Official Protocol Title:	A Phase III, Randomized, Double-Blind, Clinical Trial of Pembrolizumab (MK-3475) plus Chemotherapy (XP or FP) versus Placebo plus Chemotherapy (XP or FP) as Neoadjuvant/Adjuvant Treatment for Subjects with Gastric and Gastroesophageal Junction (GEJ) Adenocarcinoma (KEYNOTE-585)
NCT number:	NCT04882241
Document Date:	15-Mar-2024

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 10 (MK-3475-585-10) with regard to the statistical analysis plan. Additionally, the following change and clarification are added:

- In Section 4, clarify that the originally prespecified analyses (see Section 4.7) for China subpopulation (including the China participants enrolled in the global portion and the extension portion) are revised; primary efficacy analyses for China subpopulation are descriptive and will be conducted based on the same data cutoff date as the final analysis for the global population and analysis endpoints applicable to the China subpopulation are also clarified.

3 ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

Key elements of the statistical analysis plan are summarized below; the comprehensive plan is provided in provided in Sections 3.2 through 3.12.

Study Design Overview	A Phase III, Randomized, Double-blind Clinical Trial of Pembrolizumab (MK-3475) plus Chemotherapy (XP or FP) versus Placebo plus Chemotherapy (XP or FP) as Neoadjuvant/Adjuvant Treatment for Gastric and Gastroesophageal Junction (GEJ) Adenocarcinoma (KEYNOTE-585).
Treatment Assignment	Main study (XP/FP): Approximately 800 participants will be randomized in the main study in a 1:1 ratio to receive pembrolizumab (MK-3475) plus chemotherapy (XP/FP) or placebo plus chemotherapy (XP/FP). The main study (XP/FP) is double-blinded. FLOT Cohort: Approximately 200 participants will be randomized in a 1:1 ratio to receive pembrolizumab plus FLOT or placebo plus FLOT. The FLOT Cohort is double blinded. Stratification factors are: 1) Geographic regions: (Asia versus Non-Asia), 2) Tumor staging: II versus III versus IVa, and 3) Backbone chemotherapy XP/FP versus FLOT.

Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Participants as Treated (APaT)
Primary Endpoints	Main study (XP/FP): • Event-free Survival (EFS) assessed by investigators • Rate of Pathological Complete Response (pathCR) • Overall Survival (OS) FLOT Cohort: • AEs; Study treatment discontinuations due to AEs
Secondary Endpoints	Main study (XP/FP): • Disease-free Survival (DFS) assessed by investigators • AEs; Study treatment discontinuations due to AEs Main study (XP/FP) and FLOT Cohort Combined: • Overall Survival (OS) • Event-free Survival (EFS) assessed by investigators • AEs; Study treatment discontinuations due to AEs
Statistical Methods for Key Efficacy Analyses	The primary hypothesis for EFS and OS will be evaluated by comparing pembrolizumab (MK-3475) plus chemotherapy (XP/FP) to placebo plus chemotherapy (XP/FP) using a stratified log-rank test. Estimation of the hazard ratio will be done using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. Treatment comparison of the pathCR rates will be performed using the stratified Miettinen and Nurminen method in the Main Study (XP/FP). Same methods will be used for the analysis of OS and EFS in the Main Study (XP/FP) and the FLOT Cohort combined as secondary hypotheses too.
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 M&N method [1].

Interim Analysis (IA)	Three interim analyses (IA) are planned in this study. Results will be reviewed by an external DMC. Details are provided in Section 3.7. **IA1:** **Timing:** to be performed after ~ 299 EFS events have occurred in the Main Study (XP/FP) AND the enrollment of the study has completed. IA1 is expected to be ~ 41 months after first participant randomized. **Primary purpose**: final pathCR rate analysis and interim EFS and OS analyses. **IA2:** **Timing:** to be performed after ~ 336 EFS events have occurred in the Main Study (XP/FP) AND ~ 7 months after IA1. IA2 is expected to be ~ 48 months after first participant randomized. **Primary purpose:** interim EFS and OS analyses **IA3:** **Timing:** to be performed after ~ 369 EFS events have occurred in the Main Study (XP/FP) AND ~ 12 months after IA2. IA3 is expected to be ~ 60 months after first participant randomized. **Primary purpose:** final EFS analysis and interim OS analysis. **Final analysis (FA):** **Timing:** to be performed after ~ 361 OS events have occurred in the Main Study (XP/FP) AND ~ 12 months after IA3. FA is expected to be ~ 72 months after first participant randomized. **Primary purpose:** final OS analysis. Note that for IA2, IA3 and FA, if the EFS events or the OS events accrue slower than expected, the analysis may be delayed up to 3 months after the projected timing.
Multiplicity	The overall type I error over the primary and secondary endpoints (OS, EFS and pathCR rate) is strongly controlled at 2.5% (one-sided), with initially 2.45% allocated to the EFS in the Main Study (XP/FP) and 0.05% allocated to the pathCR rate in the Main Study (XP/FP). By using the graphical approach of Mauer and Bretz [2], if one hypothesis is rejected, the alpha will be shifted to other hypotheses.

Sample Size and Power	The planned sample size is approximately 800 participants in the Main Study (XP/FP), and approximately 200 in the FLOT Cohort. The final analysis of this study is driven by both a targeted number of events and a minimum duration of follow-up. For Main Study (XP/FP): For the pathCR rate, the study has ~ 99.9% power to detect a true pathCR rate improvement from 5% to 20% (i.e., a difference of 15 percentage points) at alpha = 0.05% (1-sided) with ~ 800 participants who would have been randomized at least 6 months prior to IA1 data cutoff. It is estimated that there will be ~ 369 events in the Main Study (XP/FP) participants at the final EFS analysis. With 369 EFS events, the study has ~ 87% power to demonstrate that pembrolizumab (MK-3475) with chemotherapy (XP/FP) is superior to chemotherapy alone at a 1-sided 2.45% alpha-level, if the underlying hazard ratio of EFS is 0.72. It is estimated that there will be ~ 361 deaths in the Main Study (XP/FP) participants at the final OS analysis. With 361 deaths, the study has ~ 63% power to demonstrate that pembrolizumab (MK-3475) with chemotherapy (XP/FP) is superior to chemotherapy alone at a 1-sided 0.64% alpha-level, if the underlying hazard ratio of OS is 0.74. When both EFS in the Main Study (XP/FP) and EFS in the Main Study (XP/FP) and FLOT Cohort combined are rejected, the study has 80% power to demonstrate that pembrolizumab (MK-3475) with chemotherapy (XP/FP) is superior to chemotherapy alone at a 1-sided 2.45% alpha-level, if the underlying hazard ratio of OS is 0.74.

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 or scientific review has been performed, protocol deviations have been identified, and data have been declared final and complete.

The Sponsor will generate the randomized allocation schedule(s) for study treatment assignment for this protocol, and the randomization will be implemented in IVRS.

Planned efficacy interim analyses are described in [Sec. 3.7]. Blinding to treatment assignment will be maintained at all investigational sites.

Treatment-level results of the interim analysis will be provided by the unblinded external statistician to the external DMC, which will review the safety and/or efficacy data. Limited additional Sponsor personnel may be unblinded to the treatment level results of the interim analysis (analyses), if required, in order to act on the recommendations of the DMC. The extent to which individuals are unblinded with respect to results of interim analyses will be documented in the DMC charter.

