Official Protocol Title:	A Phase 3, randomized, placebo-controlled, double-blind clinical study of pembrolizumab (MK-3475) with or without lenvatinib (E7080/MK-7902) to evaluate the safety and efficacy of pembrolizumab and lenvatinib as 1L intervention in a PD-L1 selected population of participants with recurrent or metastatic head and neck squamous cell carcinoma (R/M HNSCC) (LEAP-010)
NCT number:	NCT05523323
Document Date:	14-Jun-2023

Supplemental Statistical Analysis Plan (sSAP)

A Phase 3, randomized, placebo-controlled, double-blind clinical study of pembrolizumab (MK-3475) with or without lenvatinib (E7080/MK-7902) to evaluate the safety and efficacy of pembrolizumab and lenvatinib as 1L intervention in a PD-L1 selected population of participants with recurrent or metastatic head and neck squamous cell carcinoma (R/M HNSCC) (LEAP-010)

June 14, 2023

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

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

2 SUMMARY OF CHANGES

This sSAP aligns with the protocol amendment 04 with regard to the statistical analysis plan. In addition, the following changes were made to the sSAP:

- Description of strata collapsing strategy for stratified analyses provided in Section 3.6.1.
- Details of endpoints and analyses of patient-reported outcomes (PRO) data provided in Section 3.4.3 and Section 3.6.3.

3 ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

Key elements of the SAP are summarized below. The comprehensive plan is provided in Section 3.2 – Responsibility for Analyses/In-House Blinding through Section 3.12 – Extent of Exposure.

Study Design Overview	A Phase 3 study of pembrolizumab (MK-3475) with or without lenvatinib (E7080/MK-7902) as 1L intervention in a PD-L1 selected population with R/M HNSCC (LEAP-010)
Treatment Assignment	Approximately 500 participants will be randomized in a 1:1 ratio between two treatment groups: (1) the pembrolizumab + lenvatinib arm and (2) the pembrolizumab + placebo arm. Stratification factors are: • PD-L1 tumor expression as determined by PD-L1 immunohistochemistry (TPS <50% vs. ≥50%) • HPV status for oropharynx cancer as determined by p16 immunohistochemistry (positive vs. negative); HPV status for participants without oropharynx cancer (eg, cancers of the oral cavity, hypopharynx and larynx) is considered HPV-negative. • ECOG performance status (0 vs. 1) This is a randomized double-blind study.
Analysis Populations	Efficacy: Intent-to-Treat

	Safety: All-Participants-as-Treated
Primary Endpoints	ORR, PFS, OS
Statistical Methods for Key Efficacy Analyses	The primary hypotheses addressing PFS and OS will be evaluated by comparing the experimental group to the control group using a stratified log-rank test. The hazard ratio will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. The stratified Miettinen and Nurminen method [Miettinen, O. and Nurminen, M. 1985] with strata weighted by sample size will be used for the analysis of the primary hypothesis addressing ORR.
Statistical Methods for Key Safety Analyses	For analyses in which 95% CIs will be provided for between-treatment differences in the percentage of participants with events, these analyses will be performed using the Miettinen and Nurminen method [Miettinen, O. and Nurminen, M. 1985].
Interim Analyses	**Efficacy** Three interim analyses are planned in this study. Results will be reviewed by an external DMC. Details are provided in Section 3.7. • Interim Analysis 1 (IA1): o Timing: to be performed after 350 participants are randomized with 6 months of follow-up o Primary purpose: efficacy analysis for ORR and PFS • Interim Analysis 2 (IA2): o Timing: to be performed after both ~432 PFS events have been observed and ~ 6 months after last participant randomized o Primary purpose: efficacy analysis for PFS and OS • Interim Analysis 3 (IA3): o Timing: to be performed after both ~326 OS events have occurred and ~ 14 months after last participant randomized o Primary purpose: efficacy analysis for OS • Final Analysis (FA): o Timing: to be performed after both ~361 OS events have occurred and ~ 20 months after last participant randomized o Primary purpose: efficacy analysis for OS Note that for IA2, IA3 and the FA, if the events accrue more slowly than expected, the analysis can be delayed up to 3 months after the projected timing. **Safety:** The first safety interim analysis will be performed and reviewed by the eDMC at 6 months after the first participant is randomized, or

	when the first 60 participants have been randomized, whichever is later. Afterwards, the eDMC will review safety data periodically in the study. Details will be included in the DMC charter.
Multiplicity	The overall Type I error over the primary hypotheses is strongly controlled at 2.5% (1-sided), with 0.25% initially allocated to ORR (H1), 0.1% to PFS (H2), and 2.15% to OS (H3). By using the graphical approach of Maurer and Bretz [Maurer, W. and Bretz, F. 2013], if one hypothesis is rejected, the alpha will be shifted to other hypotheses as described in Section 3.8.
Sample Size and Power	The planned sample size is approximately 500 participants. There will be 350 participants randomized with at least 6 months of follow-up at the ORR final analysis. The study has 90.4% power for detecting a 20-percentage point difference between treatment arms with an underlying 20% response rate in the control arm at an initially assigned 0.0025 (1-sided) significance level. It is estimated that there will be ~ 432 events at the PFS final analysis (ie, IA2 of the study). With 432 PFS events, the study has >99.9% power for detecting a hazard ratio of 0.5 at an initially assigned 0.001 (1-sided) significance level. There will be ~ 361 deaths at the OS final analysis. With 361 deaths, the study has ~90.7% power for detecting a hazard ratio of 0.7 at an initially assigned 0.0215 (1-sided) significance level.

3.2 Responsibility for Analyses/In-house Blinding

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

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

The Sponsor will generate the randomized allocation schedule for study treatment assignment for this protocol and the randomization will be implemented in IVRS/IWRS.

The investigator and the study team at the Sponsor consisting of clinical, statistical, statistical programming and data management personnel will be blinded to participant-level PD-L1 biomarker score. An unblinded Sponsor statistician and unblinded Sponsor statistical programmer will have access to the participant-level PD-L1 results for the purpose of data review and will have no other responsibilities associated with the study. A summary of PD-L1 biomarker prevalence may be provided to the study team at the Sponsor by the unblinded Sponsor statistician. In addition, the independent radiologist(s) will perform the central imaging review without knowledge of treatment group assignment.

Blinding issues related to the planned interim analyses are described in Section 3.7.

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Protocol Section 3.0 – Hypotheses, Objectives and Endpoints.

3.4 Analysis Endpoints

Efficacy and safety endpoints that will be evaluated are listed below.

3.4.1 Efficacy Endpoints

Primary

- **Objective Response Rate**

The ORR is defined as the percentage of participants who achieve a confirmed complete response or partial response per RECIST 1.1 as assessed by BICR.

- **Progression-free Survival**

PFS is defined as the time from randomization to the first documented disease progression per RECIST 1.1 by BICR or death due to any cause, whichever occurs first. See Section 3.6.1 – Statistical Methods for Efficacy Analyses for definition of censoring.

- **Overall Survival**

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

Secondary

- **Duration of Response**

For participants who show confirmed CR or PR, duration of response is defined as the time from the first documented evidence of CR or PR until disease progression or death due to any cause, whichever occurs first.

3.4.2 Safety Endpoints

Safety measurements are described in Protocol Section 4.2.1.2 – Safety Endpoints and Protocol Section 8 – Study Assessments and Procedures. Safety and tolerability will be assessed by clinical review of all relevant parameters including AEs, laboratory tests and vital signs. Safety parameters to be analyzed include, but are not limited to, AEs, SAEs, fatal AEs, and laboratory changes.

