Official Protocol Title:	3475 P811-08 SSAP
NCT number:	NCT03615326
Document Date:	07-APRIL-2022

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

This supplemental SAP (sSAP) is a companion document to the protocol. In addition to the information presented in the protocol SAP which provides the principal features of confirmatory analyses for this trial, this supplemental SAP provides additional statistical analysis details/data derivations and documents modifications or additions to the analysis plan that are not "principal" in nature and result from information that was not available at the time of protocol finalization. Separate analysis plans (i.e., separate documents from this sSAP) may be developed for PK/modeling analysis, biomarker analysis, and genetic data analysis.

2 SUMMARY OF CHANGES

This sSAP aligns with the protocol amendment 08 (MK-3475-811-08) with regard to the statistical analysis plan. In addition, the following changes were made to the sSAP which were not directly related to changes required due to this protocol amendment.

- Clarified the collapsed strata (because of small number of participants/events in the strata) that decided at IA1 will be used consistently in future analyses when stratified approaches applied.

- Removed the Section 3.6.3 exploratory analysis of PFS2 as it is not a planned exploratory endpoint in the protocol.

- Clarified descriptive analysis of ORR based on all participants will be conducted at IA2. No formal hypothesis testing will be performed.

- Updated subgroup analysis, the subgroup level that has less than 20 participants are not analyzed. Clarified the subgroup analyses will based on unstratified approaches.

- Updated timepoint (at Week 24) for mean change from baseline in PRO endpoints.

3 ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

This section contains a brief summary of the statistical analyses for this trial. The Japan-specific SOX cohort will be analyzed separately. Specifically, the efficacy analysis will only include subjects from the Global Cohort. The safety analysis will be performed for the Global and Japan SOX Cohorts separately. Full detail is provided in [Sec. 3.2] to [Sec. 3.12].

Study Design Overview	This is a Phase III, randomized, double-blind trial comparing pembrolizumab and placebo, both in combination with trastuzumab plus chemotherapy as first-line treatment in participants with HER2 positive advanced gastric or GEJ adenocarcinoma
Treatment Assignment	Participants will be randomized in a 1:1 ratio to the experimental arm and the control arm. Stratification factors are in Section 6.3.1.1 of the protocol.
Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Participants as Treated (APaT)
Primary Endpoints	1) Progression-free Survival (PFS) per RECIST 1.1 assessed by BICR 2) Overall survival (OS)
Key Secondary Endpoints	1) Objective response (OR) per RECIST 1.1 assessed by BICR
Statistical Methods for Key Efficacy Analyses	The dual primary hypotheses on PFS and OS will be evaluated by comparing the experimental arm to the control arm using a stratified Log-rank test. The hazard ratio will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. The stratified Miettinen and Nurminen method with sample size weights will be used for analysis of ORR.
Statistical Methods for Key Safety Analyses	For analyses in which 95% CIs will be provided for between-treatment differences in the percentage of participants with events, these analyses will be performed using the Miettinen and Nurminen method [1].

Interim Analyses	Three interim analyses (IA) may be performed in this study based on current projection of enrollment and event accrual rates. Results will be reviewed by an external Data Monitoring Committee. Details are provided in [Sec. 3.7]. • IA1: o Timing: to be performed when ~ 260 participants have been followed up for ~ 8.5 months. o Primary purpose: efficacy analysis for ORR (hypothesis testing). • IA2[a]: o Timing: to be performed after ~ 542 PFS events have occurred AND ~ 9 months after last participant randomized. o Primary purpose: efficacy analysis for PFS and OS. • IA3[a]: o Timing: to be performed when at least 18 months after the last participant has been randomized AND at least 606 PFS events have been observed. This is the final PFS analysis. o Primary purpose: efficacy analysis for PFS and OS. • Final analysis[a]: o Timing: to be performed when at least 28 months after the last participant has been randomized AND at least ~551 deaths have occurred. o Primary purpose: efficacy analysis for OS. [a] Note for IA2, IA3, and FA, if the events accrue slower than expected, the Sponsor may conduct the analysis with up to 3 additional months of follow-up than the minimal follow-up as described above, or when the specified number of events are observed, whichever comes first.
Multiplicity	The overall type I error over the primary endpoints (PFS and OS) and the key secondary endpoint (ORR) is strongly controlled at 2.5% (one-sided), with initially 0.2% allocated to ORR, 0.3% to PFS and 2% to OS. By using the graphical approach of Maurer and Bretz, if one hypothesis is rejected, the alpha will be shifted to other hypotheses [2].
Sample Size and Power	The planned sample size is approximately 692 participants. For ORR, with sample size of ~260 at IA1, the study has ~90% power for detecting a 25% point difference in ORR (73% vs 48%) at an initially assigned 0.002 (one-sided) significance level. For PFS, there will be ~606 events at the PFS final analysis. With 606 PFS events, the study has ~95% power for detecting a HR of 0.7 at an initially assigned 0.003 (one-sided) significance level. For OS, there will be ~551 deaths at the OS final analysis. With 551 deaths, the study has ~90% power for detecting a hazard ratio (HR) of 0.75 at an initially assigned 0.020 (one-sided) significance level.

3.2 Responsibility for Analyses/In-House Blinding

The statistical analysis of the data obtained from this study will be the responsibility of the Clinical Biostatistics department of the Sponsor.

This study will be conducted as a double-blind study under in-house blinding procedures. The official, final database will not be unblinded until medical/scientific review has been performed, protocol deviations have been identified, and data have been declared final and complete.

The Clinical Biostatistics department will generate the randomized allocation schedule(s) for study treatment assignment.

Blinding issues related to the planned interim analyses are described in [Sec. 3.7].

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Protocol Section 3.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

<u>**Dual Primary**</u>

- **Progression-free survival (PFS) per RECIST 1.1 assessed by BICR**

PFS is defined as the time from randomization to the first documented disease progression per RECIST 1.1 by BICR or death due to any cause, whichever occurs first. See [Sec. 3.6.1] for the censoring rules.

- **Overall Survival (OS)**

OS is defined as the time from randomization to death due to any cause.

<u>**Secondary**</u>

- **Objective Response Rate (ORR) per RECIST 1.1 by BICR**

ORR is defined as the proportion of the subjects who have a confirmed complete response (CR) or partial response (PR). (note: only CR or PR prior to a curative surgical resection will be used among participants with curative surgical resection).

- **Duration of Response (DOR) per RECIST 1.1 by BICR**

For participants who demonstrated CR or PR, DOR is defined as the time from first response (CR or PR) to subsequent disease progression or death from any cause, whichever occurs first.

