STATFINN

STATISTICAL ANALYSIS PLAN

CELIM-NRCD-001

A PHASE 2A, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED, PARALLEL-GROUP STUDY TO EVALUATE THE EFFICACY AND SAFETY OF AMG 714 IN ADULT PATIENTS WITH CELIAC DISEASE

Version 1.0

Phase 2a study

Standard: GCP

Date of approval: 19APR2017

SPONSOR: Celimmune, LLC

Confidential

NCT Number: 02637141
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 [redacted]
Title:	Statistician
Signature:	PPD [redacted]
Date:	19APR2017

SIGNATURE PAGE: STATFINN

This Statistical Analysis Plan was reviewed/approved by:

Name of Reviewer:	PPD [redacted]
Title:	Senior Statistician
Signature:	PPD [redacted]
Date:	19APR2017

SIGNATURE PAGE: CELIMMUNE, LLC

This Statistical Analysis Plan was reviewed/approved by:

Name of Client´s Representative:	PPD, MD, PhD
Title:	CEO & Chief Medical Officer
Company:	Celimmune, LLC
Signature:	PPD
Date:	19APR2017

ABBREVIATIONS

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
Anti-DGP	Deamidated Gliadin Peptide Antibody
Anti-tTG	Tissue Transglutaminase Antibody
AR(1)	First-order auto-regressive
AST	Aspartate Aminotransferase
BID	Twice a day
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
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
HIV	Human immunodeficiency virus
icCD3+	Intra-cellular CD3-positive
IEC	Intestinal epithelial cells

IEL	Intraepithelial lymphocyte
IHC	Immunochemistry
IgA	Immunoglobulin A
IgG	Immunoglobulin G
IL-15	Interleukin 15
IL-21R	Interleukin 21 receptor
ITT	Intention to treat
LDH	Lactate Dehydrogenase
LLOQ	Lower limit of quantification
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
NRCD	Non-responsive celiac disease
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
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
SC	Subcutaneous
sCD3-	Surface CD3-negative
sCD8-	Surface CD8-negative
SD	Standard deviation
SOC	System organ class
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-NRCD-001 which is based on the final study protocol CELIM-NRCD-001 Version 3 (dated 29AUG2016). This SAP describes the statistical analyses which will be presented in the clinical study report.

2 Study objectives

The study objectives are the following:

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

- To assess the efficacy of AMG 714 in attenuating the effects of gluten exposure in adults with celiac disease.

The **primary efficacy endpoint** is:

- Attenuation of gluten-induced small intestinal mucosal morphological injury, measured morphometrically as villous height to crypt depth (VH:CD) ratio.

The primary endpoint of the study is the difference in the Baseline-to-Week-12 % change of VH:CD ratio between each of the two AMG 714 dose arms against the placebo arm.

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

- Attenuation of gluten-induced small intestinal mucosal inflammation measured as intraepithelial lymphocyte (IELs) density;
- Attenuation of gluten-induced small intestinal mucosal morphological injury using a grouped classification of Marsh score;
- Attenuation of gluten-induced serum antibodies (i.e. anti-tissue transglutaminase antibodies [anti-tTG] IgA, anti-deamidated gliadin peptide antibodies [anti-DGP] IgA and IgG);
- Attenuation of gluten-induced clinical symptoms (i.e. Bristol Stool Form Scale [BSFS], Gastrointestinal Symptom Rating Scale [GSRS] and celiac disease GSRS [CeD-GSRS]).

The secondary endpoints of Marsh score and total IEL counts will be evaluated by the change from baseline to Week 12 between the two AMG 714 dose arms and the placebo arm. Anti-tTG IgA, anti-DGP IgA and IgG will be evaluated by comparing the change from baseline in results obtained after every four weeks of two AMG 714 dose arms and placebo arm. BSFS, GSRS and CeD-GSRS will be evaluated by comparing the change from baseline in weekly scores of two AMG 714 dose arms and placebo arm.

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

- To assess the safety and tolerability of AMG 714 when administered to adult patients with celiac disease exposed to a gluten challenge.

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 treatment group, 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, including
    - Physician Global Assessment of Disease (PGA);
    - Biomarkers of disease activity;
    - Celiac Disease Patient-Reported Outcome (CeD PRO);
- Exposure/Response (PK/PD).

PK data will be tabulated by treatment group. In addition, dose proportionality, achievement of steady-state, and accumulation ratio based on C_{trough} concentrations, comparison of C_{trough} levels with corresponding values in study 20060349, and correlation of C_{trough} concentrations with biomarkers of disease activity will be evaluated. Exploratory PD endpoints are Physician Global Assessment of Disease (PGA), biomarkers of disease activity (i.e., CRP, serum IL-15 [sIL-15]) 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 arms and placebo arm. PK/PD association will mainly be assessed graphically.

PK analyses will be described in detail in a separate PK Data Analysis Plan (DAP) to be prepared by Celimmune.

In subjects with dermatitis herpetiformis (DH), optional photography will be conducted according to the study manual. Skin biopsies will be allowed per investigator's discretion but will not be part of this protocol. Any skin information on DH will be summarized by Celimmune for descriptive purposes without formal statistical analysis.

3 Study type and design

CELIM-NRCD-001 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 attenuation of the effects of gluten exposure in adult patients with celiac disease during a gluten challenge.