3.4.3 Patient-Reported Outcome Endpoints

Exploratory PRO endpoints are as follows:

- Change from baseline in
  - EORTC QLQ-C30 global health status/QoL scores, physical functioning score, and role functioning score
  - EORTC QLQ-H&N35 pain in the mouth, problems with swallowing, and problems with speech scores
  - EQ-5D-5L visual analogue scale (VAS) score, and utility scores
- Time to confirmed deterioration (TTD) is defined as the time from baseline to the first onset of a 10 or more points deterioration from baseline with confirmation at the subsequent visit of a 10 or more points deterioration from baseline, measured by
  - EORTC QLQ-C30 global health status/QoL scores, physical functioning score, and role functioning score
  - EORTC QLQ-H&N35 pain in the mouth, problems with swallowing, and problems with speech scores
- Overall improvement and overall improvement + stability in
  - EORTC QLQ-C30 global health status / QoL, physical functioning, and role functioning
  - EORTC QLQ-H&N35 pain in the mouth, problems with swallowing, and problems with speech scores.

3.4.4 Derivations of PRO Endpoints

The derivation of overall improvement and overall improvement + stability is based on the assessment for possible PRO response at a time point considering subsequent confirmation, defined as follows:

Assessment Category at a time point (one analysis visit)	**Change from baseline at a time point (one analysis visit)**	**Change from baseline at the subsequent time point (the next consecutive analysis visit)**
Improvement	score improved from baseline by ≥10 points	score improved from baseline by ≥10 points
Stability	score improved from baseline by ≥10 points	score improved or worsened from baseline by <10 points
	score improved or worsened from baseline by <10 points	score improved or worsened from baseline by <10 points
	score improved or worsened from baseline by <10 points	score improved from baseline by ≥10 points
Worsening	score worsened from baseline by ≥10 points	not required
Unconfirmed	A time point assessment that doesn't meet any of the above criteria.	

The overall improvement is defined as the best observed PRO response that is an improvement among all post-baseline assessments by timepoint. The overall improvement + stability is defined as the best observed PRO response that is an improvement or stability among all post-baseline assessments by timepoint.

Based on prior literature, Bjordal et al. (2000) [Bjordal, K., et al 2000], Osoba et al. (1998) [Osoba, D., et al 1998], and King (1996) [King, M. T. 1996], a 10-point or greater worsening from baseline for each scale represents a clinically relevant deterioration. Changes from baseline will also be interpreted according to recent subscale-specific guidelines, which indicate that clinically meaningful differences vary by scale (Cocks et al. 2012) [Cocks, K., et al 2012].

3.5 Analysis Populations

3.5.1 Efficacy Analysis Populations

The ITT population will serve as the population for the primary efficacy analyses. For the analyses of ORR, approximately the first 350 or more randomized participants will be included in the analysis population. For the analyses of PFS and OS, all randomized participants will be included in the ITT population. For the analyses of DOR, the subset of participants who show a confirmed complete response or partial response will be included in the analysis population. Participants will be included in the treatment group to which they are randomized.

3.5.2 Safety Analysis Populations

Safety analyses will be conducted in the APaT population, which consists of all randomized participants who received at least one dose of study treatment. Participants will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the APaT population. This will be the treatment group to which they are randomized except for participants who take incorrect study treatment for the entire treatment period; such participants will be included in the treatment group corresponding to the study treatment actually received.

At least 1 laboratory or vital sign measurement obtained after at least 1 dose of study treatment is required for inclusion in the analysis of each specific parameter. To assess change from baseline, a baseline measurement is also required.

3.5.3 Patient-Reported Outcome Analysis Populations

The PRO analyses are based on the PRO FAS population, defined as all randomized participants who have at least 1 PRO assessment available for the specific endpoint and have received at least 1 dose of the study intervention. Participants will be analyzed in the treatment group to which they are randomized.

3.5.4 Population Pharmacokinetic Analysis Set

No pharmacokinetic endpoints are planned for this study.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

This section describes the statistical methods that address the primary and secondary and exploratory objectives.

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

The stratification factors used for randomization (see Protocol Section 6.3.2 - Stratification) will be applied to all stratified analyses, in particular, the stratified log-rank test, stratified Cox model, and stratified Miettinen and Nurminen method [Miettinen, O. and Nurminen, M. 1985].

The efficacy analyses for ORR, DOR and PFS will include responses and documented progression events that occur prior to Second Course treatment.

Since less than 5% of the participants in the ITT population were enrolled in Russia (0% in Ukraine), it is expected to have minimal impact on the efficacy analyses.

3.6.1.1 Objective Response Rate

The stratified Miettinen and Nurminen method [Miettinen, O. and Nurminen, M. 1985] will be used for comparison of the ORR between two treatment groups. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen method with strata weighting by sample size will be reported. The stratification factors used for randomization (See Protocol Section 6.3.2 - Stratification) will be applied to the analysis. Based on a blinded review of response counts within the 8 strata prior to IA1, as the number of observed responses in one stratum was small (<5), the Miettinen and Nurminen stratified analyses will be based on the following collapsed strata levels:

1. ECOG performance status 0 + HPV status negative + PD-L1 TPS $\geq$ 50%
2. ECOG performance status 0 + HPV status negative + PD-L1 TPS < 50%
3. ECOG performance status 0 + HPV status positive + PD-L1 TPS $\geq$ 50%
4. ECOG performance status 0 + HPV status positive + PD-L1 TPS < 50%
5. ECOG performance status 1 + HPV status negative + PD-L1 TPS $\geq$ 50%
6. ECOG performance status 1 + HPV status negative + PD-L1 TPS < 50%
7. ECOG performance status 1 + HPV status positive + Any PD-L1 TPS

The same strata used for ORR analysis at IA1 will be carried in future ORR analyses, even though small strata may not occur at the time of a future analysis.

The point estimate of ORR will be provided by treatment group, together with 95% CI using exact binomial method proposed by Clopper and Pearson (1934) [Clopper, C. J. and Pearson, E. S. 1934]. Supportive analyses may be performed for comparison of ORR based on investigator's assessment.

3.6.1.2 Progression-free Survival

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. The stratification factors used for randomization (See Protocol Section 6.3.2 - Stratification) will be applied to both the stratified log-rank test and the stratified Cox model. At IA1, the same strategy of combination of small strata defined for the ORR analysis will be used for the PFS analysis. The same strata used for analysis of PFS at IA1 will be carried in future PFS analyses, even though small strata may not occur at the time of a future analysis. Supportive analyses may be performed for comparison of PFS based on investigator's assessment. Since disease progression is assessed periodically, PD can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. The true date of disease progression will be approximated by the earlier of the date of the first assessment at which PD is objectively documented per RECIST 1.1 by BICR and the date of death. Death is considered a PD event.

For the primary analysis, any participant who experiences an event (PD or death) immediately after 2 or more consecutive missed disease assessments will be censored at the last disease assessment prior to the missed visits. In addition, any participant who initiates new anticancer therapy will be censored at the last disease assessment prior to the initiation of new anticancer therapy. Participants who do not start new anticancer therapy and who do not experience an event will be censored at the last disease assessment. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied.

To evaluate the robustness of the PFS endpoint per RECIST 1.1 by BICR, two sensitivity analyses with different sets of censoring rules will be performed. The first sensitivity analysis follows the intention-to-treat principle. That is, PDs/deaths are counted as events regardless of missed study visits or initiation of new anticancer therapy. The second sensitivity analysis considers discontinuation of treatment due to reasons other than complete response or initiation of new anticancer treatment, whichever occurs later, to be a PD event for participants without documented PD or death. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied. The censoring rules for the primary and sensitivity analyses are summarized in Table 1.