Exploratory

- **Progression-free survival (PFS) per RECIST 1.1 assessed by investigator**
- **Progression-free survival (PFS) using modified RECIST 1.1 for immune- based therapeutics (iRECIST) by investigator**
- **Objective Response Rate (ORR) per RECIST 1.1 by investigator**
- **Objective Response Rate (ORR) using modified RECIST 1.1 for immune- based therapeutics (iRECIST) by investigator**

Patient Reported Outcomes (PRO) endpoints as assessed by the EORTC QLQ-C30, EORTC QLQ-STO22 and EuroQol-5D-5L. Details are provided in [Sec. 3.12].

3.4.2 Safety Endpoints

Safety measurements are described in Protocol Section 4.2.2.1.

3.5 Analysis Populations

3.5.1 Efficacy Analysis Populations

The Intention-to-Treat (ITT) population will serve as the population for primary efficacy analysis (PFS, OS, and ORR). All randomized participants, whether or not treatment was administered, will be included in this population. Any participant who receives a randomization number will be considered to have been randomized. Participants will be included in the treatment group to which they are randomized.

The ITT population excluding MSI-H participants will serve as the sensitivity analysis for the endpoints of PFS per RECIST 1.1 by BICR, OS, and ORR per RECIST 1.1 per BICR.

3.5.2 Safety Analysis Populations

The APaT population will be used for the analysis of safety data in this study. The APaT population consists of all randomized participants who received at least one dose of study treatment. Participants will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the APaT population. For most participants this will be the treatment group to which they are randomized. Participants who take incorrect study treatment for the entire treatment period will be included in the treatment group corresponding to the study treatment actually received. Any participant who receives the incorrect study medication for one or more cycles but receives the correct treatment for the remaining cycles will be analyzed according to the participant's randomized treatment group and a narrative will be provided for any events that occur during the cycle for which the participant was incorrectly dosed.

At least one laboratory or vital sign measurement obtained subsequent to at least one dose of study treatment is required for inclusion in the analysis of each specific parameter. To assess change from baseline, a baseline measurement is also required.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

In this section, for the stratified analyses, small strata may be collapsed. Please see details in the [Sec. 3.6.1.1]. Response or progression in the Second Course Phase will not count towards the PFS of the primary endpoint in this trial.

3.6.1.1 Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, HR) between the treatment arms. The HR and its 95% CI from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. Geographic region [Europe/Israel/North America/Australia, Asia, and Rest of the World (including South America], PD-L1 status (positive vs. negative), and chemotherapy regimen (FP or CAPOX) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model.

Since the number of participants in certain strata is small, stratified analyses will be based on collapsed strata by combining strata with small number of participants or events. At 1st interim analyses, the collapsed strata was based on blinded data taking into considerations of both clinical relevance and actual counts of subjects/events. As a result, the following 6 strata were planned to be used in the stratified log-rank test and the stratified Cox model:

The following pooling strategy was decided for the 1st interim analyses.

- PD-L1 negative + any region + any chemotherapy regimen
- PD-L1 positive + Asia + any chemotherapy regimen
- PD-L1 positive + Europe/Israel/North America/Australia + FP chemotherapy regimen
- PD-L1 positive + Europe/Israel/North America/Australia + CAPOX chemotherapy regimen
- PD-L1 positive + Rest of the World + FP chemotherapy regimen
- PD-L1 positive + Rest of the World + CAPOX chemotherapy regimen

Based on the blinded data monitoring before the time of IA2, the issue of small number of participants or events still exists. Therefore, the same collapsed strata that was used at IA1 will be used consistently at IA2, IA3, and FA.

Since disease progression is assessed periodically, progressive disease (PD) can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. The true date of disease progression will be approximated by the earlier of the date of the first assessment at which PD is objectively documented per RECIST 1.1 by BICR and the date of death. Sensitivity analyses will be performed for comparison of PFS based on investigator's assessment.

In order to evaluate the robustness of the PFS endpoint per RECIST 1.1 by BICR, one primary and 2 sensitivity analyses with a different set of censoring rules will be performed. For the primary analysis, if the events (PD or death) are immediately after more than one missed disease assessment, the data are censored at the last disease assessment prior to missing visits. Also data after new anti-cancer therapy are censored at the last disease assessment prior to the initiation of new anti-cancer therapy. The first sensitivity analysis follows the intention-to-treat principle. That is, PDs/deaths are counted as events regardless of missed study visits or initiation of new anti-cancer therapy. The second sensitivity analysis considers discontinuation of treatment due to reasons other than complete response or initiation of new anticancer treatment, whichever occurs later, to be a PD event for participants without documented PD or death. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied. The censoring rules for primary and sensitivity analyses are summarized in [Table 1].

In case there is an imbalance between the treatment groups on disease assessment schedules or censoring patterns, the following two additional PFS sensitivity analyses may be considered: 1) a PFS analysis using time to scheduled tumor assessment visit from randomization as opposed to actual tumor assessment time; 2) Finkelstein (1986)'s likelihood-based score test [3] for interval-censored data, which modifies the Cox proportional hazards model for interval censored data, will be used as a supportive analysis for the PFS endpoint. The interval will be constructed so that the left endpoint is the date of the last disease assessment without documented PD and the right endpoint is the date of documented PD or death, whichever occurs earlier.

Table 1 Censoring Rules for Primary and Sensitivity Analyses of PFS

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
PD or death documented after ≤ 1 missed disease assessment, and before new anti-cancer therapy[#], if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death	Progressed at date of documented PD or death

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
PD or death documented immediately after ≥ 2 consecutive missed disease assessments or after new anti-cancer therapy, if any	Censored at last disease assessment prior to the earlier date of ≥ 2 consecutive missed disease assessment and new anti-cancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death
No PD and no death; and new anticancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment	Progressed at treatment discontinuation due to reasons other than complete response; otherwise censored at last disease assessment if still on study treatment or completed study treatment.
No PD and no death; new anticancer treatment is initiated	Censored at last disease assessment before new anticancer treatment	Censored at last disease assessment	Progressed at date of initiation of new anticancer treatment or discontinuation of treatment due to reasons other than complete response, whichever occurs later

[#] New anti-cancer therapy: excluding curative surgical resections (the detailed definition in Protocol Section 6.5.3).

As imaging will continue, participants who have curative surgical resection during the study (per Protocol Section 6.5.3) will be followed for PFS events after surgery until local recurrence, distant metastasis, or death for the primary analysis of PFS. An additional sensitivity analysis will be conducted for PFS in which these participants will be censored at last disease assessment prior to the time of the curative surgical resections. Note that the curative surgical resection will be excluded from the new anti-cancer therapy under the primary analysis if it is occurred less than 10%.

