Official Protocol Title:	A Randomized, Double-Blind, Phase III Study of Carboplatin-Paclitaxel/Nab-Paclitaxel Chemotherapy with or without Pembrolizumab (MK-3475) in First Line Metastatic Squamous Non-small Cell Lung Cancer Subjects (KEYNOTE-407)
NCT number:	NCT03875092
Document Date:	29-Jan-2018

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

Listing of Tables

1 INTRODUCTION

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

2 SUMMARY OF CHANGES

This sSAP aligns with the protocol amendment v04 for the China extension study with regard to statistical analysis plan.

3 ANALYTICAL AND METHODOLOGICAL DETAILS FOR GLOBAL STUDY

3.1 Statistical Analysis Plan Summary

This section contains a brief summary of the statistical analyses for global study of this trial. Full detail is provided in Sections 3.2-3.11.

Study Design Overview	A Phase III Study of Carboplatin-Paclitaxel/Nab-Paclitaxel Chemotherapy with or without Pembrolizumab (MK-3475) in First Line Metastatic Squamous Non-small Cell Lung Cancer Subjects (KEYNOTE-407)
Treatment Assignment	Subjects will be randomized in a 1:1 ratio to receive pembrolizumab or saline placebo in combination with carboplatin and a taxane (investigators choice of paclitaxel or nab-paclitaxel). Stratification factors are in Section 5.4 of the protocol. This is a randomized double-blinded study.
Analysis Populations	Efficacy: Intent to Treat (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoints	1. Progression-free Survival (PFS) per RECIST 1.1 assessed by a blinded independent central imaging vendor 2. Overall Survival (OS)
Statistical Methods for Key Efficacy Analyses	The dual primary hypotheses on PFS and OS will be evaluated by comparing pembrolizumab to saline placebo in combination with carboplatin and a taxane using a stratified Log-rank test. The hazard ratio will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. The stratified M&N method with sample size weights will be used for analysis of ORR.
Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. There are no Tier 1 safety parameters in this trial. All safety parameters are considered either 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. The between-treatment difference will be analyzed using the Miettinen and Nurminen method. In the primary safety comparison, subjects who crossover to pembrolizumab are censored at time of crossover (i.e., AEs occurring during treatment with pembrolizumab are excluded for control-arm subjects). An exploratory safety analysis will be conducted

	for the crossover population including all safety events starting from the date of the first dose of pembrolizumab.
Interim Analyses	There are four analyses planned for this study: three interim analyses and one final analysis. Results from the first three interim analyses will be reviewed by an external data monitoring committee. Details are provided in Section 3.7. □ Interim analysis (IA) 1 o Timing: To be performed after ~200 subjects have ~28 weeks of follow-up o Purpose: To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in ORR □ Interim analysis (IA) 2 o Timing: To be performed after a target number of PFS events (~332) is observed o Purpose: 1) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in PFS; 2) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in OS □ Interim analysis (IA) 3 o Timing: To be performed after a target number of PFS events (~415) is observed o Purpose: 1) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in PFS; 2) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in OS
Multiplicity	The study uses the graphical method of Maurer and Bretz [2] to control multiplicity for multiple hypotheses as well as interim analyses. According to this approach, study hypotheses may be tested more than once, and when a particular null hypothesis is rejected, the alpha allocated to that hypothesis can be reallocated to other hypothesis tests. The overall type I error is controlled at 0.025 (one-sided) for the hypothesis testing of ORR, PFS and OS. The pre-allocated alpha is 0.005, 0.01 and 0.01 for ORR, PFS and OS, respectively. ORR may be tested at 0.005 or at 0.025 (if both PFS and OS are positive, using the p-value from IA1). PFS may be tested at 0.01 or at 0.015 (if ORR is positive but OS not positive), or at 0.02 (if OS is positive but ORR not positive) or at 0.025 (if both OS and ORR are positive). OS may be tested at 0.01 or at 0.02 (if PFS is positive but ORR not positive) or 0.025 (if both PFS and ORR are positive). A Lan-DeMets O'Brien-Fleming approximation spending function will be used for the calculation of efficacy bounds for PFS and OS.
Sample Size and Power	The final analysis occurs after ~361 deaths are observed unless the trial is terminated early. With 361 deaths, the study has ~92% power for detecting a hazard ratio (HR) of 0.7 at 0.025 (one-sided), ~90% power for detecting a HR of 0.7 at 0.02 (one-sided) and ~85% power for detecting a HR of 0.7 at 0.01 (one-sided). The planned sample size is approximately 560 subjects assuming ~15.5 months of enrollment.

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.

This trial is double blinded with a crossover phase. At the time of documented progression, subjects will have treatment assignment unblinded and be able to continue therapy in the Crossover Phase (please refer to protocol section 2.1 Trial Design for details). In addition, independent central radiologist(s) will perform the central imaging review without knowledge of treatment assignment.

An external data monitoring committee (eDMC) will be convened to review accumulating safety to provide an opportunity to terminate the study early if there are concerns regarding safety. The eDMC will also review the unblinded efficacy results at the planned interim analyses. The eDMC responsibilities and review schedules will be outlined in the eDMC charter. The recommendation of the eDMC will be communicated to an executive oversight committee of the Sponsor. In the event of a recommendation to halt the trial early due to safety concerns, the Sponsor will communicate this to the appropriate regulatory agencies. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee may be unblinded to results at the treatment level in order to act on these recommendations.

A limited number of additional SPONSOR personnel may be unblinded, if required, in order to act on the recommendations of the eDMC. The extent to which individuals are unblinded to the results 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

3.3.1 Primary Objective(s) & Hypothesis(es)

In 1L subjects with metastatic squamous non-small cell lung cancer (NSCLC) receiving investigator's choice of standard of care chemotherapy (i.e. carboplatin and a taxane):

1. **Objective:** To evaluate progression free survival (PFS) per RECIST 1.1 as assessed by a central imaging vendor in subjects treated with Pembrolizumab compared to placebo.

   Hypothesis: Pembrolizumab prolongs PFS by RECIST 1.1 as assessed by a central imaging vendor compared to placebo.

2. **Objective:** To evaluate overall survival (OS) in subjects treated with Pembrolizumab compared to placebo.

   Hypothesis: Pembrolizumab prolongs OS compared to placebo.

3.3.2 Secondary Objective(s) & Hypothesis(es)

In 1L subjects with metastatic squamous non-small cell lung cancer (NSCLC) receiving investigator's choice of standard of care chemotherapy (i.e. carboplatin and a taxane):

1. **Objective:** To evaluate the objective response rate (ORR) and duration of response (DOR) per RECIST 1.1 as assessed by a central imaging vendor in subjects treated with Pembrolizumab compared to placebo.

   Hypothesis: Pembrolizumab improves ORR per RECIST 1.1 as assessed by a central imaging vendor compared to placebo.

2. **Objective:** To evaluate the safety and tolerability profile of Pembrolizumab.

3.3.3 Exploratory Objectives

In 1L subjects with metastatic squamous non-small cell lung cancer (NSCLC) receiving investigator's choice of standard of care chemotherapy (i.e. carboplatin and a taxane):

1) Objective: Evaluate pembrolizumab compared to placebo with respect to:

a. PFS per RECIST 1.1 as assessed by investigator review in the next line of therapy (PFS2).

b. PFS per irRECIST as assessed by site investigator.

c. ORR and Duration of Response (DOR) per irRECIST as assessed by site investigator.

d. PFS and ORR per RECIST 1.1 as assessed by central imaging vendor and OS by PD-L1 status (≥1% vs. <1%) and by taxane (investigators choice of paclitaxel or nab-paclitaxel).

2) To investigate the relationship between pembrolizumab treatment and biomarkers predicting response (e.g., PD-L2, genetic variation, serum sPD-L1) utilizing newly obtained or archival FFPE tumor tissue and blood, including serum and plasma.

3) To evaluate changes in health-related quality-of-life assessments from baseline in the overall study population and by PD-L1 expression level using the EORTC QLQ-C30 and EORTC QLQ-LC13.

4) To characterize utilities in subjects treated with pembrolizumab and chemotherapy compared to saline placebo and chemotherapy using the EuroQoL(EQ)-5D.

5) To characterize the pharmacokinetic characteristics of carboplatin, paclitaxel/nab-paclitaxel treatment, and pembrolizumab.

6) To identify molecular (genomic, metabolic and/or proteomic) biomarkers that may be indicative of clinical response/resistance, safety, pharmacodynamic activity, and/or the mechanism of action of pembrolizumab and other treatments.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Dual Primary

Progression-Free Survival – RECIST 1.1 assessed by a blinded independent central imaging vendor

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded independent central imaging vendor review or death due to any cause, whichever occurs first. See Section 3.6.1 for the censoring rules.

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 analysis will be censored at the date of the last known contact.

Secondary

Objective Response Rate – RECIST 1.1 assessed by a blinded independent central imaging vendor

Objective response rate (ORR) is defined as the proportion of the subjects who have a confirmed complete response (CR) or partial response (PR). Responses are based on confirmed assessments by the blinded independent central imaging vendor per RECIST 1.1.