Table 1 Censoring Rules for Primary and Sensitivity Analyses of PFS

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
PD or death documented after ≤1 missed disease assessment, and before new anticancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death	Progressed at date of documented PD or death

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

3.6.1.3 Overall Survival

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, the hazard ratio). The hazard ratio and its 95% confidence interval from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization (See Protocol Section 6.3.2 - Stratification) will be applied to both the stratified log-rank test and the stratified Cox model. Based on a blinded review of event counts by stratum prior to IA2 (first planned analysis for OS), as the event counts in two strata were small (<5), the following collapsed strata levels will be used in the stratified log-rank test and the stratified Cox model:

1. ECOG performance status 0 + HPV status negative + PD-L1 TPS ≥ 50%
2. ECOG performance status 0 + HPV status negative + PD-L1 TPS < 50%
3. ECOG performance status 0 + HPV status positive + Any PD-L1 TPS
4. ECOG performance status 1 + HPV status negative + PD-L1 TPS ≥ 50%
5. ECOG performance status 1 + HPV status negative + PD-L1 TPS < 50%
6. ECOG performance status 1 + HPV status positive + Any PD-L1 TPS

The same strata used for the IA2 OS analysis will be carried in future OS analyses, even though small strata may not occur at the time of a future analysis.

Participants without documented death at the time of analysis will be censored at the date the participant was last known to be alive.

3.6.1.4 Duration of Response

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

Censoring rules for DOR are summarized in Table 2.

For each DOR analysis, a corresponding summary of the reasons responding participants are censored will also be provided. Responding participants who are alive, have not progressed, have not initiated new anticancer 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 participant 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 anticancer therapy initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anticancer therapy initiated	Last adequate disease assessment before new anticancer therapy initiated	Censor (non-event)
Death or progression immediately after ≥2 consecutive missed disease assessments or after new anticancer therapy, if any	Earlier date of last adequate disease assessment prior to ≥2 missed adequate disease assessments and new anticancer therapy, if any	Censor (non-event)
Death or progression after ≤ 1 missed disease assessments and before new anticancer therapy, if any	PD or death	End of response (Event)

Abbreviations: DOR = duration of response; PD = disease progression.
A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response.

3.6.1.5 Analysis Strategy for Key Efficacy Variables

A summary of the primary analysis strategy for the key efficacy endpoints is provided in Table 3.

Table 3 Analysis Strategy for Key Efficacy Variables

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Primary Analyses			
ORR per RECIST 1.1 by BICR	Testing and estimation: stratified Miettinen and Nurminen method	ITT	Participants with missing data are considered non-responders
PFS per RECIST 1.1 by BICR	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored according to rules in Table 1
OS	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at participant's last known alive date
Abbreviations: BICR = blinded independent central review; ITT = intent-to-treat; ORR = objective response rate; OS = overall survival; PFS = progression-free survival; RECIST 1.1 = Response Evaluation Criteria in Solid Tumors.			

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

3.6.2 Statistical Methods for Safety Analyses

The primary safety analyses will include only events that occur prior to Second Course Treatment.

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

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

Tier 1 Events

Safety parameters or adverse events of special interest that are identified *a priori* constitute "Tier 1" safety endpoints that will be subject to inferential testing for statistical significance. AEs that are immune-mediated or potentially immune-mediated are well documented and will be evaluated separately; however, these events have been characterized consistently throughout the pembrolizumab clinical development program, and determination of statistical significance is not expected to add value to the safety evaluation. Finally, there are no known AEs associated with participants with HNSCC for which determination of a p-value is expected

to impact the safety assessment. Thus, there are no AEs that warrant elevation to Tier 1 in this study.

Tier 2 Events

Tier 2 parameters will be assessed via point estimates with 95% CIs provided for differences in the proportion of participants with events using the Miettinen and Nurminen method, an unconditional, asymptotic method [Miettinen, O. and Nurminen, M. 1985].

Membership in Tier 2 requires that at least 10% of participants in any treatment group exhibit the event; all other AEs will belong to Tier 3. The threshold of at least 10% of participants was chosen for Tier 2 events because the population enrolled in this study is in critical condition and usually experiences various AEs of similar types regardless of treatment; events reported less frequently than 10% of participants would obscure the assessment of the overall safety profile and add little to the interpretation of potentially meaningful treatment differences. In addition, Grade 3 to 5 AEs (≥5% of participants in 1 of the treatment groups), SAEs (5% of participants in 1 of the treatment groups) and tumor hemorrhage (incidence ⩾4 of participants in one of the treatment groups) will be considered Tier 2 endpoints. Because many 95% CIs may be provided without adjustment for multiplicity, the CIs should be regarded as a helpful descriptive measure to be used in safety review, not as a formal method for assessing the statistical significance of the between-group differences.

Tier 3 Events

Safety endpoints that are not Tier 1 or 2 events are considered Tier 3 events. The broad AE categories consisting of the proportion of participants 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, discontinued due to an AE, and death that are not pre-specified as Tier 1 endpoints will be classified as belong to “Tier 3”. Laboratory test toxicity grade shift from baseline will be considered Tier 3 event. Only point estimates by treatment group are provided for Tier 3 safety parameters.

Continuous Safety Measures

Continuous measures such as changes from baseline in vital signs parameters will be considered Tier 3 event. Summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group for vital signs parameters.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	p-Value	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Specific Grade 3-5 AE (incidence ≥5% of participants in one of the treatment groups)		X	X
	Specific serious AE (incidence ≥5% of participants in one of the treatment groups)		X	X
	Specific AEs, SOCs (incidence ≥10% of participants in one of the treatment groups)		X	X
	Tumor hemorrhage (incidence ≥ 4 of participants in one of the treatment groups)		X	X
Tier 3	Any AE			X
	Any Grade 3-5 AE			X
	Any Serious AE			X
	Any Drug-Related AE			X
	Any Serious and Drug-Related AE			X
	Any Grade 3-5 and Drug-Related AE			X
	Discontinuation due to AE			X
	Dose interruption due to AE			X
	Dose reduction due to AE			X
	AEOSI, CSAE			X
	Death			X
	Specific AEs, SOCs (incidence > 0% of participants in all of the treatment groups)			X
	Change from Baseline Results (lab toxicity shift, vital signs)			X

Abbreviations: AE = adverse event; AEOSI = adverse event of special interest; CSAE = clinically significant adverse event; CI = confidence interval; SOC = system organ class.

AEs that are immune-mediated or potentially immune-mediated will be evaluated separately. These events have been characterized consistently throughout the pembrolizumab clinical development program. Point estimates and 95% CIs for between-group difference is not expected to add value to the safety evaluation, and hence only number and percentage of participants with such pembrolizumab AEOSI will be provided, as well as the number and percentage of participants with corticosteroids administration to treat an AEOSI. Summary statistics will be provided for the analysis of time from first dose to the onset of an AEOSI.

Frequency of AE by time period from first dose (e.g., 0-3, 3-6, 6-12 months) may also be provided. In each time interval, the denominator is the number of participants at risk for the event during the particular time period, defined as participants who are event-free up until the start of the interval.

To properly account for the potential difference in follow-up time between the study arms, which is expected to be longer in the pembrolizumab arm, AE incidence adjusted for treatment exposure analyses may be performed as appropriate.

Time to Grade 3-5 AE

Additional exploratory analysis may be performed on the time to the first Grade 3-5 AE. The time to the first Grade 3-5 AE is defined as the time from the first day of study drug to the first event of a Grade 3-5 AE. Summary statistics will be provided.

3.6.3 Statistical Methods for Patient-Reported Outcome Analyses

This section describes the planned analyses for the PRO endpoints.

3.6.3.1 PRO Scoring Algorithm

EORTC QLQ-C30 Scoring

EORTC QLQ-C30 Scoring: Each scale or item is scored between 0 and 100, according to the EORTC QLQ-C30 standard scoring algorithm by Scott et al. (2008) [Scott, N. W., et al 2008]. For global health status/quality of life and all functional scales, a higher value indicates a better level of function; for symptom scales and items, a higher value indicates increased severity of symptoms.

EORTC QLQ-H&N35 Scoring

EORTC QLQ-H&N35 Scoring: The head & neck cancer module incorporates seven multi-item scales that assess pain, swallowing, senses (taste and smell), speech, social eating, social contact and sexuality. There are also eleven single items. For all items and scales, high scores indicate more problems (i.e. there are no function scales in which high scores would mean better functioning). The scoring approach for the EORTC QLQ-H&N35 is identical in principle to that for the symptom scales/single items of the QLQ-C30.