The proportional hazards assumption of the Cox model may be examined using both graphical and analytical methods for the primary PFS analysis. The log[-log] of the survival function vs. time for PFS may be plotted for each treatment arm. If the curves are not parallel, indicating that hazards are not proportional, supportive analyses may be conducted to account for the possible non-proportional hazards effect associated with immunotherapies, for example, using Restricted Mean Survival Time (RMST) method [4], parametric method [5] etc.

The RMST is simply the population average of the amount of event-free survival time experienced during the study follow up time. This quantity can be estimated by the area under the KM curve up to the follow up time. The clinical relevance and feasibility of conducting the study should be taken into account in the choice of follow-up time to define

RMST (e.g. near the last observed event time assuming that the period of clinical interest in the survival experience is the whole observed follow-up time for the trial). The cut-off of follow-up time will be pre-specified by the Sponsor study team that is blinded to treatment results and details will be documented outside of sSAP prior to each of efficacy interim analysis except for IA1. The difference of two RMSTs for two treatment groups will be estimated and 95% confidence interval will be provided.

A sensitivity analysis may be performed based on the MaxCombo test with logrank FH (0, 1), FH (1, 1) at the final analysis of PFS to account for the potential loss of power with logrank test when the proportional hazard assumption is violated.

3.6.1.2 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, the HR). The HR and its 95% CI from the Cox model with a single treatment covariate will be reported. Geographic region, PD-L1 status and chemotherapy regimen (FP or CAPOX) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model. The same strategy of combination of small strata defined for the PFS analysis (see [Sec. 3.6.1.1]) will be used for the OS analysis.

Participants without documented death at the time of analysis will be censored at the date of last contact.

Sensitivity analyses to adjust for the effect of treatment switching to other PD-1 or other new anticancer therapies on OS may be performed as model assumptions permit. Three recognized methods may be included: 1) the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis (1989) [7]; 2) the two-stage model proposed by Latimer [8]; and 3) the Inverse-Probability-of-Censoring Weighting (IPCW) model [9] in which an examination of the appropriateness of the data to the assumptions is required by the methods.

Other sensitivity analyses described for the PFS endpoint will be applied to OS endpoint as appropriate.

3.6.1.3 Objective Response Rate (ORR)

The stratified Miettinen and Nurminen method will be used for the comparison of the ORR between the two treatment groups. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen method with strata weighting by sample size will be reported [1]. Geographic region, PD-L1 status and chemotherapy regimen (FP or CAPOX) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model. The same strategy of combination of small strata defined for the PFS analysis (see [Sec. 3.6.1.1]) will be used for the ORR analysis. The descriptive analysis of ORR based on all participants will be performed after IA1. No formal hypothesis testing will be conducted.

3.6.1.4 Duration of Response (DOR)

If sample size permits, DOR will be summarized descriptively using Kaplan-Meier medians and quartiles. Only the subset of patients who show a complete response or partial response will be included in this analysis.

For each DOR analysis, a corresponding summary of the reasons responding subjects are censored will also be provided. Responding subjects who are alive, have not progressed, have not initiated new anti-cancer treatment, have not been determined to be lost to follow-up, and have had a disease assessment within ~5 months of the data cutoff date are considered ongoing responders at the time of analysis. If a subject meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied. The censoring rules are summarized in [Table 2].

Table 2 Censoring Rules for DOR

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new anti-cancer therapy initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anti-cancer therapy# initiated	Last adequate disease assessment before new anti-cancer therapy initiated	Censor (non-event)
Death or progression immediately after ≥ 2 consecutive missed disease assessments or after new anti-cancer therapy#, if any	Earlier date of last adequate disease assessment prior to ≥ 2 missed adequate disease assessments and new anti-cancer therapy, if any	Censor (non-event)
Death or progression after ≤ 1 missed disease assessments and before new anti-cancer therapy#, if any	PD or death	End of response (Event)
A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response. #new anti-cancer therapy: excluding curative surgical resections (the detailed definition in Protocol Section 6.5.3).		

3.6.1.5 Analysis Strategy for Key Efficacy Endpoints

[Table 3] summarizes the primary analysis approach for key efficacy endpoints.

Table 3 Analysis Strategy for Key Efficacy Endpoints

Endpoint	Statistical Method[†]	Analysis Population	Missing Data Approach
Primary Endpoints			
PFS per RECIST 1.1 by BICR	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	• Primary censoring rule • Sensitivity analysis 1 • Sensitivity analysis 2 (More details are provided in [Table 1] Censoring Rules for Primary and Sensitivity Analyses of PFS)
OS	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at the last known alive date
Key Secondary Endpoint			
ORR per RECIST 1.1 by BICR	Test and Estimation: Stratified M&N method with sample size weights[††]	ITT	Participants without assessments are considered non-responders and conservatively included in the denominator

PFS = Progression-free survival; OS = Overall survival; ORR = Objective response rate; ITT = Intention to treat.

[†] Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization (Protocol Section 6.3.1.1) will be applied to the analysis. Small strata will be combined in a way specified by a blinded statistician prior to the analysis.

[††] Miettinen and Nurminen method

The strategy to address multiplicity issues with regard to multiple efficacy endpoints, and interim analyses is described in [Sec. 3.7] Interim Analyses and in [Sec. 3.8] Multiplicity.

3.6.2 Statistical Methods for Safety Analyses

Safety and tolerability will be assessed by clinical review of all relevant parameters including adverse experiences (AEs), laboratory tests and vital signs.

The analysis of safety results will follow a tiered approach as shown in [Table 4] The tiers differ with respect to the analyses that will be performed. Adverse events (specific terms as well as system organ class terms) are either prespecified as "Tier 1" endpoints, or will be classified as belonging to "Tier 2" or "Tier 3" based on the observed proportions of participants with an event.

Safety parameters or AEs of interest that are identified *a priori* constitute "Tier 1" safety endpoints that will be subject to inferential testing for statistical significance. There are no Tier 1 events for this protocol. Based on a review of historic chemotherapy data and data from ongoing pembrolizumab clinical studies in gastric cancer, there are no AEs of interest that warrant inferential testing for comparison between treatment arms in this study.

Tier 2 parameters will be assessed via point estimates with 95% CIs provided for between-group comparisons.