Duration of Response (DOR) - RECIST 1.1 assessed by a blinded independent central imaging vendor

For subjects who demonstrated CR or PR, duration of response (DOR) is defined as the time from first documented evidence of CR or PR until disease progression or death. Response duration for subjects who have not progressed or died at the time of analysis will be censored at the date of their last tumor assessment. Response duration will be calculated for RECIST 1.1 based on blinded independent radiologists' review.

3.4.2 Safety Endpoints

Safety measurements are described in Protocol Section 4.2.3.4.

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.

Approximately 200 randomized subjects are planned to be included in the first interim analysis. Based on actual enrollment, 204 subjects who were randomized on or prior to April 11 2017 will be included in the ITT population in the first interim analysis. Seventy subjects who failed screening on or prior to Apr 11 2017 in addition to the 204 randomized subjects will be included in the Screening Population.

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 subject's randomized treatment group and a narrative will be provided for any events that occur during the cycle for which the subject was incorrectly dosed.

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

For the first interim analysis, among the ITT population of 204 subjects, 203 of them received at least one dose of study treatment and thus will be included in the ASaT population.

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.

All statistical tests, unless otherwise specified, will be stratified for treatment and stratification factors.

3.6.1.1 Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test (based on the stratification factors defined in Protocol Section 5.4). 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 same stratification factors used for randomization will be applied to both the stratified log-rank test and the stratified Cox model. For the first interim analysis, due to the small sample size in the stratum with TPS < 1% and nab-paclitaxel and East Asia, this stratum will be combined with the stratum with TPS < 1% and paclitaxel and East Asia in stratified analyses.

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 a blinded independent central imaging vendor, 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 a blinded independent central imaging vendor, 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 new anticancer treatment, whichever occurs later, to be a PD event for subjects without documented PD or death. The censoring rules for primary and sensitivity analyses are summarized in Table 1. In case there is an imbalance between the treatment groups on disease assessment schedules or censoring patterns, we will also perform the following two additional PFS sensitivity analyses: 1) a PFS analysis using time to scheduled tumor assessment visit from randomization as opposed to actual tumor assessment time; 2) Finkelstein's likelihood-based score test for interval-censored data [4] which modifies the Cox proportional hazard model for interval censored data, will be used as a supportive analysis for the PFS endpoint. The interval will be constructed so that the left endpoint is the date of the last disease assessment without documented PD and the right endpoint is the date of documented PD or death, whichever occurs earlier. In case the proportional hazards assumption doesn't hold, Fleming and Harrington's weighted log-rank test, Restricted Mean Survival Time (RMST) method or other methods, as appropriate, may be conducted.

Table 1 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 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 after ≥2 missed disease assessments	Progressed at date of documented PD or death	Censored at last disease assessment prior to the ≥2 missed disease assessment	Progressed at date of documented PD or death

3.6.1.2 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (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 for randomization will be applied to both the stratified log-rank test and the stratified Cox model. For the first interim analysis, due to the small sample size in the stratum with TPS < 1% and nab-paclitaxel and East Asia, this stratum will be combined with the stratum with TPS < 1% and paclitaxel and East Asia in stratified analyses. The Restricted Mean Survival Time (RMST) method may be conducted for OS to account for the possible non-proportional hazards effect and to estimate the absolute benefit of experimental treatment. A cure rate model may be applied to estimate the long-term effect.

Since subjects in the control arm are allowed to switch to the pembrolizumab treatment after progressive disease, adjustment for the effect of crossover on OS may be performed based on recognized methods, e.g., a two-stage method or the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis [3], based on an examination of the appropriateness of the data to the assumptions required by the methods.

3.6.1.3 Objective Response Rate (ORR) and Duration of Response (DOR)

The stratified Miettinen and Nurminen's method [5] with weights proportional to the stratum size will be used for comparison of the ORR between the treatment arms. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen method with strata weighting by sample size with a single treatment covariate will be reported. The stratification factors used for randomization (See Protocol Section 5.4) will be applied to the analysis. For the first interim analysis, due to the small sample size in the stratum with TPS < 1% and nab-paclitaxel and East Asia, this stratum will be combined with the stratum with TPS < 1% and paclitaxel and East Asia in stratified analyses.

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

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

Table 2 Censoring Rules for DOR

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new anti-cancer therapy initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anti-cancer therapy initiated	Last adequate disease assessment before new anti-cancer therapy initiated	Censor (non-event)
Death or progression 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.		

3.6.1.4 Analysis Strategy for Key Efficacy Endpoints

Table 3 summarizes the primary analysis approach for primary and key secondary efficacy endpoints. Sensitivity analysis methods are described above for each endpoint as applicable.

The strategy to address multiplicity issues with regard to multiple efficacy endpoints, 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
Dual Primary Endpoints			
PFS per RECIST 1.1 by blinded independent central imaging vendor	Test: Stratified Log-rank test to assess the treatment difference Estimation: Stratified Cox model with Efron's tie handling method to assess the magnitude of treatment difference	ITT	• Primary censoring rule • Sensitivity analysis 1 • Sensitivity analysis 2
OS	Test: Stratified Log-rank test to assess the treatment difference Estimation: Stratified Cox model with Efron's tie handling method to assess the magnitude of treatment difference	ITT	Model based (censored at the last date the subject was known to be alive)
Key Secondary Endpoints			
ORR per RECIST 1.1 by blinded independent central imaging vendor	Test and Estimation: Stratified M&N method with sample size weights††	ITT	Subjects without assessments are considered non-responders and conservatively included in denominator
DOR per RECIST 1.1 by blinded independent central imaging vendor	Descriptive statistics for range and Kaplan-Meier estimate of median	Patients in ITT population with an objective response	

† Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization (Section 5.4) will be applied to the analysis.

†† Miettinen and Nurminen method

3.6.1.5 Exploratory Analyses

An exploratory analysis of PFS2, defined as the time from randomization to subsequent disease progression after initiation of new anti-cancer therapy, or death from any cause, whichever first, may be carried out. Patients alive and for whom a disease progression following initiation of new anti-cancer treatment has not been observed will be censored at the last time the subject was known to be alive and without disease progression

3.6.2 Statistical Methods for Safety Analyses

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

Adverse Events

Adverse events (AEs) will be coded using the standard MedDRA and grouped system organ class. Adverse events (AEs) will be graded by the investigator according to the Common Terminology Criteria for Adverse Events (CTCAE), version 4.0.

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 will be subject to inferential testing for statistical significance with p-values and 95% confidence intervals provided for between-group comparisons. For this protocol, there are no Tier 1 AEs. 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) 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 values and vital signs, and ECG parameters 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.

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, an analysis of Grade 3-5 AEs will be based on the time to first event using the time-to-event analysis methods (i.e., the log-rank test will be used for testing the time to AEs, and the Cox model with Efron's tie handling method will be used for estimating the hazard ratio and its 95% confidence interval). For other AEs with potentially differential follow-up time, such analysis may also be explored.

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 unstratified Miettinen and Nurminen method [5], an unconditional, asymptotic method.

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 Grade 3-5 AE		X	X
	Any Serious AE		X	X
	Onset of First Grade 3-5 AE		X	X
	Any Drug-Related AE		X	X
	Any Serious and Drug-Related AE		X	X
	Any Grade3-5 and Drug-Related AE		X	X
	Dose Modification Due to AE		X	X
	Discontinuation Due to AE		X	X
	Death		X	X
	Specific AEs, SOCs (including ≥4 of subjects in one of the treatment groups)		X	X
Tier 3	Specific AEs, SOCs (incidence <4 of subjects in all of the treatment groups)			X
	Change from Baseline Results (Labs, ECGs, Vital Signs)			X
There are no Tier 1 AEs pre-specified in this protocol.				

3.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

The comparability of the treatment groups for each relevant baseline 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 randomized and the primary reason for discontinuation will be displayed. Demographic variables (such as age) and baseline characteristics will be summarized by treatment either by descriptive statistics or categorical tables.