EQ-5D Scoring

The EQ-5D-5L is a 5-level standardized instrument for use as a measure of health outcome. Applicable to a wide range of health conditions and treatments, it provides a simple descriptive profile and a single index value for health status [EuroQol Group, 2023] [EuroQol Research Foundation 2021]. The EQ-5D-5L consists of 2 assessments, a descriptive system and a visual analogue scale (VAS). The descriptive system is a is a brief self-reported measure of generic health comprises of the following 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 severity levels: no problems, slight problems, moderate problems, severe problems, and extreme problems.

This health state classifier can describe 3125 unique health states that are often reported as vectors ranging from 11111 (full health) to 55555 (worst health). Societal value sets are derived from population-based valuation studies around the world, when applied to the health state

vector, result in a preference- based score, called health utility index, that typically ranges from states worse than dead (0) to 1 (full health), anchoring dead at 0.

The "crosswalk" mapping function estimated by Ben Van Hout method (2012) [van Hout, B., et al 2012] will be used to score the responses of the health state classifier into EQ-5D-5L health utility index.

The VAS records the respondent's self-rated health on a 0-100 vertical scale with endpoints labeled "the best health you can imagine" and "the worst health you can imagine", with higher scores corresponding to a better health state. This information is used as a quantitative measure of health as judged by the individual respondents.

3.6.3.2 PRO Completion and Compliance Summary

Completion and compliance of EORTC QLQ-C30, EORTC QLQ-H&N35, and EQ-5D-5L by visit and by treatment will be described. Numbers and percentages of complete and missing data at each visit will be summarized.

Completion rate of treated participants (CR-T) at a specific visit for a given instrument is defined as the number of treated participants who complete at least one item on that PRO instrument over the number of treated participants in the PRO analysis population.

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

The completion rate is expected to decrease at later visits during study period for reasons such as study design (e.g., PROs not required following progression), patient discontinuation, etc. Therefore, the compliance rate (CR-E) will also be presented in addition to completion rate. CR-E is defined as the number of treated participants who complete at least one item of the instrument over number of participants who are expected to complete the PRO assessment at that visit, excluding participants missing by design such as death, progression, translation not available.

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

The completion and compliance status will be summarized as below:

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

The reasons for non-completion as scheduled of these measures are collected using "miss_mode" forms filled by site personnel and will be summarized in a table format. The schedule (study visits and estimated study times) and mapping of study visit to analysis visit for PRO data collection is provided in Table 5.

Table 5 PRO Data Collection Schedule and Mapping of Study visit to Analysis Visit

Treatment Week	**Week 0 (Baseline)**	**Week 3**	**Week 6 to Week 21 (Every 3 weeks)**	**Week 24**	**Week 30 to Week 42 (Every 6 weeks)**	**Week 48**	**Week 57 and Every 9 weeks through end of treatment**
Day	1	22	Week number*7+1	169	Week number*7+1	337	Week number*7+1
Range (relative day to-first dose date)	[-7, 7]	[8, 32]	[Week number*7-9, week number*7+11]	[159, 189]	[Week number*7-20, week number*7+21]	[316, 367]	[Week number*7-31, week number*7+31]

3.6.3.3 Change from Baseline

The time point for the mean change from baseline analysis is defined as the latest time point at which CR-T ≥ 60% and CR-E ≥ 80%, and week 15 was selected based on blinded data review prior to the database lock for IA2.

To assess the treatment effects on the PRO score change from baseline in the EORTC QLQ-C30 global health status/QoL, physical functioning, role functioning, EORTC QLQ-H&N35 pain, swallowing and speech symptoms, EQ-5D VAS score, and EQ-5D-5L health utility index, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger (2000) [Liang, K-Y and Zeger, S. L. 2000] the treatment by time interaction, and stratification factors used for randomization as covariates.

The treatment difference in terms of least square (LS) mean change from baseline will be estimated from this model together with 95% CI. Model-based LS mean with 95% CI will be provided by treatment group for PRO scores at baseline and the post-baseline time point.

The technical details on the cLDA model are in the appendix of this sSAP.

Line plots for the empirical mean change from baseline in EORTC QLQ-C30 global health status/QoL, physical functioning, role functioning, EORTC QLQ- H&N35 pain, swallowing and speech symptoms over time will be provided as a supportive analysis.

In addition, the model-based LS mean change from baseline to the specified post-baseline time point together with 95% CI will be plotted in bar charts for EORTC QLQ-C30 global health status/quality of life scores, all functioning, EORTC QLQ- H&N35 pain, swallowing and speech symptoms.

3.6.3.4 Time to Confirmed Deterioration (TTD)

The Kaplan-Meier method will be used to estimate the TTD curve for each treatment group. The estimate of median time to deterioration and its 95% confidence interval will be obtained from the Kaplan-Meier estimates. The treatment difference in TTD will be assessed by the

stratified log-rank test, and two-sided nominal p-value will be reported. A stratified Cox proportional hazard model with Efron's method of tie handling and with a single treatment covariate will be used to assess the magnitude of the treatment difference (i.e, HR). The HR and its 95% CI will be reported. The same stratification factors used for randomization (See Protocol Section 6.3.2) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model.

The approach for the TTD analysis will be based on the assumption of non-informative censoring. The participants who do not have deterioration on the last date of evaluation will be censored. Table 6 provides censoring rule for TTD analysis.

Table 6 Censoring Rules for Time to First Deterioration

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

3.6.3.5 Overall Improvement and Overall Improvement and Stability

Overall improvement rate will be analyzed, which is defined as the proportion of participants who have achieved an overall improvement as defined in Section 3.4.3 PRO Endpoints. Stratified Miettinen and Nurminen's method will be used for comparison of the overall improvement rate between the treatment groups. The difference in overall improvement rate and its 95% CI from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be provided. The stratification factors used for randomization (See Protocol Section 6.3.2 - Stratification) will be applied to the analysis.

The point estimate of overall improvement rate will be provided by treatment group, together with 95% CI using exact binomial method by Clopper and Pearson (1934) [Clopper, C. J. and Pearson, E. S. 1934].

The same method will be used to analyze overall improvement + stability rate, which is defined as the proportion of participants who have achieved overall improvement + stability as defined in Section 3.4.3 PRO Endpoints.

3.6.3.6 Analysis Strategy for Key PRO Endpoints

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Mean change from baseline in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning EORTC QLQ-H&N35 • Pain • Swallowing • Speech problems EQ-5D VAS EQ-5D health utility index	cLDA model	PRO FAS	Model-based.
TTD in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning EORTC QLQ-H&N35 • Pain • Swallowing • Speech problems	Stratified log-rank test and HR estimation using stratified Cox model with Efron's tie handling method	PRO FAS	Censored according to rules in Table 6.
Overall improvement, and overall improvement + stability in EORTC QLQ-C30 • Global health status / QoL • Physical functioning. • Role functioning EORTC QLQ-H&N35 • Pain • SwallowingSpeech problems	Stratified Miettinen and Nurminen method	PRO FAS	Participants with missing data are considered not achieving improvement/stability at that timepoint and included in the total number of participants.
Abbreviations: cLDA = constrained longitudinal data analysis, FAS = full analysis set, TTD=time to first deterioration, HR = hazard ratio, QoL = quality of life.			

It was noted that a translation error was identified prior to any PRO analysis for Question 52 (Q52) in the 'Trouble with Social Eating' EORTC QLQ-H&N35 subscale. This occurred only to Japanese participants, where the translation for Q52 used the opposite meaning intended in English, i.e.:

- English: Q52: "Have you had trouble enjoying our meals?"
- Japanese translation: Q52: 'Could you enjoy meals?'

Per the EORTC recommendation, the derivation will exclude Q52 for Japanese participants only from the 4-question composite scale of Social Eating. The results for Q52 in Japanese

participants will be considered missing (i.e. set to "Null" in the SDTM QS dataset) to retrigger the macro calculating this score without Q52 for those participants in ADPRO.

