Official Protocol Title:	A Phase III Randomized Open-Label Study of Single Agent Pembrolizumab vs. Physicians' Choice of Single Agent Docetaxel, Paclitaxel, or Irinotecan in Subjects with Advanced/Metastatic Adenocarcinoma and Squamous Cell Carcinoma of the Esophagus that have Progressed after First-Line Standard Therapy (KEYNOTE-181)
NCT number:	NCT02564263
Document Date:	28-Feb-2018

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

1. INTRODUCTION

This 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 sSAP 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. There will be a separate pharmacokinetic data analysis plan as well as a biomarker analysis plan.

2. SUMMARY OF CHANGES

Section Number (s)	Section Title(s)	Description of Change (s)	Rationale
3.6.1	Statistical Methods for Efficacy Analyses	Note that in addition to a positive test for the treatment difference for OS in all subjects using the stratified max-combo test, the upper bound of the stratified Cox HR should be <1.1.	To allow for a robust assessment of the significance of a positive assessment of the treatment effect on OS based only on the stratified max-combo test.
3.6.1	Statistical Methods for Efficacy Analyses	'A sensitivity analysis, which tests the hypothesis of treatment difference for OS in all subjects using the stratified log-rank test, will be conducted.' was replaced by 'Due to historical precedent, the log-rank test as an alternative to the max combo test for the overall population will also be evaluated, including applying the log-rank in the multiplicity scheme in the same fashion as if it were the primary testing method.' In Overall Survival and Progression-Free	Due to the historical precedent for the log-rank test, it will be evaluated along with the as max combo test in the same fashion.

Section Number (s)	Section Title(s)	Description of Change (s)	Rationale
		analysis.	

3. ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

The key elements of the statistical analysis plan that are summarized below are applicable to the Global Cohort; the comprehensive plan is provided in Sections 3.2 through 3.12. Statistical analysis plan for the China Cohort is provided in Section 4.

Table 1 Statistical Analysis Plan

Study Design Overview	A Phase III Randomized Open-Label Study of Single Agent Pembrolizumab vs. Physicians' Choice of Single Agent Docetaxel, Paclitaxel, or Irinotecan in Subjects with Advanced/Metastatic Adenocarcinoma and Squamous Cell Carcinoma of the Esophagus that have Progressed after First-Line Standard Therapy (KEYNOTE-181)
Treatment Assignment	Subjects will be randomized in a 1:1 ratio to receive pembrolizumab or investigator's choice of paclitaxel, docetaxel, or irinotecan (Control Arm). Stratification factors are in Section 5.4. This is an open-label study.
Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoints/Hypotheses	1. Overall Survival (OS) in subjects with squamous cell carcinoma of the Esophagus. 2. Overall Survival (OS) in subjects with PD-L1 CPS≥10 3. Overall Survival (OS) in all subjects.
Statistical Methods for Key Efficacy Analyses	The primary hypotheses will be evaluated by comparing pembrolizumab to the control on OS in subjects with squamous cell carcinoma of the esophagus and OS in subjects in ESCC subjects and with PD-L1 CPS≥ 10 using a stratified Log-rank test. The primary hypotheses on OS in all subjects will be evaluated using a stratified maximum weighted log-rank test (max-combo test) [9] [10]. 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. "Tier 1" safety endpoints 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. There are no Tier 1 events in this trial. The between-treatment difference will be analyzed using the Miettinen and Nurminen method [1].

Interim Analyses	One interim efficacy analysis will be performed in this study. Results will be reviewed by an external data monitoring committee. Details are provided in Section 8.7. • Interim Efficacy Analysis o Timing: To be performed after (1) enrollment is completed, (2) approximately 251 OS events and 385 OS events have been observed among subjects with squamous cell carcinoma of the esophagus and all subjects, respectively, and (3) 8 months after last subject randomized. In addition, if there are fewer than 172 OS events among subjects with PD-L1 CPS≥10 at the time, the interim efficacy analysis may be delayed for up to 2 months or when the target number of OS events in subjects with PD-L1 CPS≥10 is reached, whichever occurs first. o Primary Purpose: Interim efficacy analysis for OS • Final analysis o Timing: after approximately 310 OS events and 473 OS events have been observed among subjects with squamous cell carcinoma of the esophagus and all subjects, respectively , and 16 months after last subject randomized
Multiplicity	The multiplicity strategy specified in this section will be applied to the three primary hypotheses (superiority of pembrolizumab on OS in subjects with squamous cell carcinoma of the esophagus, or subjects with PD-L1 CPS≥10% or in all subjects) and two secondary hypotheses (superiority of pembrolizumab on PFS in all subjects or ORR in all subjects). The overall Type-I error is strongly controlled at 2.5% (one-sided), with initially 0.8% allocated to OS hypothesis in subjects with squamous cell carcinoma of the esophagus, 0.9% allocated to OS hypothesis in subjects with PD-L1 CPS≥10 and 0.8% allocated to the OS hypotheses in all subjects, and 0% to the PFS and ORR hypotheses. By using the graphical approach of Maurer and Bretz [2], if OS hypothesis in subjects with squamous cell carcinoma of the esophagus is rejected, the corresponding alpha level can be shifted to OS hypothesis in all subjects. If OS hypothesis in subjects with PD-L1 CPS≥10 is rejected, the corresponding alpha level can also be shifted to OS hypotheses in all subjects. The secondary hypotheses of PFS and ORR will be tested only if pembrolizumab arm is superior to the control in OS in all subjects. If OS hypothesis in all subjects is rejected, the corresponding alpha level can be shifted by half to PFS in all subjects and by half to ORR in all subjects, respectively.
Sample Size and Power	For the hypotheses in all subjects, the sample size is approximately 600. Among all subjects, it is expected about 400 subjects with squamous cell carcinoma of the esophagus will be enrolled. For the hypotheses in subjects with PD-L1 CPS≥10, the sample size is approximately 280 (based on an observed prevalence rate of ~47% from KN180). For the primary endpoint, OS in subjects with squamous cell carcinoma of the esophagus, with 310 OS events, the trial has 91.3% power to demonstrate that pembrolizumab is superior to the control at a one-sided 0.8% alpha-level, if the underlying hazard ratio of OS is 0.65. For the primary endpoint, OS in subjects with PD-L1 CPS≥10, with 213 OS events, the trial has 90.9% power to demonstrate that pembrolizumab is

