Official Protocol Title:	A Phase III, Randomized, Open-label Clinical Trial of Pembrolizumab (MK-3475) versus Paclitaxel in Asian Subjects with Advanced Gastric or Gastroesophageal Junction Adenocarcinoma who Progressed after First-Line Therapy with Platinum and Fluoropyrimidine
NCT number:	NCT03019588
Document Date:	1 NOV 2019

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

LIST OF TABLES

Confidential

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 1) provides additional statistical analysis details/data derivations and 2) 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.

2 SUMMARY OF CHANGES

Section Number (s)	Section Title(s)	Description of Change(s)	Rationale
3.10	Subgroup Analyses and Effect of Baseline Factors	Added a paragraph to specify the unstratified Cox model and the unstratified log-rank test will be used for subgroup analyses.	To ensure the subgroup (s) with small group size could be analyzed properly.

3 ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

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

Table 1 Statistical Analysis Plan

Study Design Overview	A Phase III, Randomized, Open-label Clinical Trial of Pembrolizumab (MK-3475) versus Paclitaxel in Asian Subjects with Advanced Gastric or Gastroesophageal Junction Adenocarcinoma who Progressed after First-Line Therapy with Platinum and Fluoropyrimidine
Treatment Assignment	Approximately 360 subjects will be randomized in a 1:1 ratio to receive pembrolizumab or paclitaxel. Stratification factors are time to progression (TTP) on first-line therapy (< 6 months vs. ≥ 6 months) and ECOG PS (0 vs. 1). This is an open-label study.
Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoints	1. Overall Survival (OS) 2. Progression-free Survival (PFS) per RECIST 1.1 by blinded central radiologists' review
Statistical Methods for Key Efficacy Analyses	The primary hypotheses will be evaluated by comparing pembrolizumab to paclitaxel on PFS per RECIST 1.1 by blinded central radiologist review and OS 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.

Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. The tiers differ with respect to the analyses that will be performed. There are no Tier 1 events in this trial. Tier 2 parameters will be assessed via point estimates with 95% confidence intervals provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters. The between-treatment difference will be analyzed using the Miettinen and Nurminen method [1].
Interim Analysis	There is no interim analysis for PFS. One interim efficacy analysis is planned for OS. Results will be reviewed by an external data monitoring committee. The interim analysis is summarized below. Details are provided in Section 3.7. The interim analysis of OS will be performed at the time of final (only) PFS analysis o Timing: To be performed after: (1) enrollment is completed (2) approximately both 235 PFS events and 190 OS events have been observed. o Purpose: final PFS analysis and interim analysis of OS Final analysis (event driven trial) o Timing: at least 290 OS events have been observed, estimated to be 36 months after study start o Purpose: Final analysis of OS
Multiplicity	The overall type I error for the multiple endpoints will be strongly controlled by the Bonferroni procedure at 2.5% (one-sided) with initially 0.35% allocated to PFS and 2.15% allocated to OS hypotheses, and 0% to the ORR hypothesis. If the PFS null hypothesis is rejected, the corresponding alpha level can be shifted to the hypotheses for the OS endpoint using the graphical approach of Maurer and Bretz [2]. The secondary hypothesis of ORR will be tested only if pembrolizumab arm is superior to the control for OS.
Sample Size and Power	The planned sample size is approximately 360 subjects. The final analysis of this study is event driven (i.e., follow-up time is subject to change but the approximate number of events is not) and will complete after at least 290 OS events have been observed. For the PFS primary endpoint , the trial has >99% (>90%) power to demonstrate that pembrolizumab is superior to paclitaxel at a one-sided 0.35% alpha-level, if the underlying hazard ratio of PFS is 0.5 (0.6). For the OS primary endpoint, the trial has 91% power to demonstrate that pembrolizumab is superior to paclitaxel at a one-sided 2.15% alpha-level, if the underlying hazard ratio of OS is 0.67.

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.

The SPONSOR will generate the randomized allocation schedule(s) for study treatment assignment. The algorithm for the randomized allocation of subjects will be implemented in an interactive voice response system (IVRS).

