethnicity, weight, height, BMI, BSA, urine drug and alcohol screen, 12-lead ECG, medical history, and primary diagnosis (including celiac serology history). Compliance to pre-screening fasting will also be investigated.

Demographic and baseline characteristics are assessed at screening. Weight is additionally measured throughout the study at time points specified in Appendix 1 and BMI is calculated based on these weight measurements and screening heights.

Demographic and baseline characteristic will be presented by summary statistics and tabulated by treatment group, as well as listed. Medical history will be additionally broken down by system organ class (SOC) and preferred term (PT). 12-lead ECG, primary diagnosis, urine drug and alcohol screen and fasting compliance will only be listed.

Both ITT and PP populations will be used for the analysis of demographic and baseline characteristics.

8.5 Extent of exposure and compliance

Extent of exposure will be summarized showing:

- Number of subjects exposed to placebo or AMG 714 at each visit.

Extent of exposure will be tabulated by treatment group and visit using the ITT analysis population. Total exposure and number of doses will also be listed cumulatively for all subjects. Start and end dates and times of IV infusions along with duration of infusion will only be listed.

8.6 Analysis of efficacy

8.6.1 Primary efficacy variable

Primary efficacy endpoint of the study is:

- Immunological Response 1: Relative (%) change from baseline to Week 12 in the % of aberrant intestinal intraepithelial lymphocytes (IELs) vs total IELs as assessed by flow-cytometry.

The primary endpoint will be analysed using analysis of covariance (ANCOVA), where the baseline % aberrant IELs vs total IELs will be included as a covariate and treatment group as a fixed effect in the statistical model.

The analysis of the primary endpoint will be carried out using the PP population. Subjects having Week 6 biopsy data who are lost to follow-up for further visits will have their Week 6 value carried forward for the final analysis (more details available in *Chapter 8.2.2 Missing values*).

The following model will be fitted:

$$Y_{ij} = \mu + \beta \cdot baseline\ \%\ IEL\ vs\ total\ IEL_{ij} + \alpha_j + \varepsilon_{ij},$$

where

Y_{ij} is the relative (%) change from baseline in the % of aberrant IELs vs total IELs for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2$),
μ is the overall mean,
β is the parameter estimate of baseline % IELs vs total IELs,

α_j is the fixed effect due to treatment j,
ε_{ij} is the random error for subject i from treatment group j; $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

The following SAS code will be used to fit the above-specified model (NOTE: SAS codes provided in this document may be modified based on statistical considerations, without requiring SAP amendment):

```
proc glm data=IELt alpha=0.1 outstat=F_tests;
   class trt;
   model IELt_change = bl_IELt trt / clparm solution SS1 SS2;
   estimate "AMG 714 vs Placebo" trt 1 -1;
   lsmeans trt /cl stderr alpha=0.05;
   ods output ParameterEstimates=Par_est LSMeanCL=LS_meanCL
              LSMeans=LS_mean OverallANOVA=ANOVA Estimates=Eff_est;
run;
```

The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal (i.e. least squares) means estimates and results of the check of assumptions) will be tabulated and 90% confidence intervals added, where appropriate.

As a secondary assessment, the same model will be fitted using the absolute change in % of aberrant IELs vs total IELs as a response variable. The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal (i.e. least squares) means estimates and results of the check of assumptions) will be tabulated and 90% confidence intervals added, where appropriate.

A frequency table by treatment group will be used to summarize the percentage of patients who reach normalization of aberrant IEL counts by flow cytometry measured as aberrant IELs < 20% of total IEL based on the second biopsy after Week 6. Additionally, a frequency table of subjects with at least 20% reduction from baseline in the % aberrant IELs vs total IELs by treatment arm will also be created, if applicable.

The baseline and post-baseline values along with the change from baseline values (both relative and absolute) will also be described by summary statistics and tabulated by treatment group. Baseline and post baseline values will also be listed.

8.6.2 Secondary efficacy variables

The secondary efficacy endpoints of the study are:

- **Immunological Response 2**: Relative (%) change from baseline in the % of aberrant IELs vs intestinal epithelial cells (IECs)
- **Histological Response:** Relative (%) change from baseline in small intestinal villous height to crypt depth (VH:CD) ratio, Marsh score and total IEL counts
- **Clinical response:** Change from baseline in clinical symptoms
    - Bristol Stool Form Scale (BSFS)
    - Gastrointestinal Symptom Rating Scale (GSRS), including the celiac disease GSRS (CeD-GSRS)

Immunological Response 2: Relative (%) change from baseline in the % of aberrant IELs vs intestinal epithelial cells (IECs)

Immunological response 2 will be calculated as

$$(\%\ \textit{of aberrant IELs by flow-cytometry}) \cdot (\%\ \textit{total IEL vs IEC by immunochemistry}).$$

Immunological response 2 will be analysed using the same method as for the primary endpoint, i.e. the following model will be fitted:

$$Y_{ij} = \mu + \beta \cdot baseline\ \%\ IEL\ vs\ IEC_{ij} + \alpha_j + \varepsilon_{ij},$$

where

Y_{ij} is the relative (%) change from baseline in the % of aberrant IELs vs intestinal epithelial cells for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2$),
μ is the overall mean,
β is the parameter estimate of baseline % IELs vs IEC by flow-cytometry,
α_j is the fixed effect due to treatment j,
ε_{ij} is the random error for subject i from treatment group j; $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

The following SAS code will be used to fit the above-specified model:

```
proc glm data=IEL alpha=0.05 outstat=F_tests;
   class trt;
   model IEL_change = bl_IEL trt / clparm solution SS1 SS2;
   estimate "AMG 714 vs Placebo" trt 1 -1;
   lsmeans trt /cl stderr;
   ods output ParameterEstimates=Par_est LSMeanCL=LS_meanCL
              LSMeans=LS_mean OverallANOVA=ANOVA Estimates=Eff_est;
run;
```

The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal (i.e. least squares) means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate.

As a secondary assessment, the same model will be fitted using the absolute change in % of aberrant IELs vs intestinal epithelial cells as a response variable. The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal (i.e. least squares) means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate.

The baseline and post-baseline values along with the change from baseline values (both relative and absolute) will also be described by summary statistics and tabulated by treatment group. Baseline and post baseline values will also be listed.

A frequency table by treatment group will be used to summarize the percentage of patients with at least 20% reduction from baseline in the % of aberrant IELs vs intestinal epithelial cells based on the second biopsy after Week 6.

PP population will be used for the analysis.

<u>Histological Response: Relative (%) change from baseline in small intestinal villous height to crypt depth (VH:CD) ratio</u>

Percent change from baseline in small intestinal villous height to crypt depth (VH:CD) ratio will be analysed using the same method as for the primary endpoint, i.e. the following model will be fitted:

$$Y_{ij} = \mu + \beta \cdot baseline\; VH{:}\,CD_{ij} + \alpha_j + \varepsilon_{ij},$$

where

Y_{ij} is the relative (%) reduction from baseline in the VH:CD ratio for subject i ($i = 1, \ldots, n_j$) from treatment group j ($j = 1, 2$),
μ is the overall mean,
β is the parameter estimate of baseline VH:CD ratio,
α_j is the fixed effect due to treatment j,
ε_{ij} is the random error for subject i from treatment group j; $\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon}^{2})$.