	superior to the control at a one-sided 0.9% alpha-level, if the underlying hazard ratio of OS is 0.6. For the primary endpoint, OS in all subjects, with 473 OS events, the trial has 92.6% power to demonstrate that pembrolizumab is superior to the control at a one-sided 0.8% alpha-level, if the underlying hazard ratio of OS is 0.7.

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 for this protocol, and the randomization will be implemented in IVRS/IWRS.

Although the trial is open label, analyses or summaries generated by randomized treatment assignment, actual treatment received will be limited and documented. In addition, the central imaging vendor will perform the central imaging review without knowledge of treatment group assignment.

The eDMC will serve as the primary reviewer of the unblinded results 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 eDMC, the Sponsor may prepare a regulatory submission. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee and limited additional Sponsor personnel may be unblinded to results at the treatment level in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of interim analyses will be documented. Additional logistical details, revisions to the above plan and data monitoring guidance will be provided in the eDMC Charter.

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Protocol Section 3.0.

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.

Secondary

Progression-free survival (PFS) – RECIST 1.1 by central imaging vendor review in all subjects;

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on central imaging vendor review or death due to any cause, whichever occurs first. See Section 8.6.1 for definition of censoring;

Objective Response Rate (ORR) – RECIST 1.1 by central imaging vendor review in all subjects;

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) – RECIST 1.1 by investigator assessment and irRECIST assessed by central imaging vendor

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. See Section 3.6.1 for definition of censoring.

Objective Response Rate (ORR) – RECIST 1.1 by investigator assessment

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).

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.

The China Cohort

After the sample size required for the Global Cohort is reached, the study will continue to randomize subjects in China until the sample size for the Chinese subjects meets the required target for China. The Chinese subjects randomized after the enrollment of the Global Cohort is closed will not be included in the above primary efficacy analysis population which is based on the Global Cohort. The China Cohort will also be analyzed separately per local regulatory requirement.

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 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.

The China Cohort

The Chinese subjects randomized and treated in the China extension enrollment period will not be included in the above primary safety analysis population. The China Cohort will also be analyzed separately per local regulatory requirement.

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 may be computed for other efficacy analyses, but should be interpreted with caution due to potential issues of multiplicity.

Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to estimate 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 hypotheses of treatment difference for OS in subjects with squamous cell carcinoma of the esophagus and OS in subjects with PD-L1 CPS≥10 will be tested using the stratified log-rank test. The hypotheses of treatment difference for OS in all subjects will be tested using the stratified max-combo test. The stratification factors used for randomization will be applied to both the stratified log-rank test, stratified max-combo test and the stratified Cox model if applicable. Note that in addition to a positive test for the treatment difference for OS in all subjects using the stratified max-combo test, the upper bound of the stratified Cox HR should be <1.1.

The max-combo test statistic is the maximum of the log-rank test statistic and a weighted log-rank variation of the Fleming-Harrington test statistic; $Z_m = \max(Z_1, Z_2)$, where Z_1 and Z_2 are

the test statistics from the FH (0, 0) and FH (0, 1) family of test statistics, respectively. FH (0, 0) corresponds to the log-rank test, while FH (0, 1) is more sensitive to late-difference alternatives.

Due to historical precedent, the log-rank test as an alternative to the max combo test for the overall population will also be evaluated, including applying the log-rank in the multiplicity scheme in the same fashion as if it were the primary testing method.

Subjects in the control arm are expected to discontinue treatment earlier compared to subjects in the pembrolizumab arm, and may switch to another anti PD-1 treatment. 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) [6] or a two stage model [7], based on an examination of the appropriateness of the data to the assumptions required by the methods.

The RPSFT model assumes:

- 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.
- An equal treatment effect for subjects after switching to a treatment as for those initially allocated to receive it.

The RPSFT method uses the Accelerated Failure Time (AFT) model to link the observed survival time (T) and the counterfactual survival time (S) that would be observed if no treatment were received. Patients randomized to the control arm who never switch to treatment with an anti-PD-1 or anti-PD-L1 therapy don't need any adjustment. Patients randomized to the control arm who switch to treatment with an anti-PD-1 or anti-PD-L1 agent need the adjustment for their survival time after crossover using the Acceleration Factor. To identify the acceleration factor, patients who are randomized to treatment arm need to be adjusted by acceleration factor (unresolved) first for all their observed survival time to counterfactual survival time. Then a grid search method is used to get the best estimate of the Acceleration Factor such that the counterfactual survival life of treatment arm and control arm would be as close as possible. More detailed steps to implement the RPSFT method will be provided in the Programming Requirement Specification (PRS) for the macro implementing the RPSFT method.

The two stage method is based on a modified iterative parametric estimation [7]. It assumes:

- There are no unmeasured confounders at the secondary baseline time-point (disease progression),
- Treatment switching only happens after progression, and happens soon after progression.

