

PATIENT REPORTED OUTCOME STATISTICAL ANALYSIS PLAN

Version 2.0

02-MAY-2023

A Phase 3, Global, Multi-Center, Double-Blind, Randomized, Efficacy Study of
Zolbetuximab (IMAB362) Plus mFOLFOX6 Compared with Placebo Plus mFOLFOX6 as
First-line Treatment of Subjects with Claudin (CLDN)18.2-Positive, HER2-Negative,
Locally Advanced Unresectable or Metastatic Gastric or Gastroesophageal Junction (GEJ)
Adenocarcinoma
Spotlight

ISN/Protocol 8951-CL-0301

Sponsor:
Astellas Pharma Global Development, Inc. (APGD)
1 Astellas Way
Northbrook, IL 60062

This document contains confidential information which is the intellectual property of Astellas. By accepting or reviewing this document, you agree to hold this information in confidence and not copy or disclose it to others or use it for unauthorized purposes except (1) as otherwise agreed to in writing; (2) where required by applicable law; (3) where disclosure is directly related to the care and safety of the research participant; and (4) where disclosure of such information is made to a member of the investigator's team who agrees to hold this information in confidence.

Table of Contents

1	INTRODUCTION	5
1.1	Objectives of the COA Analysis	5
1.2	Study Objectives	6
1.2.1	Primary Objective	6
1.2.2	Secondary Objectives	6
1.2.3	Exploratory Objectives	6
1.3	Study Design	6
1.4	COA Instruments	8
1.4.1	EORTC QLQ-C30	9
1.4.2	EORTC QLQ-OG25	9
1.4.3	GP	9
1.4.4	EQ-5D-5L	9
1.4.5	Assessment Schedule	9
2	ANALYSIS SETS	10
3	ANALYSIS VARIABLES	10
3.1	General Variables and Derivations	10
3.1.1	Study Day	10
3.1.2	Baseline	10
3.1.3	Derived Timepoints	11
3.1.4	Other Derivations	11
3.2	PRO Variables	11
3.2.1	Variables generated from QLQ-C30	11
3.2.2	Variables Generated From QLQ-OG25	14
3.2.3	Variables Generated From GP	16
3.2.4	Variables Generated From EQ-5D-5L	17
3.3	Time to PRO Deterioration	18
3.3.1	Time To First Clinically Meaningful Deterioration (TTFD)	18
3.3.2	Time To First Confirmed Clinically Meaningful Deterioration (TTFCD)	19
3.3.3	Time To Definitive Clinically Meaningful Deterioration (TTDD)	20
3.3.4	Summary Of Time to PRO Deterioration Analysis	20
3.4	Subgroups of Interest	21
4	STATISTICAL ANALYSES	21
4.1	General Considerations	21

4.2	Handling of Missing Data	21
4.2.1	Missing Items	21
4.2.2	Missing Forms	21
4.3	Study Participants	22
4.3.1	Participant Disposition	22
4.3.2	PRO Completion	22
4.3.3	Demographic and Other Baseline Characteristics	22
4.4	Descriptive Analyses	23
4.4.1	Item Level	23
4.4.2	Domain and Overall Score Level	23
4.5	Longitudinal Analysis of Change From Baseline	23
4.5.1	Mixed Model Repeated Measures	23
4.6	Responder Analysis	25
4.7	Time to Event Analysis	25
4.7.1	Sensitivity Analyses	26
4.8	Subgroup Analyses	27
5	SUPPORTING DOCUMENTATION	28
5.1	Appendix A List of Abbreviations	28
5.2	Appendix B QLQ-C30	29
5.3	Appendix C QLQ-OG25	33
5.4	Appendix D GP	36
5.5	Appendix E EQ-5D-5L	37
6	REFERENCES	39
7	SIGNATURE	41

PRO SAP Version History Summary

The changes from the prior approved SAP that impact analyses are listed with the rationale in the table below.

PRO SAP Version	Approval Date	SAP Section(s)	Change	Rationale
1.0	09-MAR-2022		Not Applicable	Original Version
2.0	02-MAY-2023	1.4.1 3.1.4 3.2.1 3.2.2 3.2.4 4.5 4.6 4.7 6 and pg 16, 18, and 19 Appendix F	Added note on EORTC QLQ-C30 recall period Removed analysis visit window section EORTC QLQ-C30 threshold correction EORTC QLQ-OG25 threshold correction Updated EQ-5D-5L mapping function Updated MMRM Updated responder analyses Retained and specified one-sided test for the 3 key scores in time to event analyses Minor corrections to references Deleted	1.4.1 Correction 3.1.4 and 4.7 Alignment with the clinical SAP 3.2.1 and 3.2.2 Update to the thresholds 3.2.4 Updated EQ-5D-5L utility index mapping function based on NICE recommendation 4.5 and 4.6 Updated based on sample size and predictions at Cycle 9 Day 1 6 Reference corrections Appendix F Removed based on section 3.2.4 edits

1 INTRODUCTION

This document describes the rules and conventions to be used in the presentation and analysis of clinical outcome assessment (COA) data for Protocol 8951-CL-0301. It describes the data to be summarized and analyzed, including specifics of the statistical analyses to be performed.

This statistical analysis plan (SAP) is based on protocol version 6.0 dated 29 Sep 2021 and amendment 5.

This plan may be revised during the study to accommodate protocol amendments and/or to make changes to adapt to unexpected issues in study execution and/or data that affect planned analyses. The final plan, if revised, will document all changes and be issued prior to database lock.

1.1 Objectives of the COA Analysis

The aim of this project is to perform in-depth statistical analyses on the patient-reported outcomes (PROs) data collected in the 8951-CL-0301 study. This is to address protocol-specific objective COA-related as specified in [section 3.1](#).

The objective of the COA efficacy analysis is to enhance the understanding of the benefits of zolbetuximab when combined with mFOLFOX6 compared to placebo and mFOLFOX6 on disease-related symptoms, pain, and health related quality of life (HRQoL) in patients with Claudin (CLDN) 18.2-positive, HER2-negative, locally advanced unresectable or metastatic gastric or Gastroesophageal Junction (GEJ) adenocarcinoma. Treatment effects on symptoms, pain, HRQoL, function, and health status as measured by the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire EORTC QLQ-C30, EORTC Quality of Life Questionnaire - Oesophago-Gastric Module EORTC QLQ-OG25, Global Pain (GP), and the EuroQoL 5 dimension-5 level (EQ-5D-5L) instruments will be further evaluated by examining:

- Longitudinal change from baseline
- Time to first clinically meaningful deterioration (TTFD) in symptoms prior to disease progression
- Time to confirmed clinically meaningful deterioration (TTCD) in symptoms prior to disease progression
- Time to definitive clinically meaningful deterioration (TTDD) in symptoms prior to disease progression
- Proportion of patients with improvement/no change/deterioration in symptoms and HRQoL

Analysis will be performed for all scales of the above PRO instruments, except for certain analysis on physical functioning (PF) and Global Health Status (GHS)/Quality of Life (QoL) from EORTC QLQ-C30, and the EORTC QLQ-OG25 Pain which are analyzed in the clinical

SAP (Version 2.0, dated 18Nov2021). These exceptions will be specified in the corresponding sections.

1.2 Study Objectives

1.2.1 Primary Objective

The primary objective is to evaluate the efficacy of zolbetuximab plus mFOLFOX6 compared with placebo plus mFOLFOX6 (as first-line treatment) as measured by progression free survival (PFS) in subjects with CLDN 18.2-positive, HER2-negative locally advanced unresectable or metastatic gastric and GEJ adenocarcinoma.

1.2.2 Secondary Objectives

The secondary objectives are:

- To evaluate efficacy as measured by overall survival (OS) as a key secondary objective
- To evaluate the physical function (PF), OG25-Pain and GHS/QoL scores as measured by EORTC as a key secondary objective
- To evaluate efficacy as measured by objective response rate (ORR)
- To evaluate efficacy as measured by duration of response (DOR)
- To evaluate safety and tolerability of zolbetuximab
- To further evaluate other health related quality of life (HRQoL) using additional parameters as measured by EORTC QLQ-C30, QLQ-OG25, Global Pain (GP), and the EuroQoL 5 dimension-5 level (EQ-5D-5L) questionnaires
- To evaluate the pharmacokinetics of zolbetuximab
- To evaluate the immunogenicity profile of zolbetuximab.

1.2.3 Exploratory Objectives

The exploratory objectives are:

- To evaluate efficacy as measured by time to progression (TTP)
- To evaluate PFS following subsequent anti-cancer treatment (PFS2)
- To evaluate disease control rate (DCR)
- To evaluate potential genomic and /or other biomarkers that may correlate with treatment outcome to zolbetuximab and mFOLFOX6
- To evaluate health resource utilization (HRU).