The DMC will serve as the primary reviewer of the interim analyses and will make recommendations for discontinuation of the study or modification to an executive oversight committee of the Sponsor. Depending on the recommendation of the DMC, the Sponsor may prepare a regulatory submission. If the DMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee may be unblinded to results at the treatment level in order to act on these recommendations. Additional logistical details, revisions to the above plan and data monitoring guidance will be provided in the DMC Charter.

Treatment-level results (only safety data) from the FLOT Safety Cohort will be provided by the unblinded external statistician to the siDMC, which will review the data to determine if the FLOT regimen may be incorporated as one of the standards of care chemotherapy backbones in the main study. As above, limited additional Sponsor personnel may be unblinded to the treatment-level results of these analyses if required, in order to act on the recommendations of the siDMC. The extent to which individuals are unblinded with respect to results of these analyses will be documented. Further details on the processes by which recommendations and decisions are reached and communicated will be documented in the siDMC charter.

As of Amendment 06, the FLOT Safety Cohort will be expanded to 200 participants to further characterize the safety profile of the combination of FLOT with pembrolizumab and will be designated the FLOT Cohort. Future interim analyses of the FLOT Cohort data will be monitored by the external DMC. As above, limited additional Sponsor personnel may be unblinded to the treatment level results of the interim analysis (analyses), if required, in order to act on the recommendations of the DMC. The extent to which individuals are unblinded with respect to results of interim analyses will be documented in the DMC charter.

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.

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Protocol Section 4.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Primary Endpoints

- **Event-free survival (EFS) based on RECIST 1.1 as assessed by the investigator**

  EFS is defined as the time from randomization to the first of the following events: radiographic disease progression per RECIST 1.1, local or distant recurrence as assessed by CT scan or biopsy if indicated (for participants who are disease free after surgery), clinical progression as evidenced by peritoneal carcinomatosis confirmed by pre-operative laparoscopy or laparotomy (for participants who are confirmed to be free of peritoneal involvement by laparoscopy at screening), and death due to any cause.

  A second primary malignancy is not considered as an EFS event. Radiographic PD during the neoadjuvant phase that does not preclude successful surgery (ie, disease free after surgery) is not considered as an EFS event.

  For participants with documented peritoneal carcinomatosis found prior to surgery (as evidenced by pre-operative laparoscopy or laparotomy) but without radiographic PD during the neoadjuvant phase and do not undergo screening laparoscopy, EFS data will be censored as of the time of the pre-operative laparoscopy or laparotomy. See [Sec. 3.6.1.1] for the definition of censoring.

- **Pathological Complete Response (pathCR) Rate based on central review**

  pathCR rate is defined as the proportion of participants having pathCR. Pathologic complete response (pathCR) is defined as no invasive disease within an entirely submitted and evaluated gross lesion, and histologically negative nodes.

- **Overall Survival (OS)**

  OS is defined as the time from randomization to death due to any cause. Participants without documented death at the time of analysis will be censored at the date of last known alive.

Secondary

- **Disease free survival (DFS) based on RECIST 1.1 as assessed by investigator**

  DFS is defined as the time from post-surgery baseline scan until the first occurrence of: local or distant recurrence, death from any cause.

Exploratory

- **Event-free survival (EFS) based on RECIST 1.1 as assessed by BICR**
- **Disease free survival (DFS) based on RECIST 1.1 as assessed by BICR**
- **R0 resection rate based on local review**
- **pathCR rate based on local review**

3.4.2 Safety Endpoints

Safety and tolerability will be assessed by clinical review of all relevant parameters including AEs, AEs leading to discontinuation, SAEs, fatal AEs, physical examination, vital signs, ECOG performance status, and lab values. Furthermore, specific events will be collected and designated as ECIs as described in Protocol Section 9.3.7.

3.4.3 Patient-Reported Outcome Endpoints

The PRO endpoints include results from the EORTC QLQ-C30, EORTC QLQ-STO22, and EQ-5D-5L questionnaires.

The EORTC QLQ-C30 is the most widely used cancer specific HRQoL 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 quality of life (QoL) scale [3]. Thus, the exploratory objective is to assess mean changes from baseline in the global health status/QoL scale from the EORTC QLQ-C30.

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 QLG (Quality of Life Group) designed for use in clinical trials, to be administered in addition to the QLQ-C30 to assess disease-specific treatment measurements. It contains 22 items (5 scales and 4 single items): 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 [4].

The EuroQoL-5D-5L (EQ-5D-5L) is a standardized instrument for use as a measure of health outcome and will provide data for use in economic models and analyses including developing health utilities or quality adjusted life years. The 5 health state dimensions in this instrument 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 (extreme problem). The EQ-5D-5L also includes a graded (0 to 100) vertical visual analog scale 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 [5].

Key PRO Endpoint:

Mean change in Score at primary PRO analyses time point

The time point for the mean change from baseline is defined as the latest time point at which completion rate of treated participants ≥ 60% and Week 39 was selected based on blinded data review prior to the database lock for any PRO analysis in the adjuvant phase. Note, for the analysis in the neoadjuvant phase, the only post-baseline time point will be Week 6 by design.

The mean score changes from baseline to the primary analysis time point for:

- The QLQ-C30 global health status/QoL score.
- The QLQ-C30 physical functioning scale.
- The QLQ-C30 symptom scale nausea/vomiting.
- The QLQ-C30 single item appetite loss.
- The QLQ-STO22 symptom scale pain.
- The EQ-5D-5L VAS.

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 change in score in the positive direction from baseline at any time during the study and confirmed by a 10- point or more improvement at the next consecutive PRO assessment timepoint [6]. Overall improvement will be assessed in the adjuvant phase given that there is only one post-baseline PRO assessment timepoint in the neoadjuvant phase.

- **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 mean score changes from baseline to the primary analysis time point for:

- The QLQ-C30 global health status/QoL score.

- The QLQ-C30 physical functioning scale.
- The QLQ-C30 symptom scale nausea/vomiting.
- The QLQ-C30 single item appetite loss.
- The QLQ-STO22 symptom scale pain.

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

3.5 Analysis Populations

3.5.1 Efficacy Analysis Populations

The Intention-to-Treat (ITT) population will serve as the population for OS, EFS and pathCR rate. All randomized participants will be included in this population. Participants will be included in the treatment group to which they are randomized. ITT population consists of all randomized participants whether or not treatment was administered. Any participant who receives a randomization number will be considered to have been randomized. For the PathCR rate, the ITT population will only include participants randomized at least 6 months prior to the data cutoff of IA1.

DFS analysis population will be those participants who are disease free at the post-surgery baseline scan.

Separate efficacy analyses will be provided for Main Study ITT, FLOT ITT and Main Study and FLOT combined ITT.

Details on the approach to handling missing data are provided in [Sec. 3.6], Statistical Methods.

3.5.2 Safety Analysis Populations

The All Participants as Treated (APaT) population will be used for the analysis of safety data. The APaT population consists of all randomized participants who received 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 1 cycle but receives the correct treatment for all other cycles will be analyzed according to the correct treatment group and a narrative will be provided for any events that occur during the cycle for which the participant is 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.

Details on the approach to handling missing data for safety analyses are provided in [Sec. 3.6] Statistical Methods.