Although the trial is open label, analyses or summaries generated by randomized treatment assignment or actual treatment received will be limited and documented prior to the final unblinding for reporting purposes. In addition, the independent radiologist(s) will perform the blinded central radiologist review without knowledge of treatment group assignment.

Planned interim analyses are described in Section 3.7. Interim Analysis will be performed when enrollment is completed. Access to the allocation schedule for this study will be restricted to an external unblinded statistician and, as needed, a scientific programmer performing the analysis, who will have no other responsibilities associated with the study.

Treatment-level results of the interim analyses will be provided by the external unblinded statistician to the external Data Monitoring Committee (eDMC). Limited additional Sponsor personnel may be unblinded to the treatment level results of the interim analyses, if required, in order to act on the recommendations of the eDMC or facilitate regulatory filing after an interim analysis. The extent to which individuals are unblinded with respect to results of interim analyses will be documented.

The eDMC will serve as the primary reviewer of the unblinded results of the interim analysis and will make recommendations for discontinuation of the study or protocol modifications to the Executive Oversight Committee (EOC) of this SPONSOR. Depending on the recommendation of the eDMC, the Sponsor may prepare a regulatory submission. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, the EOC may be unblinded to results at the treatment level in order to act on these recommendations. Additional logistical details will be provided in the DMC Charter. In addition, there will be an unblinding plan to maintain information about unblinding of Sponsor Personnel prior to full unblinding at Sponsor. Key aspects of the interim analyses are described in Section 3.7.

The personnel who have access to allocation schedule at what time will be documented in Appendix 4 Interim Analysis Data Sources Memo in DMC charter.

3.3 Hypotheses/Estimation

In PD-L1 positive subjects with advanced gastric or GEJ adenocarcinoma who have progressed on one previous line of therapy.

3.3.1 Primary Objective(s) & Hypothesis(es)

(1) **Objective**: To compare OS.

Hypothesis: Pembrolizumab prolongs OS compared to paclitaxel.

(2) **Objective**: To compare PFS per RECIST 1.1 by blinded central radiologists' review.

Hypotheses: Pembrolizumab prolongs PFS per RECIST 1.1 by blinded central radiologists' review compared to paclitaxel.

The study is considered to have met its primary objective if pembrolizumab is superior to paclitaxel either in OS (interim or final analysis) or in the final PFS analysis.

3.3.2 Secondary Objective(s) & Hypothesis(es)

(1) To evaluate the Objective Response Rate (ORR) per RECIST 1.1 assessed by blinded central radiologists' review.

Hypotheses: Pembrolizumab improves ORR per RECIST 1.1 assessed by blinded central radiologists' review compared to paclitaxel.

(2) **Objective**: Evaluate the safety and tolerability profile of pembrolizumab compared to paclitaxel.

3.3.3 Exploratory Objectives

(1) **Objective:** To evaluate PFS per irRECIST by blinded central radiologists' review among subjects when treated with pembrolizumab compared to paclitaxel.

(2) **Objective:** To evaluate the Time to Progression (TTP) and Duration of Response (DOR) per RECIST 1.1 by blinded central radiologists' review among subjects when treated with pembrolizumab compared to paclitaxel.

(3) **Objective**: To evaluate score change of health related quality of Life using the EORTC QLQ-C30 and the EORTC QLQ-STO22 from baseline among subjects when treated with pembrolizumab compared to paclitaxel.

(4) **Objective**: To characterize utilities using EuroQol EQ-5D among subjects when treated with pembrolizumab compared to paclitaxel.

(5) **Objective**: To explore the relationship between genetic variation and response to the treatment(s) administered. Genomic variability will be analyzed for association with clinical data collected in this study.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Primary

Overall Survival

Overall Survival (OS) is defined as the time from randomization to death due to any cause. Subjects without documented death at the time of the final analysis will be censored at the date of the last follow-up.

C Confidential

Progression-free survival (PFS) – RECIST 1.1 by blinded central radiologists' review

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression or death due to any cause, whichever occurs first. See Section 3.6.1 for the definition of censoring.