1.3 Study Design

This is a global, multi-center, double-blind, 1:1 randomized, phase 3 study evaluating the efficacy of zolbetuximab plus mFOLFOX6 versus placebo plus mFOLFOX6 as first-line

treatment in subjects with CLDN 18.2-positive, HER2-negative locally advanced unresectable or metastatic gastric and GEJ adenocarcinoma.

Approximately 550 subjects will be 1:1 randomized into 1 of 2 treatment arms:

- Arm A (mFOLFOX6 chemotherapy in combination with zolbetuximab)
- Arm B (mFOLFOX6 chemotherapy in combination with placebo).

The randomization will be stratified by:

- Region (Asia vs Non-Asia)
- Number of metastatic sites (0 to 2 vs ≥ 3)
- Prior gastrectomy (Yes or No).

Subjects will be treated with either zolbetuximab (Arm A) or placebo (Arm B) on Days 1 and 22 starting at Cycle 1 Day 1 (C1D1) until the subject meets study treatment discontinuation criteria. Subjects will also receive 12 treatments of mFOLFOX6 (or components of mFOLFOX6 if some components are discontinued due to toxicity) over 4 cycles (1 cycle = 42 days) on Days 1, 15, and 29 of each cycle. After 12 mFOLFOX6 treatments, subjects may continue to receive 5-FU (fluorouracil) and folinic acid on Days 1, 15 and 29 of each cycle at the investigator's discretion until the subject meets study treatment discontinuation criteria.

If a subject discontinues mFOLFOX6 (or its components) due to any reason other than disease progression as confirmed by independent review committee (IRC), they may continue on zolbetuximab/placebo and continue to follow the study treatment period schedule of assessments at the discretion of the investigator provided that the following have been met:

- the subject completed at least 1 cycle (42 days) of mFOLFOX6 treatment;
- the subject will not receive another systemic chemotherapy, immunotherapy, radiotherapy or other treatment intended for antitumor activity; and
- in the investigator's opinion the subject continues to derive clinical benefit with acceptable toxicity.

Following discontinuation from zolbetuximab/placebo, subjects will have a study treatment discontinuation visit, and 30-day and 90-day safety follow-up visits following their last dose of zolbetuximab/placebo. Additionally, if mFOLFOX6 (all components) is discontinued on a different day than zolbetuximab/placebo, subjects will also have a study treatment discontinuation visit, and 30-day and 90-day safety follow-up visits following the last dose of mFOLFOX6.

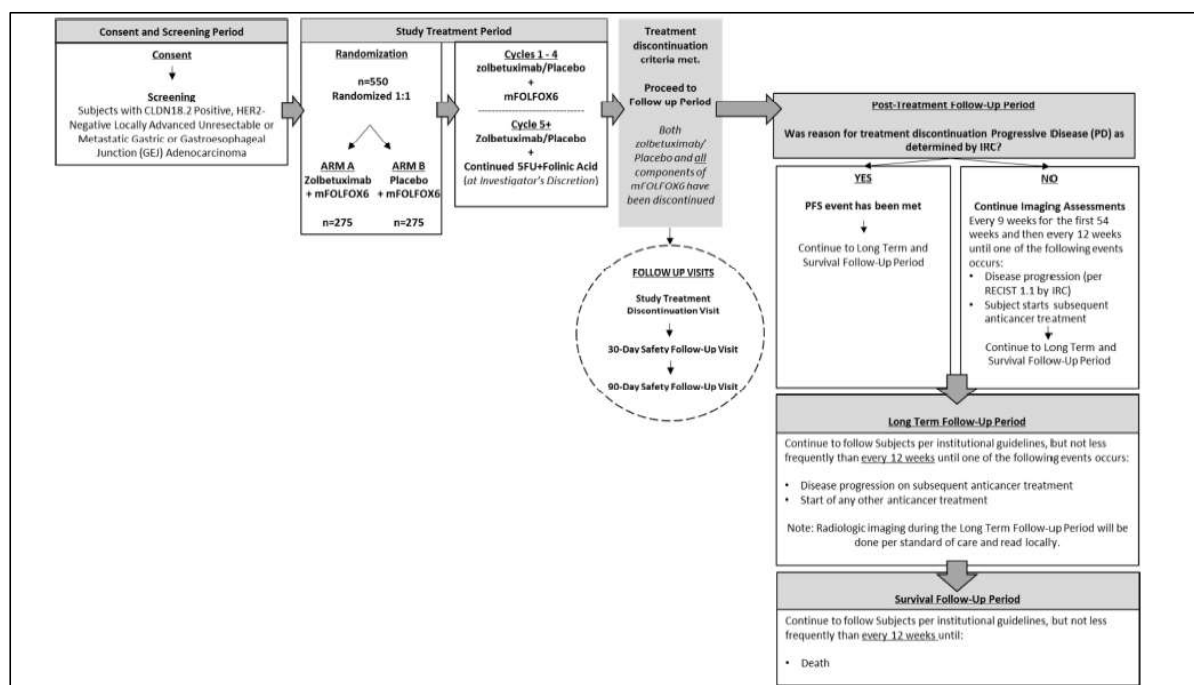
If a subject discontinues all study treatments (zolbetuximab/placebo and all components of mFOLFOX6) prior to disease progression, the subject will enter the post-treatment follow-up period and continue to undergo scheduled imaging assessments every 9 weeks until radiologic disease progression or until the subject starts any other anti-cancer treatment, whichever occurs earlier.

If study treatment (zolbetuximab/placebo and all components of mFOLFOX6) is discontinued due to disease progression, the subject will enter the long-term and survival follow-up period. Survival follow-up period will continue until death from any cause.

HRQoL and HRU will be assessed during the visit (or up to 48 hours) before any antiemetic or drug treatment(s) administration and before the disease status is discussed with the subject. Assessments will be collected at Screening (except HRU), every 3 weeks, at study treatment discontinuation and 30 and 90 days post-zolbetuximab/placebo treatment.

The study schema is provided in Figure 1.

Figure 1 Study Schema



Notes: 5-FU: fluorouracil; CLDN: Claudin; HER2: human epidermal growth factor receptor 2; IRC: independent review committee; mFOLFOX6: 5-fluorouracil, folinic acid and oxaliplatin; RECIST: Response Evaluation Criteria in Solid Tumors.

1.4 COA Instruments

Patient-reported outcomes will be assessed through four different instruments:

- European Organization for Research and Treatment of Cancer Quality of Life Core 30: EORTC QLQ-C30
- EORTC Quality of Life Questionnaire - Oesophago-Gastric Module: EORTC QLQ-OG25
- Global Pain (GP)
- EuroQoL 5 dimension-5 level (EQ-5D-5L).

1.4.1 EORTC QLQ-C30

The EORTC QLQ-C30 instrument ([Aaronson et al, 1993](#)) is a generic questionnaire consisting of 30 items developed to assess symptoms and functioning of cancer patients. The instrument yields 5 functional scales, 4 symptom scales, 1 global health status (GHS) / quality of life (QoL) scale and 1 financial impact score. Most items are scored 1 (“not at all”) to 4 (“very much”) except for the items contributing to the global health status/QoL, which are scored 1 (“very poor”) to 7 (“excellent”). The recall period for each question is “during the past week” (note a recall period is not indicated for questions 1-5). An outcome variable consisting of a score from 0 to 100 is derived for each of the scales. Higher scores on symptoms indicate a worse health state. Higher scores on the global health status and functioning scales indicate better health status/function.

The QLQ-C30 is presented in [Appendix B QLQ-C30](#).

1.4.2 EORTC QLQ-OG25

The QLQ-OG25 is a 25-item instrument that evaluates gastric and GEJ cancer-specific symptoms such as stomach discomfort, difficulties eating and swallowing, and indigestion. This module consists of 6 scales: dysphagia, eating restrictions, reflux, odynophagia, pain and discomfort, and anxiety, as well as 10 single items. Each item is rated on a four-point Likert scale ranging from 1 = “not at all” to 4 = “very much”.

The QLQ-OG25 is presented in [Appendix C QLQ-OG25](#).

1.4.3 GP

The GP is a single assessment of overall pain on a scale from 0 (no pain) to 10 (pain as bad as you can imagine).

The GP is presented in [Appendix D GP](#).

1.4.4 EQ-5D-5L

The EQ-5D-5L is a quality of life (QoL) instrument for self-reported assessment of 5 domains of health: mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each domain is rated by selecting 1 of 5 standardized categorizations ranging from ‘no problem’ to ‘extreme problem’. The final question is a visual analogue scale (VAS) to rank health status from best health imaginable (100) to worst health imaginable (0).

The EQ-5D-5L is presented in [Appendix E EQ-5D-5L](#) and Appendix F Utility Index calculation (EQ-5D-5L).

1.4.5 Assessment Schedule

The assessment schedule for the PRO instruments is provided in [Table 1](#).