The following SAS code will be used to fit the above-specified model:

```
proc glm data=IEL alpha=0.05 outstat=F_tests;
   class trt;
   model VHCD_change = bl_VHCD trt / clparm solution SS1 SS2;
   estimate "AMG 714 vs Placebo" trt 1 -1;
   lsmeans trt /cl stderr;
   ods output ParameterEstimates=Par_est LSMeanCL=LS_meanCL
              LSMeans=LS_mean OverallANOVA=ANOVA Estimates=Eff_est;
run;
```

The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate.

As a secondary assessment, the same model will be fitted using the absolute change in VH:CD as a response variable. The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal (i.e. least squares) means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate.

The baseline and post-baseline values along with the change from baseline values (both relative and absolute) will also be described by summary statistics and tabulated by treatment group. Baseline and post baseline values will also be listed.

The analysis will also include a frequency table by treatment group to summarize the percentage of patients with at least 30% improvement from baseline in VH:CD based on the second biopsy after Week 6.

PP population will be used for the analysis.

Histological Response: Change from baseline in Marsh score

Marsh-Oberhuber classification (Marsh, 1992; Oberhuber, 2000) i.e. the Marsh score, a commonly used histological score with possible values 0, 1, 2, 3a, 3b, 3c with 0 being the best and 3c the worst, will be assessed at screening and Week 12 biopsies.

The Marsh scores will be analysed using a simple logistic regression model, where improvement from baseline in the Marsh score is used as dependent variable and treatment group will be included in the model as explanatory variable. For modelling purposes, a binary variable with values 1 (in case improvement in Marsh scores, i.e. any decrease in score from baseline to Week 12, was observed) and 0 (no improvement in Marsh scores) will be used as a response. Logit function will be used as the link function. The following model will be fitted:

$$\log\left(\frac{p_j}{1-p_j}\right) = \mu + \alpha_j,$$

where

$Y_j \sim Bin(p_j, n_j)$ – is the number of subjects with improvement in Marsh scores by Week 12 in treatment group j $(j = 1,2)$,
p_j is the probability of improvement in Marsh score for subjects in treatment group j by Week 12,
$\mu$is the overall mean (on logit scale),
α_j is the fixed effect due to treatment j on the logit scale.

The odds ratio for AMG 714 vs placebo groups will be reported and the following formula will be used for obtaining the ratio:

$$OR = \exp(\alpha_{AMG\ 714} - \alpha_{Placebo}),$$

where $\alpha_{AMG\ 714}$ denotes the treatment effect of AMG 714 on logit scale and $\alpha_{Placebo}$ denotes the placebo effect on logit scale.

The following SAS code will be used to fit the above-specified model:

```
proc genmod data=marsh descending;
      class trt;
      model marsh_imp = trt / dist=bin link=logit lrci type3;
      estimate 'OR AMG vs PLA ' trt 1 -1 / exp ;
      lsmeans trt /cl ilink;
      ods output ParameterEstimates=Par_est Estimates=Eff_est
                 Type3=Type3
                 LSMeans=Ls_mean;
run;
```

The modelling results (parameter estimates, estimated odds ratio of AMG 714 vs placebo, marginal (i.e. least squares) means estimates for treatment groups) will be tabulated and 95% confidence intervals added, where appropriate. Odds ratio and marginal means estimates for the probabilities will be presented on the original (probability) scale.

The baseline and post-baseline frequencies of all categories of the Marsh score scale, along with change from baseline frequencies will be tabulated by treatment group. Baseline and post baseline values will also be listed.

The analysis will also include the percentage of subjects with complete remission of the histological abnormalities according to the Marsh score, i.e., the % of subjects with Marsh scores 0-1. Frequency table by treatment group will be used for summarizing these results.

PP population will be used for the analysis.

Histological Response: Relative (%) change from baseline in total IEL counts

Change from baseline in total IEL counts (i.e., density of IELs by IHC) will be analysed using the same method as for the primary endpoint, i.e. the following model will be fitted:

$$Y_{ij} = \mu + \beta \cdot baseline\ IEL_{ij} + \alpha_j + \varepsilon_{ij},$$

where

Y_{ij} is the relative (%) reduction from baseline in the total IEL counts for subject i ($i = 1, \ldots, n_j$) from treatment group j ($j = 1, 2$),
μ is the overall mean,
β is the parameter estimate of baseline total IEL count,
α_j is the fixed effect due to treatment j,
ε_{ij} is the random error for subject i from treatment group j; $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$.

The following SAS code will be used to fit the above-specified model:

```
proc glm data=IELc alpha=0.05 outstat=F_tests;
   class trt;
   model IELc_change = bl_IELc trt / clparm solution SS1 SS2;
   estimate "AMG 714 vs Placebo" trt 1 -1;
   lsmeans trt /cl stderr;
   ods output ParameterEstimates=Par_est LSMeanCL=LS_meanCL
              LSMeans=LS_mean OverallANOVA=ANOVA Estimates=Eff_est;
run;
```

The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate.

As a secondary assessment, the same model will be fitted using the absolute change in total IEL counts as a response variable. The modelling results (ANOVA table, parameter estimates, estimated treatment effect, marginal (i.e. least squares) means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate.

The baseline and post-baseline values along with the change from baseline values both relative and absolute) will also be described by summary statistics and tabulated by treatment group. Baseline and post baseline values will also be listed.

PP population will be used for the analysis.

Change from baseline in clinical symptoms: Bristol Stool Form Scale (BSFS)

The Bristol Stool Form Scale is a pictorial aid to help subjects identify the shape and consistency of their bowel movements during the study (Riegler *et al* 2001).

Subjects will be asked to complete this form daily using an electronic diary at the time of each bowel movement from randomization through the Final Study Visit (Visit 9; Week 16/Day 112). If no bowel movements were experienced by the subject on any given day, the subject should document this using the electronic diary.

BSFS will be described by calculating daily and weekly number and type of bowel movements.

The total weekly bowel movement counts will be analysed using generalized linear mixed models with subject as a random effect. The statistical model will include as fixed effects treatment group, time (week) and their interaction. Poisson distribution with log-link will be used for modelling the counts. The following model will be fitted:

$$\log(Y_{ijk}) = \mu + \alpha_j + \gamma_k + \lambda_{jk} + \eta_i + \varepsilon_{ijk},$$

where

Y_{ijk} is the bowel movement count for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2$) at week k ($k = 0,1, \dots,16$),
μ is the overall mean on log-scale,
α_j is the fixed effect due to treatment j on log-scale,
γ_k is the fixed effect due to week k on log-scale,
λ_{jk} is the fixed interaction effect due to treatment j and week k on log-scale,

η_i is the random effect due to subject i,
ε_{ijk} is the random error for subject i from treatment group j for week k.