3.5.3 PRO Analysis Populations

PRO analyses are based on the PRO Full Analysis Set (FAS) population, defined as randomized subjects who have at least one PRO assessment available and have received at least one dose of study treatment.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

3.6.1.1 Event-free Survival (EFS)

The non-parametric Kaplan Meier method will be used to estimate the EFS curve in each treatment group. The treatment difference in EFS will be assessed by the stratified log rank test and the p-value will be provided. 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 groups. The HR and its 95% CI from the stratified Cox model for the treatment covariate will be reported. Geographic region (Asia versus Non-Asia) and Tumor staging (II versus III versus IVa) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model in the Main Study. Stratified analyses will be based on collapsed strata by combining strata with small number of participants or events. The collapsed strata will be based on blinded data taking into considerations of both clinical relevance and actual counts of subjects/events. As a result, the following 5 strata will be used:

- Asia + Stage II
- Asia + Stage III
- Non-Asia + Stage II
- Non-Asia + Stage III
- Any Region + Stage IVa

Geographic region (Asia versus Non-Asia), Tumor staging (II versus III versus IVa) and Backbone chemotherapy XP/FP versus FLOT will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model in the Main Study and FLOT Cohort combined. The following 8 strata will be used:

- Asia + Stage II + any Backbone Chemotherapy
- Asia + Stage III + any Backbone Chemotherapy
- Asia + Stage IVa + any Backbone Chemotherapy
- Non-Asia + Stage II + Backbone Chemotherapy FP/XP No
- Non-Asia + Stage II + Backbone Chemotherapy FP/XP Yes
- Non-Asia + Stage III + Backbone Chemotherapy FP/XP No
- Non-Asia + Stage III + Backbone Chemotherapy FP/XP Yes
- Non-Asia + Stage IVa + Any Backbone Chemotherapy

For the primary analysis of EFS as assessed by investigators, the true date of disease progression/recurrence will be approximated by the date of the first assessment at which event is objectively documented, regardless of discontinuation of study drug, missed study visits, or initiation of new anti-cancer therapy. Participants who do not experience an event at the time of analysis will be censored at the last disease assessment. As a sensitivity analysis, this same approach will also be applied to analysis of EFS as assessed by BICR.

In order to evaluate the robustness of the EFS endpoint, we will perform a sensitivity analysis with a different set of censoring rules. The sensitivity analysis is the same as the primary analysis except that a participant with initiation of new anti-cancer therapy will be censored at the last disease assessment prior to initiation of new anti-cancer therapy if a participant 1) has developed documented disease progression, recurrence, or death immediately after new anti-cancer therapy or 2) has not developed documented disease progression, recurrence 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].

Table 1 Censoring Rules for Primary and Sensitivity Analyses of EFS

Situation	Primary Analysis	Sensitivity Analysis
PD, recurrence, or death documented before new anti-cancer therapy, if any	EFS event at date of documented PD, recurrence, or death	EFS event at date of documented PD, recurrence, or death
PD, recurrence or death documented immediately after new anti-cancer therapy, if any	EFS event at date of documented PD, recurrence, or death	Censored at last disease assessment before new anti-cancer treatment
No PD, no recurrence and no death; and new anti-cancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment
No PD, no recurrence and no death; new anti-cancer treatment is initiated	Censored at last disease assessment	Censored at last disease assessment before new anti-cancer treatment

EFS=event-free survival; PD=progressive disease.

The proportional hazards assumption of the Cox model will be examined using both graphical and analytical methods for the primary EFS analysis. The log[-log] of the survival function vs. time for EFS will be plotted for the comparison between the pembrolizumab arm and the control arm in each stratum. 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 [7], parametric method [8] 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 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 [9] with logrank FH (0, 1), FH (1, 1) at the final analysis of EFS to account for the potential loss of power with logrank test when the proportional hazard assumption is violated.

3.6.1.2 Pathological Complete Response (pathCR) Rate

The stratified Miettinen and Nurminen's method with strata weighted by sample size will be used for the comparison of the difference in pathCR rates between pembrolizumab with chemotherapy (XP/FP) and chemotherapy alone. The stratification factors used for randomization (see Protocol Section 7.3.1) will be applied. The same strategy of combination of small strata defined for the EFS analysis (see [Sec. 3.6.1.1]) will be used for the pathCR analysis.

Sensitivity analyses will be performed for pCR rates using Cochran-Mantel-Haenszel test. Associated odds ratios and 95% CIs will be calculated. Additional supportive unstratified analyses may also be provided.

The primary pathCR analysis is for the ITT population. Two sensitivity analyses will be considered as follows: 1) the modified ITT population: to exclude participants who went through on-study surgery with no local pathCR result or positive local pathCR while no central pathCR result from the ITT population; and 2) the modified surgery population: to exclude randomized participates without on-study surgeries and to exclude participants with no local pathCR result or positive local pathCR while no central pathCR result among randomized participants with on-study surgeries.

3.6.1.3 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 measured by the hazard ratio (HR). The HR and its

95% CI from the Cox model for the treatment covariate will be reported. The stratification factors used for randomization (See Protocol Section 7.3.1) will be applied to both the stratified log-rank test and the stratified Cox model. The same strategy of combination of small strata defined for the EFS 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 anti-cancer therapies on OS may be performed as model assumptions permit. Three recognized methods will be included: 1) the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis (1989) [10]; 2) the two-stage model proposed by Latimer [11]; and 3) the Inverse-Probability-of-Censoring Weighting (IPCW) model [12] in which an examination of the appropriateness of the data to the assumptions is required by the methods.

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

3.6.1.4 Disease-free Survival (DFS)

The non-parametric Kaplan Meier method will be used to estimate the DFS curve for each treatment group.

In order to evaluate the robustness of the DFS endpoint, A sensitivity analysis with a different set of censoring rules will be considered when new anti-cancer therapy is initiated. The censoring rules for primary and sensitivity analyses are summarized in [Table 2].

Table 2 Censoring Rules for Primary and Sensitivity Analyses of DFS

Situation	Primary Analysis	Sensitivity Analysis
recurrence, or death documented before new anti-cancer therapy, if any	DFS event at date of recurrence, or death	DFS event at date of recurrence, or death
recurrence or death documented immediately after new anti-cancer therapy, if any	DFS event at date of recurrence, or death	Censored at last disease assessment before new anti-cancer treatment
no recurrence and no death; and new anti-cancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment
no recurrence and no death; new anti-cancer treatment is initiated	Censored at last disease assessment	Censored at last disease assessment before new anti-cancer treatment
DFS=disease free survival.		

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			
EFS per RECIST 1.1 by investigator	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	• Primary censoring rule • Sensitivity analysis (More details are provided in [Table 1] Censoring Rules for Primary and Sensitivity Analyses of EFS)
pathCR rate	Test and Estimation: Stratified M&N method with sample size weights††	ITT*	Participants with relevant data missing are considered non responders
OS	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at the last known alive date

EFS=event-free survival; ITT=intention to treat; OS=overall survival

†Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization (See Protocol Section 7.3.1 – Stratification) will be applied to the analysis model. The strategy of combination of small strata is defined in [Sec. 3.6.1.1].

†† Miettinen and Nurminen method.

*For IA1 in [Sec. 3.7] pathCR rate analysis, participants randomized at least 6 months before the data cutoff will be included.

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 AEs, laboratory tests, vital signs, etc.

Adverse Events

Adverse events will be coded using the standard Medical Dictionary for Regulatory Activities (MedDRA) and grouped system organ class. AEs will be graded by the investigator according to the CTCAE, version 4.0.

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. 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 groups in this study. Based on the safety knowledge accumulated across the program to date, there are no events that warrant consideration for Tier 1.

Tier 2 parameters will be assessed via point estimates with 95% CIs provided for differences in the proportion of participants with events using the Miettinen and Nurminen method, an unconditional, asymptotic method [1].