Table 1 PRO Instruments Assessment Schedule

Study Day	Screening	Study Treatment Period (Each Cycle = 42 days)								Follow-up Period ^b				
		Cycles 1 to 4 Zolbetuximab/Placebo + mFOLFOX6				Cycle 5+ Zolbetuximab/Placebo + 5FU+Folinic Acid				Study Discontinuation	30-Day Follow-up Visit	90-Day Follow-up Visit	Post Treatment Follow-up	Long Term and Survival Follow-up ^b
		1	15	22	29	1	15	22	29					
Window (Days)	-45 to -1	0	+5	+5	+5	+5	+5	+5	+5	+7	+7	+7	±7	±14
PRO Assessments ^a	X	X		X		X		X		X	X	X		

^a PRO instruments are to be administered on IMAB362/placebo visit days before any antiemetic or drug treatment or other scheduled assessments are conducted and before the disease status is discussed with the subject.

^b If a subject discontinues all study treatments (IMAB362/placebo and all components of mFOLFOX6) prior to IRC confirmed disease progression, the subject will enter the post-treatment follow-up period and continue to undergo imaging assessment every 9 weeks (or every 12 weeks if subjects has been on study over 54 weeks) until radiologic disease progression (i.e., PFS) or the subject starts subsequent anti-cancer treatment, whichever occurs earlier. If study treatment (IMAB362/placebo and all components of mFOLFOX6) is discontinued due to PD, the subject will enter the Long term and survival follow-up period.

2 ANALYSIS SETS

All PRO analyses described in this SAP will be performed on the Full Analysis Set (FAS) as defined in the clinical SAP (Version 2.0, dated 18Nov2021), i.e. all subjects who are randomized to 1 of the treatment arms. All subjects will be analyzed as randomized (not by actual treatment received). FAS in this study is identical to intent-to-treat (ITT) set.

3 ANALYSIS VARIABLES

3.1 General Variables and Derivations

3.1.1 Study Day

Study day as defined in the clinical SAP (Version 2.0, dated 18Nov2021) will be used. The study day will be calculated in reference to the date of the first dose of study drug. Treatment Day 1 corresponds to the date the subject received the first dose of study drug. For assessments conducted on or after the date of the first dose of study drug, study day will be calculated as (assessment date - date of first dose of study drug) + 1.

3.1.2 Baseline

The baseline measurement is the last measurement taken prior to initial study drug administration (i.e., Cycle 1 Day 1 pre-dose assessment). Both date and time of drug administration and measurement should be considered to identify the baseline value. If the

time is not available, then only the date will be used, and it will be assumed that assessments on day 1 were administered prior to dosing.

The post-baseline value is defined as a measurement taken after initial study drug administration.

Change from baseline is defined as (post baseline value - baseline value).

3.1.3 Derived Timepoints

All timepoints will be used as in the ADaM datasets received by the sponsor.

3.1.4 Other Derivations

No other derivations besides PRO endpoints will be described in this analysis plan.

3.2 PRO Variables

3.2.1 Variables generated from QLQ-C30

The EORTC QLQ-C30 scale scores will be calculated using the EORTC QLQ-C30 Scoring Manual ([Fayers et al, 2001](#)). The instrument yields the following scales: Global health status/Quality of Life (QL2; 2 items; score range 0-100).

- Functional scales:
 - Physical functioning (PF2; 5 items; score range 0-100)
 - Role functioning (RF2; 2 items; score range 0-100)
 - Emotional functioning (EF; 4 items; score range 0-100)
 - Cognitive functioning (CF; 2 items; score range 0-100)
 - Social functioning (SF; 2 items; score range 0-100)
- Symptom scales/items:
 - Fatigue (3 items; score range 0-100)
 - Nausea and vomiting (2 items; score range 0-100)
 - Pain (2 items; score range 0-100)
 - Dyspnea (1 item; score range 0-100)
 - Insomnia (1 item; score range 0-100)
 - Appetite loss (1 item; score range 0-100)
 - Constipation (1 item; score range 0-100)
 - Diarrhea (1 item; score range 0-100)
- Financial difficulties (1 item; score range 0-100)

Although not included in the original scoring manual ([Fayers et al, 2001](#)), it has been suggested in the literature that a single summary score can also be calculated for this instrument ([Giesinger et al, 2016](#)):

- Physical Functioning + Role Functioning + Social Functioning + Emotional Functioning + Cognitive Functioning + (100-Fatigue) + (100-Pain) + 100-(Nausea and Vomiting) + (100-Dyspnoea) + (100-Sleeping Disturbances) + (100-Appetite Loss) + (100-Constipation) + (100-Diarrhoea))/13.

The QLQ-C30 domain and single-item scores will be calculated according to the EORTC scoring manual presented in [Appendix B QLQ-C30](#). The principle for scoring is the same for all scales. Briefly, outcome scores are computed by standardizing the average of the items (i.e., a raw score) making up the scale. Outcome scores are computed using a linear transformation of the raw score such that scores range from 0 to 100. A higher score represents a higher ("better") level of functioning, or a higher ("worse") level of symptoms, i.e., a high score for a functional scale represents a high/healthy level of functioning, a high score for the GHS represents a high QoL, but a high score for a symptom scale/item represents a high level of symptomatology/problems. Note that the global health status scale is based on only the 2 specific HRQoL items and not the entire questionnaire.

If at least half the items of a scale are present for a timepoint then the score will be calculated using the average of all items answered; otherwise the score will be set to missing. For single measures, if the item is missing the scale score is set to missing.

Physical and role functioning, as well as the global QoL revised scales are those that have been changed since version 1.0, and their short names are indicated a suffix "2" – for example, PF2, according to the instrument's manual ([Fayers et al, 2001](#)).

All the below definitions apply to all domains, except for the FI that will not be analyzed.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment for each scale and item.

Post-baseline item scores will be classified as according to the following item response categories:

- Worsening 3 points compared to baseline
- Worsening 2 points compared to baseline
- Worsening 1 point compared to baseline
- Stable
- Improved 1 point compared to baseline
- Improved 2 points compared to baseline
- Improved 3 points compared to baseline

Change in symptoms/functioning/global health status from baseline will be categorized as improvement/stable/deterioration using threshold values for change scores that connote clinically meaningful changes for patients.

Two sets of values will be used. For QL2 and PF, the primary threshold for deterioration will be derived by anchor-based analysis of the trial's data before data base lock occurs, as described in the 8951-CL-0303 study analysis plan (Version 1.0, dated 09Mar2022). For the rest of the domains, the primary threshold values will be based on values developed by [Cocks et al. 2012](#) and shown in [Table 3](#).

Table 2 Primary responder definition for EORTC QLQ-C30

Score	Deterioration	Stable	Improvement
Global health status/QoL	< -10	-10 to +8	>+8
Financial difficulties*	>+10	-10 to +3	<-3
Functional scales			
Physical functioning	< -10	-10 to +7	>+7
Role functioning	<-14	-14 to +12	>+12
Cognitive functioning	<-7	-7 to +7	>+7
Emotional functioning	<-12	-12 to +9	>+9
Social functioning	<-11	-11 to +8	>+8
Symptom scales			
Fatigue	>+10	+10 to -9	<-9
Pain	>+11	+11 to -9	<-9
Nausea and vomiting	>+11	+11 to -9	<-9
Diarrhea	>+15	+15 to -11	<-11
Constipation	>+15	+15 to -10	<-10
Appetite loss	>+14	+14 to -13	<-13
Dyspnea	>+11	+11 to -9	<-9
Insomnia	>+9	+9 to -9	<-9

*FI will not be analyzed in this analysis plan.

For the sensitivity threshold as there were no other threshold values connoting clinically meaningful within-patient change identified in the literature, it will be based on the next highest value that will provide a different classification for deterioration and improvement to the primary one. The reason of this approach is that EORTC values are discrete in nature due to the transformation of raw scores to 0-100, therefore not all values are possible and certain threshold values will result in the same categorization for patients, e.g., for a single-item domain such as appetite loss, the possible values are 0, 33.3, 66.7 and 100, therefore a applying a threshold of 15 and 30 will result in the same classification of patients into responder categories. The sensitivity values for all domains are shown in [Table 4](#).