Membership in Tier 2 requires that at least 10% of participants in any treatment group exhibit the event; all other adverse experiences and predefined limits of change will belong to Tier 3. The threshold of at least 10% of participants was chosen for Tier 2 event because the population enrolled in this study are in critical conditions and usually experience various adverse events of similar types regardless of treatment, events reported less frequent than 10% of participants would obscure the assessment of overall safety profile and add little to the interpretation of potentially meaningful treatment differences. In addition, Grade 3 to 5 AEs (≥5% of participants in one of the treatment groups) and SAEs (≥5% of participants in one of the treatment groups) will be considered Tier 2 endpoints. Because many 95% confidence intervals may be provided without adjustment for multiplicity, the confidence intervals should be regarded as a helpful descriptive measure to be used in review, not a formal method for assessing the statistical significance of the between-group differences. These analyses will be performed using the Miettinen and Nurminen method, an unconditional, asymptotic method [1].

Safety endpoints that are not Tier 1 or 2 events are considered Tier 3 events. Only point estimates are provided for Tier 3 safety parameters. For continuous measures such as vital signs, summary statistics for baseline and on-treatment will be provided by treatment group in tabular format.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	**Safety Endpoint**	**95% CI for Treatment Comparison**	**Descriptive Statistics**
Tier 2	AEs (≥10% of participants in one of the treatment groups)	X	X
	Grade 3-5 AEs (≥5% of participants in one of the treatment groups)	X	X
	SAEs (5% of participants in one of the treatment groups)	X	X
	AEs (<10% of participants in one of the treatment groups)		X
Tier 3	Discontinuation due to AE		X
	Change from Baseline Results (laboratory test toxicity grade)		X

X = results will be provided.

In addition to tiered approach, exploratory analysis will be performed on time to first Grade 3-5 AE. Time to first Grade 3-5 AE is defined as the time from the first day of study drug to the first event of Grade 3-5 AE. For patients without a Grade 3-5 AE, the time to first Grade 3-5 AE is censored at minimum of 30 days post last study dose or date of data cutoff. The Kaplan-Meier method will be used to estimate the curve of time to first Grade 3-5 AE. The treatment difference in time to first Grade 3-5 AE will be assessed by the log-rank test. A

Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (i.e., the hazard ratio). The hazard ratio and its 95% confidence interval from the Cox model with a single treatment covariate will be reported.

3.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

The comparability of the treatment groups for each relevant characteristic will be assessed by the use of tables and/or graphs. No statistical hypothesis tests will be performed on these characteristics. The number and percentage of participants screened, randomized, the primary reasons for screening failure, and the primary reason for discontinuation will be displayed.

Demographic variables (eg, age), baseline characteristics, primary and secondary diagnoses, and prior and concomitant therapies will be summarized by treatment either by descriptive statistics or categorical tables.

3.7 Interim Analysis

An external Data Monitoring Committee (DMC) will serve as the primary reviewer of the results of the interim analysis (analyses) of the study and will make recommendations for discontinuation of the study or protocol modifications to an executive committee of the SPONSOR. If the DMC recommends modifications to the design of the protocol or discontinuation of the study, this executive committee (and potentially other limited SPONSOR personnel) may be unblinded to results at the treatment level in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of interim analyses will be documented. Additional logistical details will be provided in the DMC Charter.

Treatment-level results from the interim analysis will be provided to the DMC by the unblinded statistician. Prior to final study unblinding, the unblinded statistician will not be involved in any discussions regarding modifications to the protocol, statistical methods, identification of protocol deviations, or data validation efforts after the interim analyses.

3.7.1 Efficacy Interim Analysis

Three interim analyses may be performed in this study based on current projection of enrollment and event accrual rates. The timing and the purpose of each analysis are summarized in [Table 5]. Type I error control for the efficacy analyses as well as efficacy bounds are described in [Sec. 3.8] – Multiplicity.

Table 5 Summary of Interim and Final Analyses Strategy

Analyses	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	The first 260 participants with at least 8.5 months follow-up.	~22.5 months	• Efficacy analysis of ORR (hypothesis testing)
IA2 [a]	At least 542 PFS events have occurred and ~ 9 months after the last participant has been randomized.	~37 months	• Efficacy analysis for PFS and OS
IA3 [a]	At least 18 months after the last participant has been randomized AND at least 606 PFS events have been observed. This is final PFS analysis.	~46 months	• Efficacy analysis for PFS and OS
Final Analyses [a]	Final OS analysis to be performed until at least 28 months after the last participant has been randomized AND at least ~551 deaths have occurred.	~56 months	• Efficacy analysis for OS

[a] Note for IA2, IA3, and FA, if the events accrue slower than expected, the Sponsor may conduct the analysis with up to 3 additional months of follow-up than the minimal follow-up as described above, or when the specified number of events are observed, whichever comes first.

ORR = Objective Response Rate; OS= Overall Survival; PFS = Progression-free Survival.

3.7.2 Safety Interim Analysis

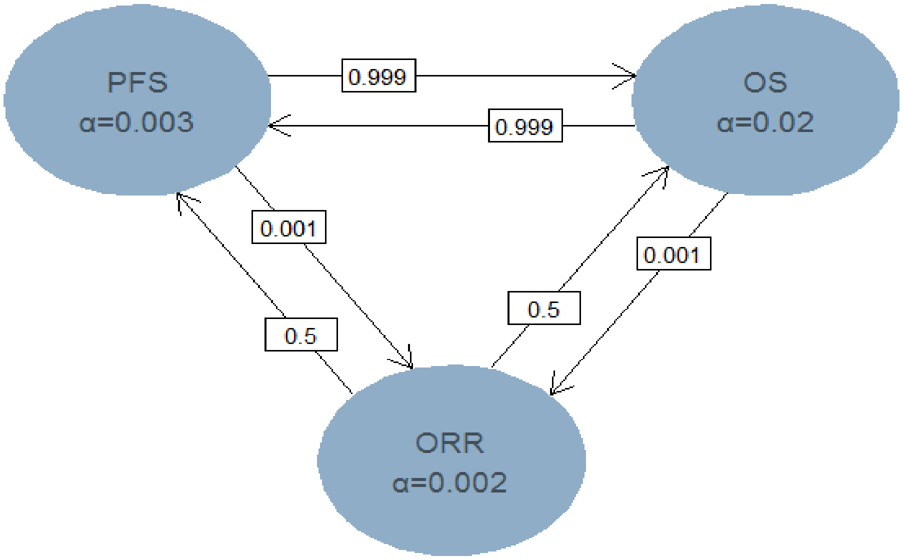
The DMC will be responsible for periodic interim safety reviews as specified in the DMC charter. Interim safety analyses will also be performed at the time of interim efficacy analyses.