At Stage 1, the date of disease progression is used as a secondary baseline for subjects who have a documented progression in the standard of care 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 in order to approximate their event time had they not switched treatment, 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.

A supportive analysis, which includes PD-L1 status (CPS≥10 vs. CPS<10) and the stratification factors used for randomization as the factors in the stratified log-rank test and Cox model, may be performed for the all comer population analysis.

Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to estimate 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 hypotheses of treatment difference for PFS in subjects with squamous cell carcinoma of the esophagus and PFS in subjects with PD-L1 CPS≥10 will be tested using the stratified log-rank test. The hypotheses of treatment difference for PFS in all subjects will be tested using the stratified max-combo test. The stratification factors used for randomization will be applied to both the stratified log-rank test, stratified max-combo test and the stratified Cox model if applicable.

Due to historical precedent, the log-rank test as an alternative to the max combo test for the overall population will also be evaluated, including applying the log-rank in the multiplicity scheme in the same fashion as if it were the primary testing method.

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, for the subjects who have PD, 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 review, regardless of discontinuation of study drug. Death is always considered as a confirmed PD event. 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 review, we will perform two sensitivity analyses with a different set of censoring rules. The first sensitivity analysis is the same as the primary analysis except that it censors at the last disease assessment without PD when PD or death is documented after more than one missed disease assessment. 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. If there is an imbalance between the treatment groups on disease assessment schedules or censoring patterns, we will also perform additional PFS sensitivity analysis: a PFS analysis using scheduled 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
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 missed disease assessment	Progressed at date of documented PD or death	Progressed at date of documented PD or death	Progressed at date of documented PD or death
PD or death documented at any time after ≥ 2 consecutive missed	Progressed at date of documented PD or death	Censored at last disease assessment prior to the ≥ 2 consecutive missed disease	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 to PFS will be plotted for the comparison between pembrolizumab and the control 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 the Restricted Mean Survival Time (RMST) method [3] or a parametric method [4].

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 for a treatment group by the area under the KM curve up to a specified follow up time. The difference of two RMSTs between the two treatment groups will be estimated along with its 95% confidence interval. A series of different cutoff time (8 months, 12 months etc.) will be pre-specified prior to unblinding of the study by a Merck team who are blinded to the treatment group assignment.

For the parametric method, accelerated failure time model and negative binomial cure rate model [8] using an underlying Weibull distribution (using gamlss.cens package in R) may be conducted.

One assumption for the 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 [5], 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.

A supportive analysis, which includes PD-L1 status (CPS≥10 vs. CPS<10) and the stratification factors used for randomization as the factors in the stratified log-rank test and Cox model, may be conducted in all subjects for the primary analysis of all comer population only.

Objective Response Rate (ORR)

Stratified Miettinen and Nurminen's method [1] will be used for comparison of the objective response rates between the treatment arms. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be reported. The stratification factors used for randomization will be applied to the analysis if applicable.

Table 3 summarizes the primary analysis approach 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, multiple populations, and interim analyses 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 in subjects with squamous cell carcinoma of the Esophagus.	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT in subjects with squamous cell carcinoma of the Esophagus.	Censored at last known alive date
Primary Hypothesis #2			
OS in subjects with PD-L1 CPS≥10.	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT in subjects with PD-L1 CPS≥10	Censored at last known alive date
Primary Hypothesis #3			
OS in all subjects	Test: Stratified Max-combo Estimation: Stratified Cox model with Efron's tie handling method	ITT in all subjects	Censored at last known alive date
Key Secondary Endpoints			
PFS per RECIST 1.1 by central imaging vendor review in all subjects	Test: Stratified Max-combo Estimation: Stratified Cox model with Efron's tie handling	ITT in all subjects	• Primary censoring rule • Sensitivity analysis 1 • Sensitivity analysis 2

Endpoint/Variable (Description, Time Point)	† Statistical Method	Analysis Population	Missing Data Approach
	method		(More details are in Table 2)
ORR per RECIST 1.1 by central imaging vendor review in all subjects	Test: Stratified M & N method‡	ITT in all subjects	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 (See Section 5.4) will be applied to the analysis model if applicable.

‡ Miettinen and Nurminen method [1]

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.

AEs (specific terms as well as system organ class terms) 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 AEs 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 AEs 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.

Based on the mechanism of action of pembrolizumab and safety data observed in historic pembrolizumab trials to date, there are no events of interest that warrant classification as Tier I events for this protocol. In addition, 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) [1], an unconditional, asymptotic method.

Detailed kinetics and characteristics of immune mediated AEs will be summarized in this study.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint†	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

Time to Grade 3-5 AE

In addition to the tiered approach, an exploratory analysis may be performed for time to first Grade 3-5 AE. Time to first Grade 3-5 AE is defined as the time from the first day of study drug to the first event of Grade 3-5 AE. For patients without a Grade 3-5 AE, the time to first Grade 3-5 AE is censored at 30 days post last study dose. The Kaplan-Meier method will be used to estimate the curve of time to first Grade 3-5 AE.

3.6.3 Statistical Methods for the Exploratory Analyses

PFS per RECIST 1.1 by investigator assessment and per irRECIST by central imaging vendor review

The primary analysis and censoring rule for PFS per RECIST 1.1 by central imaging vendor review will also be applied to the analysis for PFS per RECIST 1.1 by investigator assessment, and PFS per irRECIST by central imaging vendor review. For PFS per irRECIST by central imaging vendor review, if there is no confirmation scan available after the initial PD scan, then it is considered as a PFS event at the initial PD scan time point.