Membership in Tier 2 requires that at least 10% of participants in any treatment group exhibit the event. The threshold of at least 10% of participants was chosen for Tier 2 events because the population enrolled in this study is in critical condition and usually experiences various AEs of similar types regardless of treatment; events reported less frequently than 10% of participants would obscure the assessment of the 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 1 of the treatment groups) and SAEs (≥2% of participants in 1 of the treatment groups) will be considered Tier 2 endpoints. Because many 95% CIs may be provided without adjustment for multiplicity, the CIs should be regarded as a helpful descriptive measure to be used in safety review, not as a formal method for assessing the statistical significance of the between-group differences.

Safety endpoints that are not Tier 1 or 2 events are considered Tier 3 events. Only point estimates by treatment group are provided for Tier 3 safety parameters.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	p-Value	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 (≥2% of participants in one of the treatment groups)		X	X
Tier 3	Any AE			X
	Any Serious AE			X
	Any Grade 3-5 AE			X
	Any Drug related AE			X
	Any Serious and Drug related AE			X
	Any Grad 3-5 and Drug related AE			X
	Discontinuation due to AE			X
	Any AE leading to death			X
	Specific AEs, SOCs (incidence <10% of participants in all of the treatment groups)			X
	Change from baseline results (laboratory test toxicity grade, vital signs)			X

AE=adverse event; CI=confidence interval; SAE=serious adverse event; SOC=system organ class; X = results will be provided.

Neoadjuvant phase begins with the first dose of neoadjuvant treatment until the start of adjuvant treatment and includes the surgery. Adjuvant phase begins with the first dose of adjuvant treatment. Adverse events are summarized for neoadjuvant phase, adjuvant phase, and neoadjuvant phase and adjuvant phase combined. Laboratory results, change from baseline results in ECGs and vital signs are summarized in neoadjuvant phase and adjuvant phase combined only.

3.6.3 Statistical Methods for Patient-Reported Outcomes (PRO) Analyses

3.6.3.1 Scoring Algorithm

EORTC QLQ-C30 Scoring

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

All of the scales and single-item measures will follow a standardization procedure prior to analysis so that scores range from 0 to 100. A high scale score represents a higher response level. Thus a high score for a functional scale represents a high / healthy level of functioning; a high score for the global health status / QoL represents a high QoL; but a high score for a symptom scale / item represents a high level of symptomatology / problems.

According to the EORTC QLQ-C30 Scoring Manual [13], 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.

Specifically, if items $I_1, I_2, ..., I_n$ are included in a scale, the scoring procedure is as follows:

1. Compute the raw score: $RS = (I_1 + I_2 + ... + I_n)/n$

2. Linear transformation to obtain the score S:

Function scales: $S = \left(1 - \frac{RS - 1}{Range}\right) \times 100$

Symptom scales / items: $S = \frac{RS - 1}{Range} \times 100$

Global health status / QoL: $S = \frac{RS - 1}{Range} \times 100$

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. 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.

EORTC QLQ-STO22 Scoring

The scoring approach for the QLQ-STO22 is identical in principle to that for the function and symptom scales / single items of the QLQ-C30. A linear transformation will be applied to standardize the scores between 0 and 100 as described above for the EORTC QLQ-C30 scoring.

EQ-5D-5L Scoring

EQ-5D-5L utility score will be calculated based on the European algorithm [14]. 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.

The Schedule for PRO Data Collection:

Table 5 PRO Data Collection Schedule in Neoadjuvant Phase

Treatment	Week	
	0	6
SOC+MK-3475	C1	C3
SOC+PBO	C1	C3
C: Cycle; D: Day Each cycle is 3 weeks for MK-3475.		

The general rule of mapping relative day in the neoadjuvant phase to analysis visit is provided in [Table 6].

Table 6 Mapping Relative Day to Analysis Visit in Neoadjuvant Phase

	Week	
	0	6
Day	1	43
Range	-28-7	8-53

Table 7 PRO Data Collection Schedule in Adjuvant Phase

Treatment	Week						Discontinuation Visit	Follow-up Visit	Survival Follow-up Visit
	0	6	9	18	27	39			
SOC+MK-3475	C1	C3	C4	C7	C10	C14	X	X	X
SOC+PBO	C1	C3	C4	C7	C10	C14	X	X	X
C: Cycle; D: Day Each cycle is 3 weeks for MK-3475.									

The general rule of mapping relative day in the adjuvant phase to analysis visit is provided in [Table 8].

Table 8 Mapping Relative Day to Analysis Visit in Adjuvant Phase

	Week					
	0	6	9	18	27	39
Day	1	43	64	127	190	274
Range	-7-7	8-53	54-95	96-158	159-232	233-316

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.

3.6.3.2 Patient Reported Outcome (PRO) Score Analysis

To assess the treatment effects on the PRO score change from baseline, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [15] 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:

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

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.

3.6.3.3 Summary of Completion and Compliance

Completion and compliance of QLQ-C30, QLQ-STO22 and EQ-5D-5L by visit and treatment will be described based on the PRO FAS population.

Completion Rate is defined as the percentage of subjects who completes at least one score/item over the number of subjects in the PRO FAS population at each time point.

The completion rate is expected to shrink in the later visits due to early discontinuations. Therefore, another measurement, Compliance Rate, defined as the percentage of subjects who completes at least one score/item over the number of eligible subjects who are expected to complete the PRO assessment (not including the subjects missing by design (such as death, discontinuation, translation not available, etc.), will be employed as a supportive measure.

3.6.3.4 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 EFS analysis [Sec. 3.6.1.1] will be used for the PRO analysis.

3.6.4 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

Both efficacy and safety results will be reviewed by an external DMC. The timing of safety reviews will be documented in the DMC Charter based on consultation with external DMC members. The study has 3 planned interim efficacy analyses in addition to the final efficacy analysis. Interim Analysis 1 (IA1) is pathCR rate, EFS and OS analysis. Interim Analysis 2 (IA2) and Interim Analysis 3 (IA3) are EFS/OS analyses. The timing for IA2 and IA3 are driven by EFS event counts as well as a minimum follow-up time since last analysis. Since some of the participants may achieve long-term response, event accumulation may be slow later in the study. Consequently, the IAs are estimated to be ~7 and ~12 months apart respectively as shown in [Table 9]. Note that for IA2, IA3 and FA, if the EFS events or the OS events accrue slower than expected, the analysis may be delayed up to 3 months after the projected timing. Details of the multiplicity strategy for group sequential design are provided in [Sec. 3.8].

Table 9 Summary of Interim and Final Analysis Strategy (Main Study (XP/FP) and/or Main Study (XP/FP) and FLOT Combined)

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	EFS OS pathCR*	Enrollment is complete AND ~ 299 EFS events have occurred in main study (XP/FP)	~ 41 months	• Final pathCR rate analysis • Interim EFS/OS** analyses
IA2	EFS OS	~ 336 EFS events have occurred in main study (XP/FP) AND ~ 7 months after IA1	~ 48 months	• Interim EFS/OS** analyses
IA3	EFS OS	~ 369 EFS events have occurred in main study (XP/FP) AND ~ 12 months after IA2	~ 60 months	• Final EFS analysis/interim OS** analysis
FA	OS	~ 361 OS events have occurred in main study (XP/FP) AND ~ 12 months after IA3	~ 72 months	• Final OS** analysis

EFS=event-free survival; FA=final analysis; HR=hazard ratio; IA=interim analysis; OS=overall survival; pathCR=pathological complete response.