Table 3 Sensitivity responder Definition for EORTC QLQ-C30

Score	Deterioration	Stable	Improvement
Global health status/QoL	< -17	-17 to +8	>+9
Financial difficulties*	>+34	-34 to +34	<-34
Functional scales			
Physical functioning	< -14	-14 to +14	>+14
Role functioning	<-17	-17 to +17	>+17
Cognitive functioning	<-17	-17 to +17	>+17
Emotional functioning	<-17	-17 to +17	>+17
Social functioning	<-17	-17 to +17	>+17
Symptom scales			
Fatigue	>+12	+12 to -12	<-12
Pain	>+17	+17 to -17	<-17
Nausea and vomiting	>+17	+17 to -17	<-17
Diarrhea	>+34	+34 to -34	<-34
Constipation	>+34	+34 to -34	<-34
Appetite loss	>+34	+34 to -34	<-34
Dyspnea	>+34	+34 to -34	<-34
Insomnia	>+34	+34 to -34	<-34

3.2.2 Variables Generated From QLQ-OG25

There will be six domains and ten single-item scores calculated for QLQ-OG25:

- Domain Scale Scores
 - Dysphagia (3 items; score range 0-100)
 - Eating restrictions (4 items; score range 0-100)
 - Reflux (2 items; score range 0-100)
 - Odynophagia (2 items; score range 0-100)
 - Pain and discomfort (2 items; score range 0-100)
 - Anxiety (2 items; score range 0-100)
- Single Items Scores
 - Eating in front of others (1 item; score range 0-100)
 - Dry mouth (1 item; score range 0-100)
 - Trouble with taste (1 item; score range 0-100)
 - Body image (1 item; score range 0-100)
 - Trouble swallowing saliva (1 item; score range 0-100)
 - Choked when swallowing (1 item; score range 0-100)

- Trouble with coughing (1 item; score range 0-100)
- Trouble talking (1 item; score range 0-100)
- Weight loss (1 item; score range 0-100)
- Hair loss (1 item; score range 0-100)

The QLQ-OG25 domain and single-item scores will be calculated according to the EORTC scoring manual presented in [Appendix C QLQ-OG25](#). Raw scores will be transformed into a linear scale ranging from 0 to 100, with a high score for representing a higher (“worse”) level of symptomatology/problems.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment for each scale and item.

Post-baseline item scores will be classified as according to the same item response categories described in [section 3.1.2](#).

Post-baseline multi-item and single-item scale scores will be classified as improvement/no change/deterioration according to baseline scores at each assessment using a threshold to connote important change to subjects as follows:

- Improvement: a change from baseline \leq -threshold points;
- Deterioration: a change from baseline \geq threshold points
- No change: a change from baseline between (–threshold to threshold).

Literature reporting clinically important thresholds were not found. Two sets of values will be used. The primary threshold values will be obtained using the baseline distributions of QLQ-OG25 scores (e.g., a change equal to or greater than one-half the baseline standard deviation (SD) pooled over the two treatment arms) ([Norman et al 2003](#), [Sloan et al 2005](#)) or a 1-category change in 1 item in the scale, whichever is larger. Sensitivity analysis will be performed using the next higher threshold that results in a different classification of patients into improvement, stable and deterioration categories as compared to the primary threshold, as for EORTC QLQ-C30.

The thresholds to be used for QLQ-OG25 scores are presented in [Table 5](#). Number in parenthesis for primary threshold represents how change in 1 item corresponds to the 0-100 score of the scale.

Table 4 Predefined threshold values for QLQ-OG25

QLQ-OG25 outcome	Primary threshold	Sensitivity threshold
<ul style="list-style-type: none"> • Dysphagia (3 items) • Eating restrictions (4 items) • Reflux (2 items) • Odynophagia (2 items) • Pain and discomfort (2 items) • Anxiety (2 items) • Eating in front of others (1 item) 	<ul style="list-style-type: none"> • 4-item scale: max (pooled baseline ½ SD, 8.3) • 3-item scale: max (pooled baseline ½ SD, 11.1) • 2-item scale: max (pooled baseline ½ SD, 16.7) • 1-item scale: max (pooled baseline ½ SD, 33.3) 	Next higher threshold that results in a different classification as compared to the primary threshold

QLQ-OG25 outcome	Primary threshold	Sensitivity threshold
<ul style="list-style-type: none"> • Dry mouth (1 item) • Trouble with taste (1 item) • Body image (1 item) • Trouble swallowing saliva (1 item) • Choked when swallowing (1 item) • Trouble with coughing (1 item) • Trouble talking (1 item) • Weight loss (1 item) • Hair loss (1 item) 		

3.2.3 Variables Generated From GP

The GP is a single assessment of overall pain on a scale ranging from 0 (no pain) to 10 (pain as bad as you can imagine). The score will be the numeric endorsement on the scale.

A higher score indicates a higher degree of pain.

Pain severity at baseline is defined as described in [Table 6](#).

Table 5 Pain severity at baseline definition

Severity level	Definition
Asymptomatic subjects	Subjects with a score of 0 at the baseline visit
Mildly symptomatic subjects	Subjects with a score of 1 to 4 at the baseline visit
Moderate symptomatic subjects	Subjects with a score of 5 or 6 at the baseline visit
Severely symptomatic subjects	Subjects with a score of 7 to 10 at the baseline visit

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment for each scale and item.

Post-baseline GP score will be classified as improvement/no change/deterioration according to baseline scores at each assessment using a threshold to connote important change to subjects as follows:

- Improvement: a change from baseline \leq -threshold points;
- Deterioration: a change from baseline \geq threshold points
- No change: a change from baseline between (–threshold to threshold).

Cut-off values of 2-point or greater change on a 11-point numerical pain rating scale have been proposed in the literature to detect clinically important changes in adults ([Farrar, et al. 2000](#), [Kendrick, et al. 2005](#)). However, the use of these values in defining pain progression on the GP in subjects with CLDN18.2-positive, HER2-negative locally advanced unresectable or metastatic gastric and GEJ adenocarcinoma has not yet been validated. Therefore, a sensitivity analysis will be performed using a threshold obtained using the

baseline distributions of GP score (e.g., a change equal to or greater than one-half the baseline SD pooled over the two treatment arms) ([Norman et al 2003](#); [Sloan et al 2005](#)).

The thresholds to be used for GP scores are presented in [Table 7](#).

Table 6 Predefined thresholds for GP

GP outcome	Primary threshold	Sensitivity threshold
Pain item	2 points	Pooled baseline ½ SD

3.2.4 Variables Generated From EQ-5D-5L

A unique EQ-5D-5L health state is defined by combining 1 level from each of the 5 dimensions: this defines a profile that is primarily reported as a 5-digit number, for instance 11221. A total of 3125 possible health states are defined in this way. For example, state 11111 indicates no problems on any of the 5 dimensions, while state 12345 indicates no problems with mobility, slight problems with washing or dressing, moderate problems with doing usual activities, severe pain or discomfort and extreme anxiety or depression.

The instrument was specifically designed to provide an overall single number, called a weighted index, for each of the health states resulting from the combination of item responses ([Dolan 1997](#)). The weighted index constitutes a measure of utility, an economics term used to describe consumer preferences or in the present case patient preferences for different HRQoL states. The weighted index can be only derived from patients who have provided a complete 5-response profile. A higher index indicates better QoL.

The mapping (crosswalk) function developed by Hernández Alava et al ([2017](#)) will be used as UK's National Institute of Health and Care Excellence in a position statement ([NICE 2019](#)) does not recommend to use the EQ-5D-5L value set for England published by Devlin et al ([2017](#)) to derive utility values for their evidence submissions and according to the latest NICE health technology evaluations manual ([NICE 2022](#)): “The mapping function developed by the Decision Support Unit ([Hernández Alava et al 2017](#)), using the ‘EEPRU dataset’ ([Hernández Alava et al, 2020](#)), should be used for reference-case analyses.”

A higher score for the domains indicates worse QoL. In contrast, a higher score for EQ-5D VAS indicates better QoL.

Change from baseline defined as post-baseline value minus baseline value will be calculated for each assessment for the EQ-5D-5L VAS and utility index scores as well as for the individual domain items.

Post-baseline individual domain scores will be classified as according to the following item response categories:

- Worsening ≥ 3 points compared to baseline
- Worsening 2 points compared to baseline
- Worsening 1 point compared to baseline

- Stable
- Improved 1 point compared to baseline
- Improved 2 points compared to baseline
- Improved ≥ 3 points compared to baseline.

Post-baseline scores for EQ-5D-5L VAS and utility index will be classified as improvement/no change/deterioration according to baseline scores at each assessment using thresholds denoting clinically meaningful change to subjects as follows:

- Improvement: a change from baseline \geq threshold points;
- Deterioration: a change from baseline \leq - threshold points
- No change: a change from baseline between ($-\text{threshold}$ to $+\text{threshold}$).

Pickard et al (2007) reported a clinically meaningful change range of 7 to 10 for EQ-5D VAS score. The value 7 will be used in the primary analysis and the value of 10 will be used in a sensitivity analysis.

To be noted that no thresholds denoting clinically meaningful change to subjects have been reported in the literature for EQ-5D-5L utility index. Therefore, for each utility index score, a change threshold will be used based on the baseline distributions of EQ-5D-5L utility index score: a change equal to or greater than one-half the baseline SD (pooled over the two treatment arms) will be considered clinically meaningful, a definition that is well supported and accepted by the existing literature as a difference that is robust and likely significant to subjects (Norman et al, 2003; Sloan et al, 2005).