3.6.4 Demographic and Baseline Characteristics

The comparability of the treatment groups for each relevant demographic and 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 participants screened and randomized and the primary reasons for screening failure and discontinuation will be displayed. Demographic variables, baseline characteristics, primary and secondary diagnoses, and prior and concomitant therapies will be summarized by treatment either by descriptive statistics or categorical tables.

3.7 Interim Analyses

Access to the allocation schedule for summaries or analyses for presentation to the eDMC will be restricted to an unblinded statistician and an unblinded scientific programmer performing the interim analysis, who will have no other responsibilities associated with the study.

The eDMC will serve as the primary reviewer of the results of the interim analyses and will make recommendations for discontinuation of the study or modification to the executive oversight committee of the Sponsor. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee and potentially other limited Sponsor personnel may be unblinded to results at the treatment level in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of interim analyses will be documented by the unblinded team. Additional logistic details will be provided in the eDMC Charter.

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

3.7.1 Efficacy Interim Analysis

Three interim analyses are planned in addition to the final analysis for this study. Timing of Interim Analysis 1 (IA1) will occur when at least 350 participants have been randomized with 6 months of follow-up, i.e. 6 months after the ~350th participant randomized. At IA1, the ORR analysis population will be these ~350 participants and the PFS analysis population will be all randomized participants, regardless of length of follow-up. The other interim analyses and final analysis will include all randomized participants. Results of the interim analyses will be reviewed by the DMC. Details of the boundaries for establishing statistical significance with regards to efficacy are discussed further in Section 3.8.

The analyses planned, endpoints evaluated, and drivers of timing are summarized in Table 7.

Table 7 Summary of Interim and Final Analyses Strategy

Analyses	Key Endpoints	Timing[a]	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	ORR PFS	At least 350 participants randomized with ~ 6 months of follow-up	~ 24 months	• Final ORR analysis • Interim PFS analysis
IA2	PFS OS	Both ~ 432 PFS events have occurred and ~ 6 months after last participant randomized	~ 30 months	• Final PFS analysis • Interim OS analysis
IA3	OS	Both ~ 326 deaths have occurred and ~ 14 months after last participant randomized	~ 38 months	• Interim OS analysis
FA	OS	Both ~ 361 deaths have occurred and ~ 20 months after last participant randomized	~ 44 months	• Final OS analysis

Abbreviations: FA = final analysis; IA = interim analysis; ORR = objective response rate; OS = overall survival; PFS = progression-free survival.

a: For IA2, IA3 and FA, if the events accrue slower than expected, the analysis can be delayed up to 3 months after the projected timing, ie, the Sponsor may conduct IA2, IA3 and FA when all participants have been followed up for 9, 17 months and 23 months, respectively.

3.7.2 Safety Interim Analysis

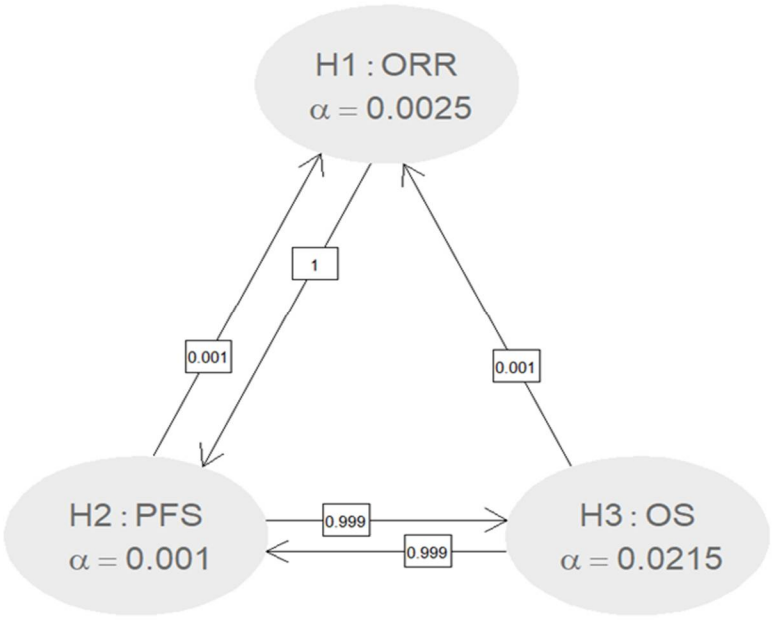
The eDMC will be responsible for periodic interim safety reviews as specified in the DMC charter. Interim safety analyses will also be performed at the time of interim efficacy analyses. Details will be included in the DMC charter.

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [Maurer, W. and Bretz, F. 2013] 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 α allocated to that hypothesis can be reallocated to other hypothesis tests. Figure 1 shows the initial 1-sided α allocation for each hypothesis in the ellipse representing the hypothesis. The weights for re-allocation from each hypothesis to the others are shown in the boxes on the lines connecting hypotheses.

The overall Type-I error is strongly controlled at 0.025 (1-sided) for the 3 primary endpoints ORR, PFS and OS. The initial α assigned to ORR, PFS, and OS will be 0.0025, 0.001 and 0.0215, respectively. If the PFS hypothesis is rejected, the corresponding alpha can be reallocated to OS. If the OS hypothesis is rejected, the corresponding alpha can be reallocated to PFS. If ORR hypothesis is rejected, the corresponding alpha can be reallocated to PFS. If both the OS and PFS hypotheses are rejected, the corresponding alpha can be reallocated to ORR.

Figure 1 Multiplicity Diagram for Type I Error Control



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

3.8.1 Objective Response Rate

The study will test ORR only once at IA1, at an initial α level of 0.0025 (Figure 1). Note that if superiority for both the PFS and OS hypotheses is declared at a future planned analysis, α = 0.0225 will be rolled over to the hypothesis for ORR, then the test statistics previously computed at IA1 for the ORR hypothesis will be used for inferential testing with an updated alpha level of 0.025.

Based on the first 350 randomized participants with at least 6 months of follow-up, power at the possible α-levels as well as the approximate treatment difference required to reach the bound (ΔORR) are shown in, assuming underlying 20% and 40% response rates in the control and experimental groups, respectively.

Table 8 Possible α Levels (One-Sided) and Approximate ORR Difference Required to Demonstrate Efficacy for Objective Response at IA1

α (One-Sided)	~Δ ORR	Power
0.0025	0.1327	0.904
0.025	0.0901	0.985
Abbreviation: ORR = objective response rate.		

3.8.2 Progression-free Survival

The study will test PFS at IA1 and IA2 only. Following the multiplicity strategy as outline in Figure 1, the PFS hypothesis may be tested at α=0.001 (initially allocated α), at α= 0.0035 (if the ORR null hypothesis is rejected but not the OS hypothesis), at α=0.0225 (if the OS null hypothesis is rejected but not the ORR hypothesis), or at α=0.025 (if both the OS and ORR null hypotheses are rejected). For the superiority hypothesis, a Lan-DeMets O'Brien-Fleming alpha spending function is used to construct group sequential boundaries to control the Type-I error rate. Table 9 shows the boundary properties for each of these α levels for the PFS analysis. Note that the final row indicates the total power to reject the null hypothesis for PFS at each α level. Also, note that if the OS null hypothesis is rejected at IA3 or FA of the study, the previously computed PFS test statistics may be used for inferential testing with its updated bounds, considering the α reallocation from the OS hypothesis.

For the PFS hypothesis H2, IA2 will be the final PFS analysis with a target number of events of 432. The bounds provided in Table 9 are based on the assumption that the number of events at IA1 and IA2 are 350 and 432, respectively. At the time of an analysis, the observed number of events may differ substantially from the expected. To avoid overspending at IA1 and keep reasonable alpha for the IA2, the minimum alpha spending strategy will be adopted. Technical details on the strategy are described at the end of Section 3.8.3 Overall Survival. At IA1, the information fraction used in Lan-DeMets spending function to determine the alpha spending at the IA will be based on the minimum of the expected information fraction and the actual information fraction at the analysis.