After signing consent subjects will be screened for the study. All subjects who meet the study entry criteria will be randomized at a 1:1:1 ratio to receive 150 mg or 300 mg AMG 714 or placebo once every two weeks for 10 weeks. Randomization will be stratified by site and sex. The study drug (AMG 714 or placebo) will be administered at the clinical site in a double-blind fashion via subcutaneous (SC) injection. Subjects will remain confined to the study site for a minimum of 1 hour after each administration of study medication. During this time the Investigator and study site staff will assess the subject for AEs and collect the required PK samples and post dose vital signs as outlined in the study schedule of events in Appendix 1.

In addition to receiving study medication (AMG 714 or placebo), all subjects will be required to consume either placebo gluten or active gluten administered in a single-blind fashion. Beginning on the day following Visit 1 (i.e. Week 0/Day 1, since Visit 1 is defined as Week 0/Day 0) through Visit 2 (Week 2/Day 14), all randomized subjects will consume placebo gluten, in the form of gluten-free foodstuffs (e.g. rusks) provided by the Sponsor, twice a day (BID) at the time of two regularly consumed gluten-free meals (chosen by the subject). Starting at Visit 2, subjects will switch to consume a daily total of approximately 2-4g of active gluten, provided in the form of foodstuff identical in appearance to that provided during the two-week single-blind placebo-gluten period. These gluten-containing foodstuffs, supplied by the sponsor, will be consumed in approximately 1-2g doses BID at the time of two regularly consumed gluten-free meals (as chosen by the subject) from Visit 2 (Week 2/Day 14) through the end of the treatment period at Visit 7 (Week 12/Day 84). Consumption of additional products containing gluten will be prohibited and subjects will be expected to continue total adherence to their former GFD from the time of screening until the follow-up study visit (Visit 8, Week 16/Day 112).

Subjects' adherence to GFD and consumption of provided gluten will be periodically assessed via stool and urine sample testing using the iVYLISA GIP stool and urine gluten tests. Subjects with known or suspected GFD transgressions or those suspected of non-compliance to protocol specified gluten dosing will be counselled and allowed to continue in the study.

A study site staff member will contact each subject by telephone both one day and one week after the first study drug administration in order to assess for AEs. Subjects will return to the clinic every two weeks for study drug administration and/or efficacy and safety assessments as indicated in the study schedule of events in Appendix 1. The final study drug administration will occur at Visit 6 (Week 10/Day 70). An end-of-study efficacy assessment

will be collected at Visit 7 (Week 12/Day 84). A final study visit will be conducted six weeks after the last administration of study drug at Week 16 (Visit 8/Day 112).

Subjects who meet all other study entry criteria will undergo upper gastrointestinal endoscopy and biopsy prior to baseline (Visit 1, Week 0/Day 0) and at the end of the 12-week randomized period while still on the gluten challenge within five days before Visit 7 (Week 12/Day 84) in order to assess changes from baseline in VH:CD ratio, IELs, and Marsh score.

If the screening biopsy shows villous atrophy with a VH:CD less than 1.5 (1.4 or lower), and after notification to the sponsor, the subject may be allowed to stay in the study but the gluten challenge will not be started (or will be interrupted if the patient has been started in the challenge) and the subject will be analysed separately for the main efficacy assessments (primary and secondary endpoints). Subjects with VH:CD 1.5 or above will receive the gluten challenge.

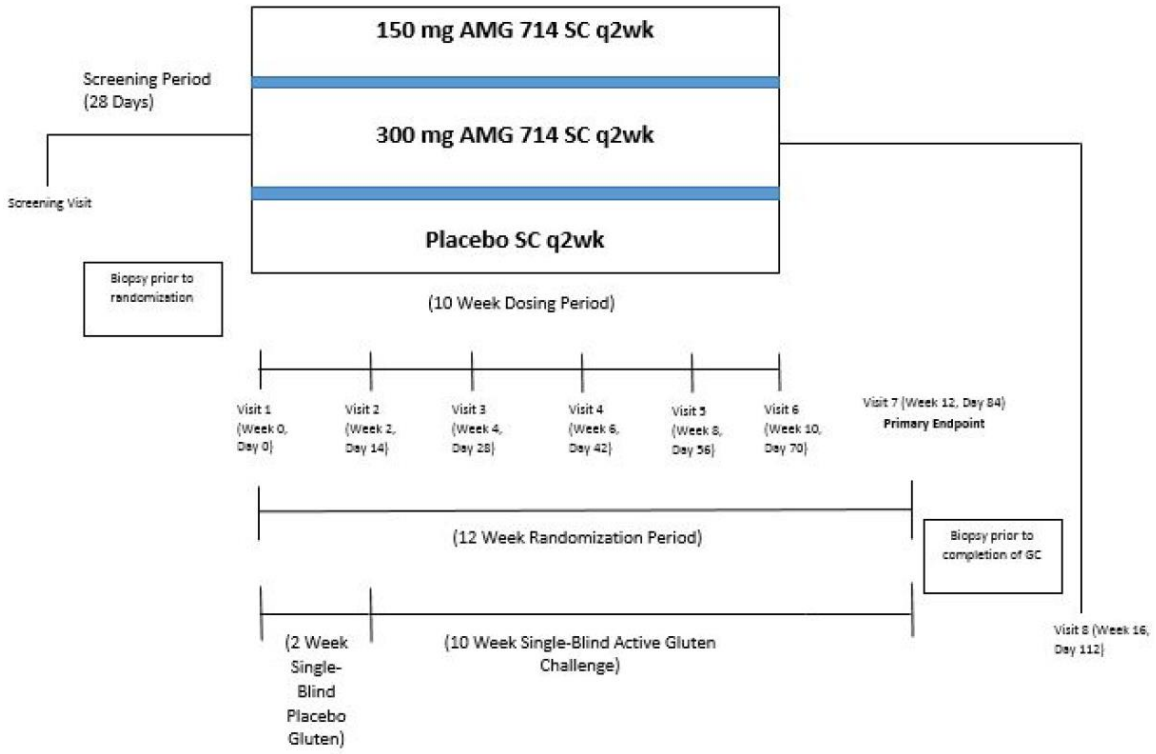
The threshold of VH:CD to receive gluten challenge was initially 2 (protocol version 1) and subsequently 1.8 (protocol version 2) before the final threshold of 1.5 VH:CD was adopted in protocol version 3. The population of subjects who received gluten challenge and had pre/post biopsies, a majority of the patients, will represent the per protocol population 1 (PP1) and will be analysed for the primary and secondary endpoints separately from the subjects who did not receive gluten challenge and had pre/post biopsies, a small fraction of the subjects who will represent the per protocol population 2 (PP2). All primary and secondary endpoints will be analysed in PP1 and PP2 separately.