Due to the repeated structure of the data, the measurements within a subject will be correlated. Therefore, it is assumed that the overall covariance matrix of the response is block-diagonal. In order to determine the covariance structure that best fits the data, the same model with different covariance structures will be fitted. Initially, models will be fit assuming unstructured (UN) variance-covariance structure. Common within and between treatment variance components (compound symmetry (CS), first-order autoregressive (AR(1)) and Toeplitz covariance structures) will be further explored to increase sensitivity of statistical tests. The best model will be chosen by comparing the Akaike Information Criterion (AIC) of the models with different covariance structures.

The following SAS code will be used to fit the above-specified model:

```
proc glimmix data=bsfs ic=q;
  class subjid trt week (ref=FIRST);
  model bsfs = trt week trt*week / solution dist=POISSON;
  random _residual_ / subject=subjid type=UN;
  estimate "AMG 714 vs Placebo" trt 1 -1 /alpha=0.05 cl ilink;
  estimate "Week 0: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1
            /alpha=0.05 cl ilink;
  estimate "Week 1: AMG 714 vs Placebo" trt 1 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
            /alpha=0.05 cl ilink;
  ...
  estimate "Week 16: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0
            /alpha=0.05 cl ilink;
  lsmeans trt /alpha=0.05 cl ilink;
  lsmeans trt*week /alpha=0.05 cl ilink;
  ods output CovParms=Cov_Par LSMeans=LS_est Estimates=Eff_est
             FitStatistics=Fit_Stat ParameterEstimates=Par_est
             Tests3=Type_3;
run;
```

Note that the value of `type` will be changed according to which covariance structure is being fitted.

The modelling results (parameter estimates, estimated ratio of the bowel movement rates in AMG 714 and placebo groups, estimated weekly ratios of the bowel movement rates in AMG 714 and placebo groups, marginal (i.e. least squares) means estimates and results of the check of assumptions, i.e. the over dispersion) will be tabulated and 95% confidence intervals added, where appropriate.

Weekly bowel movements and their BSFS types will be tabulated by treatment group and week and summarized by descriptive statistics, absolute numbers and percentages. Also, percentage of subjects with diarrhoea (at least 1 BSFS >= 6 for the week) and percentage of subjects with constipation (at least 3 days without bowel movement for the week) by week and by treatment group will be presented. Mean weekly bowel movement counts will also be presented graphically by time point and treatment group.

As an exploratory assessment, averages by treatment for proportion of subjects showing BSSF >= 6 by time will be plotted. The differences in treatment AUCs will be explored using one-way analysis of variance. These analyses will be conducted internally by Celimmune.

The daily bowel movement counts and their BSFS scores will only be listed.

ITT population will be used for the analysis as long as there is a baseline and post-treatment measure (i.e. not lost to follow-up).

<u>Change from baseline in clinical symptoms: Gastrointestinal Symptom Rating Scale (GSRS)</u>

The GSRS is a 15-question 7-scale questionnaire used to assess five dimensions of gastrointestinal syndromes: diarrhea, indigestion, constipation, abdominal pain and reflux (Svedlund *et al* 1988). While not specific for celiac disease, the GSRS is widely used in gastroenterology and has been used in several clinical trials of experimental medications in celiac disease, thus becoming a very useful tool with abundant existing reference data (Kelly *et al* 2013; Lähdeaho *et al*, 2011; Leffler *et al*, 2015).

Subjects will be asked to complete this questionnaire weekly, using an electronic diary, from the day of randomization through the Final Study Visit (i.e. on Visits 1 to 9).

The total GSRS score will be calculated as the mean of the scores of all 15 questions, with the scores for the individual questions between 1 (No discomfort at all) and 7 (Very severe discomfort). Therefore, the smaller the total GSRS score, the milder the symptoms of the subject.

The change from baseline in total GSRS score will be analysed using a linear mixed effects repeated measures model (MMRM) with the baseline value, treatment group, time point and a time point-by-treatment group interaction term as fixed effects with an underlying correlation structure between the time points that results in the best fit for the model. Subject will be included as a random effect. The following model will be fitted:

$$Y_{ijk} = \mu + \alpha_j + \gamma_k + \lambda_{jk} + \beta \cdot baseline\ GSRS_{ij} + \eta_i + \varepsilon_{ijk},$$

where

Y_{ijk} is the absolute change from baseline in total GSRS score for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2$) at week k ($k = 1, 2, \dots, 16$),

μ is the overall mean,

α_j is the fixed effect due to treatment *j*,
γ_k is the fixed effect due to week *k*,
λ_{jk} is the fixed interaction effect due to treatment *j* and week *k*,
β is the parameter estimate of the baseline total GSRS score,
η_i is the random effect due to subject *i*,
ε_{ijk} is the random error for subject *i* from treatment group *j* for week *k*.

Due to the repeated structure of the data, the measurements within a subject will be correlated. Therefore, it is assumed that the overall covariance matrix of the response is block-diagonal. In order to determine the covariance structure that best fits the data, the same model with different covariance structures is fitted. Initially, models will be fit assuming unstructured (UN) variance-covariance structure. Common within and between treatment variance components (compound symmetry (CS), first-order autoregressive (AR(1)) and Toeplitz covariance structures) will be further explored to increase sensitivity of statistical tests. The best model will be chosen by comparing the Akaike Information Criterion (AIC) of the models with different covariance structures.

The following SAS program will be used for carrying out the analysis:

```
proc glimmix data=gsrs;
  class subjid trt week (ref=FIRST);
  model gsrs = gsrs_bl trt week trt*week / solution;
  random _residual_ / subject=subjid type=UN;
  estimate "AMG 714 vs Placebo" trt 1 -1 /alpha=0.05 cl;
  estimate "Week 1: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 /alpha=0.05 cl;
  estimate "Week 2: AMG 714 vs Placebo" trt 1 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 /alpha=0.05 cl;
  ...
  estimate "Week 16: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 /alpha=0.05 cl;
  lsmeans trt /alpha=0.05 cl;
  lsmeans trt*week /alpha=0.05 cl;
  ods output CovParms=Cov_Par LSMeans=LS_est Estimates=Eff_est
             FitStatistics=Fit_Stat ParameterEstimates=Par_est
             Tests3=Type_3;
run;
```

Note that the value of `type` will be changed according to which covariance structure is being fitted.

The modelling results (parameter estimates, covariance parameter estimates, estimated treatment effect, marginal means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate. If the parametric assumptions are not met, then in addition to the above-specified model, a generalized estimating equation (GEE) approach will be used as well.

The baseline and post-baseline total GSRS scores along with the change from baseline values will also be described by summary statistics and tabulated by treatment group. In addition to that, the baseline and post-baseline scores of all 15 questions and for the mean total scores of five dimensions of gastrointestinal syndromes (diarrhoea (questions 11, 12 and 14), indigestion (questions 6-9), constipation (questions 10, 13 and 15), abdominal pain (questions 1, 4 and 5) and reflux (questions 2 and 3)) along with the change from baseline values will be described by summary statistics and tabulated by treatment group. Both the scores of individual questions and the total GSRS scores will also be listed. Mean total GSRS scores as well as the scores of the individual questions, will also be presented graphically by time point and treatment group.

ITT population will be used for the analysis as long as there is a baseline and post-treatment measure (i.e., not lost to follow-up).

Change from baseline in clinical symptoms: the celiac disease GSRS (CeD-GSRS)

The CeD-GSRS is formed by the subset of questions from GSRS questionnaire (questions 1, 4-9, 11, 12 and 14).