* pathCR analysis will use participants in the Main Study (XP/FP) randomized at least 6 months prior to IA1 data cutoff. It is estimated that ~ 800 participants in the Main Study (XP/FP) will be randomized at least 6 months before the IA1 data cutoff and will be included for pathCR analyses.

** The estimated number of OS events in the Main Study (XP/FP) at IA1, IA2 and IA3 are ~ 256, 296 and 337, respectively.

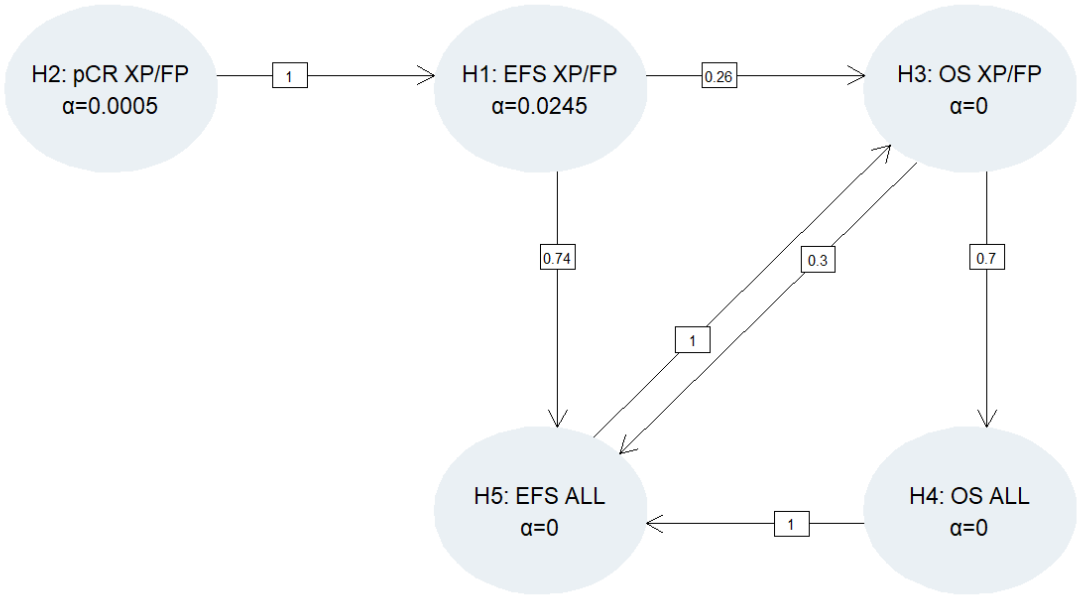
3.8 Multiplicity

The multiplicity strategy specified in this section will be applied to the primary and secondary hypotheses including EFS in the Main Study (XP/FP) (H1), pathCR rate in the Main Study (XP/FP) (H2), OS in the Main Study (XP/FP) (H3), OS in Main Study (XP/FP) and FLOT Cohort combined (H4), and EFS in the Main Study (XP/FP) and FLOT Cohort combined (H5), respectively. The graphical approach of Maurer and Bretz [2] will be taken to strongly control the overall Type I error rate for testing of multiple endpoints at 2.5% 1-sided. See [Figure 1] for the multiplicity strategy diagram of the study. The weights for re-allocation from each hypothesis to the others are shown in the boxes on the lines connecting hypotheses. For EFS and OS hypotheses, the Lan-DeMets O'Brien-Fleming approximation alpha spending function is used to control multiplicity for the interim analyses and final analysis.

The siDMC analyses performed on the FLOT Cohort will assess safety data only, as essentially no efficacy data (eg, EFS/OS) will be available. A Haybittle-Peto type adjustment

with p<0.0001 is applied to preserve the type-1 error level. It essentially does not impact the boundary for future analyses.

Figure 1 Multiplicity Strategy



To account for any multiplicity concerns raised by the DMC review of unplanned efficacy data when prompted by safety concerns, a sensitivity analysis for OS will be pre-specified in the sSAP. This analysis will be performed if requested by the DMC. However, DMC review of OS data beyond the planned efficacy analysis to assess the overall risk:benefit to study participants will not require multiplicity assessment typically associated with a planned efficacy interim analysis because these analyses are not to declare a positive efficacy finding.

3.8.1 Event-free Survival

[Figure 1] shows that the EFS hypothesis in the Main Study (XP/FP) (H1) may be tested at α=0.0245 (initially allocated α) or at α=0.025 (if H2 is rejected). The EFS hypothesis in the Main Study (XP/FP) and FLOT Cohort combined (H5) may be tested for example at α=0.0181 (if H1 is rejected), or at α=0.02 (if both H1 and H3 are rejected) and etc. [Table 10] shows the boundary properties for EFS hypothesis testing in the Main Study (XP/FP) under the initially assigned alpha derived using a Lan-DeMets O'Brien-Fleming spending function. [Table 11] shows the boundary properties for EFS hypothesis testing in the Main Study (XP/FP) and FLOT Cohort combined (H5) under two alphas, alpha passed from the immediately preceding hypothesis if H1 is rejected and alpha under both H1 and H3 are rejected. Cumulative alpha-spending at IAs is based on the Lan-DeMets O'Brien-Fleming spending function evaluated with the spending fraction and the alpha-allocated to the hypothesis. At the final analysis, the cumulative spending will be set to the total alpha allocated to the hypothesis for testing, regardless if the targeted event count is achieved. Bounds will be based on these cumulative spending values plus correlations as determined by actual event counts at interim and final analysis for that hypothesis.

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 IA1 for the Main Study (XP/FP), if 303 EFS events have occurred (ie, more than the expected 299 EFS events), the alpha spending at IA1 will be according to the expected information fraction (299/369=81%) instead of the actual information fraction (303/369 = 82%).

Table 10 Efficacy Boundaries and Properties for Event-free Survival Analyses in Main Study (XP/FP)

Analysis	**Value**	**α = 0.0245**
IA1: 81%* N = 800 Events: 299 Month: 41	Z	2.2429
	p (1-sided)[a]	0.0125
	HR at bound[b]	0.7714
	P(Cross) if HR=1[c]	0.0125
	P(Cross) if HR=0.72[d]	0.7237
IA2: 91%* N: 800 Events: 336 Month: 48	Z	2.1724
	p (1-sided)[a]	0.0149
	HR at bound[b]	0.7888
	P(Cross) if HR=1[c]	0.0184
	P(Cross) if HR=0.72[d]	0.8125
IA3: Final N: 800 Events: 369 Month: 60	Z	2.0844
	p (1-sided)[a]	0.0186
	HR at bound[b]	0.8049
	P(Cross) if HR=1[c]	0.0245
	P(Cross) if HR=0.72[d]	0.8720

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated approximately.
*Percentage of the target number of events at final analysis anticipated at interim analysis.
[a]*p* (1-sided) is the nominal α for testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P(Cross if HR=0.72) is the probability of crossing a bound under the alternative hypothesis.