All subjects will complete the BSFS at the time of each bowel movement from baseline (Visit 1, Week 0/Day 0) up to the final study visit (Visit 8, Week 16/Day 112). Subjects will complete the CeD PRO daily from baseline up to the final study visit. Subjects will also complete the GSRS at screening and, thereafter, weekly from baseline through the final study visit. The BSFS, GSRS and CeD PRO will be completed using a handheld electronic diary at times specified in Appendix 1.

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

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

Figure 1 CELIM-NRCD-001 study schematic



4 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 randomisation expert and the unblinded statistician will have access to. Once the data base is locked and the treatment code open, the randomisation documents will be moved to the non-restricted area.

Subjects will be randomized at a 1:1:1 allocation to receive 150 mg AMG 714 or 300 mg AMG 714, or placebo once every two weeks for a total of six administrations over ten weeks. Randomization will be stratified by site and sex.

5 Statistical hypotheses

The primary endpoint of the study is the difference in the Baseline-to-Week-12 % change of VH:CD ratio between each of the two AMG 714 dose arms against the placebo arm.

The primary endpoint will be tested for each of the two dose levels, 150 mg or 300 mg, separately as follows:

$$H_{10}: \mu_{AMG\ 714\ (300mg)} = \mu_{Placebo}$$

against the alternative

$$H_{11}: \mu_{AMG\ 714\ (300mg)} \neq \mu_{Placebo}$$

and

$$H_{20}: \mu_{AMG\ 714\ (150mg)} = \mu_{Placebo}$$

against the alternative

$$H_{21}: \mu_{AMG\ 714\ (150mg)} \neq \mu_{Placebo}$$

where $\mu_{AMG\ 714\ (300mg)}$, $\mu_{AMG\ 714\ (150mg)}$ and $\mu_{Placebo}$ denote the mean Baseline-to-Week-12 % change of VH:CD ratio in the high dose, low dose and placebo arm, respectively. Each of the two hypotheses will be tested using a two-sided, 0.05 level of significance without any adjustments for multiple comparisons.

6 Estimation of sample size

The sample size of approximately 63 subjects (21 subjects/arm) has been calculated to achieve at least ~ 80% power in the primary endpoint, the difference in the Baseline-to-Week-12 %-change of VH:CD ratio between the two AMG 714 arms and the placebo arm. This sample size provides close to 90% (88.8%) power to detect a 40-point difference between the placebo arm and the 300mg high-dose arm, and close to 80% (79.3%) power to detect a 35-point difference between the placebo arm and the 150mg low-dose arm.

This sample size calculation is based on the following assumptions:

- Two-sided type one error rate α = 0.05.
- Power 1-β = 80%.
- Analysis method: one-way ANOVA by SAS® (proc power).
- Common SD = 36 for the baseline to Week 12 %-change in VH:CD.
- 45%-change (decrease) in the placebo group, %-change (decrease) of 10% in the 150mg and 5% in the 300mg arms.
- Drop-out rate ~19% (assuming non-evaluable drop-outs plus DH <=15% and non-evaluable due to technical issues or villous atrophy at baseline <5%).

These assumptions are based on the observations made in a similar 12-week gluten challenge study done with 3-5 grams of gluten per day (Lähdeaho *et al*, 2011). According to this sample size calculation (including a 19% drop-out rate), 17 completed evaluable subjects/arm (51 subjects in total) will be needed for the primary analysis.

7 Statistical methods

7.1 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 populations (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. Exclusion of subjects from the PP set will be reviewed, documented and approved before the study is unblinded to the study Sponsor. There will be two PP populations, PP1 (including subjects who received gluten challenge for at least a week) and PP2 (including subjects who did not receive gluten challenge, or received gluten challenge for less than a week).

Non-evaluable subjects are pre-specified before database lock and will include subjects dropping out of the study before Week 6 (insufficient duration of the gluten challenge and/or no exit biopsy).

Major protocol deviations will be examined by the Sponsor prior to database lock and the decision to include or exclude subjects from PP population based on the protocol deviations will be documented.

Sensitivity analyses may be conducted, if deemed necessary, for protocol deviations such as:

1. 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 had a significant effect on the efficacy data obtained. For example, analgesics/anti-diarrheals which may affect symptoms;
2. Significant deviations from the inclusion/exclusion criteria or study conduct., including, but not limited to, subjects who missed one or more doses of study drug, were not fully compliant with the gluten challenge, or had dermatitis herpetiformis.

ITT population: this population consists of all randomized subjects who have received at least one dose of the study drug. Subjects will be analyzed in the treatment group to which they are randomized. 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 analyses will be performed on an as treated basis.