- In the scenario that the event accumulation is faster than expected, ie, if the number of observed events exceeds the expected number of events at IA1, then the information fraction will be calculated as the expected number of events at IA1 over the target number of events at IA2.

- In the scenario that the event accumulation is slower than expected and number of events is less than the expected number of events in the table when IA1 is conducted, the information fraction will be calculated as the actual number of events at IA1 over the target number of events at IA2.

The final PFS analysis will use the remaining Type I error that was not spent at the earlier analysis. The p-value bound at the final PFS analysis will be calculated by considering the correlation between the test statistics as determined by the actual number of PFS events at IA1 and IA2 of the study.

Table 9 Efficacy Boundaries and Properties for Progression-free Survival Analyses

Analysis	Value	α=0.001	α=0.0035	α=0.0225	α=0.025
IA1: 81%* N = 500 Events: 350 Month: 24	Z	3.4757	3.0428	2.2833	2.2344
	p (1-sided) [a]	0.0003	0.0012	0.0112	0.0127
	HR at bound[b]	0.6893	0.7221	0.7832	0.7873
	P(Cross) if HR=1[c]	0.0003	0.0012	0.0112	0.0127
	P(Cross) if HR=0.5[d]	0.9988	0.9997	1.0000	1.0000
IA2 N = 500 Events: 432 Month: 30	Z	3.1149	2.7327	2.0698	2.0276
	p (1-sided) [a]	0.0009	0.0031	0.0192	0.0213
	HR at bound[b]	0.7408	0.7686	0.8193	0.8226
	P(Cross) if HR=1[c]	0.0010	0.0035	0.0225	0.0250
	P(Cross) if HR=0.5[d]	1.0000	1.0000	1.0000	1.0000

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated approximately.
*Percentage of the target number of events at final analysis anticipated at interim analysis
[a]*p* (1-sided) is the nominal α for testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P(Cross if HR=0.5) is the probability of crossing a bound under the alternative hypothesis.

3.8.3 Overall Survival

The study will test OS at IA2, IA3 and FA. Following the multiplicity strategy as outlined in Figure 1, the OS hypothesis may be tested at α=0.0215 (initially allocated α), or α=0.0225 (if the PFS null hypothesis is rejected but not the ORR hypothesis), or α=0.025 (if both the PFS and ORR null hypotheses are rejected). For the superiority hypothesis, a Lan-DeMets O'Brien-Fleming alpha spending function is used to construct group sequential boundaries to control the Type I error rate. Table 10 shows the boundary properties for each of these α levels for the OS analysis. The bounds provided in the table are based on the assumption that the number of events at IA2, IA3 and FA are 258, 326 and 361, respectively. At the time of an analysis, the observed number of events may differ substantially from the expected. To avoid overspending at an interim analysis and keep reasonable alpha for the final analysis, the minimum alpha spending strategy will be adopted. At an IA, the information fraction used in Lan-DeMets spending function to determine the alpha spending at the IA will be based on the minimum of the expected information fraction and the actual information fraction at each analysis.

- In the scenario that the event accumulation is faster than expected, ie, if the number of observed events exceeds the expected number of events at an interim analysis, then the information fraction will be calculated as the expected number of events at the interim analysis over the target number of events at FA.

- In the scenario that the event accumulation is slower than expected and number of events is less than the expected number of events in the table when an interim analysis is conducted, the information fraction will be calculated as the actual number of events at the interim analysis over the target number of events at FA.

The final OS analysis will use the remaining Type I error that has not been spent at the earlier analyses. The event counts for all analyses will be used to compute correlations.

Of note, while the information fraction used for alpha spending calculation will be the minimum of the actual information fraction and the expected information fraction, the correlations required for deriving the bounds will still be computed using the actual information fraction based on the observed number of events at each analysis over the target number of events at FA.

Table 10 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	**Value**	**α=0.0215**	**α=0.0225**	**α=0.025**
IA2: 71%* N = 500 Events: 258 Month: 30	Z	2.4857	2.4636	2.4120
	p (1-sided)[a]	0.0065	0.0069	0.0079
	HR at bound[b]	0.7334	0.7355	0.7403
	P(Cross) if HR=1[c]	0.0065	0.0069	0.0079
	P(Cross) if HR=0.7[d]	0.6476	0.6557	0.6747
IA3: 90%* N: 500 Events: 326 Month: 38	Z	2.2091	2.1901	2.1457
	p (1-sided)[a]	0.0136	0.0143	0.0159
	HR at bound[b]	0.7828	0.7845	0.7884
	P(Cross) if HR=1[c]	0.0155	0.0163	0.0183
	P(Cross) if HR=0.7[d]	0.8509	0.8553	0.8655
FA N: 500 Events: 361 Month: 44	Z	2.1297	2.1119	2.0702
	p (1-sided)[a]	0.0166	0.0173	0.0192
	HR at bound[b]	0.7991	0.8006	0.8042
	P(Cross) if HR=1[c]	0.0215	0.0225	0.0250
	P(Cross) if HR=0.7[d]	0.9070	0.9100	0.9169

Analysis	Value	α=0.0215	α=0.0225	α=0.025
Abbreviations: FA = final analysis; HR = hazard ratio; IA = interim analysis.				
The number of events and timings are estimated approximately.				
*Percentage of the target number of events at final analysis anticipated at interim analysis.				
[a]*p* (1-sided) is the nominal α for testing.				
[b]HR at bound is the approximate HR required to reach an efficacy bound.				
[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.				
[d]P(Cross if HR=0.7) is the probability of crossing a bound under the alternative hypothesis.				

The minimum spending approach assumes timing is not based on any observed Z-value and thus the Z test statistics used for testing conditioned on timing are multivariate normal. Given the probabilities derived with the proposed spending method, the correlations based on actual event counts are used to compute bounds that control the Type I error at the specified alpha level for a given hypothesis conditioned on the interim analysis timing. Since this is true regardless of what is conditioned on, the overall Type I error for a given hypothesis unconditionally is controlled at the specified level. By using more conservative spending early in the study, power can be retained to detect situations where the treatment effect may be delayed.

3.8.4 Safety Analyses

The DMC has responsibility for assessment of overall risk/benefit. When prompted by safety concerns, the DMC can request corresponding efficacy data. DMC review of efficacy data to assess the overall risk/benefit to study participants will not require a multiplicity adjustment typically associated with a planned efficacy interim analysis.

3.9 Sample Size and Power Calculations

The study will randomize 500 participants in a 1:1 ratio into the pembrolizumab + lenvatinib arm and pembrolizumab + placebo arm. ORR, PFS and OS are primary endpoints for the study.

Based on the 350 participants with at least 6 months of follow-up, the power of the ORR testing at the initially allocated α=0.0025 (1-sided) is approximately 90.4% to detect a 20-percentage point difference between an underlying 20% response rate in the control arm and a 40% response rate in the experimental arm.

For the PFS endpoint, based on a target number of 432 events at the final analysis and 1 interim analysis at approximately 81% of the target number of events, the study has approximately >99.9% power to detect a hazard ratio of 0.5 at the initially allocated α=0.001 (1-sided).

For the OS endpoint, based on a target number of 361 events and 2 interim analyses at approximately 71% and 90% of the target number of events, the study has approximately 90.7% power to detect a hazard ratio of 0.7 at the initially allocated α=0.0215 (1-sided).

Based on [Burtness, B., et al 2018], the above sample size and power calculations for PFS and OS assume the following:

- PFS follows an exponential distribution with a median of 3 months for the control group.
- OS follows an exponential distribution with a median of 12 months for the control group.
- Enrollment period of 24 months with enrollment ramp-up over first 6 months.
- An annual dropout rate of 5% for PFS and OS.
- A follow-up period of 6 and 20 months for PFS and OS, respectively, after the last participant is randomized.