Objective Response Rate (ORR) per RECIST 1.1 by investigator assessment

The analysis for ORR per RECIST 1.1 by central imaging vendor review will also be applied to the analysis for ORR per RECIST 1.1 by investigator assessment.

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.

3.6.5 Statistical methods for patient-reported outcomes (PRO) endpoints

3.6.5.1 Statistical methods for PRO endpoints

The patient-reported outcomes are exploratory objectives in KN181, and thus no formal hypotheses were formulated. Nominal p-values to compare the pembrolizumab arm to the control arm may be provided as appropriate.

The PRO instruments are EORTC QLQ-C30, EORTC QLQ-OES18 and EuroQol-5D (EQ-5D).

PRO Endpoints:

- The mean score changes from baseline to Week 9 as measured by the EORTC QLQ-C30 global health status/quality of life scale.
- The mean score change from baseline to Week 9 for all QLQ-C30 sub-scales/items. The QLQ-C30 includes five functional dimensions (physical, role, emotional, cognitive, and social), three symptom scales (fatigue, nausea/vomiting, and pain), and six single item measures (dyspnea, sleep disturbance, appetite loss, constipation, diarrhea, and financial difficulties).
- The mean score change from baseline to Week 9 for all QLQ-OES18 sub-scales/items. The QLQ-OES18 contains 22 items with symptoms of dysphagia (three items), pain (four items), reflux symptoms (three items), eating restrictions (four items), anxiety (three items), dry mouth, taste, body image, and hair loss.
- The mean score change from baseline to Week 9 for EQ-5D VAS and utility score.
- The number and proportions of deterioration/ stable/improvement from baseline to Week 9 for all QLQ-C30 sub-scales/items.
- The time to deterioration (TTD) for QLQ-OES18 dysphagia (three items), pain (four items) and reflux symptoms (three items).

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

Scoring Algorithm:

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

According to the QLQ-C30 Manuals, if items $I_1, I_2,...,I_n$ are included in a scale, the linear transformation 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:

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

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

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

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

QLQ-OES18 scoring: QLQ-OES18 contains 18 items with symptoms of dysphagia (three items), pain (four items), reflux symptoms (three items), eating restrictions (four items), anxiety (three items), dry mouth, taste, body image, and hair loss [14].

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

The schedule for PRO data collection:

Table 5 provides the schedule for PRO data collection.

Table 5 PRO Data Collection Schedule

Treatment	Week									Discontinuation Visit	Follow-up Visit
	0	2/3/4	6	9	12	18	27	36	45		
MK-3475	C1	C2	C3	C4	C5	C7	C10	C13	C16	X	X
Paclitaxel	C1D1	C2D1	C2D15	C3D8	C4D1	C5D15	C7D22	C10D1	C12D8	X	X
Docetaxel	C1	C2	C3	C4	C5	C7	C10	C13	C16	X	X
Irinotecan	C1D1	C2 D1	C4D1	C5D8	C7D1	C10D1	C14D8	C19D1	C23D8	X	X
C: Cycle; D: Day Each cycle is 3 weeks for MK-3475. Each cycle is 4 weeks for Paclitaxel. Each cycle is 3 weeks for Docetaxel. Each cycle is 4 weeks for Irinotecan.											

The general rule of mapping relative day to analysis visit is provided in Table 6.

Table 6 Mapping Relative Day to Analysis Visit

	Week								
	0	2/3/4	6	9	12	18	27	36	45
Day	1	21	42	63	84	126	189	252	315
Range	-7-1	2-31	32-52	53- 73	74-105	106-157	158 -220	221-283	284-

At each scheduled visit, three instruments, EORTC QLQ-C30, EORTC QLQ-OES18 and EQ-5D, will be collected. If a patient does not complete the PRO instruments, the site staff will record the reason for missingness 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.

Analysis Populations

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

Statistical Methods

This section describes the planned analyses for the PRO endpoints. Descriptive statistics (mean, 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-OES18 will be plotted.

Table 7 gives an overview of the analyses planned for all PRO endpoints.

Table 7
Planned Statistical Analysis

Endpoint	Analysis	Primary Statistical Method	Report
Score change from baseline	Treatment effect estimation/comparison	Mixed effect model based on the missing at random (MAR) assumption.	lsmean score (95% CI) by treatment group and visit, lsmean score change (95% CI) from baseline by treatment group and visit, between-group difference in score change from baseline (95% CI, nominal p-value).
Proportion of deterioration/ stable/improvement	Treatment effect estimation/comparison	Summary with multiple imputation based on the MAR assumption	Proportion (95% CI) by treatment group at Week 9
Time to deterioration	Treatment effect estimation/comparison	Stratified log rank test Stratified Cox proportional hazard model Kaplan-Meier plot	Hazard Ratio (95% CI, p-value)

3.6.5.2 Treatment effect on PRO

To assess the treatment effects based on the PRO, for each continuous endpoint defined, a constrained longitudinal data analysis (cLDA) model will be used as the primary analysis method, with the PRO score as the response variable, and treatment by study visit interaction, and stratification factors as covariates.

The cLDA model is specified as follows:
$E(Y_{ijt}) = \gamma_0 + \gamma_{jt} I(t > 0) + \beta_{it} X_i$, $j = 1, 2$, $t = 0,1,2,3,...$, where Y_{ijt} is the PRO score for subject i, with treatment assignment j, at visit t, γ_0 is the baseline mean for both treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t, X_i is the stratification factor vector for this patient, and β_{it} is the coefficient vector for stratification factor at time t.

Treatment effect on PRO score change from baseline will be evaluated at Week 9. Treatment comparison will be performed and the differences in the lsmean change from baseline will be reported, together with 95% C.I. and nominal p-value at the primary analysis time points.