The primary and secondary endpoints as well as exploratory an PD variables will be analysed using the two PP populations. 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	Secondary variables	Exploratory/PD variables	Safety variables	Demographic and baseline variables
PP*	PP*	PP*	ITT	ITT PP*

*PP population includes PP1 and PP2 populations.

7.2 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%.

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 0.05 will be considered statistically significant, if not stated otherwise. In addition to the inferential statistics, 95% confidence intervals will be constructed. For the primary endpoint for each of the dose levels a 95% confidence interval of the treatment difference will be constructed. Ninety-five percent (95%) confidence intervals will be constructed for the individual treatment point estimates.

Various covariates (baseline values, site and sex) will be included in the models to increase sensitivity of the tests and will be retained in the respective models regardless of their significance or lack thereof.

7.2.1 Definition of derived variables

Geometric mean will be calculated as:

$$\text{mean}_{\text{geo}} = \sqrt[n]{x_1 x_2 \cdots x_n} = \exp\{\frac{1}{n}\sum_{i=1}^{n} \ln x_i\}.$$

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

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

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

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:

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

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

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

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, SD_{geo}	2 more than in original data
Frequencies (%)	1 decimal place
$CV\%_{geo}$	1 decimal place

7.2.2 Missing values

For the primary biopsy driven endpoint, 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. 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. For assay results over the upper limit of quantification (ULOQ) a value of ULOQ will be assigned in mean calculations.

7.2.3 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 populations 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 the 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 populations).

Subjects will be considered study completers at Visit 7 (Week 12/Day 84), regardless of whether or not the Final Study Visit (Visit 8; Week 16/Day 112) is attended.

7.3 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 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.

7.4 Demographic and baseline characteristics

Demographic and baseline characteristics collected and presented for this study, include: age, sex, race, ethnicity, weight, height, BMI, BSA, 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 height.

Demographic and baseline characteristics 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 and fasting compliance will only be listed.

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

7.5 Extent of exposure and compliance

Extent of exposure will be summarized showing:

- Number of subjects exposed to placebo, 150 mg AMG 714 or 300 mg 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. Dates and times of subcutaneous (SC) injection along with locations will only be listed.

7.6 Analysis of efficacy

7.6.1 Primary efficacy variable

Primary efficacy endpoint of the study is:

- Relative (%) change from baseline to Week 12 in villous height to crypt depth (VH:CD) ratio.

The primary endpoint will be analysed using analysis of covariance (ANCOVA), where the baseline VH:CD ratio, site and sex will be included as covariates and treatment group as a fixed effect in the statistical model.

The analysis of the primary endpoint will be carried out using the two PP populations.

The following model will be fitted:

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

where

Y_{ijkl} is the relative (%) change from baseline in the VH:CD for subject i ($i = 1, \dots, n_{jkl}$) from site l ($l = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group j ($j = 1, 2, 3$),
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,
φ_k is the fixed effect due to sex k,
τ_l is the fixed effect due to site l,
ε_{ijkl} is the random error for subject i from site l ($l = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group j; $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$.

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=VHCD alpha=0.05 outstat=F_tests;
   class trt site sex;
   model VHCD_change = bl_VHCD trt site sex / clparm solution SS1 SS2;
   estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1;
   estimate "300 mg AMG 714 vs Placebo" trt 0 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, i.e. the check of common variance assumption) will be tabulated and 95% confidence intervals added, where appropriate.

As a secondary approach, 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.

7.6.2 Secondary efficacy variables

The secondary efficacy endpoints of the study are:

- Attenuation of gluten-induced small intestinal mucosal inflammation measured as intraepithelial lymphocyte (IELs) density;
- Attenuation of gluten-induced small intestinal mucosal morphological injury using a grouped classification of Marsh score;
- Attenuation of gluten-induced serum antibodies (i.e. anti-tissue transglutaminase antibodies [anti-tTG] IgA, anti-deamidated gliadin peptide antibodies [anti-DGP] IgA and IgG);
- Attenuation of gluten-induced clinical symptoms (i.e. Bristol Stool Form Scale [BSFS], Gastrointestinal Symptom Rating Scale [GSRS] and celiac disease GSRS [CeD-GSRS]).

Attenuation of gluten-induced small intestinal mucosal inflammation measured as intraepithelial lymphocyte (IELs) density

Relative (%) change in IELs density will be analysed using the same method as for the primary endpoint, i.e. the following model will be fitted:

$$Y_{ijkl} = \mu + \beta \cdot baseline\ IEL_{ijkl} + \alpha_j + \varphi_k + \tau_l + \varepsilon_{ijkl},$$

where

Y_{ijkl} is the relative (%) change from baseline in IEL density for subject i ($i = 1, \dots, n_{jkl}$) from site l ($l = 1,2,3$), sex k ($k = 1,2$) and treatment group j ($j = 1,2,3$),
μ is the overall mean,
β is the parameter estimate of baseline IEL,
α_j is the fixed effect due to treatment j,
φ_k is the fixed effect due to sex k,
τ_l is the fixed effect due to site l,
ε_{ijkl} is the random error for subject i from site l ($l = 1,2,3$), sex k ($k = 1,2$) and treatment group j; $\varepsilon_{ijkl} \sim N(0, \sigma_\varepsilon^2)$.

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 site sex;
   model IEL_change = bl_IEL trt site sex / clparm solution SS1 SS2;
   estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1;
   estimate "300 mg AMG 714 vs Placebo" trt 0 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 approach, the same model will be fitted using the absolute change in IEL density 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, i.e. the check of common variance assumption) 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 populations will be used for the analysis.