The sample size and power calculations were performed using R ("gsDesign" package) and SAS 9.4.

3.10 Subgroup Analyses

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

- Based on eCRF value:
    - PD-L1 tumor expression as determined by PD-L1 immunohistochemistry (TPS <50% vs. ≥ 50%)
    - HPV status for oropharynx cancer as determined by p16 immunohistochemistry (positive vs. negative); HPV status for participants without oropharynx cancer (eg, cancers of the oral cavity, hypopharynx and larynx) is considered HPV-negative.
    - ECOG performance status (0 vs. 1)
- Age category (<65 years, ≥65 years)
- Sex (female, male)
- Race (white, all others)
- Geographic region (North America, European Union, Asia, Rest of the World)
- PD-L1 tumor expression as determined by PD-L1 immunohistochemistry (CPS <50 vs. ≥50)
- Baseline disease presentation (recurrent, metastatic, recurrent and metastatic)

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

A forest plot will be produced, which provides the estimated point estimates and confidence intervals for the treatment effect across the categories of subgroups listed above. If the number of participants in a category of a subgroup variable is less than 10% of the ITT population, the subgroup analysis will not be performed for this category of the subgroup variable, and this

subgroup variable will not be displayed in the forest plot. The subgroup analyses for PFS and OS will be conducted using an unstratified Cox model, and the subgroup analyses for ORR will be conducted using the unstratified Miettinen and Nurminen method.

3.11 Compliance (Medication Adherence)

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

3.12 Extent of Exposure

Extent of Exposure for a participant is defined as the number of cycles and number of days in which the participant receives the study intervention. Summary statistics will be provided on the extent of exposure for the overall study intervention, and for pembrolizumab and lenvatinib separately for the APaT population.

4 STATISTICAL ANALYSIS PLAN FOR EXTENTSION PER COUNTRY-SPECIFIC REQUIREMENT

The study plans to enroll approximately 150 participants in China (30% of the global sample size) to meet China local registration needs. If this sample size could not be achieved before the completion of enrollment of the global portion, the study may remain open to enrollment in China in an extension portion.

After the global portion enrollment is closed, participants from China will continue to be enrolled in a 1:1 ratio between 2 treatment groups: (1) the pembrolizumab + lenvatinib arm and (2) the pembrolizumab + placebo arm until the sample size in the China subpopulation reaches approximately 150 in total with global and extension portions combined. The extension portion is intended to be identical to the global portion in general (e.g., inclusion and exclusion criteria, primary and secondary endpoints, study procedures), with the additional statistical analysis plan for the China subpopulation. The purpose of this extension portion is to ensure sufficient number of China participants per China local registration requirement to evaluate the consistency of efficacy and safety between the China subpopulation and the global population.

This section outlines the statistical analysis strategy and procedures for China subpopulation. Here and throughout this whole section (Section 4), **China subpopulation** and China participants refers to all China participants randomized in the global portion and the extension portion unless specifically described.

After the cut-off date of the primary analysis of the global portion (including IA and/or FA), all participants in the China subpopulation will continue their randomized treatment and continue to be followed up for China registration purpose. The extension portion will be completed at the time of the analysis for OS in the China subpopulation, which will be conducted when all 150 China participants have been enrolled and approximately a total of 109 OS events have been observed in the China subpopulation.

4.1 Statistical Analysis Plan for China Subpopulation Summary

Key elements of the statistical analysis plan for China subpopulation are summarized below. More details are provided in following sections.

Analysis Populations	Efficacy: ITT China subpopulation (including China participants randomized in the global portion and the extension portion) Safety: APaT China subpopulation (including China participants randomized in the global portion and the extension portion who received at least 1 dose of study intervention) Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary efficacy and safety analysis populations for the global portion.
Efficacy Endpoint(s)	Efficacy endpoints are the same as described in Section 3.4.1

Safety Endpoint (s)	Safety endpoints are the same as described in Section 3.4.2
Statistical Methods for Efficacy Analyses	No formal hypothesis testing is planned, and no multiplicity adjustment will be applied to the analysis for China subpopulation. The analyses in the China subpopulation will be conducted using an unstratified Cox model for PFS and OS, and an unstratified Miettinen and Nurminen method for ORR.
Statistical Methods for Safety Analyses	Safety analyses in the China subpopulation will be the same as those for the global portion as described in Section 3.6.2 if applicable.
Summaries of Baseline Characteristics and Demographics	Same approach as that for the global portion as described in Section 3.6.4 will be implemented for the China subpopulation.
Analyses Timing	The PFS analysis in the China subpopulation will be conducted when at least 100 China participants have been enrolled and approximately a total of 77 PFS events have been observed in the China subpopulation. The OS analysis in the China subpopulation will be conducted when all 150 China participants have been enrolled and approximately a total of 109 OS events have been observed in the China subpopulation. An analysis in the China subpopulation may be conducted at the same time as the primary analysis of the global portion (including IA and/or FA). More details are included in Section 4.3.
Hypotheses and Multiplicity	No hypothesis testing is planned for the China subpopulation analyses. No multiplicity adjustment will be applied to the analysis of China subpopulation.
Sample Size Calculations	After the completion of global portion enrollment, the extension portion will continue to enroll and randomize eligible participants until the sample size for the overall randomized China subpopulation reaches approximately 150. It is estimated that there will be ~ 109 deaths in the China subpopulation at the time of the OS analysis. With 109 deaths, there is ~ 85% chance to observe a point estimate of hazard ratio for OS that preserves approximately at least 50% of the empirical risk reduction from the analysis for the global portion assuming the underlying hazard ratio is 0.7.
Subgroup Analyses	Additional analyses may be considered for the China subgroup (including China participants randomized in the global portion only) based on the Sponsor's discretion and/or consultation with regulatory if the primary analysis of the global portion (including IA and/or FA) shows positive results leading to regulatory submission. Country-specific analysis may also be conducted per local regulatory requirement.

4.2 Responsibility for Analyses/In-House Blinding

For all participants in the China subpopulation, participant level treatment randomization information will be blinded to the designated team of the Sponsor responsible for China analysis, until the database lock for extension portion is achieved. The extent to which individuals are unblinded to the results will be limited and documented by the unblinded team.

4.3 Analyses Timing

The PFS analysis in the China subpopulation will be conducted when at least 100 China participants have been enrolled and approximately a total of 77 PFS events have been observed.

The OS analysis in the China subpopulation will be conducted when all 150 China participants enrolled and approximately a total of 109 OS events have been observed.

If the timing criteria for an analysis in the China subpopulation are met before the primary analysis of the global portion (including IA or the FA), the corresponding analysis for China subpopulation may occur at the same time as the IA or the FA for the global portion.

If the primary analysis of the global portion (including IA and/or FA) has demonstrated statistical significance, and it is projected that the timing criteria for an analysis in the China subpopulation will be met within ~3 months of the cut-off date of the primary analysis of the global portion, then the analysis for the China subpopulation including the extension portion may be based on the same database lock as the IA or the FA for the global portion.

4.4 Sample Size Calculations

After the completion of global portion enrollment, the extension portion will continue to enroll participants and randomize eligible participants until the sample size for the overall randomized China subpopulation reaches approximately 150. Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary analysis population for the global portion.

Sample size is determined to provide more than 80% chance of observing the point estimate of treatment effect in the China subpopulation preserves approximately at least 50% of treatment effect (e.g. empirical risk reduction) from the global primary efficacy analysis with the same assumption of treatment effect used in the sample size and power calculation for the global portion.

With 77 PFS events, the China subpopulation has approximately 97% chance to observe a point estimate of hazard ratio for PFS that preserves approximately at least 50% of the empirical risk reduction from the analysis for the global portion assuming the underlying hazard ratio is 0.5.

With 109 OS events, the China subpopulation has approximately 85% chance to observe a point estimate of hazard ratio for OS that preserves approximately at least 50% of the empirical risk reduction from the analysis for the global portion assuming the underlying hazard ratio is 0.7.