Secondary

Objective Response Rate (ORR) – RECIST 1.1 by blinded central radiologists' review

Objective response rate is defined as the proportion of the subjects in the analysis population who have a complete response (CR) or partial response (PR).

Exploratory

Progression-free survival (PFS) – irRECIST by blinded central radiologists' review

Progression-free-survival (PFS) is defined as the time from randomization to the first confirmed disease progression or death due to any cause, whichever occurs first.

Time to Progression (TTP) – RECIST 1.1 by blinded central radiologists' review

Time to Progression (TTP) is defined as the time from randomization to the first documented disease progression. If there is no documented disease progression, TTP is censored at last tumor assessment date.

Duration of Overall Response (DOR) – RECIST 1.1 by blinded central radiologists' review

For subjects who demonstrated CR or PR, response duration is defined as the time from first documented evidence of CR or PR until disease progression or death due to any cause, whichever occurs first. Censoring rules are provided in Section 3.6.3.2.

3.4.2 Safety Endpoints

Safety measurements are described in Protocol Section 7.

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. All randomized subjects will be included in this population. Subjects will be included in the treatment group to which they are randomized.

Details on the approach to handling missing data are provided in Section 3.6 Statistical Methods.

C Confidential

3.5.2 Safety Analysis Populations

The All Subjects as Treated (ASaT) population will be used for the analysis of safety data in this study. The ASaT population consists of all randomized subjects who received at least one dose of study treatment. Subjects will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the ASaT population. For most subjects this will be the treatment group to which they are randomized. Subjects 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 subject who receives the incorrect study medication for one cycle but receives the correct treatment for all other cycles will be analyzed according to the correct (randomized) treatment group and a narrative will be provided for any events that occur during the cycle for which the subject 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 Section 3.6 Statistical Methods.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

Efficacy results that will be deemed to be statistically significant after consideration of the Type I error control strategy are described in Section 3.8, Multiplicity. Nominal p-values will be computed for other efficacy analyses, but should be interpreted with caution due to potential issues of multiplicity.

3.6.1.1 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curve**s**. 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 (i.e., the hazard ratio). The hazard ratio and its 95% confidence interval from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization will be applied to both the stratified log-rank test and the stratified Cox model.

Subjects in the paclitaxel arm are expected to discontinue treatment earlier compared to subjects in the pembrolizumab arm, and may switch to another anti PD-1 treatment following the verification of progressive disease by blinded central radiologists' vendor. Exploratory analyses to adjust for the effect of crossover [to other PD-1 therapies] on OS may be performed based on recognized methods, e.g. the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis (1989) [3], two stage model [4], etc., based

on an examination of the appropriateness of the data to the assumptions required by the methods.

The RPSFT model provides a randomization-based estimate of the treatment effect corrected for bias introduced by crossover from the control arm to active treatment. The model is refereed as rank preserving as it is assumed that given two subjects i and j, if i failed before j when both were on one treatment, then i would also fail before j if both subjects took the same alternative treatment. This assumption may not be plausible with certain subjects likely to see more or less benefit than others on different types of treatments due to biological factors. However testing for any violations of this assumption in real data may not be possible. The method also assumes an equal treatment effect for subjects switching to a treatment as for those initially allocated to receive it. For the RPSFT method, time post crossover is adjusted using an accelerated failure time model, and then the resulting adjusted times to events are analyzed using the same methods as the primary analyses. The 95% confidence intervals of the hazard ratio for OS after adjustment of the cross-over effect will be provided at the final analysis. More detailed steps to implement the RPSFT method will be provided in the Programming Requirement Specification (PRS) for the macro implementing the RPSFT method.