Attenuation of gluten-induced small intestinal mucosal morphological injury using a grouped classification of 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, site and sex will be included in the model as explanatory variables. For modelling purposes, a binary variable with values 1 (in case improvement in Marsh scores was observed) and 0 (no improvement or worsening in Marsh scores) will be used as a response. Logit function will be used as the link function. The following model will be fitted:

$$\log\left(\frac{p_{jkl}}{1-p_{jkl}}\right) = \mu + \alpha_j + \varphi_k + \tau_l,$$

where

$Y_j \sim Bin(p_{jkl}, n_{jkl})$ – is the number of subjects with improvement in Marsh scores from the second biopsy in treatment group j ($j = 1,2,3$),
p_{jkl} is the probability of improvement in Marsh score for subjects in treatment group j sex k and site l by Week 12,
μ is the overall mean (on logit scale),
α_j is the fixed effect due to treatment j on the logit scale,
φ_k is the fixed effect due to sex k on the logit scale,
τ_l is the fixed effect due to site l on the logit scale.

The odds ratios for 150 mg AMG 714 vs placebo groups and 300 mg AMG 714 vs placebo groups will be reported and the following formula will be used for obtaining the ratios:

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

where $\alpha_{AMG\ 714}$ denotes the treatment effect of AMG 714 (for 150 mg AMG 714 and 300mg AMG 714 groups) 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 site sex;
      model marsh_imp = trt site sex / dist=bin link=logit lrci type3;
      estimate 'OR 150AMG vs PLA ' trt 1 0 -1 / exp ;
      estimate 'OR 300AMG vs PLA ' trt 0 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 150 mg AMG 714 vs placebo, estimated odds ratio of 300 mg AMG 714 vs placebo, marginal (i.e. least squares) means estimates for treatment groups) will be tabulated and 95% confidence intervals

added, where appropriate. Marginal means estimates for the probabilities will be presented on the original (probability) scale, odds ratios will also be presented.
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 without atrophy according to the Marsh score, i.e., the % of subjects with Marsh scores 0, 1 or 2. Frequency table by treatment group will be used for summarizing these results.

PP populations will be used for the analysis.

Attenuation of gluten-induced serum antibodies: anti-tissue transglutaminase antibodies (anti-tTG) IgA

Anti-tTG auto-antibodies, while not very responsive to modest dietary transgressions, are very specific for celiac disease activity and constitute an important tool to assess disease modification (Lähdeaho *et al*, 2011; Kelly *et al*, 2013).

The relative (%) change from baseline in anti-tTG IgA result will be analysed using a linear mixed effects repeated measures model (MMRM) with the baseline value, treatment group, site, sex 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:

$$Y_{ijklm} = \mu + \tau_j + \varphi_k + \alpha_l + \gamma_m + \lambda_{lm} + \beta \cdot baseline\ tTGIgA_{ijklm} + \eta_i + \varepsilon_{ijklm},$$

where

Y_{ijklm} is the relative (%) change from baseline in anti-tTG IgA result for subject i ($i = 1, \dots, n_{jkl}$) from site j ($j = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group l ($l = 1, 2, 3$) at visit m ($m = 3, 5, 7, 8$),
μ is the overall mean,
α_l is the fixed effect due to treatment l,
φ_k is the fixed effect due to sex k,
τ_j is the fixed effect due to site j,
γ_m is the fixed effect due to time point m,
λ_{lm} is the fixed interaction effect due to treatment l and time point m,
β is the parameter estimate of the baseline anti-tTG IgA result,
η_i is the random effect due to subject i,
ε_{ijklm} is the random error for subject i from site j ($j = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group l for time point m.

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=ttgiga;
  class subjid trt site sex time(ref=FIRST);
  model ttgiga = ttgiga_bl trt site sex time trt*time / solution;
  random _residual_ / subject=subjid type=UN;
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl;
  estimate "300 mg AMG 714 vs Placebo" trt 0 1 -1 /alpha=0.05 cl;
  estimate "Visit 3: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*time 0 0 0 1
                     0 0 0 0
                     0 0 0 -1 /alpha=0.05 cl;
  estimate "Visit 3: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*time 0 0 0 0
                     0 0 0 1
                     0 0 0 -1 /alpha=0.05 cl;
  estimate "Visit 5: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*time 1 0 0 0
                     0 0 0 0
                     -1 0 0 0 /alpha=0.05 cl;
  ...
  estimate "Visit 8: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*time 0 0 0 0
                     0 0 1 0
                     0 0 -1 0 /alpha=0.05 cl;
  lsmeans trt /alpha=0.05 cl;
  lsmeans trt*time /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, e.g. the normality assumption) 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.

As a secondary approach, the same model will be fitted using the absolute change in anti-tTG IgA density as a response variable. Modelling results (parameter estimates, covariance parameter estimates, estimated treatment effect, marginal means estimates and results of the check of assumptions, e.g. the normality assumption) 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 anti-tTG IgA results along with the change from baseline values (both relative and absolute) will also be described by summary statistics and tabulated by treatment group. Anti-tTG IgA results will also be listed. Individual and mean anti-tTG IgA results, will also be presented graphically by time point and treatment group.

PP populations will be used for the analysis.

Attenuation of gluten-induced serum antibodies: anti-deamidated gliadin peptide antibodies (anti-DGP) IgA and IgG

Anti-deamidated gliadin peptide antibodies (anti-DGP), IgA or IgG, may be positive in some patients with celiac disease who are negative for anti-tTG, due to their different kinetics and to the possibility of IgA deficiency, in which case, DGP IgG are the most reliable antibodies (Brusca, 2015).