3.8 Multiplicity

The trial uses an extension of the graphical method of Maurer and Bretz [2] to provide strong multiplicity control for multiple hypotheses while making the interim and final analysis timing be more flexible (Anderson et al, unpublished data, 2018). According to the Maurer and Bretz approach, study hypotheses may be tested in a group sequential fashion, and when a particular null hypothesis is rejected, the alpha allocated to that hypothesis can be reallocated to other hypothesis tests.

[Figure 1] shows the initial one-sided alpha allocation for each hypothesis in the ellipse representing the hypothesis. The weights for reallocation from each hypothesis to the others are represented in the boxes on the lines connecting hypotheses.

Figure 1 Type I Error Reallocation Strategy



Note: If both PFS and OS null hypotheses are rejected, the reallocation strategy allows re-testing of ORR at alpha=0.025 based on the p-value at IA1.
ORR=objective response rate; OS=overall survival; PFS=progression-free survival.

The extended graphical method spends alpha as a function of the minimum of the actual event information fraction and the expected event information fraction. This ensures that the actual spending will be no more aggressive than the planned, while at the same time ensuring that not all alpha is spent prior to final planned event counts.

3.8.1 Objective Response Rate

The study allocates alpha=0.002, one-sided, to test ORR, and ORR is tested only at the first interim analysis (IA1). However, if the test does not reach statistical significance at IA1, the p-value from IA1 can be compared to an updated alpha-level if the null hypotheses for both PFS and OS are rejected at a later time. Power at the possible alpha-levels as well as the approximate treatment difference required to reach the bound (ORR difference) are shown in [Table 6].

Table 6 Possible Alpha-levels and Approximate ORR Difference Required to Demonstrate Efficacy for ORR at IA1

Alpha	ORR difference	Power
0.002	~0.17	0.90
0.025	~0.11	0.99

3.8.2 Progression-free Survival

The initial alpha-level for testing PFS is 0.003. If the null hypothesis for ORR is rejected, [Figure 1] shows that half of its alpha, ie, 0.001, is reallocated to PFS hypothesis testing. If the null hypothesis for OS is rejected, then alpha=0.02 is essentially fully reallocated to PFS hypothesis testing. Thus, the PFS null hypothesis may be tested at alpha=0.003, alpha=0.004 (if the ORR null hypothesis is rejected but not the OS null hypothesis), alpha=0.023 (if the OS null hypothesis is rejected but not the ORR null hypothesis), or alpha=0.025 (if both the ORR and OS null hypotheses are rejected). [Table 7] shows the boundary properties for each of these alpha-levels for the interim analyses, which were derived using a Lan-DeMets O'Brien-Fleming spending function based on predicted number of events at the planned time of interim analysis. Note that the final row indicates the total power to reject the null hypothesis for PFS at each alpha-level.

If events accrue more slowly than expected or the same as expected, spending will be based on actual information fraction. If events accrue more quickly than expected, cumulative spending based on the expected information fraction will be used in order to save some alpha for analyses that will be performed with more than the originally planned maximum events. For example, at IA2, if 560 PFS events have occurred (ie, more than the expected 542 PFS events), the alpha spending at IA2 will be according to the expected information fraction (542/606=89%) instead of the actual information fraction (560/606=92%). The final analysis will use the remaining Type I error that has not been spent at the earlier analysis. The event counts for all analyses will be used to compute correlations.

Also note that if the OS or ORR null hypothesis is rejected at an interim or final analysis, each PFS interim and final analysis test may be compared to its updated bounds considering the alpha reallocation from the OS or ORR hypothesis.

Table 7 Efficacy Boundaries and Properties for Progression-free Survival Analyses

Analysis	Value	α=0.003	α=0.004	α=0.023	α=0.025
IA 2: 90%* N: 692 Events: 542 Month: 37	Z	2.927	2.826	2.137	2.1
	p (1-sided) §	0.0017	0.0024	0.0163	0.0179
	HR at bound%	0.7777	0.7849	0.8326	0.8353
	P(Cross) if HR=1†	0.0017	0.0024	0.0163	0.0179
	P(Cross) if HR=0.7#	0.8915	0.9091	0.9785	0.9804
IA 3: 100%* N: 692 Events: 606 Month: 46	Z	2.807	2.714	2.086	2.052
	p (1-sided) §	0.0025	0.0033	0.0185	0.0201
	HR at bound%	0.796	0.8024	0.8444	0.8467
	P(Cross) if HR=1†	0.003	0.004	0.023	0.025
	P(Cross) if HR=0.7#	0.9475	0.9569	0.9909	0.9917

*Percentage of expected number of events at final analysis

§p (1-sided) is the nominal alpha for testing.

%HR at bound is the approximate HR required to reach an efficacy bound

†P(Cross if HR=1) is the cumulative probability of crossing a bound under the null hypothesis

#P(Cross if HR=0.7) is the cumulative probability of crossing a bound under the alternative hypothesis

3.8.3 Overall Survival

The OS hypothesis may be tested at alpha=0.02 (initially allocated alpha), alpha=0.023 (if the PFS but not the ORR null hypothesis is rejected), alpha=0.021(if the ORR but not the PFS null hypothesis is rejected), or alpha=0.025 (if both the ORR and PFS null hypotheses are rejected). [Table 8] shows the bounds and boundary properties for OS hypothesis testing derived using a Lan-DeMets O'Brien-Fleming spending function based on the predicted number of events at the planned time of interim analysis.

If events accrue more slowly than expected or the same as expected, spending will be based on actual information fraction. If events accrue more quickly than expected, cumulative spending based on the expected information fraction will be used in order to save some alpha for analyses that will be performed with more than the originally planned maximum events. For example, at IA2, if 420 OS events have occurred (ie, more than the expected 401 OS events), the alpha spending at IA2 will be according to the expected information fraction (401/551 = 73%) instead of the actual information fraction (420/551 = 76%). The final analysis will use the remaining Type I error that has not been spent at the earlier analyses. The event counts for all analyses will be used to compute correlations.

Also note that if the PFS or ORR null hypothesis is rejected at an interim or final analysis, each OS interim and final analysis test may be compared to its updated bounds considering the alpha reallocation from the PFS or ORR hypothesis.