Table 11 Efficacy Boundaries and Properties for Event-free Survival Analyses in Main Study (XP/FP) and FLOT Cohort Combined

Analysis	**Value**	**α = 0.0181**	**α = 0.02**
IA1: 77%* N = 1000 Events: 352 Month: 41	Z	2.4537	2.4079
	p (1-sided)[a]	0.0071	0.0080
	HR at bound[b]	0.7697	0.7735
	P(Cross) if HR=1[c]	0.0071	0.0080
	P(Cross) if HR=0.73[d]	0.6816	0.6977
IA2: 89%* N: 1000 Events: 407 Month: 48	Z	2.3189	2.2779
	p (1-sided)[a]	0.0102	0.0114
	HR at bound[b]	0.7945	0.7977
	P(Cross) if HR=1[c]	0.0122	0.0137
	P(Cross) if HR=0.73[d]	0.8070	0.8182
IA3: Final N: 1000 Events: 457 Month: 60	Z	2.1933	2.1550
	p (1-sided)[a]	0.0141	0.0156
	HR at bound[b]	0.8144	0.8173
	P(Cross) if HR=1[c]	0.0181	0.0200
	P(Cross) if HR=0.73[d]	0.8840	0.8915

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated approximately.
*Percentage of the target number of events at final analysis anticipated at interim analysis.
[a]*p* (1-sided) is the nominal α for testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P(Cross if HR=0.73) is the probability of crossing a bound under the alternative hypothesis when assuming HR=0.73 in Main Study (XP/FP) and FLOT Cohort combined.

3.8.2 Overall Survival

[Figure 1] shows that the OS hypothesis in the Main Study (XP/FP) (H3) may be tested at α=0.0064 (if H1 is rejected) or at α= 0.0245 (if H1 and H5 are rejected) and etc. The OS hypothesis in the Main Study (XP/FP) and FLOT Cohort combined (H4) may be tested at for example α= 0.0045 (if H1 and H3 are rejected) or at α= 0.0245 (if H1, H3, and H5 are rejected). [Table 12] shows the boundary properties for OS hypothesis testing in the Main Study (XP/FP) with α=0.0064 and α= 0.0245, respectively. [Table 13] shows the boundary properties for OS hypothesis testing in the Main Study (XP/FP) and FLOT Cohort combined (H4) with α=0.0045 and α= 0.0245, respectively. Cumulative alpha-spending at interim analyses is based on the Lan-DeMets O'Brien-Fleming spending function evaluated with the spending fraction and the alpha-allocated to the hypothesis. At the final analysis, the cumulative spending will be set to the total alpha allocated to the hypothesis for testing, regardless of if the targeted event count is achieved. Bounds will be based on these cumulative spending values plus correlations as determined by actual event counts at interim and final analysis for that hypothesis.

The interim OS analysis timing is selected to match the EFS analysis timing. The final OS analysis timing is driven by OS events in the Main Study (XP/FP) and ~12 months after IA3. Similar as EFS analyses, 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.

Table 12 Efficacy Boundaries and Properties for Overall Survival Analyses in Main Study (XP/FP)

Analysis	Value	α = 0.0064	α = 0.0245
IA1: 71%* N = 800 Events: 256 Month: 41	Z	3.0324	2.4275
	p (1-sided)[a]	0.0012	0.0076
	HR at bound[b]	0.6844	0.7382
	P(Cross) if HR=1[c]	0.0012	0.0076
	P(Cross) if HR=0.74[d]	0.2672	0.4923
IA2: 82%* N: 800 Events: 296 Month: 48	Z	2.8418	2.2967
	p (1-sided)[a]	0.0022	0.0108
	HR at bound[b]	0.7184	0.7655
	P(Cross) if HR=1[c]	0.0026	0.0130
	P(Cross) if HR=0.74[d]	0.4126	0.6315
IA3: 93%* N: 800 Events: 337 Month: 60	Z	2.6574	2.1508
	p (1-sided)[a]	0.0039	0.0157
	HR at bound[b]	0.7484	0.7910
	P(Cross) if HR=1[c]	0.0048	0.0199
	P(Cross) if HR=0.74[d]	0.5583	0.7478
Final N: 800 Events: 361 Month: 72	Z	2.5900	2.1034
	p (1-sided)[a]	0.0048	0.0177
	HR at bound[b]	0.7612	0.8013
	P(Cross) if HR=1[c]	0.0064	0.0245
	P(Cross) if HR=0.74[d]	0.6310	0.7990

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated approximately.
*Percentage of the target number of events at final analysis anticipated at interim analysis.
[a]*p* (1-sided) is the nominal α for testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P(Cross if HR=0.74) is the probability of crossing a bound under the alternative hypothesis.

Table 13 Efficacy Boundaries and Properties for Overall Survival Analyses in Main Study (XP/FP) and FLOT Cohort Combined

Analysis	Value	α = 0.0045	α = 0.0245
IA1: 69%* N = 1000 Events: 306 Month: 41	Z	3.2266	2.4690
	p (1-sided)[a]	0.0006	0.0068
	HR at bound[b]	0.6913	0.7538
	P(Cross) if HR=1[c]	0.0006	0.0068
	P(Cross) if HR=0.75[d]	0.2334	0.5102
IA2: 80%* N: 1000 Events: 355 Month: 48	Z	3.0106	2.3278
	p (1-sided)[a]	0.0013	0.0100
	HR at bound[b]	0.7262	0.7808
	P(Cross) if HR=1[c]	0.0015	0.0119
	P(Cross) if HR=0.75[d]	0.3840	0.6555
IA3: 93%* N: 1000 Events: 412 Month: 60	Z	2.7780	2.1486
	p (1-sided)[a]	0.0027	0.0158
	HR at bound[b]	0.7605	0.8091
	P(Cross) if HR=1[c]	0.0032	0.0197
	P(Cross) if HR=0.75[d]	0.5607	0.7872
Final N: 1000 Events: 443 Month: 72	Z	2.7026	2.0998
	p (1-sided)[a]	0.0034	0.0179
	HR at bound[b]	0.7734	0.8190
	P(Cross) if HR=1[c]	0.0045	0.0245
	P(Cross) if HR=0.75[d]	0.6410	0.8370

Abbreviations: HR = hazard ratio; IA = interim analysis.

The number of events and timings are estimated approximately.

*Percentage of the target number of events at final analysis anticipated at interim analysis.

[a]*p* (1-sided) is the nominal α for testing.

[b]HR at bound is the approximate HR required to reach an efficacy bound.

[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.

[d]P(Cross if HR=0.75) is the probability of crossing a bound under the alternative hypothesis when assuming HR=0.75 in Main Study (XP/FP) and FLOT Cohort combined.

3.8.3 Pathological Complete Response

The study will test pathCR in the Main Study (XP/FP) only once at the IA1, at a one-sided alpha level of 0.0005, based on ~800 participants in the Main Study (XP/FP) randomized at least 6 months prior to the IA1 data cutoff.

3.9 Sample Size and Power Calculations

The main study (XP/FP) will randomize approximately 800 participants in a 1:1 ratio between the 2 groups while the FLOT Cohort will randomize approximate 200 participants in a 1:1 ratio as well.

Main Study (XP/FP)

PathCR rate:

The pathCR rate analysis will be performed at IA1. Participants randomized in the Main Study (XP/FP) at least 6 months before the IA1 data cutoff will be included in pathCR rate analysis.

It is estimated that ~800 participants in the Main Study will be included.

A sample size of ~800 gives ~ 99.9% power to detect a true pathCR rate difference of 15 percentage points (an improvement from 5% to 20%) at alpha = 0.05% (one-sided) significance level. The difference in the pathCR rates between treatment and control groups required to demonstrate superiority in pathCR rate of the treatment group is ~ 8%.

EFS:

The final analysis of EFS will be performed when there are ~ 369 EFS events in the Main Study (XP/FP). With 369 events, the study has ~ 87% power for EFS to detect a hazard ratio (experimental group versus control group) of 0.72 at initially assigned 2.45% (one-sided) significance level.

