Official Protocol Title:	MK-7902-ssap-p011-04-amend03_final-redaction
NCT number:	NCT03898180
Document Date:	14-Oct-2024

MK-7902 PROTOCOL 011-004

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

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

2 SUMMARY OF CHANGES

The sSAP has been amended to include the following change:

- Add second non-binding futility analysis timing and rules for ORR and PFS to Sections 3.1, 3.7.1, and 3.8.1
- Sections 3.3, 3.4.3, 3.4.4, 3.6.3, 3.6.3.1 – 3.6.3.4 narrowed the scope of ePRO's TLF's to include only those specific items which are required for Disclosure tables and clinicaltrials.gov

3 STATISTICAL ANALYSIS PLAN

This section outlines the statistical analysis strategy and procedures for the study. If, after the study has begun, but before any unblinding/final database lock, changes are made to primary and/or key secondary hypotheses, or the statistical methods related to those hypotheses, then the protocol will be amended (consistent with ICH Guideline E9). Changes to exploratory or other nonconfirmatory analyses made after the protocol has been finalized, but before unblinding/final database lock, will be documented in this supplemental statistical analysis plan (sSAP) and referenced in the clinical study report (CSR) for the study. Post hoc exploratory analyses will be clearly identified in the CSR. Other planned analyses (ie, those specific to PK data and PROs) will be documented in separate analysis plans.

3.1 STATISTICAL ANALYSIS PLAN SUMMARY

Key elements of the statistical analysis plan (SAP) are summarized below. The comprehensive plan is provided in Sections 3.2 through 3.12.

Study Design Overview	A Phase 3, Randomized, Double-blind Study to Compare the Efficacy and Safety of Pembrolizumab (MK-3475) in Combination with Lenvatinib (E7080/MK-7902) Versus Pembrolizumab and Placebo as First Line Treatment for Locally Advanced or Metastatic Urothelial Carcinoma in Cisplatin-ineligible Participants Whose Tumors Express PD-L1, and in Participants Ineligible for Any Platinum-containing Chemotherapy Regardless of PD-L1 Expression (LEAP-011)

Treatment Assignment	Approximately 694 participants will be randomized in a 1:1 ratio between 2 treatment arms: (1) pembrolizumab + lenvatinib and (2) pembrolizumab + placebo. Stratification factors are as follows: • Ineligible for any platinum-containing chemotherapy, PD-L1 CPS ≥10, ECOG PS 2 • Ineligible for any platinum-containing chemotherapy, PD-L1 CPS <10, ECOG PS 2 • Cisplatin-ineligible, PD-L1 CPS ≥10, ECOG PS 2 • Cisplatin-ineligible, PD-L1 CPS ≥10, ECOG PS 0 or 1.

Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Participants as Treated (APaT)
Primary Endpoints	• PFS per RECIST 1.1 by BICR • OS
Key Secondary Endpoint	• ORR per RECIST 1.1 by BICR
Statistical Methods for Key Efficacy Analyses	The primary hypotheses will be evaluated by comparing pembrolizumab + lenvatinib to pembrolizumab + placebo with respect to PFS and OS using a stratified log-rank test. HR 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 difference in ORR will be estimated using the stratified Miettinen and Nurminen method with strata weighted by sample size.
Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. The tiers differ with respect to the analyses that will be performed. There are no events of interest that warrant elevation to Tier 1 in this study. Tier 2 parameters will be assessed via point estimates with 95% CIs provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters. The 95% CIs for the between-treatment differences in percentages will be provided using the Miettinen and Nurminen method.

Interim Analyses	Two interim efficacy analyses and two interim futility analysis will be performed in this study. Results will be reviewed by an external DMC. These interim analyses are summarized below. Details are provided in Section 3.7.
	• Futility Analysis 1
	  ○ Timing: To be performed when 297 participants have been enrolled, and ≥151 PFS events (27% of targeted events) are observed.
	  ○ Analysis: A non-binding futility analysis on PFS and ORR will be performed.
	• Futility Analysis 2
	  ○ Timing: To be performed when 370 participants have been enrolled, and ≥191 PFS events (30% of targeted events) are observed.
	  ○ Analysis: A non-binding futility analysis on PFS and ORR will be performed.
	• Interim Analysis 1
	  ○ Timing: To be performed when enrollment is complete and ≥530 PFS events (95% of targeted events) and 386 deaths (70% of targeted events) are observed.
	  ○ Testing: Superiority analyses for PFS and OS will be provided. Superiority analysis for ORR will be provided if both PFS and OS are positive.
	• Interim Analysis 2
	  ○ Timing: To be performed when enrollment is complete and ≥558 PFS events (100% of targeted events) and 469 deaths (85%) of targeted events) are observed.
	  ○ Testing: Superiority analyses for PFS and OS will be provided.
	• Final Analysis
	  ○ Timing: To be performed when 552 deaths (100%) of targeted events) are observed.
	  ○ Testing: Superiority analysis for OS will be provided.