If there is no unmeasured confounder at the secondary baseline time-point (disease progression), treatment switching only happens after progression, and happens soon after progression, the “two-stage” approach may be appropriate. At Stage 1, the date of disease progression is used as a secondary baseline for subjects who have a documented progression in the standard treatment arm and data from these subjects beyond this time-point are considered as an observational dataset. An accelerated failure time (AFT) model including covariates for crossover and other prognostic covariates measured at the secondary baseline will be applied to this observational dataset to estimate an acceleration factor. At Stage 2, a counterfactual survival dataset will be constructed such that survival time of subjects with treatment switching will be shrunk by the inverse of the acceleration factor, while no shrinkage is performed for the survival time of subjects in the control group without treatment switching or subjects in the experimental arm. A Cox model will then be applied to the counterfactual survival dataset to estimate the HR from this two-stage method. More detailed steps to implement the two-stage method will be provided in the Programming Requirement Specification (PRS) for the macro of two-stage method.

It is very important to assess trial data, crossover mechanism, and treatment effect to determine which method is likely to be most appropriate to evaluate the cross-over effect as well as to evaluate the consistency of alternate approaches.

3.6.1.2 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 (i.e., hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be

reported. The stratification factors used for randomization will be applied to both the stratified log-rank test and the stratified Cox model.

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. For the primary analysis, the true date of disease progression will be approximated by the date of the first assessment at which PD is objectively documented per RECIST 1.1 by central imaging vendor, regardless of discontinuation of study drug or missed study visits. Death is always considered as a confirmed PD event. Subjects who do not experience a PFS event will be censored at the last disease assessment. 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 central imaging vendor, two sensitivity analyses with different sets of censoring rules will be performed. The first sensitivity analysis is the same as the primary analysis except that the data for any subject who misses more than one consecutive disease assessment (with or without a subsequent death or progression) are censored at the last disease assessment prior to missing visits. The second sensitivity analysis is the same as the primary analysis except that it considers discontinuation of treatment or initiation of an anticancer treatment subsequent to discontinuation of study-specified treatments, whichever occurs later, to be a PD event for subjects without documented PD or death. If a subject 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 2.

Additional PFS sensitivity analyses may be performed, including a PFS analysis using time to scheduled tumor assessment visit from randomization as opposed to actual tumor assessment time.

Table 2 Censoring rules for Primary and Sensitivity Analyses of PFS

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
No PD and no death; new anticancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment	Censored at last disease assessment if still on study therapy; progressed at treatment discontinuation otherwise
No PD and no death; new anticancer treatment is initiated	Censored at last disease assessment before new anticancer treatment	Censored at last disease assessment before new anticancer treatment	Progressed at date of new anticancer treatment
No PD and no death; ≥ 2 consecutive missed disease assessments	Censored at last disease assessment	Censored at last disease assessment prior to ≥2 consecutive missed visits	Censored at last disease assessment
PD or death documented after ≤ 1	Progressed at date of documented PD or death	Progressed at date of documented PD or	Progressed at date of documented PD or death

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
missed disease assessment		death	
PD or death documented at any time after ≥ 2 consecutive missed disease assessments	Progressed at date of documented PD or death	Censored at last disease assessment prior to the ≥ 2 consecutive missed disease assessment	Progressed at date of documented PD or death

The proportional hazards assumption on PFS will be examined using both graphical and analytical methods if warranted. The log[-log] of the survival function vs. time for PFS will be plotted for the comparison between pembrolizumab and the paclitaxel 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 [5], parametric method [6], 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 will be pre-specified prior to unblinding of the study by the Sponsor study team that is blinded to treatment results. The difference of two RMSTs for two treatment groups will be estimated and 95% confidence interval will be provided.

One assumption for stratified Cox proportional hazard model is that, the treatment hazard ratio (HR) is constant across the strata. If strong departures from the assumption of the HR being the same for all the strata observed (which can result in a notably biased and/or less powerful analysis), a sensitivity analysis may be performed based on a two-step weighted Cox model approach by Mehrotra 2012 [7], in which the treatment effect is first estimated for each stratum and then the stratum specific estimates are combined for overall inference using sample size weights.

3.6.1.3 Objective Response Rate (ORR)

The stratified Miettinen and Nurminen method [1] with strata weighting by sample size will be used for comparison of the objective response rates between the treatment arms. A 95% confidence interval for the difference in response rates between the pembrolizumab arm and paclitaxel arm will be provided. The stratification factors used for randomization will be applied to the analysis.