The absolute change from baseline in DGP IgA and IgG results will be analysed using a linear mixed effects repeated measures model (MMRM) with the baseline value, treatment group, site, sex, 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:

$$Y_{ijklm} = \mu + \tau_j + \varphi_k + \alpha_l + \gamma_m + \lambda_{lm} + \beta \cdot baseline\ DGPIgA_{ijklm} + \eta_i + \varepsilon_{ijklm},$$

where

Y_{ijklm} is the absolute change from baseline in anti-DGP IgA result for subject i ($i = 1, \dots, n_{jkl}$) from site j ($j = 1,2,3$), sex k ($k = 1,2$) and treatment group l ($l = 1,2,3$) at visit m ($m = 3,5,7,8$),
μ is the overall mean,
α_l is the fixed effect due to treatment l,
φ_k is the fixed effect due to sex k,
τ_j is the fixed effect due to site j,
γ_m is the fixed effect due to time point m,
λ_{lm} is the fixed interaction effect due to treatment l and time point m,
β is the parameter estimate of the baseline anti-DGP IgA result,
η_i is the random effect due to subject i,
ε_{ijklm} is the random error for subject i from site j ($j = 1,2,3$), sex k ($k = 1,2$) and treatment group l for time point m.

The same model will be fitted to anti-DGP IgG results.

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=dgpiga;
  class subjid trt site sex time(ref=FIRST);
  model dgpiga = dgpiga_bl trt site sex time trt*time / solution;
  random _residual_ / subject=subjid type=UN;
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl;
  estimate "300 mg AMG 714 vs Placebo" trt 0 1 -1 /alpha=0.05 cl;
  estimate "Visit 3: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*time 0 0 0 1
                     0 0 0 0
                     0 0 0 -1 /alpha=0.05 cl;
  estimate "Visit 3: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*time 0 0 0 0
                     0 0 0 1
                     0 0 0 -1 /alpha=0.05 cl;
  estimate "Visit 5: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*time 1 0 0 0
                     0 0 0 0
                     -1 0 0 0 /alpha=0.05 cl;
  ...
  estimate "Visit 8: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*time 0 0 0 0
                     0 0 1 0
                     0 0 -1 0 /alpha=0.05 cl;
  lsmeans trt /alpha=0.05 cl;
  lsmeans trt*time /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, e.g. the normality assumption) 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 anti-DGP IgA and IgG results along with the change from baseline values will also be described by summary statistics and tabulated by treatment

group. Anti-DGP IgA and IgG results will also be listed. Individual and mean anti-DGP IgA and IgG results, will also be presented graphically by time point and treatment group.

PP populations will be used for the analysis.

Attenuation of gluten-induced 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 8; 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, sex, site, time (week) and time point-by-treatment group interaction. Poisson distribution with log-link will be used for modelling the counts. The following model will be fitted:

$$\log(Y_{ijklm}) = \mu + \tau_j + \varphi_k + \alpha_l + \gamma_m + \lambda_{lm} + \eta_i + \varepsilon_{ijklm},$$

where

Y_{ijklm} is the total bowel movement count for subject *i* ($i = 1, \dots, n_l$) from site *j* ($j = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group l ($l = 1, 2, 3$) at visit m ($m = 3, 5, 7, 8$),
treatment group *l* ($l = 1, 2, 3$) at week *m* ($m = 0,1,\dots,16$),
μ is the overall mean on log-scale,
α_l is the fixed effect due to treatment *l* on log-scale,
φ_k is the fixed effect due to sex *k* on log-scale,
τ_j is the fixed effect due to site *j* on log-scale,γ_m is the fixed effect due to week *m* on log-scale,
λ_{lm} is the fixed interaction effect due to treatment *l* and week *m* on log-scale,
η_i is the random effect due to subject *i*,
ε_{ijklm} is the random error for subject *i* from treatment group *l* for week *m*.

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 sex site week(ref=FIRST);
  model bsfs = trt sex site week trt*week / solution dist=POISSON;
  random _residual_ / subject=subjid type=UN;
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl ilink;
  estimate "300 mg AMG 714 vs Placebo" trt 0 1 -1 /alpha=0.05 cl ilink;
  estimate "Week 0: 150 mg AMG 714 vs Placebo" trt 1 0 -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 0
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1
            /alpha=0.05 cl ilink;
  estimate "Week 0: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     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: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     0 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: 300 mg AMG 714 vs Placebo" trt 0 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 0 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. Individual and mean weekly bowel movement counts will also be presented graphically by time point and treatment group.

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

PP populations will be used for the analysis.

Attenuation of gluten-induced clinical symptoms: Gastrointestinal Symptom Rating Scale (GSRS)

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 Screening through the Final Study Visit (i.e. on Screening and Visits 1 to 9).

The total GSRS score will be calculated as the sum 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 relative (%) 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, site, sex, 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:

$$Y_{ijklm} = \mu + \tau_j + \varphi_k + \alpha_l + \gamma_m + \lambda_{lm} + \beta \cdot baseline\ GSRS_{ijklm} + \eta_i + \varepsilon_{ijklm},$$

where

Y_{ijklm} is the relative (%) change from baseline in total GSRS score for subject i ($i = 1, \ldots, n_{jkl}$) from site j ($j = 1,2,3$), sex k ($k = 1, 2$) and treatment group l ($l = 1, 2, 3$) at week m ($m = 1, \ldots,16$),
μ is the overall mean,
α_l is the fixed effect due to treatment l,
φ_k is the fixed effect due to sex k,
τ_j is the fixed effect due to site j,
γ_m is the fixed effect due to week m,
λ_{lm} is the fixed interaction effect due to treatment l and week m,
β is the parameter estimate of the baseline total GSRS score,
η_i is the random effect due to subject i,

ε_{ijklm} is the random error for subject i from site j ($j = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group l for week m.