3.7 Interim Analysis

There are three planned interim analyses (IA) in addition to the final analysis for this study. Details on the boundaries for establishing statistical significance with regard to efficacy are discussed further in Section 3.8. The trial will continue until the number of deaths (See Section

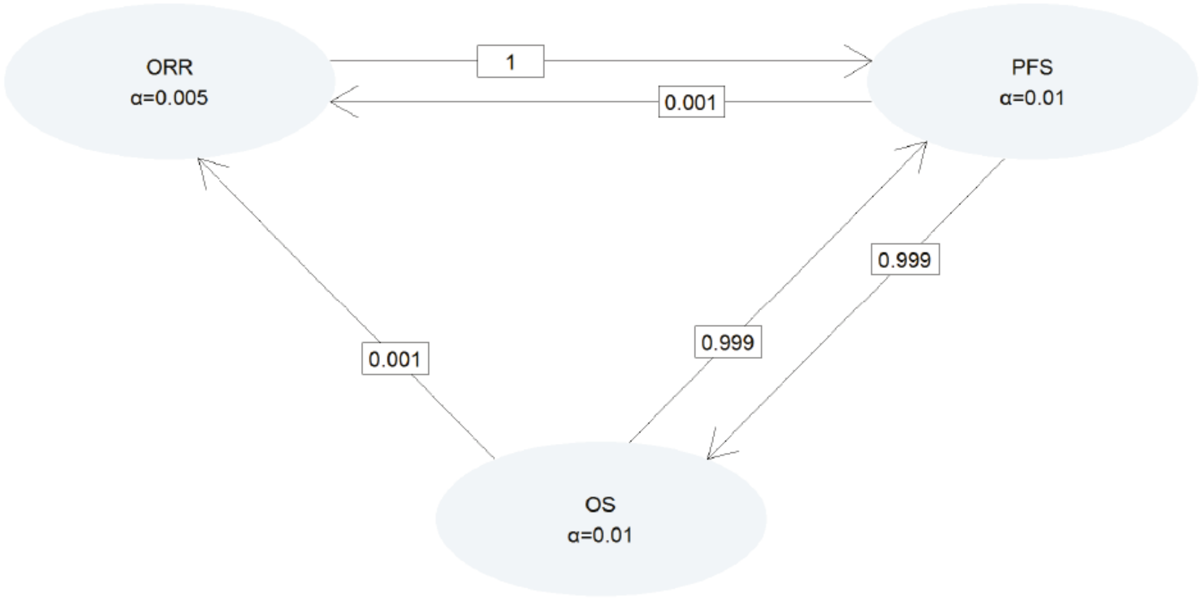
3.9) is approximately equal to the targeted number for the final analysis, irrespective of the outcome from the interim analyses. The analyses planned, endpoints evaluated, and drivers of timing are summarized in Table 5.

Table 5 Summary of Interim and Final Analyses Strategy

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	ORR	~ 200 subjects are followed for ~ 28 weeks so that each patient has at least 4 tumor assessments	~ 15 months	• Demonstrate ORR superiority
IA2	PFS OS	~ 332 PFS events have been observed.	~ 20 months	• Demonstrate PFS superiority • Demonstrate OS superiority
IA3	PFS OS	~ 415 PFS events have been observed	~ 25 months	• Demonstrate PFS superiority • Demonstrate OS superiority
Final Analysis	OS	~ 361 deaths have occurred.	~ 31 months	• Demonstrate OS superiority

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [2] to control multiplicity for multiple hypotheses as well as interim analyses. According to this approach, study hypotheses may be tested more than once, and when a particular null hypothesis is rejected, the alpha allocated to that hypothesis can be reallocated to other hypothesis tests. Figure 1 shows the initial one-sided alpha allocation for each hypothesis in the ellipse representing the hypothesis. The weights for reallocation from each hypothesis to the others are represented in the boxes on the lines connecting hypotheses.



Note: If both PFS and OS null hypotheses are rejected, the reallocation strategy allows re-testing of ORR at alpha=0.025 based on the p-value at IA1

ORR=objective response rate; OS=overall survival; PFS=progression-free survival

Figure 1 Type I Error Reallocation Strategy Following Closed Testing Principle

3.8.1 Objective Response Rate

The study allocates alpha=0.005, one-sided, to test ORR, and ORR is tested only at the first interim analysis (IA1). However, if the test does not reach statistical significance at IA1, the p-value from IA1 can be compared to an updated alpha-level if the null hypotheses for both PFS and OS are rejected at a later time. Power at the possible alpha-levels as well as the approximate treatment difference required to reach the bound (ORR difference) are shown in Table 6, assuming underlying 25% and 50% response rates in the control and experimental groups, respectively.

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

Alpha	ORR difference	Power
0.005	~ 0.18	0.84
0.025	~ 0.13	0.94

3.8.2 Progression-free Survival

The initial alpha-level for testing PFS is 0.01. If the null hypothesis for ORR is rejected, Figure 1 shows that its alpha=0.005 is fully reallocated to PFS hypothesis testing. If the null hypothesis for OS is rejected, then alpha=0.01 is essentially fully reallocated to PFS hypothesis testing. Thus, the PFS null hypothesis may be tested at alpha=0.01, alpha=0.015 (if the ORR null hypothesis is rejected but not the OS null hypothesis), alpha=0.02 (if the OS null hypothesis is rejected but not the ORR null hypothesis), or alpha=0.025 (if both the ORR and OS null hypotheses are rejected). Table 7 shows the boundary properties for each of these alpha-levels for the interim analyses, which were derived using a Lan-DeMets O'Brien-Fleming spending function. Note that the final row indicates the total power to reject the null hypothesis for PFS at each alpha-level. If the actual number of events at the PFS analyses differ from those specified in the table, the bounds will be adjusted using the Lan-DeMets O'Brien-Fleming spending function accordingly. Also note that if the OS null hypothesis is rejected at an interim or final analysis, each PFS interim and final analysis test may be compared to its updated bounds considering the alpha reallocation from the OS hypothesis.

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

Analysis	Value	α=0.01	α=0.015	α=0.02	α=0.025
IA 2: 80%* N: 560 Events: 332 Month: 20	Z	-2.6539	-2.4817	-2.3536	-2.2504
	p (1-sided)§	0.004	0.0065	0.0093	0.0122
	HR at bound%	0.7473	0.7616	0.7723	0.7811
	P(Cross) if HR=1†	0.004	0.0065	0.0093	0.0122
	P(Cross) if HR=0.7#	0.7243	0.7787	0.8148	0.8411
IA 3: 100%* N: 560 Events: 415 Month: 25	Z	-2.3737	-2.2244	-2.1138	-2.025
	p (1-sided)§	0.0088	0.0131	0.0173	0.0214
	HR at bound%	0.7921	0.8038	0.8126	0.8197
	P(Cross) if HR=1†	0.01	0.015	0.02	0.025
	P(Cross) if HR=0.7#	0.9	0.9243	0.9392	0.9494

*Percentage of expected number of events at final analysis required at interim analysis
§p (1-sided) is the nominal alpha for testing.
%HR at bound is the approximate HR required to reach an efficacy bound
†P(Cross if HR=1) is the cumulative probability of crossing a bound under the null hypothesis
#P(Cross if HR=0.7) is the cumulative probability of crossing a bound under the alternative hypothesis

3.8.3 Overall Survival

The OS hypothesis may be tested at alpha=0.01 (initially allocated alpha), alpha=0.02 (if the PFS but not the ORR null hypothesis is rejected), or alpha=0.025 (if both the ORR and PFS null hypotheses are rejected). Table 8 demonstrates the bounds and boundary properties for OS hypothesis testing derived using a Lan-DeMets O'Brien-Fleming spending function. If the actual number of OS events at the interim and final analyses differs from those specified in the table, the bounds will be adjusted using the Lan-DeMets O'Brien-Fleming spending function accordingly.

Table 8 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	**Value**	**α=0.01**	**α=0.02**	**α=0.025**
IA 2: 59%* N: 560 Events: 212 Month: 20	Z	-3.1648	-2.8202	-2.702
	p (1-sided) [§]	0.0008	0.0024	0.0034
	HR at bound[%]	0.6474	0.6788	0.6899
	P(Cross) if HR=1[†]	0.0008	0.0024	0.0034
	P(Cross) if HR=0.7 [#]	0.2849	0.4115	0.458
IA 3: 79%* N: 560 Events: 286 Month: 25	Z	-2.6914	-2.3992	-2.2995
	p (1-sided) [§]	0.0036	0.0082	0.0107
	HR at bound [%]	0.7274	0.7530	0.7619
	P(Cross) if HR=1[†]	0.0038	0.009	0.0117
	P(Cross) if HR=0.7[#]	0.6312	0.7362	0.7684
Final N: 560 Events: 361 Month: 31	Z	-2.3742	-2.116	-2.028
	p (1-sided) [§]	0.0088	0.0172	0.0213
	HR at bound [%]	0.7789	0.8003	0.8078
	P(Cross) if HR=1[†]	0.01	0.02	0.025
	P(Cross) if HR=0.7 [#]	0.85	0.9034	0.9181

* Percentage of expected number of events at final analysis required at interim analysis

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

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

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

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

3.9 Sample Size and Power Calculations

With ~200 subjects, the study has ~ 84% power for detecting a 25% difference in ORR (50% vs 25%) or ~ 97% power for detecting a 30% difference in ORR (50% vs. 20%) at initially assigned 0.005 (one-sided) significance level. The study has ~ 94% power for detecting a 25% difference in ORR (50% vs 25%) or ~ 99% power for detecting a 30% difference in ORR (50% vs 20%) at 0.025 (one-sided) significance level.

With 415 PFS events, the study has ~ 90% power for detecting a HR of 0.7 at initially assigned 0.01 (one-sided) significance level, ~ 92% power for detecting a HR of 0.7 at 0.015 (one-sided) significance level, ~ 94% power for detecting a HR of 0.7 at 0.02 (one-sided) significance level, and ~ 95% power for detecting a HR of 0.7 at 0.025 (one-sided) significance level.

With 361 deaths, the study has ~ 85% power for detecting a hazard ratio (HR) of 0.7 at 0.01 (one-sided) significance level, ~ 90% power for detecting a HR of 0.7 at 0.02 (one-sided) significance level, and ~ 92% power for detecting a HR of 0.7 at 0.025 (one-sided) significance level.