Table 3 summarizes the primary analysis approaches for primary and secondary efficacy endpoints. Sensitivity analysis methods are described above for each endpoint.

The strategy to address multiplicity issues with regard to multiple efficacy endpoints and interim analysis is described in Section 3.7 Interim Analyses and in Section 3.8 Multiplicity.

Table 3 Analysis Strategy for Key Efficacy Endpoints

Endpoint/Variable (Description, Time Point)	Statistical Method†	Analysis Population	Missing Data Approach
Primary Hypothesis #1			
OS	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at last known alive date
Primary Hypothesis #2			
PFS per RECIST 1.1 by blinded central radiologists' review	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 in Table 2)
Secondary Hypothesis			
ORR per RECIST 1.1 by blinded central radiologists' review	Stratified M & N method‡	ITT	Subjects with missing data are considered non-responders

† Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization (time to progression on first-line therapy (< 6 months vs. ≥ 6 months) and ECOG PS (0 vs. 1)), will be applied to the analysis.

‡ Miettinen and Nurminen method.

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, vital signs, etc.

Tiered Approach

The analysis of safety results will follow a tiered approach (Table 4). The tiers differ with respect to the analyses that will be performed. "Tier 1" safety endpoints that will be subject to inferential testing for statistical significance with p-values and 95% confidence intervals provided for between-group comparisons. Other safety parameters will be considered Tier 2 or Tier 3. Tier 2 parameters will be assessed via point estimates with 95% confidence

intervals provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters.

Adverse experiences (specific terms as well as system organ class terms) and predefined limits of change in laboratory, vital signs, that are not pre-specified as Tier-1 endpoints will be classified as belonging to "Tier 2" or "Tier 3", based on the number of events observed. Membership in Tier 2 requires that at least 4 subjects 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 4 events was chosen because the 95% confidence interval for the between-group difference in percent incidence will always include zero when treatment groups of equal size each have less than 4 events and thus would add little to the interpretation of potentially meaningful differences. 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 in adverse experiences and predefined limits of change.

Continuous measures such as changes from baseline in laboratory, vital signs, that are not pre-specified as Tier-1 endpoints will be considered Tier 3 safety parameters. Summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group in table format.

There are no events of interest that warrant elevation to Tier 1 in this study. The broad clinical and laboratory AE categories consisting of the percentage of subjects with any AE, any drug related AE, any Grade 3-5 AE, any serious AE, any AE which is both drug-related and Grade 3-5, any AE which is both serious and drug-related, dose modification due to AE, and who discontinued due to an AE, and death will be considered Tier 2 endpoints. 95% confidence intervals (Tier 2) will be provided for between-treatment differences in the percentage of subjects with events; these analyses will be performed using the Miettinen and Nurminen method (1985), an unconditional, asymptotic method.

To properly account for the potential difference in follow-up time between the study arms, which is expected to be longer in the pembrolizumab arm, AE incidence density adjusted for treatment exposure analyses may be performed as appropriate. Based on emerging external data, the supportive analysis strategy for safety parameters may be modified to improve the integrity and efficiency of the design. Should this happen, the change will be documented in supplemental SAP, if not in a protocol amendment, at the earliest time before any unblinding of the data.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint†	p-Value	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Any AE		X	X
	Any Serious AE		X	X
	Any Grade 3-5 AE		X	X
	Any Drug-Related AE		X	X
	Any Serious and Drug-Related AE		X	X
	Any Grade 3-5 and Drug-Related AE		X	X
	Dose Modification due to AE		X	X
	Discontinuation due to AE		X	X
	Death			
	Specific AEs, SOCs, or PDLCs(incidence ≥4 of subjects in one of the treatment groups)		X	X
Tier 3	Specific AEs, SOCs or PDLCs (incidence <4 of subjects in all of the treatment groups)			X
	Change from Baseline Results (Labs, ECGs, Vital Signs)			X

† Adverse Experience references refer to both Clinical and Laboratory AEs.
Note: SOC=System Organ Class; PDLC=Pre-Defined Limit of Change; X = results will be provided.