Most of the patients without disease progression are expected to have complete data up to 9 weeks. Patients with disease progression confirmed or feeling worse due to drug-related AE may have missing PRO assessments. The missing data must be handled accordingly to obtain valid statistical inference. The cLDA model implicitly treats missing data as missing at random

(MAR), i.e. missingness may depend on observed outcomes. Sensitivity analyses may be conducted in case the robustness of MAR assumption is questionable.

3.6.5.3 Analysis of the Proportions of Deterioration/Stable/Improvement

Patient's post-baseline PRO score will be classified as "improved" "stable" or "deteriorated" according to a change from baseline ≥10 points for each of the instrument/scale, as this magnitude of change is perceived by patients as being clinically significant [12].

The number and proportion of patients who were "improved", "stable", or "deteriorated", from baseline will be summarized by treatment group at 9 weeks based on MAR imputation (i.e. model based) of missing data.

3.6.5.4 Analysis of the Time to Deterioration

The true time-to-deterioration is defined as the time to first onset of a decrease of ≥10 points from baseline with confirmation under right-censoring rule (the last observation). The non-parametric Kaplan-Meier method will be used to estimate the deterioration curve in each group. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (hazard ratio) between treatment arms.

Compliance Summary

Completion and compliance of QLQ-C30, QLQ-OES18 and EQ-5D by visit and by treatment will be described based on PRO FAS population. Numbers and percentages of complete and missing data at each visit will be summarized for each of the treatment groups.

Completion rate in the FAS population is defined as the percentage of number of subjects who complete at least one item over the number of subjects in the FAS population at each time points.

$$Completion\ Rate = \frac{Number\ of\ Subjects\ who\ Complete\ at\ least\ one\ Item}{Number\ of\ Randomized\ Subjects}$$

The completion rate is expected to shrink in the later visit during study period due to the subjects who discontinued early. Therefore, another measurement, Compliance Rate, defined as the percentage of observed visit over 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)) will be employed as the support for completion rate.

$$Compliance\ Rate = \frac{Number\ of\ Subjects\ who\ Complete\ at\ least\ one\ Item}{Number\ of\ Eligible\ Subjects\ who\ are\ Expected\ to\ Complete}$$

Reason for non-completion will be summarized. An instrument is considered complete as at least one valid score available according to the missing item rules outlined in the EORTC QLQ-C30 Manual for each functional and symptoms scale.

3.7 Interim Analyses

There is one planned interim efficacy analysis in this trial. Results will be reviewed by an external data monitoring committee (eDMC).

The primary purpose of the interim efficacy analysis is to evaluate superiority of pembrolizumab in OS. In order to account for potential delayed treatment effects that have been observed with immunotherapy, the interim efficacy analysis will be performed after: (1) enrollment is completed, (2) approximately 251 OS events and 385 OS events have been observed among subjects with squamous cell carcinoma of the Esophagus and all subjects, respectively, and (3) 8 months after last subject randomized. In addition, if there are fewer than 172 OS events among subjects with PD-L1 CPS≥10 at the time, the interim efficacy analysis may be delayed for up to 2 months or when the target number of OS events in subjects with PD-L1 CPS≥10 is reached, whichever occurs first. Thus, adequate follow-up time is incorporated into the trial to ensure that the interim efficacy analysis is conducted at an appropriate time to characterize the potential benefit of immunotherapy. The boundary for the final analysis will be adjusted according to the actual alpha spent at IA and the actual number of events at IA and FA.

For the OS hypothesis, Lan-DeMets O'Brien-Fleming alpha spending function with specified calendar time fraction (0.76) [11] will be used to construct group sequential boundaries to control the Type-I error.

$$Calendar\ Time\ Fraction = \frac{Interim\ Analysis\ Time\ (\sim 25\ \ months\ after\ first\ subject\ randomized)}{Final\ Analysis\ Time\ (\sim 33\ \ months\ after\ first\ subject\ randomized)} = 0.76$$

The actual boundaries for interim analysis will be determined from the number of OS events observed at the time of the interim efficacy analysis using the alpha-spending function. The actual boundaries for final analysis will be determined from the number of OS events observed at the time of the interim efficacy analysis and final analysis using the alpha-spending function.

Table 8 summarizes the timing, sample size and decision guidance of the interim analysis and final analysis. Bounds are based on estimated number of events and will be updated at times of analyses using spending functions as noted above.

Table 8 Summary of Timing, Sample Size and Decision Guidance of Interim Analysis and Final analysis

Analysis	Criteria for Conduct of Analysis	Endpoint	Value	Efficacy
Interim Efficacy Analysis	~ 25 months after first subject randomized Approximately 251 OS events and 385 OS events have been observed among subjects with squamous cell carcinoma of the esophagus and all subjects, respectively, and 8 months after last subject randomized. If there are fewer than 172 OS events among subjects with PD-L1 CPS≥10 at the time, the interim efficacy analysis may be delayed for up to 2 months or when the target number of OS events in subjects with PD-L1 CPS≥10 is reached, whichever occurs first. OS events among subjects with squamous cell carcinoma of the esophagus: ~251 OS events among subjects with PD-L1 CPS≥10: ~172 OS events among all subjects: 385	OS in subjects with squamous cell carcinoma of the esophagus	p value (1-sided) at boundary ~ HR at boundary	≤0.0023 0.70
		OS in subjects with PD-L1 CPS≥10	p value (1-sided) at boundary ~ HR at boundary	≤0.0027 0.65
		OS in all subjects	p value (1-sided) at boundary ~ HR at boundary	≤0.0023 0.75
Final Analysis	~ 33 months after first subject randomized Approximately 310 OS events and 473 OS events have been observed among subjects with squamous cell carcinoma of the esophagus and all subjects, respectively, and 16 months after last subject randomized. OS events among subjects with squamous cell carcinoma of the esophagus: ~310 OS events among subjects with PD-L1 CPS≥10: ~213 OS events among all subjects: 473	OS in subjects with squamous cell carcinoma of the esophagus	p value (1-sided) at boundary ~ HR at boundary	≤0.0075 0.76
		OS in subjects with PD-L1 CPS≥10	p value (1-sided) at boundary ~ HR at boundary	≤0.0084 0.72
		OS in all subjects	p value (1-sided) at boundary ~ HR at boundary	≤0.0075 0.80