The planned sample size is approximately 560 subjects assuming: (1) the enrollment period is 15.5 months and the ramp-up period of enrollment is 7 months; (2) median PFS is 6 months in the control group and the true hazard ratio is 0.7; (3) median OS is 12 months in the control group and the true hazard ratio is 0.7; (4) the annual dropout rate is 3% for PFS and 1% for OS; (5) the number of events and alpha levels of interim analyses and final analysis are as specified in Section 3.7 and Section 3.8.

3.10 Subgroup Analyses and Effect of Baseline Factors

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

- Age category (< 65 vs. ≥ 65 years, and < 65 vs. 65-74 vs. 75-84 vs. ≥ 85 years)
- Sex (female vs. male)
- Race (white vs. non-white)
- ECOG status (0 vs. 1)
- Geographic region of enrolling site (East Asia vs. Non-East Asia; US vs. Ex-US; EU vs. Non-EU)
- Smoking status (never vs. former/current)
- Brain metastasis status at baseline (yes vs. no)
- PD-L1 status (TPS < 1% vs. TPS ≥ 1%, TPS < 50% vs. TPS ≥ 50%, TPS < 1% vs. TPS 1 – 49% vs .TPS ≥ 50%)
- Taxane chemotherapy (paclitaxel vs. nab-paclitaxel)

Additional subgroup analyses such as China vs. non-China may be conducted per China local regulatory needs. 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. For the first interim analysis, the subgroup analysis for ORR will be conducted using unstratified Miettinen and Nurminen method.

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

3.11 Extent of Exposure

The extent of exposure will be summarized as duration of treatment in cycles.

4 STATISTICAL ANALYSIS PLAN FOR EXTENSION

4.1 Introduction

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

After the enrollment for the global study 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 for the overall Chinese subpopulation reaches approximately 120.

After the cut-off date for the primary analyses of the global study (including interim and final analyses), all Chinese subjects, including subjects enrolled in the global study and the extension study , will continue their randomized treatment and continue to be followed up for PFS and OS events for China registration purpose. The extension study will be completed after target number of deaths has been observed between the two arms in the Chinese subpopulation. The expected timing of the analysis for the subpopulation is around 3.3 years from the time when the first subject from China is enrolled in the global study. However, if the target number of PFS events or deaths in the Chinese subpopulation is reached before an IA for the global study, the corresponding analysis for Chinese subpopulation will occur at the same time as the global IA or the final analysis (global study).

4.2 Responsibility for Analyses/In-House Blinding

The trial is double-blinded, analyses or summaries generated by randomized treatment assignment and actual treatment received will be limited and documented. Subjects randomized in the extension study will not be included in the database lock and primary analysis for the global study.

For all Chinese subjects, including subjects randomized in the global study and the extension study, patient level treatment randomization information will be blinded to the statistician(s)/programmer(s) responsible for the analysis of China extension study until the extension study data base lock is achieved. The extent to which individuals are unblinded to the results will be limited. Blinded and unblinded members will be clearly documented with blinding status along with time information.

4.3 Hypotheses/Estimation

No hypothesis testing is planned for the China extension study.

After succeeding in the global study, 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 from the global primary efficacy analyses (based on point estimates). Sample size is designed to provide about 80% chance of observing the point estimate of Chinese subpopulation preserves ≥ approximately 50% of empirical risk reduction from the global primary efficacy analysis assuming the same hazard ratio used in the sample size and power calculation for the global study.

4.4 The Analysis Endpoints

4.4.1 Efficacy Endpoints

Dual Primary

Overall survival (OS) is defined as the time from randomization to death due to any cause, the same as described in Section 3.4.1.

Progression-free survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded independent radiologists' assessment or death due to any cause, whichever occurs first, the same as described in Section 3.4.1.

Secondary

Objective response rate (ORR) based upon blinded independent central imaging vendor's assessed RECIST 1.1 as described in Section 3.4.1.

Duration of response (DOR) based upon blinded independent central imaging vendor's assessed RECIST 1.1 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

Efficacy analysis will be carried out in the entire intention-to-treat (ITT) population. This population will include all subjects who are randomized in the global study and all subjects who are randomized in the extension study. Chinese subpopulation will include all Chinese subjects in this population.

4.5.2 Safety Analysis Populations

Safety analysis will be carried out in the entire All Subjects as Treated (ASaT) population, i.e., all randomized subjects (in the global study and extension study) who received at least 1 dose of study treatment. Chinese subpopulation will include all Chinese subjects in this population.

4.6 Statistical Methods

Regarding the analysis for extension, no hypothesis testing is planned. There is no plan of interim analysis. No multiplicity adjustment will be applied to the analysis for extension.

4.6.1 Statistical Methods for Efficacy Analyses

4.6.1.1 Overall Survival (OS)

Analysis of OS for extension is the same to that for the global study if 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 study will be used. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for the Chinese subgroup analysis.

As an exploratory analysis, recognized methods, e.g., the Rank Preserving Structural Failure Time (RPSFT) model, two-stage method, etc., will be used to adjust for the effect of crossover on OS based upon the appropriateness of the data to the assumption required by the methods. The RPSFT model provides a randomization-based estimate of treatment effect (RBEE) corrected for the bias induced by crossover. The 95% confidence intervals of the hazard ratio for OS after adjustment of the effect of crossover will be provided. To further account for the possible confounding effect, a sensitivity analysis of OS that censors subjects at the time of initiation of new therapy will be performed and an OS analysis that treats initiation of new therapy as a time-dependent binary covariate will also be conducted.

Consistency of efficacy will be evaluated using the percentage of risk reduction preserved in the Chinese subpopulation from the empirical risk reduction from the global primary efficacy analyses (based on point estimates). Sample size is designed to provide about 80% chance of observing the point estimate of Chinese subpopulation preserves ≥ approximately 50% of empirical risk reduction from the global primary efficacy analysis assuming the same hazard ratio used in the sample size and power calculation for the global study.

In addition, supportive analyses on the entire ITT population will be provided with the data pooling global study (prior to data cutoff for the primary analysis) and China extension study together. Accordingly, non-Chinese subjects will be censored at last known alive date (this can be cutoff date if some assessment happens to be on that day or there's assessment beyond the cutoff date) which is consistent with the primary analysis in the global study if subjects are still alive at primary analysis time for global study. The primary analysis for OS will be conducted in the Chinese subpopulation when approximately 75 OS events have been collected.

4.6.1.2 Progression-Free Survival (PFS)

Analysis of PFS for extension is the same to that for the global study if applicable.

In detail, the Kaplan-Meier method will be used to estimate the survival curves. For the whole population, 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. The same stratification factors used in the global study will be used. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for Chinese subgroup analysis. In case the proportional hazards assumption doesn't hold, Fleming and Harrington's weighted log-rank test, Restricted Mean Survival Time (RMST) method or other methods, as appropriate, may be conducted.

Consistency in PFS will be evaluated similarly as that in OS. The primary analysis for PFS will be conducted in the Chinese subpopulation when approximately 75 PFS events have been collected.

4.6.1.3 Objective Response Rate (ORR) and Duration of Response (DOR)

Stratified Miettinen and Nurminen's method with weights proportional to the stratum size will be used for comparison of the ORR between the treatment arms. A 95% CI for the difference in response rates between the pembrolizumab arm and the control arm will be provided. The same stratification factors used in the global study will be used. For the Chinese subpopulation analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for Chinese subgroup analysis.

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

4.6.1.4 Exploratory Analyses

Exploratory Analyses for extension is the same to that for the global study (if applicable).

4.6.2 Statistical Methods for Safety Analyses

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

4.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

They are the same for extension to that for the global study as described in Section 3.10.

4.7 Interim Analysis & Final analysis

The primary analysis for PFS will be conducted in the Chinese subpopulation when approximately 75 PFS events have been collected. OS will also be analyzed.

The primary analysis for OS will be conducted in the Chinese subpopulation when approximately 75 OS events have been collected.

4.8 Multiplicity

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

4.9 Sample Size and Power Calculations

After the enrollment of global study completes, the extension study will continue to randomize subjects in a 1:1 ratio into the pembrolizumab arm and the placebo arm in China until the sample size for the overall Chinese subjects (including those enrolled in the global study) reaches approximately 120. The extension study population, i.e., those Chinese subjects randomized after the close of enrollment for the global study, 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 Chinese subpopulation assuming the underlying hazard ratio for OS 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 50% of the empirical risk reduction from the global analysis in the Chinese subpopulation assuming the underlying hazard ratio is 0.70. The same consideration applies to PFS.

The above calculations for the consistency evaluation in PFS and OS are based on the same assumptions on the corresponding median OS/PFS and the true hazard ratio respectively.

4.10 Subgroup Analyses and Effect of Baseline Factors

All subgroup analysis defined in Section 3.10 will be repeated for the entire population and Chinese subpopulation if applicable. In addition, results for Chinese subpopulation vs. non-Chinese subpopulation will be provided. Country-specific analysis may also be conducted per local regulatory requirement.