The total CeD-GSRS score is calculated as the mean of scores of all 10 questions in CeD-GSRS questionnaire, with the scores of the questions between 1 (No discomfort at all) and 7 (Very severe discomfort). Therefore, the smaller the total CeD-GSRS score, the milder the symptoms of the subject.

The change from baseline in CeD-GSRS will be analysed using a linear mixed effects repeated measures model (MMRM) with the baseline value, treatment group, time point and a time point-by-treatment group interaction term as fixed effects with an underlying correlation structure between the time points that results in the best fit for the model. Subject will be included as a random effect. The following model will be fitted:

$$Y_{ijk} = \mu + \alpha_j + \gamma_k + \lambda_{jk} + \beta \cdot baseline\ CeD\ GSRS_{ij} + \eta_i + \varepsilon_{ijk},$$

where

Y_{ijk} is the absolute change from baseline in total CeD-GSRS score for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2$) at week k ($k = 1, 2, \dots, 9$),
μ is the overall mean,
α_j is the fixed effect due to treatment j,
γ_k is the fixed effect due to time point k,
λ_{jk} is the fixed interaction effect due to treatment j and time point k,
β is the parameter estimate of the baseline total CeD-GSRS score,
η_i is the random effect due to subject i,
ε_{ijk} is the random error for subject i from treatment group j for time point k.

Due to the repeated structure of the data, the measurements within a subject will be correlated. Therefore, it is assumed that the overall covariance matrix of the response is block-diagonal. In order to determine the covariance structure that best fits the data, the same model with different covariance structures is fitted. Initially, models will be fit assuming unstructured (UN) variance-covariance structure. Common within and between treatment variance components (compound symmetry (CS), first-order autoregressive (AR(1)) and Toeplitz covariance structures) will be further explored to increase sensitivity of statistical tests. The best model will be chosen by comparing the Akaike Information Criterion (AIC) of the models with different covariance structures.

The following SAS program will be used for carrying out the analysis:

```
proc glimmix data=ced_gsrs;
  class subjid trt week (ref=FIRST);
  model ced_gsrs = ced_gsrs_bl trt week trt*week / solution;
  random _residual_ / subject=subjid type=UN;
  estimate "AMG 714 vs Placebo" trt 1 -1 /alpha=0.05 cl;
  estimate "Week 1: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 /alpha=0.05 cl;
  estimate "Week 2: AMG 714 vs Placebo" trt 1 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 /alpha=0.05 cl;
  ...
  estimate "Week 16: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 /alpha=0.05 cl;
  lsmeans trt /alpha=0.05 cl;
  lsmeans trt*week /alpha=0.05 cl;
ods output CovParms=Cov_Par LSMeans=LS_est Estimates=Eff_est
             FitStatistics=Fit_Stat ParameterEstimates=Par_est
             Tests3=Type_3;
run;
```

Note that the value of type will be changed according to which covariance structure is being fitted.

The modelling results (parameter estimates, covariance parameter estimates, estimated treatment effect, marginal means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate. If the parametric assumptions are not met, then in addition to the above-specified model, a generalized estimating equation (GEE) approach will be used as well.

The baseline and post-baseline total weekly CeD-GSRS scores along with the change from baseline values will also be described by summary statistics and tabulated by treatment group. The total weekly CeD-GSRS scores will also be listed. Mean total weekly CeD-GSRS will also be plotted by time point and treatment group.

ITT population will be used for the analysis.

8.7 Exploratory analysis

The exploratory endpoints of the study are:

- Reduction in aberrant and abnormal IELs by flow cytometry, immunohistochemistry and TcR clonality analyses (including % of patients with normalization in IELs);
- PGA and PtGA;
- Quality of Life Assessments:
    - SF-12 v. 2;
    - EQ-5D;
- Biomarkers of disease activity;
- PK, and PK/PD correlations;
- CeD PRO;
- Assessment of the gluten-free diet (iVYLISA GIP stool test and dietician consultation).

8.7.1 Reduction in aberrant and abnormal IELs by flow cytometry. immunohistochemistry and T cell receptor (TcR) clonality analyses

Aberrant IELs were narrowly defined in Section 2 for the purpose of primary and secondary endpoint analysis. In addition to these aberrant IELs, found by definition in the intestinal epithelium, cells with identical characteristics can be measured in the *lamina propria* of the mucosa ("aberrant IELs in the *lamina propria*") and -should the aberrant IELs leave the gut mucosa to spread to other tissues- in peripheral blood also ("aberrant IELs in blood"). These cells will be assessed.

In addition to the frequency of aberrant IELs, multiple markers will be studied in IELs which may identify other abnormal IELs, defined as IELs with an inflammatory (e.g., Granzyme B-positive) or otherwise abnormal phenotype or a phenotypically normal population with abnormal prevalence in blood or biopsy tissue. These phenotypes and abnormalities may also be assessed in aberrant IELs.

IEL abnormalities may be detected by several methods, such as flow cytometry, immunohistochemistry, or with TcR clonality analyses (a method to detect abnormal and also aberrant IELs by quantifying the characteristic T cell Receptor, which is monoclonal in aberrant IELs).

While aberrant IELs are pre-defined in primary and secondary endpoints based on surface and intracellular CD3 expression by flow cytometry (and CD8 expression by IHC), at this time it is not possible to predict which other populations of IELs may be abnormal and may require analysis, which will be done ad-hoc in an exploratory fashion.

Percentages and counts of aberrant and abnormal IELs by flow cytometry, immunochemistry or TcR clonality, and changes from baseline will be presented using summary statistics by treatment group and visit. Specifically, the % of Granzyme B-positive aIELs and IELs, as well as the mean fluorescence intensity of the Granzyme B expression in those and other cell types, will be summarized and analysed by treatment group.

PP population will be used for summarizing these results.

These analyses will be done internally at Celimmune.

8.7.2 Physician Global Assessment of Disease (PGA) and Patient Global Assessment of Disease (PtGA)

The PGA is designed to be used by the Investigator or qualified designee to assess the subjects' disease activity at the time points specified in the study schedule of events (Appendix 1). An attempt should be made to use the same assessor at each specified time point. Part B of the PGA is designed to assess physician perception of change in disease activity. Part B should be completed as per the schedule of events (Appendix 1) beginning after Visit 1, Week 0/Day 0.

The PtGA is designed to be used by the subject to assess perception of disease activity at the time points listed in Appendix 1. Part B of the PtGA, is designed to assess subject perception of change in disease activity and should be completed as per the schedule of events (Appendix 1) beginning after Visit 1, Week 0/Day 0.

Both PGA and PtGA are filled in on paper.

PGA and PtGA will be tabulated by treatment group, visit and part (PGA/PtGA part A, PGA/PtGA part B). Change from baseline in PGA and PtGA part A will be presented by treatment group and visit. All PGA and PtGA results will also be listed. Mean PGA and PtGA results will also be plotted by visit and treatment group.

As an exploratory assessment, Week 12 PGA scores will be dichotomized for an outcome endpoint where scores ≤ 2 will be considered treatment success and scores > 2 will be considered treatment failure. Assessment of differences in treatment proportions will be conducted.