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 site sex week(ref=FIRST);
  model gsrs = gsrs_bl trt site sex week trt*week / solution;
  random _residual_ / subject=subjid type=UN;
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl;
  estimate "300 mg AMG 714 vs Placebo" trt 0 1 -1 /alpha=0.05 cl;
  estimate "Week 1: 150 mg AMG 714 vs Placebo" trt 1 0 -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
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 /alpha=0.05 cl;
  estimate "Week 1: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 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 -1 /alpha=0.05 cl;
  estimate "Week 2: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     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 /alpha=0.05 cl;
  ...
  estimate "Week 16: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 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 -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, e.g. the normality assumption) 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.

As a secondary approach, the same model will be fitted using the absolute change in GSRS as a response variable. 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 (both relative and absolute) 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 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 (both relative and absolute) 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. Individual and mean total GSRS scores as well as the scores of the individual questions, will also be presented graphically by time point and treatment group.

PP populations will be used for the analysis.

Attenuation of gluten-induced 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 sum 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 absolute change from baseline in CeD-GSRS will be analysed using a linear mixed effects repeated measures model (MMRM) with the baseline value, treatment group, site, sex, 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:

$$Y_{ijklm} = \mu + \tau_j + \varphi_k + \alpha_l + \gamma_m + \lambda_{lm} + \beta \cdot baseline\ CeD\ GSRS_{ijklm} + \eta_i + \varepsilon_{ijklm},$$

where

Y_{ijklm} is the absolute change from baseline in total CeD-GSRS score for subject i ($i = 1, \dots, n_{jkl}$) from site j ($j = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group l ($l = 1, 2, 3$) at week m ($m = 1, \dots, 16$),

μ is the overall mean,

α_l is the fixed effect due to treatment l,
φ_k is the fixed effect due to sex k,
τ_j is the fixed effect due to site j,
γ_m is the fixed effect due to time point m,
λ_{lm} is the fixed interaction effect due to treatment l and time point m,
β is the parameter estimate of the baseline total CeD-GSRS score,
η_i is the random effect due to subject i,
ε_{ijklm} is the random error for subject i from site j ($j = 1, 2, 3$), sex k ($k = 1, 2$) and treatment group l for time point m.

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 site sex week(ref=FIRST);
  model ced_gsrs = ced_gsrs_bl trt site sex week trt*week / solution;
  random _residual_ / subject=subjid type=UN;
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl;
  estimate "300 mg AMG 714 vs Placebo" trt 0 1 -1 /alpha=0.05 cl;
  estimate "Week 1: 150 mg AMG 714 vs Placebo" trt 1 0 -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
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 /alpha=0.05 cl;
  estimate "Week 1: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 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 -1 /alpha=0.05 cl;
  estimate "Week 2: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     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 /alpha=0.05 cl;
  ...
  estimate "Week 16: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 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 -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, e.g. the normality assumption) 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. Individual and mean total weekly CeD-GSRS will also be plotted by time point and treatment group.

PP populations will be used for the analysis.

7.7 Exploratory analysis

The exploratory endpoints of the study are:

- PGA;
- Biomarkers of disease activity;
- PK, and PK/PD correlations;
- CeD PRO;
- Assessment of the gluten-free diet (iVYLISA GIP stool test and urine gluten test).

7.7.1 Physician Global Assessment of Disease (PGA)

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. Assessments should be made prior to study drug administration using all tools available to the assessor, including laboratory test values.

PGA will be tabulated by treatment group and visit. Change from baseline will also be presented by treatment group and visit. All PGA results will also be listed. Individual PGA results will additionally be plotted by visit and treatment group.

PP population will be used for summarizing PGA results.

7.7.2 Biomarkers of disease activity

Several biomarkers of disease activity will be analysed in serum and in biopsy tissue at the time points specified in Appendix 1. The biomarkers of interest are:

- Serum IL-15 (pg/mL);
- Serum CRP (mg/L).

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

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. Individual and mean curves of biomarkers of disease activity will also be plotted by visit and treatment group, if applicable.

PP analysis population will be used for summarizing biomarker results.

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

PK data will be tabulated by treatment group. In addition, dose proportionality, achievement of steady-state, and accumulation ratio based on C_{trough} concentrations, comparison of C_{trough} levels with corresponding values in study 20060349, and correlation of C_{trough} concentrations with biomarkers of disease activity will be evaluated. PK analyses will be further described in the PK analysis plan. PK/PD association will mainly be assessed graphically.

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 groups. Levels of AMG 714 in serum will also be listed and both the individual and mean results plotted by time point.

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 may be explored:

- Primary endpoint (VH:CD);
- Select secondary endpoints (IELs density);
- Select PROs (CeD-PRO);
- Select biomarkers of disease activity (e.g., 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.