5 REFERENCES

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

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

3. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Commun Stat-Theor M 1991;20(8): 2609-31.

4. Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics 1986;42:845-54.

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

Revision History

Date	Summary of Change
09JUN2017	Original Document
23OCT2017	Version 02
15NOV2017	Version 03
29JAN2018	Version 04

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

Listing of Tables

1 INTRODUCTION

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

2 SUMMARY OF CHANGES

This sSAP aligns with the protocol amendment v04 for the China extension study with regard to statistical analysis plan.

3 ANALYTICAL AND METHODOLOGICAL DETAILS FOR GLOBAL STUDY

3.1 Statistical Analysis Plan Summary

This section contains a brief summary of the statistical analyses for global study of this trial. Full detail is provided in Sections 3.2-3.11.

Study Design Overview	A Phase III Study of Carboplatin-Paclitaxel/Nab-Paclitaxel Chemotherapy with or without Pembrolizumab (MK-3475) in First Line Metastatic Squamous Non-small Cell Lung Cancer Subjects (KEYNOTE-407)
Treatment Assignment	Subjects will be randomized in a 1:1 ratio to receive pembrolizumab or saline placebo in combination with carboplatin and a taxane (investigators choice of paclitaxel or nab-paclitaxel). Stratification factors are in Section 5.4 of the protocol. This is a randomized double-blinded study.
Analysis Populations	Efficacy: Intent to Treat (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoints	1. Progression-free Survival (PFS) per RECIST 1.1 assessed by a blinded independent central imaging vendor 2. Overall Survival (OS)
Statistical Methods for Key Efficacy Analyses	The dual primary hypotheses on PFS and OS will be evaluated by comparing pembrolizumab to saline placebo in combination with carboplatin and a taxane using a stratified Log-rank test. The hazard ratio will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. The stratified M&N method with sample size weights will be used for analysis of ORR.
Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. There are no Tier 1 safety parameters in this trial. All safety parameters are considered either 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. The between-treatment difference will be analyzed using the Miettinen and Nurminen method. In the primary safety comparison, subjects who crossover to pembrolizumab are censored at time of crossover (i.e., AEs occurring during treatment with pembrolizumab are excluded for control-arm subjects). An exploratory safety analysis will be conducted

	for the crossover population including all safety events starting from the date of the first dose of pembrolizumab.
Interim Analyses	There are four analyses planned for this study: three interim analyses and one final analysis. Results from the first three interim analyses will be reviewed by an external data monitoring committee. Details are provided in Section 3.7. □ Interim analysis (IA) 1 o Timing: To be performed after ~200 subjects have ~28 weeks of follow-up o Purpose: To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in ORR □ Interim analysis (IA) 2 o Timing: To be performed after a target number of PFS events (~332) is observed o Purpose: 1) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in PFS; 2) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in OS □ Interim analysis (IA) 3 o Timing: To be performed after a target number of PFS events (~415) is observed o Purpose: 1) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in PFS; 2) To demonstrate superiority of pembrolizumab in combination with carboplatin and a taxane in OS
Multiplicity	The study uses the graphical method of Maurer and Bretz [2] to control multiplicity for multiple hypotheses as well as interim analyses. According to this approach, study hypotheses may be tested more than once, and when a particular null hypothesis is rejected, the alpha allocated to that hypothesis can be reallocated to other hypothesis tests. The overall type I error is controlled at 0.025 (one-sided) for the hypothesis testing of ORR, PFS and OS. The pre-allocated alpha is 0.005, 0.01 and 0.01 for ORR, PFS and OS, respectively. ORR may be tested at 0.005 or at 0.025 (if both PFS and OS are positive, using the p-value from IA1). PFS may be tested at 0.01 or at 0.015 (if ORR is positive but OS not positive), or at 0.02 (if OS is positive but ORR not positive) or at 0.025 (if both OS and ORR are positive). OS may be tested at 0.01 or at 0.02 (if PFS is positive but ORR not positive) or 0.025 (if both PFS and ORR are positive). A Lan-DeMets O'Brien-Fleming approximation spending function will be used for the calculation of efficacy bounds for PFS and OS.
Sample Size and Power	The final analysis occurs after ~361 deaths are observed unless the trial is terminated early. With 361 deaths, the study has ~92% power for detecting a hazard ratio (HR) of 0.7 at 0.025 (one-sided), ~90% power for detecting a HR of 0.7 at 0.02 (one-sided) and ~85% power for detecting a HR of 0.7 at 0.01 (one-sided). The planned sample size is approximately 560 subjects assuming ~15.5 months of enrollment.

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.

This trial is double blinded with a crossover phase. At the time of documented progression, subjects will have treatment assignment unblinded and be able to continue therapy in the Crossover Phase (please refer to protocol section 2.1 Trial Design for details). In addition, independent central radiologist(s) will perform the central imaging review without knowledge of treatment assignment.

An external data monitoring committee (eDMC) will be convened to review accumulating safety to provide an opportunity to terminate the study early if there are concerns regarding safety. The eDMC will also review the unblinded efficacy results at the planned interim analyses. The eDMC responsibilities and review schedules will be outlined in the eDMC charter. The recommendation of the eDMC will be communicated to an executive oversight committee of the Sponsor. In the event of a recommendation to halt the trial early due to safety concerns, the Sponsor will communicate this to the appropriate regulatory agencies. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee may be unblinded to results at the treatment level in order to act on these recommendations.

A limited number of additional SPONSOR personnel may be unblinded, if required, in order to act on the recommendations of the eDMC. The extent to which individuals are unblinded to the results 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

3.3.1 Primary Objective(s) & Hypothesis(es)

In 1L subjects with metastatic squamous non-small cell lung cancer (NSCLC) receiving investigator's choice of standard of care chemotherapy (i.e. carboplatin and a taxane):

1. **Objective:** To evaluate progression free survival (PFS) per RECIST 1.1 as assessed by a central imaging vendor in subjects treated with Pembrolizumab compared to placebo.

   Hypothesis: Pembrolizumab prolongs PFS by RECIST 1.1 as assessed by a central imaging vendor compared to placebo.

2. **Objective:** To evaluate overall survival (OS) in subjects treated with Pembrolizumab compared to placebo.

   Hypothesis: Pembrolizumab prolongs OS compared to placebo.

3.3.2 Secondary Objective(s) & Hypothesis(es)

In 1L subjects with metastatic squamous non-small cell lung cancer (NSCLC) receiving investigator's choice of standard of care chemotherapy (i.e. carboplatin and a taxane):

1. **Objective:** To evaluate the objective response rate (ORR) and duration of response (DOR) per RECIST 1.1 as assessed by a central imaging vendor in subjects treated with Pembrolizumab compared to placebo.

   Hypothesis: Pembrolizumab improves ORR per RECIST 1.1 as assessed by a central imaging vendor compared to placebo.

2. **Objective:** To evaluate the safety and tolerability profile of Pembrolizumab.

3.3.3 Exploratory Objectives

In 1L subjects with metastatic squamous non-small cell lung cancer (NSCLC) receiving investigator's choice of standard of care chemotherapy (i.e. carboplatin and a taxane):

1) Objective: Evaluate pembrolizumab compared to placebo with respect to:

a. PFS per RECIST 1.1 as assessed by investigator review in the next line of therapy (PFS2).

b. PFS per irRECIST as assessed by site investigator.

c. ORR and Duration of Response (DOR) per irRECIST as assessed by site investigator.

d. PFS and ORR per RECIST 1.1 as assessed by central imaging vendor and OS by PD-L1 status (≥1% vs. <1%) and by taxane (investigators choice of paclitaxel or nab-paclitaxel).

2) To investigate the relationship between pembrolizumab treatment and biomarkers predicting response (e.g., PD-L2, genetic variation, serum sPD-L1) utilizing newly obtained or archival FFPE tumor tissue and blood, including serum and plasma.

3) To evaluate changes in health-related quality-of-life assessments from baseline in the overall study population and by PD-L1 expression level using the EORTC QLQ-C30 and EORTC QLQ-LC13.

4) To characterize utilities in subjects treated with pembrolizumab and chemotherapy compared to saline placebo and chemotherapy using the EuroQoL(EQ)-5D.

5) To characterize the pharmacokinetic characteristics of carboplatin, paclitaxel/nab-paclitaxel treatment, and pembrolizumab.

6) To identify molecular (genomic, metabolic and/or proteomic) biomarkers that may be indicative of clinical response/resistance, safety, pharmacodynamic activity, and/or the mechanism of action of pembrolizumab and other treatments.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Dual Primary

Progression-Free Survival – RECIST 1.1 assessed by a blinded independent central imaging vendor

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded independent central imaging vendor review or death due to any cause, whichever occurs first. See Section 3.6.1 for the censoring rules.

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 analysis will be censored at the date of the last known contact.

Secondary

Objective Response Rate – RECIST 1.1 assessed by a blinded independent central imaging vendor

Objective response rate (ORR) is defined as the proportion of the subjects who have a confirmed complete response (CR) or partial response (PR). Responses are based on confirmed assessments by the blinded independent central imaging vendor per RECIST 1.1.