ITT population will be used for summarizing PGA and PtGA results.

8.7.3 Quality of Life Assessments

The SF-12 v. 2 Health Survey is a shorter version of the SF-36 v. 2 Health Survey which uses just 12 questions to measure functional health and well-being from the patient's point of view. The SF-12 v. 2 covers the same eight health domains as the SF-36 v. 2 with one or two questions per domain (Ware, *et al*. 2009). The eight health domains covered by SF-12 v. 2 are: general health perceptions (GH), physical functioning (PF), role limitations due to physical health (RP), role limitations due to emotional problems (RE), bodily pain (BP), mental health (MH), vitality (VT), and social functioning (SF). Physical component summary (PCS) score and mental component summary (MCS) score are also calculated based on the individual answers.

The EQ-5D is a simple, brief and standardized instrument for use as a measure of health outcome (http://www.euroqol.org/about-eq-5d.html). The 5L version will be used in this study.

Subject are asked to fill in SF-12 v. 2 and EQ-5D quality of life surveys on paper on the following visits: Visit 1 (Week 0), Visit 4 (Week 4), Visit 8 (Week 12) and Visit 9, (Week 16).

For the calculation of SF-12 v. 2 health domain scores and summary scores, QualityMetric Health Outcomes™ Scoring Software 4.5.1 is used.
The baseline and post-baseline SF-12 v. 2 health domain and summary scores along with the change from baseline values will be described by summary statistics and presented by treatment group and visit. Individual health domain scores, summary scores and the scores of the individual questions in the survey will also be listed by visit and treatment group. Individual and mean health domain and summary scores will also be plotted by visit and treatment group.

Change from baseline in EQ-5D will be summarized by frequencies and percentages and tabulated by treatment group and question. The health assessments on visual analogue scale (VAS) with 0 being the worst imaginable health state and 100 being the best imaginable health state, will be described by summary statistics, the change from baseline values will also be summarized by treatment group and visit. The answers to all questions in the survey will be listed by visit. Mean EQ-5D assessments on VAS scale will also be plotted by visit and treatment group.

ITT population will be used for summarizing SF-12 v. 2 and EQ-5D results.

8.7.4 Biomarkers of disease activity

Several biomarkers of disease activity may be analysed in serum and in biopsy tissue at the time points specified in Appendix 1, if available and deemed appropriate by the Sponsor. The biomarkers of interest include:

- Serum IL-15 (pg/mL);
- Serum Granzyme B (pg/mL);
- Serum CRP (mg/L);
- Serum albumin (g/dL);
- Flow NKG2D (% of positive cells; mean fluorescence intensity, MFI);
- Flow GzmB (% of positive cells; MFI);
- Flow CD122 (% of positive cells; MFI).

In addition, other exploratory biomarkers may be analysed, if deemed appropriate.

The biomarkers of disease activity will be described by summary statistics, along with the change from baseline values and tabulated by treatment group and visit, if applicable. All individual results will be listed by visit. Mean curves of biomarkers of disease activity will also be plotted by visit and treatment group, if applicable. Individual curves of relevant biomarkers may also be created, if deemed appropriate.

Flow cytometry biomarkers will be analysed internally by Celimmune.

ITT and PP analysis populations will be used for summarizing biomarker results depending on the type of marker (biopsy-derived, PP; continuous variable, ITT).

8.7.5 Pharmacokinetics (PK)and Exposure/Response (PK/PD)

Data for all subjects that receive at least one dose of AMG 714 and provide at least one quantifiable concentration value will be used in the PK analysis. Effects of major covariates (e.g., weight or BMI, BSA, sex, serum albumin and disease characteristics at baseline) on AMG 714 disposition will also be evaluated via modelling. Individual exposures at steady-state (AUCss and Ctrough,ss) will be predicted from the developed population PK model and used for PK/PD assessments. These results will be provided in an independent PK Report.

In addition, AMG 714 concentrations in serum will be summarized. Summary statistics for concentrations will be calculated and the results tabulated by time point for AMG 714 treatment group. Levels of AMG 714 in serum will also be listed and mean results plotted by time point.

Finally, exposure/response (PK/PD) relationships will be investigated graphically.

For PK/PD assessments, individual patients' exposure measures obtained from the PK analysis will be graphically assessed with select PD endpoints and if associations are observed will be further elucidated with modelling and/or summaries by quartiles of exposure.

The following PD variables will be explored:

- Primary endpoint (Immunological Response 1: aberrant IELs by flow cytometry);
- Select secondary endpoints (Immunological Response 2: aberrant IELs by immunochemistry; Histological Response: VH:CD, total IEL count);
- Select PROs (CeD-PRO);
- Select biomarkers of disease activity (Serum IL-15).

Change from baseline to Week 12 for these variables will be plotted against individual predictions of steady-state exposure. For subjects with missing values at Week 12, earlier post-dose values will be used (if available), but those data points will be marked on the plots.

8.7.6 Celiac Disease Patient Reported Outcome (CeD PRO)

The CeD PRO questionnaire was developed to assess symptom severity in clinical trials in subjects with celiac disease. It is not validated in RCD-II. Items in the questionnaire were formulated based on one-on-one interviews with patients with celiac disease, thus they reflect the symptoms that patients consider part of their celiac disease experience. The questionnaire is designed as a self-administered daily diary, to be completed at the same time each day, and requires less than 10 minutes to complete. It includes nine items asking participants about the severity of celiac disease symptoms they may experience each day.

Participants are asked to rate their symptom severity on an 11-point, 0 to 10 scale; from “not experiencing the symptom” to “the worst possible symptom experience”. Symptoms include abdominal cramping, abdominal pain, bloating, gas, diarrhoea, loose stool, nausea, headache and tiredness.

Subjects will be asked to maintain a daily e-diary for the CeD PRO instrument from baseline to final study visit.

The total CeD PRO score is calculated as the sum of scores of all nine questions in CeD PRO questionnaire. The smaller the total CeD PRO score, the milder the symptoms of the subject. For modelling purposes, the total CeD PRO score on weekly level is also calculated (the weekly total score is calculated as the mean of total daily scores of a given subject).

The CeD PRO will be analysed using a linear mixed effects repeated measures model (MMRM) with the baseline value, treatment group, time point and a time point-by-treatment group interaction term as fixed effects with an underlying correlation structure between the time points that results in the best fit for the model. Subject will be included as a random effect.

Two separate models fill be fitted: one for daily level data with time (days from baseline) as continuous variable and another for weekly level data with time (weeks from baseline) as a categorical variable.

For the daily level data, the following model will be fitted:

$$Y_{ijk} = \mu + \alpha_j + \beta \cdot baseline\ CeD\ PRO_{ij} + \gamma \cdot day_{ijk} + \lambda_{jk} + \eta_i + \varepsilon_{ijk},$$

where

Y_{ijk} is the total CeD PRO score for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2$) at day k ($k = 0,1, \dots,112$),
μ is the overall mean,
α_j is the fixed effect due to treatment j,
β is the parameter estimate of the baseline total CeD PRO score,
γ is the parameter estimate of day,
λ_{jk} is the fixed interaction effect due to treatment j and day k,
η_i is the random effect due to subject i,
ε_{ijk} is the random error for subject i from treatment group j for day k.