3.3 Time to PRO Deterioration

Time to clinically meaningful symptom worsening or HRQoL deterioration (PRO deterioration) will be analysed for each domain separately as appropriate and as described in section 4.7. For convenience, a generic term “time to first clinically meaningful deterioration” (TTFD) will be used both for symptom worsening and HRQoL deterioration, with an understanding of a specific meaning depending on the domain analysed. Three definitions of TTFD will be explored: time to first clinically meaningful deterioration (TTFD), time to first confirmed clinically meaningful deterioration (TTFCDD), and time to definitive clinically meaningful deterioration (TTDD).

3.3.1 Time To First Clinically Meaningful Deterioration (TTFD)

TTFD will be defined as the duration of time from the date of randomization to the date of the first deterioration in PRO scores of at least one threshold unit as compared to the baseline score.

For those subjects who experienced a first clinically meaningful deterioration, TTFD will be computed as follows and then converted to months:

TTFD = Date of assessment when first clinically meaningful deterioration of at least one threshold unit was observed – Date of randomization + 1

Patients who did not experience clinically meaningful deterioration prior to the end of follow-up, radiographic progression, or death (if not progressed before death) will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment, patients with no post-baseline assessments, or patients whose baseline scores do not allow for further deterioration will be censored at the date of randomization. Death or progression will not be considered deterioration events.

In addition, the above definition will be repeated where death (due to any cause) will be counted as a TTFD event if the subject does not experience PRO deterioration prior to death and where death occurs within 2 scheduled assessments (e.g., 42 days) after the last available PRO assessment; progression will not be considered deterioration event. Analyses will be also repeated by using the sensitivity clinically meaningful threshold as defined in [section 3.2](#).

3.3.2 Time To First Confirmed Clinically Meaningful Deterioration (TTFCD)

Time to first confirmed clinically meaningful deterioration (TTFCD) will be defined as the duration of time from the date of randomization to the date of the first clinically meaningful deterioration in PRO scores of at least one threshold unit as compared to the baseline score which is

- confirmed at the next consecutive scheduled visit or
- followed by drop out, resulting in monotone missing data.

For those patients who experienced a first confirmed clinically meaningful deterioration, TTFCD will be computed as follows and then converted to months:

$$\text{TTFCD} = \text{Date of assessment when first confirmed clinically meaningful deterioration was observed} - \text{Date of randomization} + 1$$

Patients who did not experience a confirmed clinically meaningful deterioration prior to the end of follow-up, radiographic progression, or death (if not progressed before death) will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment, patients with no post-baseline assessments, or patients whose baseline scores do not allow for further deterioration will be censored at the date of randomization. Death or progression will not be considered deterioration events.

In addition, the above definition will be repeated where death (due to any cause) will be counted as a TTFCD event if the subject does not experience PRO deterioration prior to death and where death occurs within 2 scheduled assessments (e.g., 42 days) after the last available PRO assessment; progression will not be considered deterioration event.

Analyses will be also repeated by using the sensitivity clinically meaningful threshold as defined in [section 3.2](#).

3.3.3 Time To Definitive Clinically Meaningful Deterioration (TTDD)

TTDD will be defined as the duration of time from the date of randomization to the date of the first deterioration in PRO scores of at least one threshold unit as compared to the baseline score if the deterioration of at least one threshold unit as compared to the baseline score is:

- also observed at all time points thereafter (e.g., after the first deterioration is observed) or
- the patient dropped out after deterioration, resulting in missing data.

For those patients who experienced a definitive meaningful deterioration, TTDD will be computed as follows and then converted to months:

TTDD = Date of assessment when definitive clinically meaningful deterioration was observed – Date of randomization + 1

Patients who did not experience definitive clinically meaningful deterioration prior to the end of follow-up, radiographic progression, or death (if not progressed before death) will be censored at the date of the last available PRO assessment (i.e., date of the last non-missing value). Patients with no baseline assessment, patients with no post-baseline assessments, or patients whose baseline scores do not allow for further deterioration will be censored at the date of randomization. Death or progression will not be considered deterioration events.

In addition, the TTTD definition will be repeated where death (due to any cause) will be counted as an event if the subject does not experience PRO deterioration prior to death and where death occurs within 2 scheduled assessments (e.g., 42 days) after the last available PRO assessment; progression will not be considered deterioration event.

Analyses will be also repeated by using the sensitivity clinically meaningful threshold as defined in [section 3.2](#) above.

3.3.4 Summary Of Time to PRO Deterioration Analysis

A summary of all the definitions that will be explored is provided in the following table:

Table 7 Summary of time to PRO deterioration analysis

Definition #	Description	Threshold	Death
1	First deterioration	Primary	Excluding
2	First deterioration	Primary	Including
3	First deterioration	Sensitivity	Excluding
4	First deterioration	Sensitivity	Including
5	Confirmed deterioration	Primary	Excluding
6	Confirmed deterioration	Primary	Including
7	Confirmed deterioration	Sensitivity	Excluding
8	Confirmed deterioration	Sensitivity	Including
9	Definitive deterioration	Primary	Excluding
10	Definitive deterioration	Primary	Including
11	Definitive deterioration	Sensitivity	Excluding
12	Definitive deterioration	Sensitivity	Including

3.4 Subgroups of Interest

The following subgroups may be considered:

1. Age category 1 (65 years or younger vs older than 65 years)
2. Region (Asia vs Non-Asia)

4 STATISTICAL ANALYSES

4.1 General Considerations

Continuous data will be described by the number of observations (N), the number of missing observations (Nmiss), mean, SD, median, first quartile (Q1), third quartile (Q3), minimum (min), maximum (max). Categorical data will be described by the number (n) and percentage (%) of patients in each category. Missing and invalid observations will be tabulated as separate categories. The calculation of proportions will not include the missing/invalid category.

For continuous data, the mean, median, Q1 and Q3 will be rounded to 1 additional decimal place compared to the original data. The SD will be rounded to 2 additional decimal places compared to the original data. Minimum and maximum will be displayed with the same accuracy as the original data.

For categorical data, percentages will be rounded to 2 decimal places.

For the PRO analyses, all summaries will be presented by treatment group, unless specified otherwise. Statistical comparisons will be made using two-sided tests at the $\alpha=0.05$ significance level unless stated otherwise. Due to the exploratory nature of the PRO analyses, adjustments for multiple comparisons will not be made.

Unless otherwise specified, all summaries will be presented by treatment group.

All data processing, summarization, and analyses will be performed using SAS Version 9.4 or higher (SAS Institute, North Carolina).

4.2 Handling of Missing Data

4.2.1 Missing Items

In case of missing items, the domain and total scores will be calculated as indicated in the scoring manuals for each of the PRO instruments.

4.2.2 Missing Forms

It is anticipated that the great majority of missing data in this study will have a monotone pattern, meaning that once a patient has missing data at one visit, data will be missing at all subsequent visits. There may be some small amount of intermittent (non-monotone) missing data (when patient skips intermediate visits but return for evaluations at subsequent visits).

The number and percentage of patients for each of the missing data patterns (no missing data, monotone missing data, and intermittent missing data) will be presented by treatment group.

Tabular summaries for the percentage of patients by the reason for discontinuation of study treatment, as well as for withdrawal from the study, will be provided in the clinical study report and will not be repeated herein. A plot of the mean score for the PRO scores over time by selected categories of discontinuation (including completers) will be provided. The reasons of discontinuation will be grouped as follows:

- Death
- Adverse event
- Disease relapse, lack of efficacy, and progressive disease
- Other.

4.3 Study Participants

4.3.1 Participant Disposition

The subject disposition by treatment group for all PRO assessment time-points (e.g., analysis visits) will be provided:

- The number of subjects with PRO assessment expected
- The number and % of subjects with PRO assessment not expected due to progression
- The number and % of subjects with PRO assessment not expected due to death
- The number and % of subjects with PRO assessment not expected due to other reasons.

A PRO assessment is expected as long as the subject is alive and on treatment.

The subject disposition by treatment group per analysis visit will also be provided graphically by means of a stacked bar chart.

4.3.2 PRO Completion

The PRO completion (unadjusted, i.e. over the FAS population) and compliance rates (adjusted, i.e. among expected patients at each assessment) will be presented in the clinical study report and will not be repeated here.

4.3.3 Demographic and Other Baseline Characteristics

An extensive list of demographics and baseline disease characteristics is presented in the clinical SAP (Version 2.0, dated 18Nov2021), therefore these will not be repeated here.

4.4 Descriptive Analyses

4.4.1 Item Level

For all items from all four instruments (QLQ-C30, QLQ-OG25, GP and EQ-5D-5L), the following will be provided:

- A table with the distribution of responses by treatment group at each assessment;
- A stacked column chart of the distribution of responses by treatment group at each assessment;
- A table with the distribution of change in response categories from baseline to each assessment by treatment group;
- A stacked column chart of the distribution of change in response categories from baseline to each assessment by treatment group.