3.6.3 Statistical Methods for Exploratory Analyses

3.6.3.1 Time to Progression (TTP)

The non-parametric Kaplan-Meier method will be used to estimate the TTP curve in each treatment group. The treatment difference in TTP 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 (i.e., hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. The stratification factors used for randomization will be applied to both the stratified log-rank test and the stratified Cox model. The censoring rules for analysis of TTP are summarized in Table 5.

Table 5 Censoring rules for Analyses of TTP

Situation	Analysis
No PD;new anticancer treatment is not initiated	Censored at last disease Assessment
No PD;new anticancer treatment is initiated	Censored at last disease,assessment before new anticancer treatment
PD	Progressed at date of documented PD

3.6.3.2 Duration of Response (DOR)

If sample size permits, response duration will be summarized descriptively using Kaplan-Meier method. Only the subset of subjects who show a complete response or partial response will be included in this analysis.

Censoring rules for DOR are summarized in Table 6. 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.

Table 6 Censoring Rule of 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 after ≥ 2 consecutive missed disease assessments	Last adequate disease assessment prior to ≥ 2 missed adequate disease assessments	Censor (non-event)
Death or progression after ≤ 1 missed disease assessments	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.		

3.6.3.3 Progression-Free Survival 2 (PFS2)

An exploratory analysis of PFS2, defined as the time from randomization to disease progression on the next line of therapy post study drug, or death from any cause, whichever first, may be carried out.

The analysis of PFS2 will be conducted using the same statistical methods as the primary analysis of PFS and OS, for example, the stratified log-rank test and Cox model.

3.6.4 Summaries of Demographic and Baseline Characteristics

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 subjects screened, randomized, the primary reasons for screening failure, and the primary reason for discontinuation will be displayed. Demographic variables (e.g., age), baseline characteristics, primary and secondary diagnoses, and prior and concomitant therapies will be summarized by treatment either by descriptive statistics or categorical tables.

Statistical testing and inference for safety analyses are described in Section 3.6.2. Efficacy results that will be considered to be statistically significant after consideration of the strategy for controlling the type I error are described in Section 3.8– Multiplicity. Nominal p-values may be computed for other efficacy analyses as a measure of strength of association between the endpoint and the treatment effect rather than formal tests of hypotheses.

3.7 Interim Analyses

There is no interim efficacy analysis planned in this study for PFS.

There is one interim efficacy analysis planned in this study for OS. For OS, a Hwang-Shih-DeCani alpha-spending function with the gamma parameter (-4) is constructed to implement group sequential boundaries that control the Type-I error rate. The actual boundaries will be determined from the number of OS events observed at the time of the interim analysis, using the alpha-spending function.

The interim OS will be performed at the time of final PFS analysis: (1) enrollment is completed; (2) approximately 235 PFS events and 190 OS events have been observed. The final analysis (FA) for OS will be performed when at least 290 OS events have been observed (~36 months after trial starts).

The secondary hypothesis of ORR will be tested only if pembrolizumab arm is superior to the control in OS. The information fraction for the group sequential boundaries and alpha spending function of ORR will be defined by the proportion of subjects whose randomization dates are at least 6 months before the data cutoff date of the analysis. Only these eligible subjects can be included into the ORR analysis. It is projected that there will be at least 270 eligible subjects at the interim analysis time point. The nominal Type I error rates for the interim analysis and final analysis that will allow tight control of the overall Type I error for testing the ORR hypothesis will be derived using the alpha-spending function approach. The group sequential testing of the ORR hypothesis will be conducted with an efficacy boundary only. The efficacy boundary for the ORR will be set using an Exponential spending function $f(t) = \alpha^{t^{-\nu}}$ [8] with parameter ν=0.25, which yields a Pocock-like boundary.