Table 8 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	Value	α=0.02	α=0.021	α=0.023	α=0.025
IA 2: 73%* N: 692 Events: 401 Month: 37	Z	2.493	2.47	2.426	2.385
	p (1-sided) §	0.0063	0.0068	0.0076	0.0085
	HR at bound %	0.7794	0.7815	0.7849	0.7881
	P(Cross) if HR=1†	0.0063	0.0068	0.0076	0.0085
	P(Cross) if HR=0.75#	0.6513	0.6598	0.6757	0.6902
IA 3: 89%* N: 692 Events: 488 Month: 46	Z	2.272	2.252	2.213	2.178
	p (1-sided) §	0.0115	0.0122	0.0134	0.0147
	HR at bound %	0.814	0.8158	0.8186	0.8212
	P(Cross) if HR=1†	0.0134	0.0142	0.0157	0.0172
	P(Cross) if HR=0.75#	0.8263	0.8316	0.8413	0.85
Final N: 692 Events: 551 Month: 56	Z	2.152	2.133	2.097	2.064
	p (1-sided) §	0.0157	0.0165	0.018	0.0195
	HR at bound %	0.8325	0.834	0.8366	0.8389
	P(Cross) if HR=1†	0.02	0.021	0.023	0.025
	P(Cross) if HR=0.75 #	0.9001	0.9035	0.9097	0.9151

* Percentage of expected number of events at final analysis
§p (1-sided) is the nominal α for testing.
%HR at bound is the approximate HR required to reach an efficacy bound
†P(Cross if HR=1) is the cumulative probability of crossing a bound under the null hypothesis
#P(Cross if HR=0.75) is the cumulative probability of crossing a bound under the alternative hypothesis

3.9 Sample Size and Power Calculations

In the Global Cohort, approximately 692 participants will be randomized in a 1:1 ratio between the 2 arms.

3.9.1 Objective Response Rate

With sample size of ~260 at IA1, the study has ~90% power for detecting a 25% difference in ORR (73% versus 48%) at an initially assigned 0.002 (one-sided) significance level.

3.9.2 Progression-free Survival

There will be ~606 events at the PFS final analysis. With 606 PFS events, the study has ~95% power for detecting a HR of 0.7 at an initially assigned 0.003 (one-sided) significance level.

3.9.3 Overall Survival

There will be ~551 deaths at the OS final analysis. With 551 deaths, the study has ~90% power for detecting an HR of 0.75 at an initially assigned 0.020 (one-sided) significance level.

The sample size (number of participants) calculations are based on the following assumptions: (1) the enrollment period is 28 months and the ramp-up period of enrollment is 6 months; (2) the duration of PFS and OS follow exponential distribution; (3) median PFS is 6.7 months in the control group and the true HR is 0.7; (4) median OS is 13.8 months in the control group and the true HR ratio is 0.75 [16]. Power and interim analysis calculations were performed using the gsDesign R package.

3.10 Subgroup Analyses and Effect of Baseline Factors

To determine whether the treatment effect is consistent across various subgroups, the estimate of the between-group treatment effect (with a nominal 95% CI) for the primary endpoints will be estimated and plotted within each category of each subgroup. The following are examples of classification variables:

- Age category: (<65 versus ≥65 years)
- Sex: (female, male)
- Race: (Asian versus non-Asian)
- Region: Europe/Israel/North America/Australia versus Asia versus Rest of World (including South America)
- PD-L1: Positive versus Negative
- MSI status
- Primary location: Stomach versus GEJ
- Histological subtype: Diffuse versus intestinal versus indeterminate
- Tumor Burden: ≥ median versus <median
- Number of Metastases: ≤2 versus ≥3
- Prior Gastrectomy/Esophagectomy: yes versus no
- Baseline ECOG : 0 versus 1
- Region: US versus ex-US
- Chemotherapy regimen: FP or CAPOX

The consistency of the treatment effect will be assessed descriptively via summary statistics by category for the classification variables listed above. If any level of a subgroup variable has fewer than 20 participants, above analysis may not be performed for this level of the subgroup variable. The subgroup analyses for PFS and OS will be conducted using an unstratified Cox model, and the subgroup analyses for ORR will be conducted using the unstratified Miettinen and Nurminen method.

3.11 Extent of Exposure

The extent of exposure will be summarized as duration of treatment in months and number of cycles or administrations as appropriate.

3.12 Statistical Considerations for Patient-Reported Outcomes (PRO)

The patient-reported outcomes are exploratory objectives in KEYNOTE 811, and thus no formal hypotheses were formulated. This sSAP will focus on PRO endpoints as measured by the EORTC QLQ-C30, EORTC QLQ-STO22 and EuroQol-5D-5L. Note: The PRO will not be analyzed at IA1 because the statistical inferences at IA1 will be based on a subset of the total study population.

EORTC QLQ-C30 is the most widely used cancer specific health related quality of life (QoL) instrument, which contains 30 items and measures 5 functional dimensions (physical, role, emotional, cognitive, and social), 3 symptom items (fatigue, nausea/vomiting, and pain), 6 single items (dyspnea, sleep disturbance, appetite loss, constipation, diarrhea, and financial impact), and a global health and QoL scale [10]. The EORTC QLQ-C30 is a psychometrically and clinically validated instrument appropriate for assessing QoL in oncology studies [11].

The EORTC QLQ-STO22 is a disease-specific questionnaire developed and validated to address measurements specific to gastric cancer. It is one of multiple disease-specific modules developed by the EORTC QoL Group designed for use in clinical studies, to be administered in addition to the EORTC QLQ-C30 to assess disease-specific treatment measurements. It contains 22 items with symptoms of dysphagia (3 items), pain or discomfort (4 items), reflux symptoms (3 items), eating restrictions (4 items), anxiety (3 items), dry mouth, hair loss, taste, and body image.

The EuroQoL-5D-5L (EQ-5D-5L) is a standardized instrument for use as a measure of health outcome and will provide data to develop health utilities for use in health economic analyses [12]. The 5 health state dimensions in the EQ-5D-5L include the following: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension is rated on a 5-point scale from 1 (no problem) to 5 (unable to/extreme problems). The EQ-5D-5L also includes a graded (0 to 100) vertical visual analog scale (VAS) on which the participant rates his or her general state of health at the time of the assessment. This instrument has been used extensively in cancer studies and published results from these studies support its validity and reliability [13].

3.12.1 Analysis Population

The primary analysis approach for the pre-specified exploratory PRO endpoints will be based on a quality of life related full analysis set (FAS) population following the intention-to-treat (ITT) principle and ICH E9 guidelines. This population consists of all randomized patients who have received at least one dose of study medication and have completed at least one PRO assessment. The PRO analysis will be conducted in the Global Cohort only.