7.7.4 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. 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, site, sex, 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:

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

where

Y_{ijklm} is the total CeD PRO score for subject i ($i = 1, \dots, n_j$) from treatment group j ($j = 1, 2, 3$), sex l ($l = 1, 2$) and site m ($m = 1, 2, 3$) at day k ($k = 0, 1, 2, \dots, 112$),
μ is the overall mean,
α_j is the fixed effect due to treatment j,
φ_l is the fixed effect due to sex l,
τ_m is the fixed effect due to site m,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_{ijklm} is the random error for subject i from treatment group j, sex l and site m 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 site sex;
  model cedpro_d = cedpro_d_bl trt site sex day trt*day / solution;
  random _residual_ / subject=subjid type=AR(1);
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl;
  estimate "300 mg AMG 714 vs Placebo" trt 0 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:

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

where

Y_{ijklm} is the total CeD PRO score for subject i ($i = 1, \ldots, n_j$) from treatment group j ($j = 1, 2, 3$), sex l ($l = 1, 2$) and site m ($m = 1, 2, 3$) at week k ($k = 0, 1, \ldots, 16$),
μ is the overall mean,
α_j is the fixed effect due to treatment j,
φ_l is the fixed effect due to sex l,
τ_m is the fixed effect due to site m,
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_{ijklm} is the random error for subject i from treatment group j, sex l and site m 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 site sex week(ref=FIRST);
  model cedpro_w = cedpro_w_bl trt site sex week trt*week / solution;
  random _residual_ / subject=subjid type=AR(1);
  estimate "150 mg AMG 714 vs Placebo" trt 1 0 -1 /alpha=0.05 cl;
  estimate "300 mg AMG 714 vs Placebo" trt 0 1 -1 /alpha=0.05 cl;
  estimate "Week 0: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*week 0 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 0 0
                     0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 /alpha=0.05 cl;
  estimate "Week 0: 300 mg AMG 714 vs Placebo" trt 0 1 -1
            trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     0 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: 150 mg AMG 714 vs Placebo" trt 1 0 -1
            trt*week 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
                     0 0 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 0 /alpha=0.05 cl;
  ...
```

```
  estimate "Week 16: 300 mg AMG 714 vs Placebo" trt 0 1 -1
           trt*week 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 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 -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, e.g. the normality assumption) 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. Individual and 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.

PP populations will be used for summarizing CeD PRO results.

7.7.5 Assessment of the gluten-free diet (iVYLISA GIP stool test and urine gluten test)

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 the evaluation of a potential therapeutic effect of any experimental

celiac-related medication, since histologic, serologic and clinical endpoints are heavily influenced by the presence of gluten in the diet.

In addition to measuring symptoms through patient reported diaries, Celimmune plans to use the iVYLISA GIP-S stool and urine gluten tests, gluten assays developed to detect inadvertent gluten consumption by measuring gluten immunogenic peptides (GIP) in faeces (Comino *et al*, 2012) and urine (Moreno *et al*, 2015).

The stool test detects gluten for up to seven days after consumption, and the urine test for 1-2 days. Testing will be done every two weeks – subjects should provide a urine and optional stool sample collected up to three days before or after 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 stool test is considered negative when the average amount of gluten of a stool sample is <300 ng GIP/g stool sample.

iVYLISA GIP stool test and urine gluten test results (positive/negative) will be presented by visit and treatment group as well as listed.

PP analysis sets will be used.

7.8 Analysis of safety and tolerability

7.8.1 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 to study drug 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) and gluten challenge (defined as before gluten challenge and on or after gluten challenge). In addition, the number of events and their proportion (%) of the total number of events can 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.

7.8.2 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 2 Clinical laboratory tests

Haematology	Clinical Chemistry	Urinalysis
Basophils (% and absolute) Eosinophils (% and absolute) Haematocrit Hemoglobin Lymphocytes (% and absolute) Monocytes (% and absolute) Neutrophils (% and absolute) Platelet Count Red Blood Cell (RBC) count White Blood Cell (WBC) count	Alanine Aminotransferase (ALT) Albumin Alkaline Phosphatase Aspartate Aminotransferase (AST) Bilirubin (Total) Blood Urea Nitrogen (BUN) Calcium Chloride Creatinine Glucose Lactate dehydrogenase (LDH) Potassium Protein (Total) Sodium	Blood Cells (Erythrocytes, Leukocytes) Glucose Ketones Microscopic evaluation Protein/Albumin

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 Specimen Collection and Processing Manual or equivalent document supplied by the Sponsor.

Laboratory test values 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 individual and mean curves. ITT analysis population will be used.

7.8.3 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.

7.8.4 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. Individual and 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.

7.8.5 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.

7.8.6 Other safety variables

7.8.6.1 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.

7.8.6.2 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.

7.9 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 (insufficient or excessive gluten consumption) based on serial iVYLISA GIP testing;
- Sex;
- Site;
- Baseline histology and serology;
- Expression of certain biomarkers at baseline, such as CD122, Granzyme B and IL-21R in IELs, or IL-15 in serum.

These additional sensitivity analyses may be carried out only for 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, 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.

7.10 Execution of statistical analyses

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

8 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.

9 References

1. Brusca I. Overview of biomarkers for diagnosis and monitoring of celiac disease. *Adv Clin Chem*. 2015;68:1-55.
2. 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.
3. 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.
4. 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.
5. 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.
6. Oberhuber G. Histopathology of celiac disease. *Biomed Pharmacother*. 2000; 54(7):368-72.
7. 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.
8. SAS, Institute Inc., Cary, NC, USA.
9. 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

10 Appendix

Appendix 1. Study schedule

Appendix 2. List of tables and figures

Appendix 3. List of subject data listings

Appendix 4. Table templates

11 Document history

Version number	Version date	Status	Author
1.0	19APR2017	Final	Marju Valge