Duration of Response (DOR) - RECIST 1.1 assessed by a blinded independent central imaging vendor

For subjects who demonstrated CR or PR, duration of response (DOR) is defined as the time from first documented evidence of CR or PR until disease progression or death. Response duration for subjects who have not progressed or died at the time of analysis will be censored at the date of their last tumor assessment. Response duration will be calculated for RECIST 1.1 based on blinded independent radiologists' review.

3.4.2 Safety Endpoints

Safety measurements are described in Protocol Section 4.2.3.4.

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.

Approximately 200 randomized subjects are planned to be included in the first interim analysis. Based on actual enrollment, 204 subjects who were randomized on or prior to April 11 2017 will be included in the ITT population in the first interim analysis. Seventy subjects who failed screening on or prior to Apr 11 2017 in addition to the 204 randomized subjects will be included in the Screening Population.

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 subject’s randomized treatment group and a narrative will be provided for any events that occur during the cycle for which the subject was incorrectly dosed.

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

For the first interim analysis, among the ITT population of 204 subjects, 203 of them received at least one dose of study treatment and thus will be included in the ASaT population.

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.

All statistical tests, unless otherwise specified, will be stratified for treatment and stratification factors.

3.6.1.1 Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test (based on the stratification factors defined in Protocol Section 5.4). 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 same stratification factors used for randomization will be applied to both the stratified log-rank test and the stratified Cox model. For the first interim analysis, due to the small sample size in the stratum with TPS < 1% and nab-paclitaxel and East Asia, this stratum will be combined with the stratum with TPS < 1% and paclitaxel and East Asia in stratified analyses.

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 a blinded independent central imaging vendor, 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 a blinded independent central imaging vendor, 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 new anticancer treatment, whichever occurs later, to be a PD event for subjects without documented PD or death. The censoring rules for primary and sensitivity analyses are summarized in Table 1. In case there is an imbalance between the treatment groups on disease assessment schedules or censoring patterns, we will also perform the following two additional PFS sensitivity analyses: 1) a PFS analysis using time to scheduled tumor assessment visit from randomization as opposed to actual tumor assessment time; 2) Finkelstein's likelihood-based score test for interval-censored data [4] which modifies the Cox proportional hazard model for interval censored data, will be used as a supportive analysis for the PFS endpoint. The interval will be constructed so that the left endpoint is the date of the last disease assessment without documented PD and the right endpoint is the date of documented PD or death, whichever occurs earlier. In case the proportional hazards assumption doesn't hold, Fleming and Harrington's weighted log-rank test, Restricted Mean Survival Time (RMST) method or other methods, as appropriate, may be conducted.

Table 1 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 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 after ≥2 missed disease assessments	Progressed at date of documented PD or death	Censored at last disease assessment prior to the ≥2 missed disease assessment	Progressed at date of documented PD or death

3.6.1.2 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (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 for randomization will be applied to both the stratified log-rank test and the stratified Cox model. For the first interim analysis, due to the small sample size in the stratum with TPS < 1% and nab-paclitaxel and East Asia, this stratum will be combined with the stratum with TPS < 1% and paclitaxel and East Asia in stratified analyses. The Restricted Mean Survival Time (RMST) method may be conducted for OS to account for the possible non-proportional hazards effect and to estimate the absolute benefit of experimental treatment. A cure rate model may be applied to estimate the long-term effect.

Since subjects in the control arm are allowed to switch to the pembrolizumab treatment after progressive disease, adjustment for the effect of crossover on OS may be performed based on recognized methods, e.g., a two-stage method or the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis [3], based on an examination of the appropriateness of the data to the assumptions required by the methods.

3.6.1.3 Objective Response Rate (ORR) and Duration of Response (DOR)

The stratified Miettinen and Nurminen's method [5] with weights proportional to the stratum size will be used for comparison of the ORR between the treatment arms. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen method with strata weighting by sample size with a single treatment covariate will be reported. The stratification factors used for randomization (See Protocol Section 5.4) will be applied to the analysis. For the first interim analysis, due to the small sample size in the stratum with TPS < 1% and nab-paclitaxel and East Asia, this stratum will be combined with the stratum with TPS < 1% and paclitaxel and East Asia in stratified analyses.

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

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

Table 2 Censoring Rules for DOR

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new anti-cancer therapy initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anti-cancer therapy initiated	Last adequate disease assessment before new anti-cancer therapy initiated	Censor (non-event)
Death or progression 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.		

3.6.1.4 Analysis Strategy for Key Efficacy Endpoints

Table 3 summarizes the primary analysis approach for primary and key secondary efficacy endpoints. Sensitivity analysis methods are described above for each endpoint as applicable.

The strategy to address multiplicity issues with regard to multiple efficacy endpoints, 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
Dual Primary Endpoints			
PFS per RECIST 1.1 by blinded independent central imaging vendor	Test: Stratified Log-rank test to assess the treatment difference Estimation: Stratified Cox model with Efron's tie handling method to assess the magnitude of treatment difference	ITT	• Primary censoring rule • Sensitivity analysis 1 • Sensitivity analysis 2
OS	Test: Stratified Log-rank test to assess the treatment difference Estimation: Stratified Cox model with Efron's tie handling method to assess the magnitude of treatment difference	ITT	Model based (censored at the last date the subject was known to be alive)
Key Secondary Endpoints			
ORR per RECIST 1.1 by blinded independent central imaging vendor	Test and Estimation: Stratified M&N method with sample size weights††	ITT	Subjects without assessments are considered non-responders and conservatively included in denominator
DOR per RECIST 1.1 by blinded independent central imaging vendor	Descriptive statistics for range and Kaplan-Meier estimate of median	Patients in ITT population with an objective response	

† Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization (Section 5.4) will be applied to the analysis.

†† Miettinen and Nurminen method

3.6.1.5 Exploratory Analyses

An exploratory analysis of PFS2, defined as the time from randomization to subsequent disease progression after initiation of new anti-cancer therapy, or death from any cause, whichever first, may be carried out. Patients alive and for whom a disease progression following initiation of new anti-cancer treatment has not been observed will be censored at the last time the subject was known to be alive and without disease progression

3.6.2 Statistical Methods for Safety Analyses

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

Adverse Events

Adverse events (AEs) will be coded using the standard MedDRA and grouped system organ class. Adverse events (AEs) will be graded by the investigator according to the Common Terminology Criteria for Adverse Events (CTCAE), version 4.0.

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 will be subject to inferential testing for statistical significance with p-values and 95% confidence intervals provided for between-group comparisons. For this protocol, there are no Tier 1 AEs. 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) 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 values and vital signs, and ECG parameters 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.

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, an analysis of Grade 3-5 AEs will be based on the time to first event using the time-to-event analysis methods (i.e., the log-rank test will be used for testing the time to AEs, and the Cox model with Efron's tie handling method will be used for estimating the hazard ratio and its 95% confidence interval). For other AEs with potentially differential follow-up time, such analysis may also be explored.

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 unstratified Miettinen and Nurminen method [5], an unconditional, asymptotic method.

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 Grade 3-5 AE		X	X
	Any Serious AE		X	X
	Onset of First Grade 3-5 AE		X	X
	Any Drug-Related AE		X	X
	Any Serious and Drug-Related AE		X	X
	Any Grade3-5 and Drug-Related AE		X	X
	Dose Modification Due to AE		X	X
	Discontinuation Due to AE		X	X
	Death		X	X
	Specific AEs, SOCs (including ≥4 of subjects in one of the treatment groups)		X	X
Tier 3	Specific AEs, SOCs (incidence <4 of subjects in all of the treatment groups)			X
	Change from Baseline Results (Labs, ECGs, Vital Signs)			X
There are no Tier 1 AEs pre-specified in this protocol.				

3.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

The comparability of the treatment groups for each relevant baseline 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 randomized and the primary reason for discontinuation will be displayed. Demographic variables (such as age) and baseline characteristics will be summarized by treatment either by descriptive statistics or categorical tables.