3.12.2 Scoring Algorithm

QLQ-C30 Scoring: For each scale or item, a linear transformation will be applied to standardize the score as between 0 and 100, according to the corresponding scoring standard. For functioning and global health status/quality-of-life scales, a higher value indicates a better level of function; for symptom scales and items, a higher value indicates increased severity of symptoms.

According to the QLQ-C30 Manuals, if items $I_1, I_2, \ldots, I_n$ are included in a scale, the linear transformation procedure is as follows:

1. Compute the raw score: $RS = (I_1 + I_2 + \ldots + I_n)/n$
2. Linear transformation to obtain the score S:
3. Function scales: $$S = \left(1 - \frac{RS - 1}{Range}\right) \times 100$$
4. Symptom scales/items: $$S = \frac{RS - 1}{Range} \times 100$$
5. Global health status/QoL: $S = \dfrac{RS - 1}{Range} \times 100$

Range is the difference between the maximum possible value of *RS* and the minimum possible value. If more than half of the items within one scale are missing, then the scale is considered missing, otherwise, the score will be calculated as the average score of those available items [13].

QLQ-STO22 scoring: QLQ-STO22 contains 22 items with symptoms of dysphagia (3 items), pain or discomfort (4 items), reflux symptoms (3 items), eating restrictions (4 items), anxiety (3 items), dry mouth, hair loss, taste, and body image [14]. A linear transformation will be applied to standardize the scores between 0 (worst) and 100 (best) as described above for the QLQ-STO22 Scoring.

EQ-5D-5L scoring: EQ-5D-5L utility score will be calculated based on the European algorithm [15]. The five health state dimensions in this instrument include the following: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has five response levels.

3.12.3 Completion and Compliance Rate Summary for PROs

Completion and compliance of QLQ-C30, QLQ-STO22 and EQ-5D by visit and by treatment will be described. Numbers and percentages of complete and missing data at each visit will be summarized for each of the treatment groups. An instrument is considered complete if at least one valid score is available according to the missing item rules outlined in the scoring manual for the instrument.

Completion rate of treated participants (CR-T) at a specific time point is defined as the number of treated participants who complete at least one item over the number of treated participants in the PRO analysis population.

$$\text{CR-T} = \frac{\mathit{Number\ of\ treated\ participants\ who\ complete\ at\ least\ one\ item}}{\mathit{Number\ of\ treated\ participants\ in\ the\ PRO\ analysis\ population}}$$

The completion rate is expected to shrink in the later visit during study period due to the participants who discontinued early. Therefore, another measurement, compliance rate of eligible participants (CR-E) will also be employed as the support for completion rate. CR-E is defined as the number of treated participants who complete at least one item over number of eligible participants who are expected to complete the PRO assessment, not including the participants missing by design such as death, discontinuation, translation not available.

$$\text{CR-E} = \frac{\mathit{Number\ of\ treated\ participants\ who\ complete\ at\ least\ one\ item}}{\mathit{Number\ of\ eligible\ participants\ who\ are\ expected\ to\ complete}}$$

The reasons of non-completion and non-compliance will be provided in supplementary table:

- Completed as scheduled
- Not completed as scheduled
- Off-study: not scheduled to be completed.

In addition, reasons for non-completion as scheduled of these measures will be collected using “miss_mode” forms filled by site personnel and will be summarized in table format. The schedule (study visits and estimated study times) and mapping of study visit to analysis visit for PRO data collection is provided in [Table 9].

The Schedule for PRO Data Collection:

Table 9 PRO Data Collection Schedule

Treatment	Week											Discontinuation Visit	Follow-up Visit
	0	3	6	9	12	18	24	30	36	42	48		
SOC+MK-3475	C1	C2	C3	C4	C5	C7	C9	C11	C13	C15	C17	X	X
SOC+PBO	C1	C2	C3	C4	C5	C7	C9	C11	C13	C15	C17	X	X
C: Cycle; D: Day Each cycle is 3 weeks for MK-3475.													

The general rule of mapping relative day to analysis visit is provided in [Table 10].

Table 10 Mapping Relative Day to Analysis Visit

	Week										
	0	3	6	9	12	18	24	30	36	42	48
Day	1	22	43	64	85	127	169	211	253	295	337
Range	C1D1	2-32	33-53	54-74	75-106	107-148	149-190	191-232	233-274	275-316	317-

At each scheduled visit, three instruments, EORTC QLQ-C30, EORTC QLQ-STO22 and EQ-5D-5L, will be collected. If a patient does not complete the PRO instruments, the site staff will record the reason for the missing from pre-defined choices. If there are multiple PRO collections within any of the stated time windows, we use the closest collection to the target day.

3.12.4 PRO Endpoints

Mean change from baseline

- The mean score changes from baseline at 24 weeks as measured by the EORTC QLQ-C30 global health status/quality of life scale.
- The mean score change from baseline at 24 weeks for QLQ-C30 physical functioning scale
- The mean score change from baseline at 24 weeks for all QLQ-STO22 sub-scales/items.
- The mean score change from baseline at 24 weeks for EQ-5D-5L VAS.

Note: Week 24 is the latest time point where both the completion and compliance rates are acceptable (e.g. close to 60% and 80%, respectively) based on blinded data review.

Time-to-Deterioration (TTD)

- The number and proportions of deterioration/stable/improvement from baseline to the latest time point where the completion rate is still high enough based on blinded data review, the time to deterioration (TTD), and the overall improvement rate during the study. Specifically:
    - The QLQ-C30 global health status/quality of life scale (based on two items).
    - The QLQ-C30 physical functioning scale (based on five items)
    - The QLQ-C30 symptom scale nausea/vomiting (two items).
    - The QLQ-C30 single item appetite loss.
    - The QLQ-STO22 symptom scale pain (four items).

For multi-item scale(s), the analysis will focus on the subscale score rather than each single item.

Overall Improvement and Overall Improvement/Stability Rate

- **Overall improvement**

Improvement is defined as a 10-point or more increase in score (in the positive direction) from baseline at any time during the study and confirmed by a 10- point or more improvement at a visit scheduled at least 6 weeks later [6].

- **Overall improvement/stability**

Stability is defined as, when the criteria for improvement are not met, a less than 10 points worsening in score from baseline at any time during the study and confirmed by a less than 10 points worsening at a visit scheduled at least 6 weeks later. Overall improvement/stability is defined as the composite of improvement and stability.

The following domains will be analyzed:

- The QLQ-C30 global health status/quality of life scale (based on two items).
- The QLQ-C30 physical functioning scale (based on five items)
- The QLQ-C30 symptom scale nausea/vomiting (two items).
- The QLQ-C30 single item appetite loss.
- The QLQ-STO22 symptom scale pain (four items).