4.4.2 Domain and Overall Score Level

Domain and total scores will be summarized for the PRO instruments in the clinical study report and will not be repeated here.

A cumulative distribution plot showing a continuous plot of the absolute change from baseline during the study for the PRO scores on the X-axis and the cumulative percent of subjects experiencing that change on the Y-axis will be presented for the first six post-baseline visits. The presentation will be restricted to the first 6 post-baseline assessments and the following scales:

- EORTC QLQ-C30: GHS/QoL and PF and OG25-Pain.

4.5 Longitudinal Analysis of Change From Baseline

4.5.1 Mixed Model Repeated Measures

Change from baseline in PRO domain (single or multi-item) and overall scores while on study drug treatment will be further analyzed using a restricted maximum likelihood (REML)-based repeated measures approach (MMRM – Mixed Model Repeated Measures) ([Brown & Prescott, 2006](#)).

The MMRM assumes that the missing observations are missing at random (MAR). That is, MMRM assumes that, given the statistical model and given the observed values of the outcome, missingness is independent of the unobserved values. A corollary is that MAR assumes that a subject's missing values can be estimated based on similar subjects who remained in the study. This infers that withdrawals (who may not receive study medication) have similar symptoms to some who continue to be treated. While MAR's assumption of the similarity of withdrawals and those who stay in the study may not be realistic for all subjects, MAR can be justified as not being biased to an important degree in favor of the zolbetuximab arm. Given that subjects tend to have poor efficacy scores before they withdraw, MAR will

tend to impute similarly poor symptoms for the missing values: MAR will to that extent reflect that withdrawals are “different” from those who stay in the study.

The primary objective of this analysis is to compare zolbetuximab versus placebo at Cycle 9 Day 1. The Cycle 9 Day 1 timepoint was selected to minimize the impact of missing data given that median rPFS for placebo arm is 12 months (as confirmed by the 8951-CL-0301 study results).

Separate MMRM analysis will be considered for each PRO score:

- QLQ-C30: global health, physical functioning, role functioning, emotional functioning, cognitive functioning, social functioning, fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea
- QLQ-OG25: Dysphagia, eating restrictions, reflux, odynophagia, pain and discomfort, anxiety, eating in front of others, dry mouth, trouble with taste, body image, trouble swallowing saliva, choked when swallowing, trouble with coughing, trouble talking, weight loss, hair loss
- GP: Pain item
- EQ-5D-5L: EQ-5D-5L utility index (UK mapping algorithm), EQ-5D-5L VAS

The analysis will be based on observed data, i.e., data collected at each time point without carrying forward previous values. Only subjects with a baseline and at least one post-baseline score will be included in the analysis. All visits with at least 10% of patients have non-missing data in each treatment arm will be included, excluding the study discontinuation and safety follow-up visits.

The response variable will be the change from baseline to each post-baseline assessment visit for each PRO score. The model will include the following covariates:

- Fixed effects
 - Treatment arm (zolbetuximab or placebo)
 - Timepoint (categorical: visit)
 - Baseline PRO score (continuous)
 - Region (Asia vs Non-Asia)
 - Number of metastatic sites (0 to 2 vs ≥ 3)
 - Prior gastrectomy (yes vs no)
- Interactions
 - Baseline PRO score x time
 - Treatment arm x time

Both main effects and the interaction terms will remain in the model, regardless of significance. The model will present least squares (LS) mean estimates, standard errors, 95%

CI, and p-values (where applicable) for mean changes from baseline to each visit. A plot of the LS means accompanied by the 95% CI will be produced.

In addition, an overall adjusted mean estimate will be derived that will estimate the average change from baseline across all time points, giving each visit equal weight.

The standardized mean difference (SMD) including 95% CI (Hedges' g) will also be provided.

The analysis will be conducted using PROC MIXED in SAS. The model will assume unstructured covariance among the within-subject repeated measurements. If the algorithm does not converge, a heterogeneous Toeplitz (the TOEPH option in SAS PROC MIXED) will be tried first and then AR(1) as a covariance structure to achieve convergence. The Kenward-Roger approximation will be used to estimate denominator degrees of freedom.

Variables listed as categorical in the list above will be included in the CLASS statement of the procedure. The unique subject identifier will also be included as a class variable. A REPEATED statement over the visits will be included with the unique subject identifier as the SUBJECT variable in the REPEATED statement.

The normality and homoscedasticity of the residuals will be visually checked. Particularly, the scatter plot of the residuals versus the predicted endpoint values and the histograms and the normal probability plots of the residuals will be reviewed. Transformation of the raw data will be considered if the graphs of residuals clearly indicate heavily skewed and heteroscedastic distribution of errors.

4.6 Responder Analysis

The responder analysis will be performed for all multi-item and single-item scales of the PRO questionnaires. The analysis will be performed on the FAS population and include only participants who have an assessment at baseline and at least one post-baseline assessment. The following descriptive analyses will be performed:

- The proportion of participants with improvement, who were stable, or deteriorated ([section 3.2](#)) will be summarized at each post-baseline PRO assessment visit up to including all cycles with at least 10% of patients with a baseline and at least one post-baseline score (excluding the study discontinuation and safety follow-up visits) using both the primary and the sensitivity thresholds. The denominator in this descriptive analysis will be the number of participants with non-missing data at the particular visit.
- A stacked column chart for the distribution of change in response categories from baseline for each scale of both PROs for each post-baseline assessment up to Cycle 9 Day 1 by treatment arms.

4.7 Time to Event Analysis

Kaplan-Meier (KM) curves will be used to estimate the distribution of time to deterioration for the PRO total, domain, and item scores for all three definitions described in [section 3.3](#).

The 50th percentile of Kaplan-Meier estimates will be used to estimate the median duration of time to deterioration (in months). A two-sided 95% confidence interval will be provided for these estimates. Median time to deterioration will be compared using stratified log rank test adjusting for randomization stratification factors: region (Asia vs Non-Asia), number of metastatic sites (0 to 2 vs ≥ 3), and prior gastrectomy (yes vs no). A Kaplan-Meier plot by treatment group will be presented.

Additionally, the benefit of zolbetuximab + mFOLFOX6 compared to placebo + mFOLFOX6 will be evaluated by a single hazard ratio (HR) (zolbetuximab vs placebo) with its 95% confidence interval (CI) based on a stratified Cox regression model with Efron's method of tie handling with the same strata as above. The proportional hazards assumption will be tested by examining plots of complementary log(log(survival)) versus log(time). In case departures from the assumption are observed, only the KM and the quartiles of the survival distribution will be presented. For EORTC QLQ-C30 GHS/QoL, PF, and EORTC QLQ-OG25 Pain, time-to-event analysis will be conducted using one-sided tests at the $\alpha=0.025$ significance level for consistency with the clinical SAP.

For each of the four instruments (QLQ-C30, QLQ-OG25, GP, and EQ-5D-5L), the HR, p-value, and 95% CI will also be presented graphically on a forest plot.

Kaplan-Meier analysis will be performed using PROC LIFETEST (SAS procedure). Cox proportional hazard regression model will be performed using PROC PHREG (SAS procedure).

Analyses will be performed separately for the following PRO scores:

- QLQ-C30: role functioning, emotional functioning, cognitive functioning, social functioning, fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea
- QLQ-OG25: Dysphagia, eating restrictions, reflux, odynophagia, anxiety, eating in front of others, dry mouth, trouble with taste, body image, trouble swallowing saliva, choked when swallowing, trouble with coughing, trouble talking, weight loss, hair loss
- GP: Pain item
- EQ-5D-5L: EQ-5D-5L utility index (UK mapping algorithm), EQ-5D-5L VAS.

4.7.1 Sensitivity Analyses

The time to event analyses will be repeated when deterioration is defined using the sensitivity threshold (see [section 3.2](#)).

In addition, the analyses using the primary threshold to define deterioration will also be repeated where death (due to any cause) will be counted as an event.

4.8 Subgroup Analyses

Analyses on subgroups will be performed for all PRO domains will be performed to determine whether the treatment effect is concordant among subgroups. The following analyses will be performed:

- TTFD, TTCD and TTDD using the primary threshold and excluding death (i.e. 3 definitions).

The analysis will be performed as described in [section 4.7](#). However, to avoid possible issues related to small number of events, subgroup analyses will not be adjusted for the stratification factors used at randomization. Subgroups are defined in [section 3.4](#).

If there are too few events available for a meaningful analysis of a particular subgroup (it is not considered appropriate to present analyses where there are less than 20 events in a subgroup), the HR and 95% CI will not be produced. In this case, only descriptive summaries will be provided.