3.7 Interim Analysis

There are three planned interim analyses (IA) in addition to the final analysis for this study. Details on the boundaries for establishing statistical significance with regard to efficacy are discussed further in Section 3.8. The trial will continue until the number of deaths (See Section

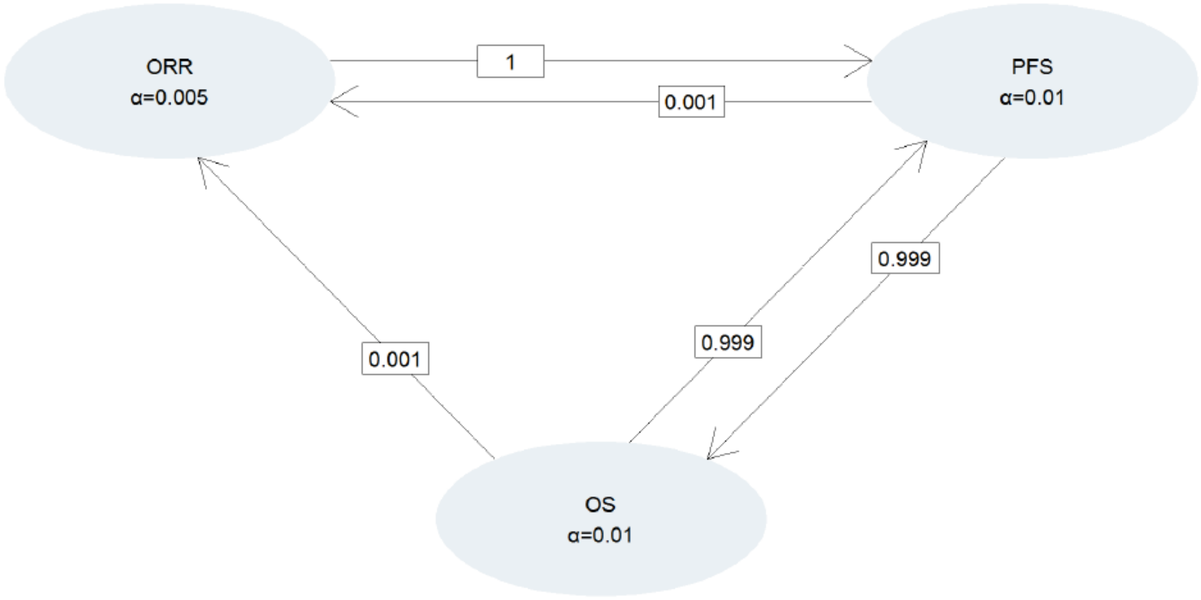
3.9) is approximately equal to the targeted number for the final analysis, irrespective of the outcome from the interim analyses. The analyses planned, endpoints evaluated, and drivers of timing are summarized in Table 5.

Table 5 Summary of Interim and Final Analyses Strategy

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	ORR	~ 200 subjects are followed for ~ 28 weeks so that each patient has at least 4 tumor assessments	~ 15 months	• Demonstrate ORR superiority
IA2	PFS OS	~ 332 PFS events have been observed.	~ 20 months	• Demonstrate PFS superiority • Demonstrate OS superiority
IA3	PFS OS	~ 415 PFS events have been observed	~ 25 months	• Demonstrate PFS superiority • Demonstrate OS superiority
Final Analysis	OS	~ 361 deaths have occurred.	~ 31 months	• Demonstrate OS superiority

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [2] to control multiplicity for multiple hypotheses as well as interim analyses. According to this approach, study hypotheses may be tested more than once, and when a particular null hypothesis is rejected, the alpha allocated to that hypothesis can be reallocated to other hypothesis tests. Figure 1 shows the initial one-sided alpha allocation for each hypothesis in the ellipse representing the hypothesis. The weights for reallocation from each hypothesis to the others are represented in the boxes on the lines connecting hypotheses.



Note: If both PFS and OS null hypotheses are rejected, the reallocation strategy allows re-testing of ORR at alpha=0.025 based on the p-value at IA1

ORR=objective response rate; OS=overall survival; PFS=progression-free survival

Figure 1 Type I Error Reallocation Strategy Following Closed Testing Principle

3.8.1 Objective Response Rate

The study allocates alpha=0.005, one-sided, to test ORR, and ORR is tested only at the first interim analysis (IA1). However, if the test does not reach statistical significance at IA1, the p-value from IA1 can be compared to an updated alpha-level if the null hypotheses for both PFS and OS are rejected at a later time. Power at the possible alpha-levels as well as the approximate treatment difference required to reach the bound (ORR difference) are shown in Table 6, assuming underlying 25% and 50% response rates in the control and experimental groups, respectively.

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

Alpha	ORR difference	Power
0.005	~ 0.18	0.84
0.025	~ 0.13	0.94

3.8.2 Progression-free Survival

The initial alpha-level for testing PFS is 0.01. If the null hypothesis for ORR is rejected, Figure 1 shows that its alpha=0.005 is fully reallocated to PFS hypothesis testing. If the null hypothesis for OS is rejected, then alpha=0.01 is essentially fully reallocated to PFS hypothesis testing. Thus, the PFS null hypothesis may be tested at alpha=0.01, alpha=0.015 (if the ORR null hypothesis is rejected but not the OS null hypothesis), alpha=0.02 (if the OS null hypothesis is rejected but not the ORR null hypothesis), or alpha=0.025 (if both the ORR and OS null hypotheses are rejected). Table 7 shows the boundary properties for each of these alpha-levels for the interim analyses, which were derived using a Lan-DeMets O'Brien-Fleming spending function. Note that the final row indicates the total power to reject the null hypothesis for PFS at each alpha-level. If the actual number of events at the PFS analyses differ from those specified in the table, the bounds will be adjusted using the Lan-DeMets O'Brien-Fleming spending function accordingly. Also note that if the OS null hypothesis is rejected at an interim or final analysis, each PFS interim and final analysis test may be compared to its updated bounds considering the alpha reallocation from the OS hypothesis.

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

Analysis	Value	α=0.01	α=0.015	α=0.02	α=0.025
IA 2: 80%* N: 560 Events: 332 Month: 20	Z	-2.6539	-2.4817	-2.3536	-2.2504
	p (1-sided)[§]	0.004	0.0065	0.0093	0.0122
	HR at bound[%]	0.7473	0.7616	0.7723	0.7811
	P(Cross) if HR=1[†]	0.004	0.0065	0.0093	0.0122
	P(Cross) if HR=0.7[#]	0.7243	0.7787	0.8148	0.8411
IA 3: 100%* N: 560 Events: 415 Month: 25	Z	-2.3737	-2.2244	-2.1138	-2.025
	p (1-sided)[§]	0.0088	0.0131	0.0173	0.0214
	HR at bound[%]	0.7921	0.8038	0.8126	0.8197
	P(Cross) if HR=1[†]	0.01	0.015	0.02	0.025
	P(Cross) if HR=0.7[#]	0.9	0.9243	0.9392	0.9494

*Percentage of expected number of events at final analysis required at interim analysis

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

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

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

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

3.8.3 Overall Survival

The OS hypothesis may be tested at alpha=0.01 (initially allocated alpha), alpha=0.02 (if the PFS but not the ORR null hypothesis is rejected), or alpha=0.025 (if both the ORR and PFS null hypotheses are rejected). Table 8 demonstrates the bounds and boundary properties for OS hypothesis testing derived using a Lan-DeMets O'Brien-Fleming spending function. If the actual number of OS events at the interim and final analyses differs from those specified in the table, the bounds will be adjusted using the Lan-DeMets O'Brien-Fleming spending function accordingly.

Table 8 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	Value	α=0.01	α=0.02	α=0.025
IA 2: 59%* N: 560 Events: 212 Month: 20	Z	-3.1648	-2.8202	-2.702
	p (1-sided)[§]	0.0008	0.0024	0.0034
	HR at bound[%]	0.6474	0.6788	0.6899
	P(Cross) if HR=1[†]	0.0008	0.0024	0.0034
	P(Cross) if HR=0.7 [#]	0.2849	0.4115	0.458
IA 3: 79%* N: 560 Events: 286 Month: 25	Z	-2.6914	-2.3992	-2.2995
	p (1-sided)[§]	0.0036	0.0082	0.0107
	HR at bound [%]	0.7274	0.7530	0.7619
	P(Cross) if HR=1[†]	0.0038	0.009	0.0117
	P(Cross) if HR=0.7[#]	0.6312	0.7362	0.7684
Final N: 560 Events: 361 Month: 31	Z	-2.3742	-2.116	-2.028
	p (1-sided)[§]	0.0088	0.0172	0.0213
	HR at bound [%]	0.7789	0.8003	0.8078
	P(Cross) if HR=1[†]	0.01	0.02	0.025
	P(Cross) if HR=0.7 [#]	0.85	0.9034	0.9181

* Percentage of expected number of events at final analysis required at interim analysis

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

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

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

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

3.9 Sample Size and Power Calculations

With ~200 subjects, the study has ~ 84% power for detecting a 25% difference in ORR (50% vs 25%) or ~ 97% power for detecting a 30% difference in ORR (50% vs. 20%) at initially assigned 0.005 (one-sided) significance level. The study has ~ 94% power for detecting a 25% difference in ORR (50% vs 25%) or ~ 99% power for detecting a 30% difference in ORR (50% vs 20%) at 0.025 (one-sided) significance level.

With 415 PFS events, the study has ~ 90% power for detecting a HR of 0.7 at initially assigned 0.01 (one-sided) significance level, ~ 92% power for detecting a HR of 0.7 at 0.015 (one-sided) significance level, ~ 94% power for detecting a HR of 0.7 at 0.02 (one-sided) significance level, and ~ 95% power for detecting a HR of 0.7 at 0.025 (one-sided) significance level.

With 361 deaths, the study has ~ 85% power for detecting a hazard ratio (HR) of 0.7 at 0.01 (one-sided) significance level, ~ 90% power for detecting a HR of 0.7 at 0.02 (one-sided) significance level, and ~ 92% power for detecting a HR of 0.7 at 0.025 (one-sided) significance level.