For multi-item scale(s), the analysis will focus on the subscale score rather than each single item.

3.12.5 Statistical Methods

This section describes statistical methods for the PRO endpoints defined in [Sec. 3.12.4]. [Table 11] presents the summary of planned statistical for PRO endpoints.

Table 11 Planned Statistical Analysis of PRO

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Mean change from baseline	cLDA model	FAS	Model-based.
TTD	stratified log-rank test and HR estimation using stratified Cox model with Efron's tie handling method	FAS	Censored according to rules in [Table 12].
Overall improvement rate and overall improvement/stability rate	Stratified Miettinen and Nurminen method	FAS	Participants with missing data are considered not achieving improvement/stability.
Abbreviations: cLDA = constrained longitudinal data analysis, FAS = full analysis set, QoL = quality of life.			

Mean change from baseline

To assess the treatment effects on the PRO score change from baseline, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [17] will be applied, with the PRO score as the response variable, and treatment, time, the treatment by time interaction, and stratification factors used for randomization as covariates. The treatment difference in terms of least square (LS) mean change from baseline will be estimated from this model together with 95% CI. Model-based LS mean with 95% CI will be provided by treatment group for PRO scores at baseline and post-baseline time point. Descriptive statistics (mean change from baseline and SE) of observed data with no imputation for missing data on global HRQoL score and/or key functional and symptom scales of the EORTC QLQ-C30 and EORTC QLQ-C22 will be plotted.

The cLDA model assumes a common mean across treatment groups at baseline and a different mean for each treatment at each of the post-baseline time points. In this model, the response vector consists of baseline and the values observed at each post-baseline time point. Time is treated as a categorical variable so that no restriction is imposed on the trajectory of the means over time. The cLDA model is specified as follows:

$$E\left(Y_{ijt}\right) = \gamma_0 + \gamma_{jt} I(t > 0) + \beta X_i, j = 1,2,,3,..,n;\ t = 0,1,2,3,..k$$

where Y_{ijt} is the PRO score for participant i, with treatment assignment j at visit t; γ_0 is the baseline mean for all treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t; X_i is the stratification factor (binary) vector for this participant, and β is the coefficient vector for stratification factors. An unstructured covariance matrix will be used to model the correlation among repeated measurements. If the unstructured covariance model fails to converge with the default algorithm, then Fisher scoring algorithm or other appropriate methods can be used to provide initial values of the covariance parameters. In the rare event that none of the above methods yield convergence, a structured covariance such as Toeplitz can be used to model the correlation among repeated measurements. In this case, the asymptotically unbiased sandwich variance estimator will be used. The cLDA model implicitly treats missing data as missing at random (MAR).

In addition, the model-based LS mean change from baseline to the specified post-baseline time point together with 95% CI will be plotted in bar charts for EORTC QLQ-C30 global health status/quality of life scores, all functioning and symptom scores, and the EORTC QLQ-STO22 Symptom Scale Pain.

Time-to-Deterioration (TTD)

The Kaplan-Meier method will be used to estimate the TTD curve for each treatment group. The estimate of median time to deterioration and its 95% confidence interval will be obtained from the Kaplan-Meier estimates. The treatment difference in TTD will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling and with a single treatment covariate will be used to assess the magnitude of the treatment difference (i.e, HR). The HR and its 95% CI will be reported. The same stratification factors used for randomization will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model.

The approach for the time-to-deterioration analysis will be based on the assumption of non-informative censoring. The participants who do not have deterioration on the last date of evaluation will be censored. [Table 12] provides censoring rule for TTD analysis.

Table 12 Censoring Rules for Time-to-Deterioration

Scenario	Outcome
Deterioration documented	Event observed at time of assessment (first deterioration)
Ongoing or discontinued from study without deterioration	Right censored at time of last assessment
No baseline assessments	Right censored at treatment start date

Overall Improvement and Overall Improvement/Stability

Stratified Miettinen and Nurminen’s method will be used for comparison of the overall improvement rate and Overall Improvement/Stability rate between the treatment groups. The difference in overall improvement rate and its 95% CI from the stratified Miettinen and Nurminen’s method with strata weighting by sample size will be provided. The stratification factors used for randomization will be applied to the analysis.

The point estimate of overall improvement rate will be provided by treatment group, together with 95% CI using exact binomial method by Clopper and Pearson.

The same strategy of combination of small strata defined for the PFS analysis (see [Sec. 3.6.1.1]) will be used for the PRO analysis.

4 REFERENCES

1. Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med 1985; 4:213-26.

2. Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. Stat Biopharm Res 2013;5(4):311-20.

3. Finkelstein DM (1986). A proportional hazards model for interval-censored failure time data. Biometrics 42,845-54.

4. Uno, H., et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol. 2014;32 (22):2380-2385.

5. Odell P.M., Anderson K.M., Kannel B.W. New models for predicting cardiovascular events. Journal of Clinical Epidemiology 1994;47(6):583-592.

6. Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16:139-44.

7. Robins, J.M., Tsiatis, A.A. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics-Theory and Methods, 1991;20(8):2609-2631.

8. Latimer N.R., Abrams K.R., Siebert U. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. BMC Medical Research Methodology. 2019; 19:69.

9. Nicholas R. Latimer, Keith R. Abrams, et al. Adjusting survival time estimates to account for treatment switching in randomized controlled trials—an economic evaluation eontext: methods, limitations, and recommendations. Med Decision Making 2014; 34:387–402.

10. Aaronson NK, Ahmedzai S, Bergman B, Bullinger M, Cull A, Duez NJ, et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 1993;85(5):365-76.

11. Rabin, R.; de Charro, F. EQ-5D: a measure of health status from the EuroQol group. Ann Med 2001;33:337-43.

12. Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes 2007;5:1-8.

13. The EORTC QLQ-C30 Manuals. Reference Values and Bibliography.

14. Blazeby JM, Conroy T, Bottomley A, et al. Clinical and Psychometric Validation of a Questionnaire Module, the EORTC QLQ-STO22, to Assess Quality of Life in Patients with Gastric Cancer. Eur J Cancer 40:2260-2268, 2004.

15. EuroQol Research Foundation. EQ-5D-5L User Guide, 2019. Available from: https://euroqol.org/publications/user-guides.

16. Bang YJ, Van Cutsem E, Feyereislova A, Chung HC, Shen L, Sawaki A, et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial. Lancet. 2010 Aug 28;376(9742):687-97.

17. Liang K, Zeger, S (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhyā: The Indian Journal of Statistics, 62 (Series B), 134-148.