Table 7 summarizes the timing, number of events and decision guidance for the PFS, OS and ORR analysis. The actual boundaries and the alpha level for the OS and ORR analyses will be determined from the actual number of events observed at the time of the analysis using the corresponding alpha-spending function.

Table 7 Summary of Timing, Sample Size and Decision Guidance of Interim Analysis and Final analysis

Analysis	Criteria for Conduct of Analysis	Endpoint	p value (1-sided) at boundary	Approx. Observed HR or ORR-Difference at Boundary
Final PFS Analysis/ Interim OS Analysis	~ 24 months after trial starts (1) enrollment is complete (2) approximately 235 PFS events and 190 OS events have been observed PFS Events: ~235 OS Events: ~190	PFS	0.0035	0.70
		OS	0.0051	0.69
		ORR†	0.0161	10.2%
Final OS Analysis	~ 36 mos after trial starts At least 290 OS events have been observed OS Events: ~290	OS	0.0198	0.78
		ORR†	0.0123	9.3%

†: The secondary hypothesis of ORR will be tested only if pembrolizumab arm is superior to the control in OS. The assumed expected ORR in pembrolizumab and control groups are 30% and 10%, respectively. Depending on the results of the OS and PFS hypothesis testing, the ORR hypothesis can be tested at Type I error levels of α=2.15% or 2.5%; this table assumes an ORR Type I error of 2.15%.

3.8 Multiplicity

The multiplicity strategy specified in this section will be applied to the two primary hypotheses (superiority of pembrolizumab for OS or PFS) and one secondary hypothesis (superiority of pembrolizumab for ORR).

The overall Type-I error is strongly controlled at 2.5% (one-sided), with 0.35% allocated to PFS and 2.15% allocated to OS hypothesis and 0% to the ORR hypothesis.

For the OS and ORR endpoints, the Type-I error rate for the interim analysis and final analysis is controlled through alpha-spending functions as described in Section 3.7 Interim Analyses.

By using the graphical approach of Maurer and Bretz [2], if the PFS hypothesis is rejected, the corresponding alpha level can be shifted to the OS hypotheses. If the OS hypothesis is rejected, the corresponding alpha level can be shifted to the ORR hypothesis.

See Figure 1 for the multiplicity strategy diagram of the study.

Confidential

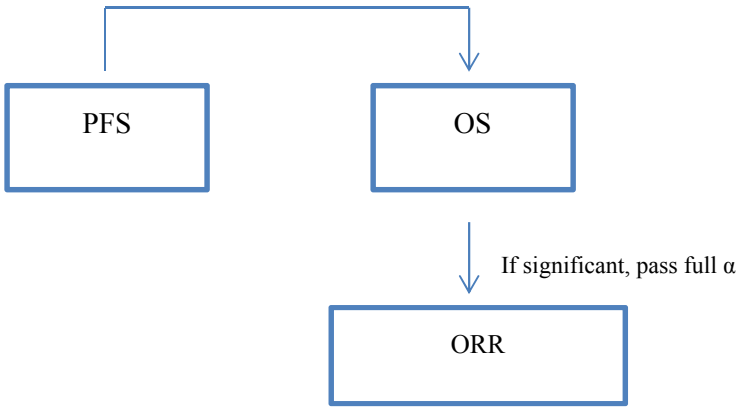

Figure 1 Multiplicity Strategy

3.9 Sample Size and Power Calculations

The study will randomize subjects in a 1:1 ratio into pembrolizumab arm and paclitaxel arm. The overall sample size will be approximately 360. However, subjects already in screening phase may be enrolled even after we have reached the maximum sample size.

The final analysis of the study is event driven (i.e., follow-up time are subject to change but number of events is not) and will complete after at least 290 OS events have been observed.

PFS analysis: the final PFS analysis will be performed after approximately 235 PFS events observed. With ~235 PFS events, this trial has overall 99% (88%) power to demonstrate that pembrolizumab is superior to paclitaxel at a one-sided 0.35% alpha-level, if the underlying hazard ratio of PFS is 0.5 (0.6). Success for PFS at the main analysis approximately corresponds to an observed hazard ratio of < 0.70.