Note that the proper functional form of the time (day) will be explored, e.g. square of day might be added.

The following SAS program will be used for carrying out the daily level analysis:

```
proc glimmix data=cedpro_d;
  class subjid trt;
  model cedpro_d = cedpro_d_bl trt day trt*day / solution;
  random _residual_ / subject=subjid type=AR(1);
  estimate "AMG 714 vs Placebo" trt 1 -1 /alpha=0.05 cl;
  lsmeans trt /alpha=0.05 cl;
  ods output CovParms=Cov_Par LSMeans=LS_est Estimates=Eff_est
             FitStatistics=Fit_Stat ParameterEstimates=Par_est
             Tests3=Type_3;
run;
```

For the weekly level data, the following model will be fitted:

$$Y_{ijk} = \mu + \alpha_j + \gamma_k + \lambda_{jk} + \beta \cdot baseline\ CeD\ PRO_{ij} + \eta_i + \varepsilon_{ijk},$$

where

Y_{ijk} is the total CeD PRO score for subject i ($i = 1, \ldots, n_j$) from treatment group j ($j = 1, 2$) at week k ($k = 0,1,\ldots,16$),
μ is the overall mean,
α_j is the fixed effect due to treatment j,
γ_k is the fixed effect due to week k,
λ_{jk} is the fixed interaction effect due to treatment j and week k,
β is the parameter estimate of the baseline total CeD PRO score,
η_i is the random effect due to subject i,
ε_{ijk} is the random error for subject i from treatment group j for week k.

The following SAS program will be used for carrying out the weekly level analysis:

```
proc glimmix data=cedpro_w;
  class subjid trt week (ref=FIRST);
  model cedpro_w = cedpro_w_bl trt week trt*week / solution;
  random _residual_ / subject=subjid type=AR(1);
  estimate "AMG 714 vs Placebo" trt 1 -1 /alpha=0.05 cl;
  estimate "Week 0: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 /alpha=0.05 cl;
  estimate "Week 1: AMG 714 vs Placebo" trt 1 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 /alpha=0.05 cl;
  ...
  estimate "Week 16: AMG 714 vs Placebo" trt 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 /alpha=0.05 cl;
  lsmeans trt /alpha=0.05 cl;
  lsmeans trt*week /alpha=0.05 cl;
  ods output CovParms=Cov_Par LSMeans=LS_est Estimates=Eff_est
             FitStatistics=Fit_Stat ParameterEstimates=Par_est
             Tests3=Type_3;
run;
```

Due to the repeated structure of the data, the measurements within a subject will be correlated. Therefore, it is assumed that the overall covariance matrix of the response is block-diagonal (e.g. the observations between different subjects are nor related, but the same correlation structure exists within subjects). In order to determine the covariance structure that best fits the data, the same model with different covariance structures is fitted. Initially, the models will be fitted assuming first-order autoregressive (AR(1)) variance-covariance structure. Common within and between treatment variance components (compound symmetry (CS), and Toeplitz covariance structures) will be further explored to increase sensitivity of statistical tests. The best model will be chosen by comparing the Akaike Information Criterion (AIC) of the models with different covariance structures.

Note that the value of `type` will be changed according to which covariance structure is being fitted.

The modelling results (parameter estimates, covariance parameter estimates, estimated treatment effect, marginal means estimates and results of the check of assumptions) will be tabulated and 95% confidence intervals added, where appropriate. If the parametric assumptions are not met, then in addition to the above-specified model, a generalized estimating equation (GEE) approach will be used as well.

The baseline and post-baseline weekly total CeD PRO scores along with the change from baseline values will also be described by summary statistics and tabulated by treatment group. Change from baseline for the weekly CeD PRO scores of individual questions (calculated as the average score of a question within a week) will also be tabulated by week and treatment group and described by summary statistics. The scores of individual questions and the total daily CeD PRO scores will be listed, as well as the total and by question weekly CeD PRO scores. Mean curves of both the total weekly and daily CeD PRO scores as well as the individual questions will also be plotted by time point and treatment group.

A tabulation of the proportion of subjects with at least a 50% change from baseline for ≥ 6 weeks will be provided. This analysis will be done internally by Celimmune.

ITT population will be used for summarizing CeD PRO results.

8.7.7 Assessment of the gluten-free diet (iVYLISA GIP stool test and dietician consultation)

In celiac disease, identification of gluten contamination is essential for the management of the disease and for the successful conduct of clinical trials. Contaminating gluten is a confounding factor in both diagnosis of RCD-II and in the evaluation of a potential therapeutic effect of any experimental medication. Histologic and clinical endpoints are heavily influenced by the presence of gluten in the diet.

Because of that the iVYLISA GIP-S gluten stool test, a gluten assay developed to detect inadvertent gluten consumption by measuring gluten immunogenic peptides (GIP) in feces (Comino et al, 2012), will be used in the CELIM-RCD-002 trial to assist with data interpretation by assessing if the subjects are compliant with the GFD before enrolment and during the study.

The test detects gluten for up to between four and seven days after consumption, and testing will be done every two weeks – subjects should provide a stool sample collected up to 3 days before the visit to the sites, in order to have a good probability of identifying dietary transgressors to enable correct data interpretation. Testing will be done at a central lab. For the purpose of the study, the test is considered negative when the average amount of gluten of a stool sample is <300 ng GIP/g stool sample.

In addition to the gluten stool test, adherence to the gluten-free diet (GFD) will be assessed periodically (see Appendix 1) by an expert dietician who will counsel the subjects. The dietician will fill out a questionnaire, provided by the Sponsor, to document the conversation and any detected dietary transgressions. This information may be analysed in an exploratory fashion.

The questionnaire will address the following items:

- Did the subject have nutritional counselling with an expert dietitian during the visit? (Yes/No)
- Was a dietary transgression detected since the last visit? (Yes/No)
- If Yes, please indicate the number of transgressions since the last visit.

Patients will be provided a calendar or similar tool to note their dietary transgressions between visits.

iVYLISA GIP stool test results (positive/negative) will be presented by visit and treatment group as well as listed. Dietician assessment of gluten-free diet (i.e. number of transgressions and counselling with expert dietician) will be tabulated by treatment group and visit as well as listed. The mean numbers of transgressions will also be plotted by visit and treatment group.

ITT analysis set will be used.

8.8 Analysis of safety and tolerability

8.8.1 Adverse events

Adverse events (AEs) and adverse drug reactions (ADRs) reported after administration of the study treatment will be classified by system organ classes (SOC) and preferred terms using the MedDRA dictionary (version 18.1). An ADR is defined as an AE to which the study treatment is assessed to be related by the investigator.

The number and proportion (%) of subjects having each AE or ADR will be given by treatment group. The numbers and proportions will be additionally broken down by severity (mild, moderate, severe) and by the causality (definitely not related, unlikely to be related, possibly related, probably related, definitely related). In addition, the number of events and their proportion (%) of the total number of events will be tabulated. All AEs and SAEs will also be listed.