The sample size (number of participants) calculations are based on the following assumptions: 1) EFS follows a Poisson mixture model with ~ 40% participants achieving excellent long-term results (ie, cure rate of 40%) and a median EFS of 30 months in the control group (See Protocol Appendix 7 – Technical Note for the Statistical Model for more details); 2) true hazard ratio for experimental group versus control group of 0.72; 3) enrollment duration of ~ 35 months with the actual monthly enrollment incorporated up to September 2020; 4) a yearly drop-out rate of 5% for both experimental group and control group.

With the Poisson mixture model and HR=0.72, it is estimated that the 3-year EFS rate is 47% for the control group and 58% for the experimental group.

OS:

If EFS in the Main Study (XP/FP) (H1) is rejected, α=0.0064 can be passed to OS in the Main Study (XP/FP) (H3). If both H1 and EFS in XP/FP and FLOT combined (H5) are rejected, α=0.0245 can be passed to H3. The final analysis of OS will be performed when there are ~ 361 OS events in the Main Study (XP/FP). With 361 events, the study has ~ 63% or ~80% power for OS to detect a hazard ratio (experimental group versus control group) of 0.74 at 0.64% or 2.45% (one-sided) significance level.

The sample size (number of participants) calculations are based on the following assumptions: 1) OS follows a Poisson mixture model with ~ 40% participants achieving excellent long-term results (ie, cure rate of 40%) and a median OS of 42 months (See Protocol Appendix 7 – Technical Note for the Statistical Model for more details); 2) true hazard ratio for experimental group versus control group is 0.74; 3) enrollment duration of ~ 35 months with the actual monthly enrollment incorporated up to September 2020; and 4) a yearly drop-out rate of 5% for both experimental group and control group.

With the Poisson mixture model and HR=0.74, it is estimated that the 5-year OS rate is 45% for the control group and 56% for the experimental group.

Main Study (XP/FP) and FLOT Cohort Combined

EFS:

If H1 is rejected, α=0.0181 can be passed to H5. Approximately 457 EFS events in the Main Study (XP/FP) and FLOT Cohort combined are expected at the time of the final analysis of EFS (IA3 of the study). With 457 EFS events, the study has ~ 88% power for EFS in the Main Study (XP/FP) and FLOT Cohort combined to detect a hazard ratio (experimental group versus control group) of 0.73 at 1.81% (one-sided) significance level.

The sample size (number of participants) calculations are based on the same assumptions as above for the Main Study (XP/FP), while for the FLOT Cohort, the following assumptions are assumed: 1) EFS follows a Poisson mixture model with ~ 40% participants achieving excellent long-term results (ie, cure rate of 40%) and a median EFS of 30 months in the control group (See Protocol Appendix 7 – Technical Note for the Statistical Model for more details); 2) true hazard ratio for experimental group versus control group of 0.78; 3) actual enrollment for the first 143 FLOT participants and then 10 participants/month; 4) a yearly drop-out rate of 5% for both experimental group and control group. An overall HR=0.73 was calculated to represent the constant hazard ratio for the sample size/power calculation of the Main Study (XP/FP) and FLOT Cohort combined (See Protocol Appendix 7 – Technical Note for the Statistical Model for more details).

With the Poisson mixture model and HR=0.73, it is estimated that the 3-year EFS rate in the Main Study (XP/FP) and FLOT Cohort combined is 47% for the control group and 58% for the experimental group.

OS:

If H1 and H3 are rejected, α= 0.0045 can be passed to OS in the Main Study (XP/FP) and FLOT Cohort combined (H4). If H1, H3 and H5 are rejected, α=0.0245 can be passed to H4. Approximately 443 OS events in the Main Study (XP/FP) and FLOT Cohort combined are expected at the time of the final analysis of OS. With 443 events, the study has ~ 64% or 84% power for OS in the Main Study (XP/FP) and FLOT Cohort combined to detect a hazard ratio (experimental group versus control group) of 0.75 at 0.45% or 2.45% (one-sided) significance level.

The sample size (number of participants) calculations are based on the same assumptions as above for the Main Study (XP/FP), while for the FLOT Cohort, the following assumptions are assumed: 1) OS follows a Poisson mixture model with ~ 42.7% participants achieving excellent long-term results (ie, cure rate of 42.7%) and a median OS of 50 months in the control group (See Protocol Appendix 7 – Technical Note for the Statistical Model for more details); 2) true hazard ratio for experimental group versus control group of 0.8; 3) actual enrollment for the first 143 FLOT participants and then 10 participants/month; 4) a yearly drop-out rate of 5% for both experimental group and control group. An overall HR=0.75 was calculated to represent the constant hazard ratio for the sample size/power calculation of the main study (XP/FP) and FLOT Cohort combined (See Protocol Appendix 7 – Technical Note for the Statistical Model for more details).

With the Poisson mixture model and HR=0.75, it is estimated that the 5-year OS rate is 46% for the control group and 56% for the experimental group.

All sample size and power calculations were performed using the R software package gsDesign.

3.10 Subgroup Analyses and Effect of Baseline Factors

To determine whether the treatment effect is consistent across various subgroups, the between-group treatment effect for OS and EFS (with a nominal 95% CI) will be estimated and plotted within each category of the following classification variables:

- Geographic region (Asia versus Non-Asia)
- Tumor staging (II versus III versus IVa)
- Tumor location (Stomach versus GEJ)
- PD-L1 (+ versus -)
- Age category (<65 versus ≥65 years)
- Gender (Male versus Female)
- Race (Asian versus Non-Asian)
- Backbone therapy (FP versus XP [versus FLOT*])
- ECOG status (0 versus 1)

- MSI status: (MSI-H versus non MSI-H)
- Histology subtype (Diffuse versus Intestinal)

Subgroup analyses will be performed on the Main Study (XP/FP) as well as on the Main Study (XP/FP) and FLOT Cohort combined. The consistency of the treatment effect will be assessed descriptively via summary statistics by category for the classification variables listed above. If the number of participants in a category of a subgroup variable is less than 10% of the ITT population, the subgroup analysis will not be performed for this category of the subgroup variable, and this subgroup variable will not be displayed in the forest plot.

*Category of FLOT for the subgroup variable of backbone therapy is only applicable to the subgroup analyses for the Main Study (XP/FP) and FLOT Cohort combined.

Subgroup analyses for efficacy endpoints will be conducted using unstratified methods.

Country and/or region-specific subgroup (eg, China, Japan, EU etc.) may also be analyzed per local registration needs.

3.11 Compliance (Medication Adherence)

Drug accountability data for study treatment will be collected during the study. Any deviation from protocol-directed administration will be reported.

3.12 Extent of Exposure

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

4 STATISTICAL ANALYSIS PLAN FOR CHINA EXTENSION

NOTE: Based on IA3, the study did not meet statistical significance for EFS in the global population; thus, OS will not be formally tested per the multiplicity strategy. KEYNOTE-585 data will not be used to support regulatory registration in China for the disease indication of this study. Original prespecified analyses (see [Sec. 4.7]) for China subpopulation (including the China participants enrolled in the global portion and the extension portion) are for reference only. Primary efficacy analyses for China subpopulation are revised accordingly that, descriptive analyses will be conducted based on the same data cutoff date (16FEB2024) as the final analysis for the global population. As all China participants were enrolled for the Main Study (XP/FP), only the endpoints relevant to the Main Study (XP/FP) are used for the China subpopulation analyses.