3.8 Multiplicity

The multiplicity strategy specified in this section will be applied to the three primary hypotheses (superiority of pembrolizumab on OS in subjects with squamous cell carcinoma of the esophagus or subjects with PD-L1 CPS≥10 or all subjects) and two secondary hypotheses (superiority of pembrolizumab on PFS in all subjects and ORR in all subjects).

The overall Type-I error is strongly controlled at 2.5% (one-sided), with initially 0.8% allocated to OS hypothesis in subjects with squamous cell carcinoma of the esophagus, 0.9% allocated to OS hypothesis in subjects with PD-L1 CPS≥10 and 0.8% allocated to OS hypothesis in all subjects, and 0% to PFS and ORR hypotheses.

Within each hypothesis, the Type-I error rate for the interim efficacy analysis and final analysis is controlled through alpha-spending functions as described in Section 8.7 Interim Analyses.

By using the graphical approach of Maurer and Bretz [2], if OS hypothesis in subjects with squamous cell carcinoma of the esophagus is rejected, the corresponding alpha level can be shifted to OS hypothesis in all subjects. If the OS hypothesis in subjects with PD-L1 CPS≥10is rejected, the corresponding alpha level can also be shifted to the OS hypothesis in all subjects.

The secondary hypotheses of PFS and ORR will be tested only if pembrolizumab arm is superior to the control on OS in all subjects. If the OS hypothesis in all subjects is rejected, the corresponding alpha level can be shifted by half to PFS in all subjects and half to ORR in all subjects. The cumulative alpha spending for PFS hypothesis is determined by the same alpha spending function (with calendar fraction=0.76) defined for OS hypotheses. If OS in all subjects is not statistically significant at the interim analysis, then ORR at the IA will be considered without any data updates if the step-down criteria allow formal testing based on FA of OS in all subjects.

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

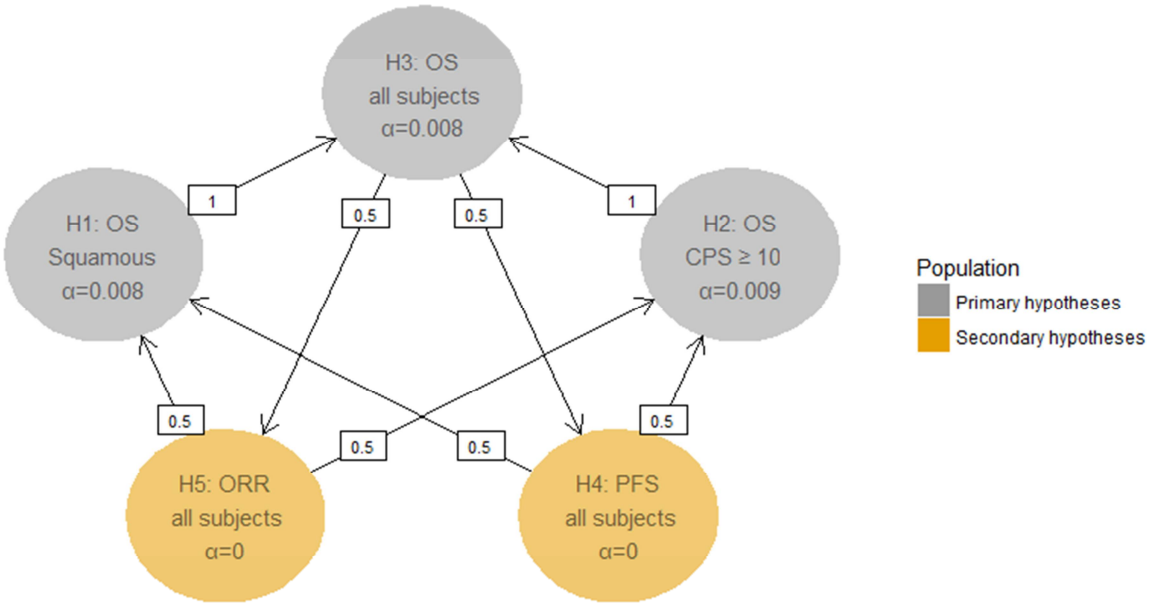

Figure 1 Multiplicity Strategy

3.9 Sample Size and Power Calculations

The study will randomize subjects in a 1:1 ratio into pembrolizumab arm and the control arm. The enrollment is driven by all subjects. The total sample size in the Global Cohort is approximately 600. It is expected that approximately 400 subjects with squamous cell carcinoma of the esophagus will be enrolled. Based on the observed preliminary prevalence of PD-L1 CPS≥10 in subjects with esophageal carcinoma of ~47% from MK3475 KN180, for the hypotheses in subjects with PD-L1 CPS≥10, the sample size is approximately 280.

The final analysis of the study will complete after approximately 310 OS events and 473 OS events have been observed among subjects with squamous cell carcinoma of the esophagus and all subjects, respectively, and 16 months after last subject randomized.