Additionally, narrative descriptions will be included in the study report for all SAEs and AEs leading to discontinuation of the treatment.

Any symptoms recorded before entry to the study, which remained unchanged or improved, will be followed and evaluated separately from the AEs. If the severity of a symptom increases during the study, the symptom will be considered an AE and it will be reported in the AE section.

8.8.2 Clinical laboratory tests

Clinical laboratory tests include haematology, clinical chemistry, and urinalysis panels. The complete list of clinical laboratory parameters is presented in Table 3 below.

Table 3 Clinical laboratory tests

Haematology	Clinical Chemistry	Urinalysis
Basophils (absolute) Basophils/Leukocytes Eosinophils (absolute) Eosinophils/Leukocytes Haematocrit Haemoglobin Lymphocytes (absolute) Lymphocytes/Leukocytes Monocytes (absolute) Monocytes/Leukocytes Neutrophils (absolute) Neutrophils/Leukocytes Platelet Count Red Blood Cell (RBC) count White Blood Cell (WBC) count Anti-tissue transglutaminase (tTG) IgA and IgG Anti-tTG6 IgA and IgG	Alanine Aminotransferase (ALT) Albumin Alkaline Phosphatase Aspartate Aminotransferase (AST) Bilirubin (Total) Calcium Chloride Creatinine Glucose Lactate dehydrogenase (LDH) Potassium Protein (Total) Sodium Urea (BUN)	Blood Cells (Erythrocytes, Leukocytes) Glucose Ketones Microscopic evaluation (i.e. Bacteria, Squamous Epithelial Cells, Mucous Fiberis Urine, Crystals) Protein

Clinical laboratory parameters will be obtained at times indicated in the study schedule (Appendix 1). Blood and urine samples collected at the Screening Visit will require a minimum 8-hour fast.

All clinically significant findings during the study should be followed until resolution or until the finding is clinically stable. Subjects may be withdrawn from study drug if the Investigator or Sponsor deems the clinically significant finding compromising to the subject's safety; however, these subjects will continue to be followed-up per protocol, unless consent is withdrawn.

Detailed information regarding the collection and handling of clinical laboratory specimens, including blood draw totals for each visit and instructions for re-testing of missing or compromised specimens, can be found in a separate Central Laboratory Manual or equivalent document supplied by the central clinical laboratory.

Laboratory test values, including anti-tTG and anti-tTG6 antibodies, will be presented by individual listings with flagging of values outside the normal ranges (normal ranges will be presented in statistical analysis report appendix). Absolute laboratory values and changes from baseline will be presented using summary statistics by treatment group and visit. Clinical laboratory variables will also be explored in mean curves. Urinalysis parameters will not be included in figures. ITT analysis population will be used.

Anti-tTG6 antibody data will be analysed internally by Celimmune.

Additionally, shift table of liver function tests (Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), Bilirubin (Total) and Alkaline Phosphatase (ALP)) will be created. Count and frequencies by treatment group will be presented in the shift table.

8.8.3 Physical examination

Physical examination will be performed at times indicated in the study schedule (Appendix 1) and includes an examination of general appearance; head, eyes, ears, nose, throat (HEENT); lymph nodes; respiratory; cardiovascular; gastrointestinal; musculoskeletal; neurological, psychological and dermatological systems.

Physical examination results will be tabulated by treatment group, visit, body system, result (normal, abnormal, not done) and clinical significance (yes, no) and listed using ITT analysis population.

8.8.4 Vital signs

Vital signs include body temperature, pulse rate, systolic blood pressure (sitting), diastolic blood pressure (sitting), and respiratory rate. BMI and BSA will be obtained from measurements of body weight and height. Vital signs will be measured at screening and all other study visits.

Vital signs will be listed and changes from baseline and absolute values will be presented using summary statistics by treatment group and visit. Mean absolute values and change from baseline values in weight will also be presented graphically by time point and treatment group.

ITT analysis population will be used for the analysis of vital signs

8.8.5 Immunogenicity

Immunogenicity, i.e. the generation of anti-drug antibodies (ADA), is a potential risk for any biologic therapeutic. Immunogenicity may lead to injection reactions and to loss of efficacy when the antibodies are neutralizing and high-titer.

A two-tiered immunogenicity testing approach will be used in order to determine if a sample contains ADAs. Samples will be initially tested in an immunoassay. Samples that test positive for binding antibodies will then be tested in an assay to detect neutralizing antibodies (NAb). Immunogenicity testing will be performed at times specified in schedule of study procedures (Appendix 1).

Immunogenicity will be tabulated by treatment group, visit, ADA test result (negative, positive) and neutralizing antibodies test result (negative, positive) and listed. ITT analysis population will be used.

8.8.6 Other safety variables

8.8.6.1 Pregnancy test

All females of child bearing potential (FOCBP) will have urine or serum pregnancy tests throughout the study as outlined in schedule of study procedures (Appendix 1). Subjects who become pregnant during the study will be withdrawn from participation and the outcome of the pregnancy followed.

Pregnancy test results will be listed by treatment group, visit, test type (serum, urine) and result (negative, positive). ITT analysis population will be used for presenting pregnancy test results.

8.8.6.2 Prior and concomitant medications

Prior and concomitant medications will be collected throughout the study and coded using the World Health Organization Drug Dictionary (WHO DD) 2014 September version.

Prior and concomitant medication will be listed for ITT analysis population.

8.8.6.3 Imaging tests

The conduct of imaging tests is acceptable if required to evaluate the status of the lymphoma at any time during the conduct of the study, including during screening for the purpose of assessing eligibility criteria (e.g., exclusion of EATL, which requires the use of the site's standard imaging techniques). These tests may include enteroscopy (videocapsule and double balloon enteroscopy), entero CT scan, MRI (for size of mesenteric lymph nodes; thickness of bowel wall) and 18FDG-PET scan.

The results of imaging tests will be listed for ITT analysis population, if applicable.

8.9 Additional analyses

In addition to the main analyses, pre-specified subgroup analyses may include the following, if reasonable distributions of subgroups are available for statistically meaningful assessments:

- Dietary transgressions (gluten consumption) based on serial iVYLISA GIP testing and on the dietician's assessment;
- Previous RCD-II treatment (e.g. previous treatment with BMT, cladribine, steroids, Hu-Mik-beta1, etc.);
- Duration of disease;
- Age of onset of RCD-II;
- Sex;
- Site;
- Expression of certain biomarkers at baseline, such as CD122, Granzyme B and IL-21R in IELs, or IL-15 in serum;
- Levels of anti-tTG antibodies, at baseline and during the study;
- Protocol deviations: missed doses of IP, incorrect IP volume for administration, etc.;
- Use of immune-suppressants: concomitant or prior to the study (cladribine, azathioprine, budesonide, stem cell transplant, etc.).

These additional sensitivity analyses may be carried out only for the key flow cytometry and histology endpoints (i.e., the primary and secondary efficacy endpoints dependent on biopsy tissue). Separate models will be fitted for males and females, compliant and non-compliant subjects, subjects with each of the previous treatments, groups of subjects with different disease durations, groups of subjects with different age of onset of disease, and for each site. In order to determine whether the effects of sex, compliance and site are statistically significant, additional three models are fitted where compliance, sex and site and their interaction terms with treatment group, respectively, are included.