5 SUPPORTING DOCUMENTATION

5.1 Appendix A List of Abbreviations

Abbreviation	Definition
ANCOVA	Analysis of covariance
CLDN	Claudin
CI	Confidence interval
DCR	Disease control rate
DOR	Duration of response
ECDF	Empirical cumulative distribution function
ECOG	Eastern cooperative oncology group
EORTC	European organization for research and treatment of cancer
EQ-5D-5L	EuroQoL group-5 dimension-5 level instrument
ES	Effect size
GEJ	Gastroesophageal Junction
GP	Global pain
HER2	Human epidermal growth factor receptor 2
HR	Hazard ratio
HRQoL	Health-related quality of life
HRU	Health resource utilization (HRU)
ICC	Intraclass coefficient
ITT	Intent-To-Treat
IRC	Independent review committee
LS	Least squares
MAR	Missing at random
MMRM	Mixed Model Repeated Measures
ORR	Objective response rate
OS	Overall survival
PFS	Progression-free survival
PGIC	Patient global impression of change
PRO	Patient-reported outcome
Q1	First quartile
QoL	Quality of life
RECIST	Response evaluation criteria in solid tumors
REML	Restricted maximum likelihood
ROC	Receiver operating characteristic
rPFS	Radiographic progression-free survival
SAP	Statistical analysis plan
SD	Standard deviation
SEM	Standard error of measurement
SRM	Standardized response mean
TTP	Time to progression
VAS	Visual analogue scale

5.2 Appendix B QLQ-C30

ENGLISH



EORTC QLQ-C30 (version 3)

We are interested in some things about you and your health. Please answer all of the questions yourself by circling the number that best applies to you. There are no "right" or "wrong" answers. The information that you provide will remain strictly confidential.

Please fill in your initials:

Your birthdate (Day, Month, Year):

Today's date (Day, Month, Year):

	Not at All	A Little	Quite a Bit	Very Much
1. Do you have any trouble doing strenuous activities, like carrying a heavy shopping bag or a suitcase?	1	2	3	4
2. Do you have any trouble taking a long walk?	1	2	3	4
3. Do you have any trouble taking a short walk outside of the house?	1	2	3	4
4. Do you need to stay in bed or a chair during the day?	1	2	3	4
5. Do you need help with eating, dressing, washing yourself or using the toilet?	1	2	3	4

During the past week:

	Not at All	A Little	Quite a Bit	Very Much
6. Were you limited in doing either your work or other daily activities?	1	2	3	4
7. Were you limited in pursuing your hobbies or other leisure time activities?	1	2	3	4
8. Were you short of breath?	1	2	3	4
9. Have you had pain?	1	2	3	4
10. Did you need to rest?	1	2	3	4
11. Have you had trouble sleeping?	1	2	3	4
12. Have you felt weak?	1	2	3	4
13. Have you lacked appetite?	1	2	3	4
14. Have you felt nauseated?	1	2	3	4
15. Have you vomited?	1	2	3	4
16. Have you been constipated?	1	2	3	4

Please go on to the next page

ENGLISH

During the past week:

	Not at All	A Little	Quite a Bit	Very Much
17. Have you had <u>diarrhea</u> ?	1	2	3	4
18. Were you tired?	1	2	3	4
19. Did pain interfere with your daily activities?	1	2	3	4
20. Have you had difficulty in concentrating on things, <u>like</u> reading a newspaper or watching television?	1	2	3	4
21. Did you feel tense?	1	2	3	4
22. Did you worry?	1	2	3	4
23. Did you feel irritable?	1	2	3	4
24. Did you feel depressed?	1	2	3	4
25. Have you had difficulty remembering things?	1	2	3	4
26. Has your physical condition or medical treatment <u>interfered</u> with your <u>family</u> life?	1	2	3	4
27. Has your physical condition or medical treatment <u>interfered</u> with your <u>social</u> activities?	1	2	3	4
28. Has your physical condition or medical treatment <u>caused</u> you financial difficulties?	1	2	3	4

For the following questions please circle the number between 1 and 7 that best applies to you

29. How would you rate your overall health during the past week?

1 2 3 4 5 6 7

Very poor

Excellent

30. How would you rate your overall quality of life during the past week?

1 2 3 4 5 6 7

Very poor

Excellent

Table 8 EORTC QLQ-C30 Scoring Guide (Fayers et al 2001)

	Abbreviation	Number of items	Item range*	Version 3.0 Item numbers
Global health status / QoL				
Global health status/QoL (revised)†	QL2	2	6	29, 30
Functional scales				
Physical functioning (revised)†	PF2	5	3	1 to 5
Role functioning (revised)†	RF2	2	3	6, 7
Emotional functioning	EF	4	3	21 to 24
Cognitive functioning	CF	2	3	20, 25
Social functioning	SF	2	3	26, 27
Symptom scales / items				
Fatigue	FA	3	3	10, 12, 18
Nausea and vomiting	NV	2	3	14, 15
Pain	PA	2	3	9, 19
Dyspnoea	DY	1	3	8
Insomnia	SL	1	3	11
Appetite loss	AP	1	3	13
Constipation	CO	1	3	16
Diarrhoea	DI	1	3	17
Financial difficulties	FI	1	3	28

* Item range is the difference between the possible maximum and the minimum response to individual items; most items take values from 1 to 4, giving range = 3.

† (revised) scales are those that have been changed since version 1.0, and their short names are indicated in this manual by a suffix “2” – for example, PF2.

Scoring algorithm

The QLQ-C30 is composed of both multi-item scales and single-item measures. These include five functional scales, three symptom scales, a global health status / QoL scale, and six single items. Each of the multi-item scales includes a different set of items - no item occurs in more than one scale.

The principle for scoring these scales is the same in all cases:

1. Estimate the average of the items that contribute to the scale; this is the raw score.
2. Use a linear transformation to standardize the raw score, so that scores range from 0 to 100; a higher score represents a higher ("better") level of functioning, or a higher ("worse") level of symptoms.

The technical details are provided below.

If items I_1, I_2, \dots, I_n are included in a scale, the procedure is as follows:

Calculate the raw score (RS) as follows: $RS = (I_1 + I_2 + \dots + I_n) / n$

Apply the linear transformation as follows:

Functional scales: Scale score = $(1 - ((RS - 1) / \text{range}) \times 100$

Symptom scales / items: Scale score = $(RS - 1) / \text{range} \times 100$

Global health status/QoL: Scale score = $(RS - 1) / \text{range} \times 100$

where range is the difference between the maximum possible value of RS and the minimum possible value. The QLQ-C30 has been designed so that all items in any scale take the same range of values. Therefore, the range of RS equals the range of the item values (see [Table 7](#)).

If at least half the components of a scale are present then the scale score will be calculated using the average of all items answered as the raw score; otherwise the score will be set to missing. For single measures, if the item is missing the scale score is set to missing.

An overall summary score has been suggested that can be calculated as ([Giesinger et al 2016](#)) the mean of all scales, except for QL2 and FI, i.e.

Physical Functioning + Role Functioning + Social Functioning + Emotional Functioning + Cognitive Functioning + (100-Fatigue) + (100-Pain) + 100-(Nausea and Vomiting) + (100-Dyspnoea) + (100-Sleeping Disturbances) + (100-Appetite Loss) + (100-Constipation) + (100-Diarrhoea))/13

5.3 Appendix C QLQ-OG25



EORTC QLQ – OG25

Patients sometimes report that they have the following symptoms or problems. Please indicate the extent to which you have experienced these symptoms or problems during the past week. Please answer by circling the number that best applies to you.

During the past week:	Not at all	A little	Quite a bit	Very much
31. Have you had problems eating solid foods?	1	2	3	4
32. Have you had problems eating liquidised or soft foods?	1	2	3	4
33. Have you had problems drinking liquids?	1	2	3	4
34. Have you had trouble enjoying your meals?	1	2	3	4
35. Have you felt full up too quickly after beginning to eat?	1	2	3	4
36. Has it taken you a long time to complete your meals?	1	2	3	4
37. Have you had difficulty eating?	1	2	3	4
38. Have you had acid indigestion or heartburn?	1	2	3	4
39. Has acid or bile coming into your mouth been a problem?	1	2	3	4
40. Have you had discomfort when eating?	1	2	3	4
41. Have you had pain when you eat?	1	2	3	4
42. Have you had pain in your stomach area?	1	2	3	4
43. Have you had discomfort in your stomach area?	1	2	3	4
44. Have you been thinking about your illness?	1	2	3	4
45. Have you worried about your health in the future?	1	2	3	4
46. Have you had trouble with eating in front of other people?	1	2	3	4
47. Have you had a dry mouth?	1	2	3	4
48. Have you had problems with your sense of taste?	1	2	3	4
49. Have you felt physically less attractive as a result of your disease or treatment?	1	2	3	4

Please go on to the next page

During the past week:	Not at all	A little	Quite a bit	Very much
50. Have you had difficulty swallowing your saliva?	1	2	3	4
51. Have you choked when swallowing?	1	2	3	4
52. Have you coughed?	1	2	3	4
53. Have you had difficulty talking?	1	2	3	4
54. Have you worried about your weight being too low?	1	2	3	4
55. Answer this question only if you lost any hair: If so, were you upset by the loss of your hair?	1	2	3	4

QLQ-OG25 Scoring algorithm

The QLQ-OG25 is composed of both multi-item domains and single-item measures. Each of the multi-item domains includes a different set of items - no item occurs in more than one domain.