OS analysis

The sample size and power calculations are based on the following assumptions: 1) Overall survival follows an exponential distribution with a median of 8 months in the control arm; 2) an enrollment period of 17 months and a minimum of 16 months follow-up after enrollment completion; 3) a yearly dropout rate of 2%.

The final OS analysis will be carried out after approximately 310 OS events and 473 OS events have been observed among subjects with squamous cell carcinoma of the esophagus and all subjects, respectively, and 16 months after last subject randomized. It is expected that approximately 213 OS events would have been observed in subjects with PD-L1 CPS≥10. With 310/213/473 OS events in subjects with squamous cell carcinoma of the esophagus/subjects with PD-L1 CPS≥10/all subjects, respectively, the trial has at least 91.3%/90.9%/92.6% power to demonstrate that pembrolizumab is superior to the control at a one-sided 0.8%/0.9%/0.8% alpha-

level, if the underlying hazard ratio of OS is 0.65/0.6/0.7. Success for OS at the final analysis approximately corresponds to an observed hazard ratio of < 0.76 for subjects with squamous cell carcinoma of the esophagus, 0.72 for subjects with PD-L1 CPS≥10, and < 0.80 for all subjects. To further investigate the impact of the delayed separation of OS curve on the actual power in all subjects, a simulation was carried out using the current study design parameters described above but with a piece-wise time varying hazard ratio: the hazard ratio was specified as 1 and 0.6 at the beginning of time intervals of Month 0 and 5 since randomization respectively. With 1,000 simulations the overall study power with 473 events at the final analysis given the hazard ratio assumption above is approximately 74.7% using log-rank test statistics and 86% using max-combo test statistics.

The China Cohort

After the enrollment for the Global Cohort has completed, the study will continue to randomize subjects in a 1:1 ratio into the pembrolizumab arm and the SOC arm in China until the sample size for the Chinese subjects overall reaches approximately 120. Chinese subjects randomized after completion of enrollment in the Global Cohort will not be included in the analyses of the Global Cohort.

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 three primary endpoints (OS) will be estimated and plotted within each category of the following classification variables:

- Age category (<65 vs. ≥65 years)
- Sex (Female vs. Male)
- Geographic region (Asia vs. Rest of the World)
- ECOG Performance Scale (0 vs. 1)
- Histological subtype (Squamous cell carcinoma vs. adenocarcinoma/Siewert type 1 adenocarcinoma of the EGJ)

For OS, the stratified Cox model will be used. The consistency of the treatment effect will be assessed descriptively via summary statistics by category for the classification variables listed above. If any level of a subgroup variable has fewer than 10% of the ITT population, above analysis will not be performed for this level of the subgroup variable. If a subgroup variable has two levels and one level of the subgroup variable has fewer than 10% of the ITT population, then this subgroup will not be displayed in the forest plot.

Asia includes China, Japan, Korea, Hong Kong, Taiwan, Malaysia, Thailand, and Singapore.

In addition to the subgroup based on Asia vs. Rest of the World, US vs. ex-US and EU vs. ex-EU will also be assessed.

The EU region includes countries from both EU member states (2016) and EFTA members.

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

3.11 Compliance (Medication Adherence)

Drug accountability data for trial 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 cycles.

4. STATISTICAL ANALYSIS PLAN FOR CHINA COHORT

4.1 Introduction

Approximately 120 subjects from China will be enrolled in the China Cohort; this will include subjects enrolled in China during the global enrollment period as well as during the China extension enrollment period. After the enrollment in the Global Cohort is closed, subjects from China will continue to be enrolled in the China extension period designed to meet local regulatory needs. The China Cohort will be identical to the Global Cohort (e.g., inclusion and exclusion criteria, study endpoints, primary and secondary objectives, study procedures) in general, with additional statistical analyses for the China Cohort.

The purpose of the China Cohort is to evaluate the consistency of efficacy and safety in the Chinese subpopulation and the global population. Country-specific analysis may also be conducted per local regulatory requirement.

After the enrollment in the Global Cohort is completed, subjects in China will continue to be enrolled in a 1:1 ratio into the pembrolizumab arm and the SOC arm until the sample size in this subpopulation reaches approximately 120.

After the cut-off date for the primary analyses based on the Global Cohort (including interim and final analyses), all Chinese subjects, including subjects enrolled in the Global Cohort and those enrolled during the China extension period, will continue their randomized treatment and continue to be followed up for OS events for China registration purpose. The China Cohort will be completed after the target number of OS events has been observed between the two arms in all Chinese subjects and 8 months after last Chinese subject randomized. The expected timing of the analysis for the China Cohort is about 23 months from when the first Chinese subject is randomized in the Global Cohort. Additional analyses may be considered for China Cohort based on Sponsor's discretion and/or consultation with regulatory if global final analysis shows positive results and leads to filing.

4.2 Responsibility for Analyses/In-House Blinding

Although the trial is open-label, analyses or summaries generated by randomized treatment assignment and actual treatment received will be limited and documented. Subjects randomized in the China extension period will not be included in the data base locked for the analysis of the

global Cohort. For all Chinese subjects, including subjects randomized in the Global Cohort and the extension period, patient level treatment randomization information, will be in-house blinded in the analysis team for statistician(s)/programmer(s) responsible for the China analysis until the data base lock for China.

4.3 Hypotheses/Estimation

No hypothesis testing is planned for the China Cohort.