8.10 Execution of statistical analyses

Statistical analysis will be performed by StatFinn Oy under the supervision of Celimmune, LLC. The PK analyses will be performed in collaboration between Celimmune and StatFinn Oy.

9 Hardware and software

Statistical analysis, tables and subject data listings will be performed with SAS® for Windows (SAS Institute Inc., Cary, NC, USA), version 9.4 will be used.

SF-12 v. 2 health domain scores and summary scores are calculated using QualityMetric Health Outcomes™ Scoring Software 4.5.1.

10 References

1. Brar P, Lee S, Lewis S, *et al*. Budesonide in the treatment of refractory celiac disease. *Am J Gastroenterol*. 2007;102:2265-9.
2. Goerres MS, Meijer JW, Wahab PJ, *et al*. Azathioprine and prednisone combination therapy in refractory coeliac disease. *Aliment Pharmacol Ther* 2003;18:487-94.
3. Kelly CP, Green PH, Murray JA, *et al*. Larazotide acetate in patients with coeliac disease undergoing a gluten challenge: a randomized placebo-controlled study. *Aliment Pharmacol Ther*. 2013;37:252-62.
4. Lähdeaho ML, Mäki M, Laurila K, *et al*. Small- bowel mucosal changes and antibody responses after low- and moderate-dose gluten challenge in celiac disease. *BMC Gastroenterol*. 2011;11:129.
5. Leffler DA, Kelly CP, Green PH, *et al*. Larazotide Acetate for Persistent Symptoms of Celiac Disease Despite a Gluten-Free Diet: A Randomized Controlled Trial. *Gastroenterology*. 2015;148:1311-19.
6. Malamut G, El Machhour R, Montcuquet N, *et al*. IL-15 triggers an antiapoptotic pathway in human intraepithelial lymphocytes that is a potential new target in celiac disease-associated inflammation and lymphomagenesis. *J Clin Invest*. 2010;120:2131-43.
7. Marsh MN. Gluten, major histocompatibility complex, and the small intestine. A molecular and immunobiologic approach to the spectrum of gluten sensitivity ('celiac sprue'). *Gastroenterology.* 1992;102(1):330-54.
8. Oberhuber G. Histopathology of celiac disease. *Biomed Pharmacother*. 2000; 54(7):368-72.
9. Riegler G, Esposito I. Bristol scale stool form. A still valid help in medical practice and clinical research. *Tech Coloproctol*. 2001 Dec;5(3):163-4.
10. SAS, Institute Inc., Cary, NC, USA.
11. Svedlund, J, Sjodin, I, Dotevall, G. GSRS-a clinical rating scale for gastrointestinal symptoms in patients with irritable bowel syndrome and peptic ulcer disease. *Dig Dis Sci* 1988;33:129-134
12. Tack GJ, Verbeek WH, Al-Toma A, *et al*. Evaluation of Cladribine treatment in refractory celiac disease type II. *World J Gastroenterol* 2011;17:506-13.
13. Tack GJ, Wondergem MJ, Al-Toma A, *et al*. Auto-SCT in refractory celiac disease type II patients unresponsive to cladribine therapy. *Bone Marrow Transplant* 2011;46:840-6.
14. Ware J, Kosinski, M, Turner-Bowker D, Gandek B. *User's manual for the SF-12v2 Health Survey*. 2009.

11 Appendix

Appendix 1. Study schedule

Appendix 2. List of tables and figures

Appendix 3. List of subject data listings

Appendix 4. Table templates

12 Document history

Version number	Version date	Status	Author
1.0	24OCT2016	Final	Marju Valge
2.0	16JUN2017	Final	Marju Valge

Changes and additions made in SAP V2.0

Additions

1. Sensitivity analysis of absolute change from baseline added to the following endpoints:
    a. Change from baseline in the % of aberrant IELs vs total IELs as assessed by flow-cytometry;
    b. Change from baseline in the % of aberrant IELs vs intestinal epithelial cells;
    c. Change from baseline in VH:CD ratio;
    d. Change from baseline in total IEL counts.
2. The following additional sensitivity analyses which might be carried out if deemed appropriate, were added:
    a. Levels of anti-tTG antibodies, at baseline and during the study;
    b. Protocol deviations: missed doses of IP, incorrect IP volume for administration, etc.;
    c. Use of immune-suppressants: concomitant or prior to the study (cladribine, azathioprine, budesonide, stem cell transplant, etc.);
    d. If a subject had a misallocated treatment on a specific visit, observed upon unblinding, a secondary sensitivity analysis will be performed on as treated basis;
    e. Sensitivity analysis excluding atypical subjects from all efficacy assessments.
3. The following exploratory analyses were also added:
    a. Averages by treatment for proportion of subjects showing BSSF >= 6 by time will be plotted. The differences in treatment AUCs will be explored using one-way analysis of variance;
    b. Week 12 PGA scores will be dichotomized for an outcome endpoint where scores ≤ 2 will be considered treatment success and scores > 2 will be considered treatment failure. Assessment of differences in treatment proportions will be conducted;

    c. A tabulation of the proportion of subjects with at least a 50% change from baseline for ≥ 6 weeks in CeD PRO will be provided.
4. The following hematology parameters were added:
    a. Basophils/Leukocytes;
    b. Eosinophils/Leukocytes;
    c. Lymphocytes/Leukocytes;
    d. Monocytes/Leukocytes;
    e. Anti -tissue transglutaminase (tTG) IgA and IgG;
    f. Anti-tTG6 IgA and IgG.
5. Serum Granzyme B (pg/mL) was added to the list of biomarkers of disease activity to be analyzed.
6. Additionally, shift table of liver function tests (Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), Bilirubin (Total) and Alkaline Phosphatase (ALP)) will be created. Count and frequencies by treatment group will be presented in the shift table.
7. Major protocol deviations will be tabulated and summarized by treatment group in addition to listing these.
8. Definition of relative change from baseline was added.

Changes

1. The total GSRS score will be calculated as the mean (not sum as specified in SAP V1.0) of the scores of all 15 questions, with the scores for the individual questions between 1 (No discomfort at all) and 7 (Very severe discomfort).
2. The total CeD-GSRS score will be calculated as the mean (not sum as specified in SAP V1.0) of scores of all 10 questions in CeD-GSRS questionnaire, with the scores of the questions between 1 (No discomfort at all) and 7 (Very severe discomfort).
3. The definition of change from baseline was given in more detail.
4. It was decided, that in addition to other imputations, all assay results over the upper limit of quantification (ULOQ) will be assigned a value of ULOQ.
5. The individual overlaid figures were deemed unnecessary and therefore were removed from SAP V2.0 with the exception of overlaid individual curves of CeD PRO.
6. The analysis of resolution of mucosal atrophy according to Marsh scores was removed.
7. Analysis of subjects with normalization of total IEL counts by immunochemistry was removed.
8. Analysis of immunohistochemistry IL-15 positive cells was removed.
9. Referral to PK Data Analysis Plan was replaced with PK Report (PK Data Analysis Plan will not be created).
10. Minor changes to wording were made (e.g. typos corrected, wording clarified).