If at least half the components of a domain are present, then the domain score will be calculated using the average of all items answered as the raw score; otherwise the score will be set to missing. For single measures, if the item is missing the domain score is set to missing.

Data are scored as follows according to the algorithm described in the EORTC QLQ-OG25 scoring manual:

Scales:

Dysphagia: $((Q31+Q32+Q33)/3-1)/3 * 100$
Eating restrictions: $((Q34+Q35+Q36+Q37)/4-1)/3 * 100$
Reflux: $((Q38+Q39)/2-1)/3 * 100$
Odynophagia: $((Q40+Q41)/2-1)/3 * 100$
Pain and discomfort: $((Q42+Q43)/2-1)/3 * 100$
Anxiety: $((Q44+Q45)/2-1)/3 * 100$

Single items:

Eating in front of others: $(Q46-1)/3 * 100$
Dry mouth: $(Q47-1)/3 * 100$
Trouble with taste: $(Q48-1)/3 * 100$
Body image: $((Q49-1)/3 * 100$
Trouble swallowing saliva: $(Q50-1)/3 * 100$
Choked when swallowing: $(Q51-1)/3 * 100$
Trouble with coughing: $(Q52-1)/3 * 100$
Trouble talking: $(Q53-1)/3 * 100$
Weight loss: $(Q54-1)/3 * 100$
Hair loss: $(Q55-1)/3 * 100$

The algorithms above transform raw scores to a linear scale ranging from 0 to 100, with a high score for scales/single items representing a high level of symptomatology/problems.

5.4 Appendix D GP

III. Global Pain (1 item)

 Global Pain

Please rate your pain by selecting the one number that best describes your pain at its worst in the last 24 hours.

0	1	2	3	4	5	6	7	8	9	10
↑										↑
No Pain										Pain as bad as you can imagine

5.5 Appendix E EQ-5D-5L

Under each heading, please tick the ONE box that best describes your health TODAY.

MOBILITY

- | | |
|---|--------------------------|
| I have no problems in walking about | <input type="checkbox"/> |
| I have slight problems in walking about | <input type="checkbox"/> |
| I have moderate problems in walking about | <input type="checkbox"/> |
| I have severe problems in walking about | <input type="checkbox"/> |
| I am unable to walk about | <input type="checkbox"/> |

SELF-CARE

- | | |
|---|--------------------------|
| I have no problems washing or dressing myself | <input type="checkbox"/> |
| I have slight problems washing or dressing myself | <input type="checkbox"/> |
| I have moderate problems washing or dressing myself | <input type="checkbox"/> |
| I have severe problems washing or dressing myself | <input type="checkbox"/> |
| I am unable to wash or dress myself | <input type="checkbox"/> |

USUAL ACTIVITIES (e.g. work, study, housework, family or leisure activities)

- | | |
|--|--------------------------|
| I have no problems doing my usual activities | <input type="checkbox"/> |
| I have slight problems doing my usual activities | <input type="checkbox"/> |
| I have moderate problems doing my usual activities | <input type="checkbox"/> |
| I have severe problems doing my usual activities | <input type="checkbox"/> |
| I am unable to do my usual activities | <input type="checkbox"/> |

PAIN / DISCOMFORT

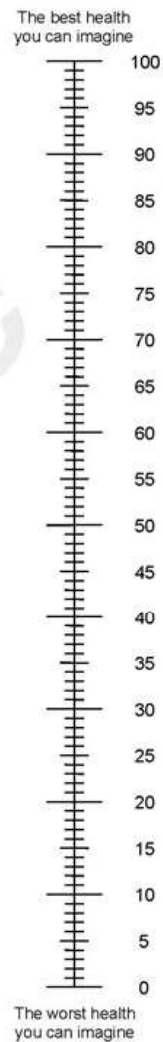
- | | |
|------------------------------------|--------------------------|
| I have no pain or discomfort | <input type="checkbox"/> |
| I have slight pain or discomfort | <input type="checkbox"/> |
| I have moderate pain or discomfort | <input type="checkbox"/> |
| I have severe pain or discomfort | <input type="checkbox"/> |
| I have extreme pain or discomfort | <input type="checkbox"/> |

ANXIETY / DEPRESSION

- | | |
|--------------------------------------|--------------------------|
| I am not anxious or depressed | <input type="checkbox"/> |
| I am slightly anxious or depressed | <input type="checkbox"/> |
| I am moderately anxious or depressed | <input type="checkbox"/> |
| I am severely anxious or depressed | <input type="checkbox"/> |
| I am extremely anxious or depressed | <input type="checkbox"/> |

- We would like to know how good or bad your health is TODAY.
- This scale is numbered from 0 to 100.
- 100 means the best health you can imagine.
0 means the worst health you can imagine.
- Mark an X on the scale to indicate how your health is TODAY.
- Now, please write the number you marked on the scale in the box below.

YOUR HEALTH TODAY =



6 REFERENCES


Aaronson et al 1993	Aaronson N et al (1993). The European Organisation for Research and Treatment of Cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. <i>Journal of the National Cancer Institute</i> , pp. 85:365-376.
Brown and Prescott 2006	Brown H and Prescott R (2006). <i>Applied Mixed Models in Medicine</i> . John Wiley and Sons Ltd, 2nd Edition.
Cocks et al, 2012	Cocks, K. & et al, 2012. Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. <i>Eur J Cancer</i> , pp. 48:1713-21.
Devlin et al, 2017	Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing healthrelated quality of Life: An EQ-5D-5L Value Set for England. <i>Health Economics</i> . 2017;1-16.
Dolan, 1997	Dolan P (1997). Modelling valuations for EuroQol health states. <i>Medical Care</i> 35: 1095-1108.
Farrar et al, 2000	Defining the clinically important difference in pain outcome measures. <i>Pain</i> , 2000. 88(3): p. 287-94.
Fayers et al, 2021	Fayers, P., Aaronson, N., Bjordal, K., Groenvold, M., Curran, D., & Bottomley, A. (2001). <i>EORTC QLQ-C30 Scoring Manual</i> (3rd ed.). EORTC.
Giesinger et al, 2016	Giesinger JM et al (2016). Replication and validation of higher order models demonstrated that a summary score for the EORTC QLQ-C30 is robust. <i>J Clin Epidemiol</i> . 2016 Jan;69:79-88. doi: 10.1016/j.jclinepi.2015.08.007
Hernández Alava et al, 2017	Hernández-Alava, M., & Pudney, S. E. (2017). Econometric modelling of multiple self-reports of health states: The switch from EQ-5D-3L to EQ-5D-5L in evaluating drug therapies for rheumatoid arthritis. <i>Journal of Health Economics</i> , 55, 139– 152.
Hernández Alava et al, 2020	Hernández Alava, M, Pudney, S. , Wailoo, A. (2020) Estimating the relationship between EQ-5D-5L and EQ-5D-3L: results from an English Population Study. Policy Research Unit in Economic Evaluation of Health and Care interventions. Universities of Sheffield and York. Report 063.
Kendrick and Strout, 2005	The minimum clinically significant difference in patient assigned numeric scores for pain. <i>Am J Emerg Med</i> , 2005. 23(7): p. 828-32.
NICE position statement, 2019	NICE position statement, available at: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisal-guidance/eq5d5l_nice_position_statement.pdf
NICE 2022	National Institute for Health and Care Excellence (2022) NICE health technology evaluations: the manual Process and methods [PMG36], https://www.nice.org.uk/process/pmg36/resources/nice-health-technology-evaluations-the-manual-pdf-72286779244741
Norman et al 2003	Norman GR, Sloan JA, Wyrwich KW (2003). Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. <i>Med Care</i> 2003; 41(5): 582–592.
Pickard et al, 2007	Pickard AS, Neary MP and Cella D. (2007). Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. <i>Health and quality of life outcomes</i> . 5:70.

Sloan et al 2005	Sloan JA, Cella D, Hays RD (2005). Clinical significance of patient-reported questionnaire data: another step toward consensus. J Clin Epidemiol 2005; 58 (12): 1217–1219.
------------------	--

7 SIGNATURE

Prepared by:  3, 2023
(DD Mmm YYYY)

Approved by:  May 3, 2023
Date (DD Mmm YYYY)

Approved by:  May 4, 2023
Date (DD Mmm YYYY)

Approved by:  May 2, 2023
Date (DD Mmm YYYY)

Approved by:  May 4, 2023
Date (DD Mmm YYYY)

Approved by:  May 3, 2023
Date (DD Mmm YYYY)