After succeeding in the global trial, the consistency of efficacy and safety in the Chinese subpopulation to the global population will be evaluated. Consistency of efficacy will be evaluated using the percentage of risk reduction preserved in the Chinese subpopulation from the empirical risk reduction based on the global primary efficacy analyses (based on point estimates). The planned sample size for the China Cohort is estimated to provide about 80% chance of that the observed point estimate in the Chinese subpopulation preserves at least approximately 50% of the empirical risk reduction based on the global primary efficacy analysis assuming the same hazard ratio used in the sample size and power calculation for the Global Cohort.

4.4 The Analysis Endpoints

4.4.1 Efficacy Endpoints

Efficacy endpoints are the same as described in section 3.4.1.

4.4.2 Safety Endpoints

Safety endpoints are the same as described in Section 3.4.2.

4.5 Analysis Populations

4.5.1 Efficacy Analysis Populations

The efficacy analysis population will include subjects based on the intention-to-treat (ITT) principle, i.e., subjects will be included in the treatment group to which they are randomized. For the China Cohort this population will include all Chinese subjects by their assigned treatment groups in this population.

4.5.2 Safety Analysis Populations

Safety analysis will be carried out in the All Subjects as Treated (ASaT) population which consists of all randomized subjects who received at least 1 dose of study treatment. The China Cohort will include all Chinese subjects in this population.

4.6 Statistical Methods

No formal testing of hypotheses is planned. Nominal p-values will be computed as noted below. However, the focus of analyses will be estimation of treatment effects, including confidence intervals and a comparison of these estimates between the China and global Cohorts.

4.6.1 Statistical Methods for Efficacy Analyses

Overall Survival (OS)

Analysis of OS is the same as that for the Global Cohort as applicable.

In detail, the Kaplan-Meier method will be used to estimate the survival curves. For the whole population, stratified log-rank will be used to assess the treatment difference and 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 same stratification factors used in the global Cohort will be used. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geographic region (Asia vs. Rest of World) will not be included in the stratified analysis for Chinese subgroup analysis.

Consistency in OS between the China and Global Cohorts will be evaluated.

In addition, supportive analyses on the entire ITT population will be provided with the data pooled the global Cohort (prior to data cutoff for the primary analysis) and Chinese subjects together. Accordingly, non-Chinese subjects will be censored at the data cut off for the primary analysis in the global study if subjects are still alive at that time.

Progression-Free Survival (PFS)

Analysis of PFS is the same as that for the global Cohort if applicable.

In detail, the Kaplan-Meier method will be used to estimate the survival curves. For the Global Cohort, stratified log-rank test will be used to assess the treatment difference and 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. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geography (Asia vs. Rest of World) will not be included in the stratified analysis for Chinese subgroup analysis.

Objective Response Rate (ORR)

Analysis of ORR is the same as that for the global Cohort if applicable.

Exploratory Analyses

Exploratory analyses are the same to that for the global study (if applicable).

4.6.2 Statistical Methods for Safety Analyses

Safety analyses are the same to that for the main study as described in Section 3.6.2.

4.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

They are the same to that for the global study as described in Section 3.6.4.

4.7 Interim Analysis

No interim analysis is planned.

4.8 Multiplicity

No multiplicity adjustment will be applied.

4.9 Sample Size and Power Calculations

After the sample size for the Global Cohort reaches approximately 600, global enrollment period will finish and the study will continue to randomize subjects in a 1:1 ratio into the Pembrolizumab arm and the SOC arm in China extension period until the sample size for the Chinese subpopulation reaches approximately 120. All Chinese subjects will be included in the China primary analysis. Whereas, the Chinese subjects enrolled in extension period, i.e., those Chinese subjects randomized after the global LPI, will not be included in the global primary analysis.

The extension study will complete after approximately 75 deaths have been observed between the two arms in the China Cohort and 8 months after last subject randomized assuming the underlying hazard ratio is 0.70. With 75 deaths and a true hazard ratio of 0.70, the extension study has >90% chance to observe a hazard ratio on OS <1 and ~80% chance to observe a point estimate that preserves approximately at least 50% of the empirical risk reduction from the global analysis in the Chinese subpopulation assuming the underlying hazard ratio is 0.70 respectively. The above calculations for the consistency evaluation are based on the same assumptions on the median OS and the true hazard ratio.

5. 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. 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.
4. Odell P.M., Anderson K.M., Kannel B.W. New models for predicting cardiovascular events. Journal of Clinical Epidemiology 1994; 47(6): 583-592.
5. Mehrotra D., Su S.C., Li X. An efficient alternative to the stratified Cox model analysis. Statist. Med. 2012, 31 1849–1856.
6. 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.
7. 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 2013.
8. De Castro, M., V.G. Cancho and J. Rodrigues, 2010. A hands-on approach for fitting long-term survival models under the GAMLSS framework. Comput. Methods Programs Biomed., 97: 168-177.
9. Lee, S.-H. 2007. On the versatility of the combination of the weighted log-rank statistics. Computational Statistics and Data Analysis 51: 6557–6564.
10. Karrison T.G, 2016. Versatile tests for comparing survival curves based on weighted log-rank statistics. The Stata Journal (2016).
11. Lan KKG, DeMets DL. Group sequential procedures: calendar versus information time: biostatics technical report. Madison (WI): University of Wisconsin Clinical Cancer Center (US); 1988 May. 16 p. Report No.:44.
12. 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.
13. The EORTC QLQ-C30 Manuals, Reference Values and Bibliography.
14. Blazeby JM, Conroy T, Hammerlid E, Fayers P, Sezer O, Koller M, et al. Clinical and psychometric validation of an EORTC questionnaire module, the EORTC QLQ-OES18, to assess quality of life in patients with oesophageal cancer. Eur J Cancer. 2003 Jul;39(10):1384-94.
15. EQ-5D-3L User Guide, Oct 2013.