The power calculation is based on the following assumptions for subjects: 1) Progression-free survival follows an exponential distribution with a median of 4.5 months in the paclitaxel arm; 2) An enrollment period of 24 months (IA is conducted after enrollment is complete); 3) A yearly drop-out rate of 5%.

OS analysis: The final OS analysis will be carried out when at least 290 OS events have occurred. For primary endpoint OS, the trial has 91% (85%) power to demonstrate that pembrolizumab is superior to paclitaxel at an one-sided 2.15% alpha-level, if the underlying hazard ratio of OS is 0.67 (0.7). Success for OS at the final analysis approximately corresponds to an observed hazard ratio of < 0.78 (approximately a 2-month improvement or greater in median OS).

The sample size and power calculation is based on the following assumptions for subjects: 1) Overall survival follows an exponential distribution with a median of 7.5 months in the

C Confidential

control arm; 2) An enrollment period of 24 months and a minimum of 12 months follow-up after enrollment completion; 3) A yearly dropout rate of 2%.

The assumptions for the median PFS of 4.5 months and the median OS of 7.5 months in the paclitaxel arm are based on estimates of median PFS and median OS from China subgroup analysis of TyTAN trial [9].

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 dual primary endpoints will be estimated and plotted within each category of the following classification variables:

- Age category (≤65 vs. >65 years)
- Sex (Female vs. Male)
- ECOG Performance Scale (0 vs. 1)
- Primary location (Stomach vs. GEJ)
- Histological subtype (Diffuse vs. intestinal vs. mixed)
- Disease Status (Locally advanced vs. Metastatic)
- Time to progression on first-line therapy (< 6 months vs. ≥ 6 months)
- Geographic region of enrolling site (China vs. ex-China)

Country specific population (e.g. Chinese, etc.) may also be analyzed per local regulatory requirements.

For analysis in subgroup of the ITT population, the unstratified Cox model will be used to estimate the OS and PFS hazard ratio between the treatment arms and the unstratified log-rank test will be used to assess the treatment difference. If any level of a subgroup variable has fewer than 20% of the ITT population, above analysis will not be performed for this level of the subgroup variable.

3.11 Compliance (Medication Adherence)

Drug accountability data for trial treatment will be collected during the study. Compliance with trial treatment administration will be measured by subjects: 1) receiving unscheduled study agent infusions/injections; 2) missing an infusion/injection. Numbers and percentages of subjects and infusion/injection visits with any deviation in these measures will be reported for the ITT population.

3.12 Extent of Exposure

The extent of exposure will be summarized as duration of treatment in cycles. Dose intensity will also be summarized as appropriate.

C Confidential

4 REFERENCES

1. Miettinen O, Nurminen M. Comparative analysis of two rates. Statistics in Medicine 1985; 4:213-226.

2. Maurer W, Bretz F. Multiple Testing in Group Sequential Trials using Graphical Approaches. Statistics in Biopharmaceutical Research 2013; 5(4): 311-320.

3. 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.

4. Latimer, N.R., Abrams, K.R., Lambert, P.C., Crowther, M.J., Wailoo, A.J., Morden,J.P. et al. Adjusting Survival Time Estimates to Account for Treatment Switching in Randomized controlled Trials - an Economic Evaluation Context: Methods, Limitations, and Recommendations. Medical Decision Making 2014 Apr;34(3):387-402.

5. 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.

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

7. Mehrotra D., Su S.C., Li X. An efficient alternative to the stratified Cox model analysis. Statist. Med. 2012, 31 1849–1856.

8. Anderson KM, Clark JB. Fitting spending functions. Stat Med 2010;29(3):321-7.

9. Satoh et al. Lapatinib Plus Paclitaxel Versus Paclitaxel Alone in the Second-Line Treatment of HER2-Amplified Advanced Gastric Cancer in Asian Populations: TyTAN - A Randomized, Phase III Study. J Clin Oncol, 2014, 32:2039-2049.

C Confidential

Revision History

Date	Summary of Change
16FEB2017	Original Document