This section outlines the statistical analysis strategy and procedures for China specific subgroup analysis which is required by local regulatory. China refers to China mainland in this section.

4.1 Introduction

After the global portion enrollment is closed, participants from China will continue to be enrolled in an extension portion designed to meet China local registration needs. The extension portion will be identical to the global portion of the Main Study (XP/FP) (e.g., inclusion and exclusion criteria, primary and secondary endpoints, study procedures) in general, with the additional statistical analysis plan for the China subpopulation. The purpose of this extension portion is to evaluate the consistency of efficacy and safety in the China subpopulation to the global population in the Main Study (XP/FP). Country-specific analysis may also be conducted per local regulatory requirement.

After the enrollment for the global portion is completed, participants in China will continue to be enrolled in a 1:1 ratio to receive pembrolizumab (MK 3475) plus chemotherapy (XP/FP) or placebo plus chemotherapy (XP/FP) until the sample size for China subpopulation reaches approximately 120 in total of the global and extension portion combined.

After the cut-off date for the primary analyses of the global portion (for the interim and/or final analysis), all China participants, including the China participants enrolled in the global portion and in the extension portion, will continue their randomized treatment and continue to be followed up for China registration purpose. The extension portion will be completed after the primary analysis for OS in the China subpopulation (including China participants enrolled in the global portion and the extension portion) has been conducted. The timing of the analyses for China subpopulation is provided in [Sec. 4.7]

4.2 Responsibility for Analyses/In-House Blinding

For all China participants, including participants randomized in the global portion and the extension portion, patient level treatment randomization information will be blinded until the extension portion database lock is achieved. The extent to which individuals are unblinded to the results will be limited. Blinded and unblinded members will be clearly documented with blinding status along with time information.

4.3 Hypotheses/Estimation

No hypothesis testing is planned for the China extension portion.

4.4 Analysis Endpoints

4.4.1 Efficacy Endpoints

Efficacy endpoints are the same as described in [Sec. 3.4.1].

4.4.2 Safety Endpoints

Safety endpoints are the same as described in [Sec. 3.4.2].

4.5 Analysis Population

4.5.1 Efficacy Analysis Populations

The Intention-to-Treat (ITT) China subpopulation will serve as the population for OS, EFS and pathCR. This population will include all China participants who are randomized in the global portion and all participants who are randomized in the extension portion. For the pathCR rate, the ITT China subpopulation will only include China participants randomized at least 6 months prior to the data cutoff as the primary analysis.

DFS analysis population will be the China participants who are disease free at the post-surgery baseline scan.

Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary efficacy analysis population for the global portion.

4.5.2 Safety Analysis Populations

Safety analysis will be carried out in the All Participants as Treated (APaT) China subpopulation, i.e., all randomized China participants (in the global portion and extension portion) who received at least one dose of study treatment.

Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary safety analysis population for the global portion.

4.6 Statistical Methods

No formal hypothesis testing is planned, and no multiplicity adjustment will be applied to the analysis for China subpopulation. The analysis methods for China subpopulation are the same as described in [Sec. 3.6] for the global portion if applicable.

4.6.1 Statistical Methods for Efficacy Analyses

Degree of consistency of efficacy will be evaluated using the percentage of treatment effect (e.g., risk reduction for EFS and OS) preserved in the China subpopulation from the empirical treatment effect from the global primary efficacy analyses (based on point estimate).

4.6.1.1 Event-free Survival (EFS)

The non-parametric Kaplan-Meier method will be used to estimate the EFS curve in each treatment group. The treatment difference in EFS will be assessed by the unstratified log-rank test. An unstratified 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 groups. The HR and its 95% CI from the unstratified Cox model for the treatment covariate will be reported.

4.6.1.2 Pathological Complete Response (pathCR) Rate

The unstratified Miettinen and Nurminen's method will be used for comparison of the difference in pathCR rates between pembrolizumab with chemotherapy (XP/FP) and chemotherapy alone.

4.6.1.3 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 unstratified log-rank test. An unstratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference measured by the hazard ratio (HR). The HR and its 95% CI from the unstratified Cox model for the treatment covariate will be reported. Participants without documented death at the time of analysis will be censored at the date the participant was last known to be alive.

4.6.1.4 Disease-free Survival (DFS)

The non-parametric Kaplan-Meier method will be used to estimate the DFS curve for each treatment group.

4.6.2 Statistical Methods for Safety Analyses

Safety analyses for China subpopulation are the same to those for the global portion as described in [Sec. 3.6.2] if applicable.

4.6.3 Summaries of Demographic and Baseline Characteristics

They are the same for China subpopulation to those of the global portion as described in [Sec. 3.6.4]

4.7 Interim Analyses and Final analysis

The primary analysis for pathCR will be conducted in the China subpopulation when there is ~6 months follow-up after the last China participant randomized in the extension portion. EFS and OS will also be analyzed at this same time.

The primary analysis for EFS will be conducted in the China subpopulation when approximately a total of 55 EFS events have been observed in the China subpopulation (including China participants enrolled in the global portion and the extension portion). OS will also be analyzed at this same time.

The primary analysis for OS will be conducted in the China subpopulation when approximately a total of 53 OS events have been observed in the China subpopulation (including China participants enrolled in the global portion and the extension portion).

However, if the criteria for conduct of the analyses in China subpopulation are met before an IA or final analysis for the global portion, the corresponding analysis for China

subpopulation may occur at the same time as the IA or the final analysis for the global portion. If the statistical significance for the global portion has been demonstrated at an IA or final analysis and it is projected that the criteria for conduct of the analysis in China subpopulation including the extension portion will be met within ~3 months after the IA or final analysis for the global portion, then the analysis for China subpopulation including the extension portion may be based on the same database lock as the IA or the final analysis for the global portion.

4.8 Multiplicity

No multiplicity adjustment will be applied to the analysis of China subpopulation.

4.9 Sample Size and Power Calculations

After the completion of global portion enrollment, the extension portion will continue to enroll participants and randomize eligible participants until the sample size for the overall randomized China subpopulation reaches approximately 120. Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary analysis population for the global portion.

There will be approximately a total of 55 EFS events in the China subpopulation at the time of the primary analysis for EFS. With 55 EFS events, the China subpopulation has ~74% chance to observe a point estimate of hazard ratio for EFS that preserves approximately at least 50% of the empirical risk reduction from the analysis for the global portion of the Main Study (XP/FP), assuming the underlying hazard ratio is 0.72.

There will be approximately a total of 53 OS events in the China subpopulation at the time of the primary analysis for OS. With 53 OS events, the China subpopulation has ~71% chance to observe a point estimate of hazard ratio for OS that preserves approximately at least 50% of the empirical risk reduction from the analysis for the global portion of the Main Study (XP/FP), assuming the underlying hazard ratio is 0.74.

The above calculations in EFS and OS are based on the same assumptions in the global portion of Main Study (XP/FP) for sample size and power evaluation as specified in [Sec. 3.9].

5 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. 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.

4. 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.

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

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. 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.

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

9. Lin, RS, Lin, J, Roychoudhury, S, et al. Alternative analysis methods for time to event endpoints under non-proportional hazards: a comparative analysis. Statistics in Biopharmaceutical Research 2020; pp.1–19.

10. 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.

11. 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.

12. 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.

13. P. Fayers, N. Aaronson, K. Bjordal, M. Groenvold, D. Curran and A. Bottomley, EORTC QLQ-C30 Scoring Manual (3rd edition), Brussels: EORTC, 2001

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

15. 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.