The above calculations in PFS and OS are based on the same assumptions of treatment effect applied in the global portion for sample size and power evaluation as specified in Section 3.9.

5 REFERENCES

[Bjordal, K., et al 2000]	Bjordal K, de Graeff A, Fayers PM, Hammerlid E, van Pottelsberghe C, Curran D, et al. A 12 country field study of the EORTC QLQ-C30 (version 3.0) and the head and neck cancer specific module (EORTC QLQ-H&N35) in head and neck patients. EORTC Quality of Life Group. Eur J Cancer. 2000 Sep;36(14):1796-807.	[040VGM]
[Burtness, B., et al 2018]	Burtness B, Harrington KJ, Greil R, Soulieres D, Tahara M, De Castro G Jr, et al. KEYNOTE-048: phase III study of first-line pembrolizumab (P) for recurrent/metastatic head and neck squamous cell carcinoma (R/MHNSCC) [abstract]. Presented at: European Society for Medical Oncology (ESMO) 2018 Congress; 2018 Oct 19-23; Munich (Germany). Ann Oncol. 2018 Oct;29(suppl 8):viii729. Abstract no. LBA8_PR.	[053ZSC]
[Clopper, C. J. and Pearson, E. S. 1934]	Clopper CJ and Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404-13.	[03Y75Y]
[Cocks, K., et al 2012]	Cocks K, King MT, Velikova G, de Castro G , Martyn St-James M, Fayers PM, et al. Evidence-based guidelines for interpreting change scores for the European organisation for the research and treatment of cancer quality of life questionnaire core 30. Eur J Cancer. 2012 Jul;48(11):1713-21.	[04SL8J]
[EuroQol Research Foundation 2021]	EQ-5D [Internet]. Rotterdam (Netherlands): EuroQol Research Foundation; c2023. EuroQoL group. EQ-5D; 2021 Nov 30 [cited 2023 Feb 28]; [about 3 screens]. Available from: https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/.	[0886KB]

[King, M. T. 1996]	King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Quality of Life Research 1996;5:555-67.	[03Q4R5]
[Liang, K-Y and Zeger, S. L. 2000]	Liang K-Y, Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankyha: The Indian Journal of Statistics 2000;62(Series B, Part 1):134-48.	[00V5V6]
[Maurer, W. and Bretz, F. 2013]	Maurer W and Bretz F. Multiple testing in group sequential trials using graphical approaches. Stat Biopharm Res 2013;5(4):311-20.	[03XQVB]
[Miettinen, O. and Nurminen, M. 1985]	Miettinen O and Nurminen M. Comparative analysis of two rates. Stat Med 1985;4:213-26.	[00VMQY]
[Osoba, D., et al 1998]	Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139-44.	[03RL72]
[Scott, N. W., et al 2008]	Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. This manual presents reference data for the QLQ-C30 based upon data provided by EORTC Quality of Life Group Members and other users of the QLQ-C30 [Internet]. Belgium: EORTC Quality of Life Group; 2008. Available from: http://groups.eortc.be/qol/sites/default/files/img/newsletter/reference_values_manual2008.pdf.	[04HN7P]
[van Hout, B., et al 2012]	van Hout B, Janssen MF, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012;15:708-15.	[058TT2]

6 APPENDIX

6.1 Technical Note for The Minimum Spending Approach

The Lan-DeMets spending function to approximate an O'Brien-Fleming bound is defined as

$$f(t;\alpha) = 2 - 2\Phi\left(\frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{\sqrt{t}}\right)$$

where t in $f(t;\alpha)$ is the spending time, which is not necessarily information fraction or actual time.

The test statistics Z_i at each analysis i is assumed to follow a multivariate normal distribution with expectations $E(Z_i) = \theta\sqrt{I_i}$ and covariances $Cov\left(Z_i, Z_j\right) = \sqrt{I_i/I_j}$ where θ is the treatment effect difference of interest and I_i is the actual statistical information available based on the actual observed event number.

To illustrate how the minimun spending approach is implemented, examples with 2 hypothetical scenarios where events accrue slower and faster than expected are given in Table 11 for the OS analyses with the total alpha of 2.15% (initially allocated). There are 3 planned analyses for OS at IA2, IA3 and FA, respectively.

IA2 boundary calculation:

For the first OS interim analysis at IA2, the p-value boundary is the same as alpha spending determined from the Lan-DeMets spending function. At the time of the analysis, 258 events are expected over the target 361 events at the FA.

- Hypothetical scenario 1 (events accrue slower than expected): 245 events are observed. The spending time is calculated as t = 245/361 = 67.9% and p-value boundary = 0.0053.
- Hypothetical scenario 2 (events accrue faster than expected): 270 events are observed. The spending time is calculated as t = 258/361 = 71%, p-value boundary = 0.0065.

IA3 boundary calculation:

The alpha spending α_2 for OS at IA3 is determined from the Lan-DeMets spending function with spending time calculated as follows:

- Hypothetical scenario 1 (events accrue slower than expected): 305 events are observed. The spending time is calculated as t = 305/361 = 84.5%.
- Hypothetical scenario 2 (events accrue faster than expected): 358 events are observed. The spending time is calculated as t = 326/361 = 90%.

The boundary (C_2) is solved from $P(Z_2 \geq C_2, Z_1 < C_1 | H_0) = \alpha_2$, with test statistics Z_1 and Z_2 being multivariate normal and correlations based on observed event numbers at each analysis.

FA boundary calculation:

The alpha spending at the FA is $\alpha - \alpha_1 - \alpha_2$. FA boundary ($C_3$) is solved from $P(Z_3 \geq C_3, Z_1 < C_1, Z_2 < C_2 | H_0) = \alpha - \alpha_1 - \alpha_2$, with test statistics Z_1, Z_2 and Z_3 being multivariate normal and correlations based on observed event numbers.

Table 11 below summarizes the boundary properties of 2 hypothetical scenarios, together with the planned scenario for OS.

Table 11 Efficacy Boundaries and Properties for Overall Survival Analyses (Planned and Hypothetical Scenarios)

Value	Planned scenario	Hypothetical scenario 1 (events accrue slower)	Hypothetical scenario 2 (events accrue faster)
IA2: Month: 30			
Events (I.F.)	258 (71%)	245 (67.9%*)	270 (71%*)
Z	2.4857	2.5575	2.4857
p (1-sided)	0.0065	0.0053	0.0065
HR at bound	0.7334	0.7212	0.7389
P(Cross) if HR=1	0.0065	0.0053	0.0065
P(Cross) if HR=0.7	0.6476	0.5943	0.6735
IA3: Month: 38			
Events (I.F.)	326 (90%)	305 (84.5%*)	358 (90%*)
Z	2.2091	2.2962	2.2194
p (1-sided)	0.0136	0.0108	0.0132
HR at bound	0.7828	0.7688	0.7909
P(Cross) if HR=1	0.0155	0.0124	0.0155
P(Cross) if HR=0.7	0.8509	0.8024	0.8827
FA: Month: 44			
Events	361	345	380
Z	2.1297	2.0886	2.1005
p (1-sided)	0.0166	0.0184	0.0178
HR at bound	0.7991	0.7986	0.8061
P(Cross) if HR=1	0.0215	0.0215	0.0215
P(Cross) if HR=0.7	0.9070	0.8970	0.9229

* Information fraction is the minimum of the expected information fraction and the actual information fraction used for the alpha spending calculation. The actual information fraction based on the observed number of events is used for computing correlations needed to derive the group sequential bounds.

I.F. = information fraction.

6.2 Technical Note for The cLDA Model in PRO Analysis

The cLDA model assumes a common mean across treatment groups at baseline and a different mean for each treatment at each of the post-baseline time points. In this model, the response vector consists of baseline and the values observed at each post-baseline time point. Time is

treated as a categorical variable so that no restriction is imposed on the trajectory of the means over time. The cLDA model is specified as follows:

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

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