The planned sample size is approximately 560 subjects assuming: (1) the enrollment period is 15.5 months and the ramp-up period of enrollment is 7 months; (2) median PFS is 6 months in the control group and the true hazard ratio is 0.7; (3) median OS is 12 months in the control group and the true hazard ratio is 0.7; (4) the annual dropout rate is 3% for PFS and 1% for OS; (5) the number of events and alpha levels of interim analyses and final analysis are as specified in Section 3.7 and Section 3.8.

3.10 Subgroup Analyses and Effect of Baseline Factors

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

- Age category (< 65 vs. ≥ 65 years, and < 65 vs. 65-74 vs. 75-84 vs. ≥ 85 years)
- Sex (female vs. male)
- Race (white vs. non-white)
- ECOG status (0 vs. 1)
- Geographic region of enrolling site (East Asia vs. Non-East Asia; US vs. Ex-US; EU vs. Non-EU)
- Smoking status (never vs. former/current)
- Brain metastasis status at baseline (yes vs. no)
- PD-L1 status (TPS < 1% vs. TPS ≥ 1%, TPS < 50% vs. TPS ≥ 50%, TPS < 1% vs. TPS 1 – 49% vs .TPS ≥ 50%)
- Taxane chemotherapy (paclitaxel vs. nab-paclitaxel)

Additional subgroup analyses such as China vs. non-China may be conducted per China local regulatory needs. 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. For the first interim analysis, the subgroup analysis for ORR will be conducted using unstratified Miettinen and Nurminen method.

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

3.11 Extent of Exposure

The extent of exposure will be summarized as duration of treatment in cycles.

4 STATISTICAL ANALYSIS PLAN FOR EXTENSION

4.1 Introduction

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

After the enrollment for the global study 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 for the overall Chinese subpopulation reaches approximately 120.

After the cut-off date for the primary analyses of the global study (including interim and final analyses), all Chinese subjects, including subjects enrolled in the global study and the extension study , will continue their randomized treatment and continue to be followed up for PFS and OS events for China registration purpose. The extension study will be completed after target number of deaths has been observed between the two arms in the Chinese subpopulation. The expected timing of the analysis for the subpopulation is around 3.3 years from the time when the first subject from China is enrolled in the global study. However, if the target number of PFS events or deaths in the Chinese subpopulation is reached before an IA for the global study, the corresponding analysis for Chinese subpopulation will occur at the same time as the global IA or the final analysis (global study).

4.2 Responsibility for Analyses/In-House Blinding

The trial is double-blinded, analyses or summaries generated by randomized treatment assignment and actual treatment received will be limited and documented. Subjects randomized in the extension study will not be included in the database lock and primary analysis for the global study.

For all Chinese subjects, including subjects randomized in the global study and the extension study, patient level treatment randomization information will be blinded to the statistician(s)/programmer(s) responsible for the analysis of China extension study until the extension study data base lock is achieved. The extent to which individuals are unblinded to the results will be limited. Blinded and unblinded members will be clearly documented with blinding status along with time information.

4.3 Hypotheses/Estimation

No hypothesis testing is planned for the China extension study.

After succeeding in the global study, 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 from the global primary efficacy analyses (based on point estimates). Sample size is designed to provide about 80% chance of observing the point estimate of Chinese subpopulation preserves ≥ approximately 50% of empirical risk reduction from the global primary efficacy analysis assuming the same hazard ratio used in the sample size and power calculation for the global study.

4.4 The Analysis Endpoints

4.4.1 Efficacy Endpoints

Dual Primary

Overall survival (OS) is defined as the time from randomization to death due to any cause, the same as described in Section 3.4.1.

Progression-free survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded independent radiologists' assessment or death due to any cause, whichever occurs first, the same as described in Section 3.4.1.

Secondary

Objective response rate (ORR) based upon blinded independent central imaging vendor's assessed RECIST 1.1 as described in Section 3.4.1.

Duration of response (DOR) based upon blinded independent central imaging vendor's assessed RECIST 1.1 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

Efficacy analysis will be carried out in the entire intention-to-treat (ITT) population. This population will include all subjects who are randomized in the global study and all subjects who are randomized in the extension study. Chinese subpopulation will include all Chinese subjects in this population.

4.5.2 Safety Analysis Populations

Safety analysis will be carried out in the entire All Subjects as Treated (ASaT) population, i.e., all randomized subjects (in the global study and extension study) who received at least 1 dose of study treatment. Chinese subpopulation will include all Chinese subjects in this population.

4.6 Statistical Methods

Regarding the analysis for extension, no hypothesis testing is planned. There is no plan of interim analysis. No multiplicity adjustment will be applied to the analysis for extension.

4.6.1 Statistical Methods for Efficacy Analyses

4.6.1.1 Overall Survival (OS)

Analysis of OS for extension is the same to that for the global study if 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 study will be used. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for the Chinese subgroup analysis.

As an exploratory analysis, recognized methods, e.g., the Rank Preserving Structural Failure Time (RPSFT) model, two-stage method, etc., will be used to adjust for the effect of crossover on OS based upon the appropriateness of the data to the assumption required by the methods. The RPSFT model provides a randomization-based estimate of treatment effect (RBEE) corrected for the bias induced by crossover. The 95% confidence intervals of the hazard ratio for OS after adjustment of the effect of crossover will be provided. To further account for the possible confounding effect, a sensitivity analysis of OS that censors subjects at the time of initiation of new therapy will be performed and an OS analysis that treats initiation of new therapy as a time-dependent binary covariate will also be conducted.

Consistency of efficacy will be evaluated using the percentage of risk reduction preserved in the Chinese subpopulation from the empirical risk reduction from the global primary efficacy analyses (based on point estimates). Sample size is designed to provide about 80% chance of observing the point estimate of Chinese subpopulation preserves ≥ approximately 50% of empirical risk reduction from the global primary efficacy analysis assuming the same hazard ratio used in the sample size and power calculation for the global study.

In addition, supportive analyses on the entire ITT population will be provided with the data pooling global study (prior to data cutoff for the primary analysis) and China extension study together. Accordingly, non-Chinese subjects will be censored at last known alive date (this can be cutoff date if some assessment happens to be on that day or there’s assessment beyond the cutoff date) which is consistent with the primary analysis in the global study if subjects are still alive at primary analysis time for global study. The primary analysis for OS will be conducted in the Chinese subpopulation when approximately 75 OS events have been collected.

4.6.1.2 Progression-Free Survival (PFS)

Analysis of PFS for extension is the same to that for the global study if applicable.

In detail, the Kaplan-Meier method will be used to estimate the survival curves. For the whole population, 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. The same stratification factors used in the global study will be used. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for Chinese subgroup analysis. In case the proportional hazards assumption doesn't hold, Fleming and Harrington's weighted log-rank test, Restricted Mean Survival Time (RMST) method or other methods, as appropriate, may be conducted.

Consistency in PFS will be evaluated similarly as that in OS. The primary analysis for PFS will be conducted in the Chinese subpopulation when approximately 75 PFS events have been collected.

4.6.1.3 Objective Response Rate (ORR) and Duration of Response (DOR)

Stratified Miettinen and Nurminen's method with weights proportional to the stratum size will be used for comparison of the ORR between the treatment arms. A 95% CI for the difference in response rates between the pembrolizumab arm and the control arm will be provided. The same stratification factors used in the global study will be used. For the Chinese subpopulation analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for Chinese subgroup analysis.

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

4.6.1.4 Exploratory Analyses

Exploratory Analyses for extension is the same to that for the global study (if applicable).

4.6.2 Statistical Methods for Safety Analyses

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

4.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

They are the same for extension to that for the global study as described in Section 3.10.

4.7 Interim Analysis & Final analysis

The primary analysis for PFS will be conducted in the Chinese subpopulation when approximately 75 PFS events have been collected. OS will also be analyzed.

The primary analysis for OS will be conducted in the Chinese subpopulation when approximately 75 OS events have been collected.

4.8 Multiplicity

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

4.9 Sample Size and Power Calculations

After the enrollment of global study completes, the extension study will continue to randomize subjects in a 1:1 ratio into the pembrolizumab arm and the placebo arm in China until the sample size for the overall Chinese subjects (including those enrolled in the global study) reaches approximately 120. The extension study population, i.e., those Chinese subjects randomized after the close of enrollment for the global study, 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 Chinese subpopulation assuming the underlying hazard ratio for OS 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 50% of the empirical risk reduction from the global analysis in the Chinese subpopulation assuming the underlying hazard ratio is 0.70. The same consideration applies to PFS.

The above calculations for the consistency evaluation in PFS and OS are based on the same assumptions on the corresponding median OS/PFS and the true hazard ratio respectively.

4.10 Subgroup Analyses and Effect of Baseline Factors

All subgroup analysis defined in Section 3.10 will be repeated for the entire population and Chinese subpopulation if applicable. In addition, results for Chinese subpopulation vs. non-Chinese subpopulation will be provided. Country-specific analysis may also be conducted per local regulatory requirement.

5 REFERENCES

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

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

3. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Commun Stat-Theor M 1991;20(8): 2609-31.

4. Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics 1986;42:845-54.

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

Revision History

Date	Summary of Change
09JUN2017	Original Document
23OCT2017	Version 02
15NOV2017	Version 03
29JAN2018	Version 04