Multiplicity	The type I error rate over the multiple endpoints will be controlled by the Lan-DeMets [Ref. 5.4: 03P3QC] and O'Brien-Fleming [Ref. 5.4: 00VPPS] group sequential methods using the graphical approach of Maurer and Bretz [Ref. 5.4: 045MYM].
Sample Size and Power	The planned sample size is 694 participants. For PFS, the study has 94.7% power to demonstrate that pembrolizumab + lenvatinib is superior to pembrolizumab + placebo at an overall 1-sided α level of 0.005, if the underlying treatment comparison in HR in PFS is 0.7. For OS, the study has 90.1% power to demonstrate that pembrolizumab + lenvatinib is superior to pembrolizumab + placebo at an overall 1-sided α level of 0.02, if the underlying treatment comparison in HR in OS is 0.75.

3.2 RESPONSIBILITY FOR ANALYSES/IN-HOUSE BLINDING

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

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

The Clinical Biostatistics Department will generate the randomized allocation schedule(s) for study intervention assignment.

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

3.3 HYPOTHESES/ESTIMATION

The objectives and endpoints apply to a study population of male or female participants at least 18 years of age with a histologically confirmed diagnosis of advanced/unresectable or metastatic urothelial carcinoma (UC), who are cisplatin-ineligible and whose tumors express programmed death ligand 1 (PD-L1) (combined positive score [CPS] ≥10), or who are medically ineligible to receive any platinum-based chemotherapy.

This study will be considered to have met its primary objective if pembrolizumab +lenvatinib is superior to pembrolizumab + placebo for either primary endpoint.

Progression-free survival (PFS), objective response rate (ORR), duration of response (DOR), and disease control rate (DCR) will be assessed per Response Evaluation Criteria in Solid Tumors Version 1.1 (RECIST 1.1) modified to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ.

Primary Objectives	Primary Endpoints
• To compare pembrolizumab + lenvatinib to pembrolizumab + placebo with respect to PFS per Response Evaluation Criteria in Solid Tumors Version 1.1 (RECIST 1.1) by blinded independent central review (BICR). • Hypothesis 1**:** Pembrolizumab + lenvatinib is superior to pembrolizumab + placebo with respect to PFS per RECIST 1.1 by BICR.	• PFS, defined as the time from randomization to the first documented progressive disease (PD) or death from any cause, whichever occurs first.
• To compare pembrolizumab + lenvatinib to pembrolizumab + placebo with respect to overall survival (OS). • Hypothesis 2: Pembrolizumab + lenvatinib is superior to pembrolizumab + placebo with respect to OS.	• OS, defined as the time from randomization to the date of death from any cause.
Secondary Objectives	Secondary Endpoints
To compare pembrolizumab + lenvatinib to pembrolizumab + placebo with respect to objective response rate (ORR) per RECIST 1.1 by BICR. • Hypothesis 3: Pembrolizumab + lenvatinib is superior to pembrolizumab + placebo with respect to ORR per RECIST 1.1 by BICR.	• Objective response (OR), defined as a confirmed complete response (CR) or partial response (PR).
To evaluate the safety and tolerability of treatment with pembrolizumab + lenvatinib versus pembrolizumab + placebo.	• Adverse events (AEs) and discontinuations due to AEs.
To evaluate pembrolizumab + lenvatinib and pembrolizumab + placebo with respect to duration of response (DOR) per RECIST 1.1 by BICR.	• DOR, defined as the time from the first documented evidence of CR or PR to the earliest date of PD or death due to any cause, whichever comes first, for individuals with a confirmed CR or PR.
To evaluate pembrolizumab + lenvatinib and pembrolizumab + placebo with respect to disease control rate (DCR) per RECIST 1.1 by BICR.	• Disease control, defined as a confirmed response of CR or PR or stable disease (SD).

Primary Objectives	Primary Endpoints
To evaluate changes in patient-reported outcomes (PROs) from baseline, and to evaluate time to deterioration (TTD) in European Organization for Research and Treatment of Cancer EORTC QLQ-C30 global health status/QoL score.	• Change from baseline in EORTC QLQ-C30 global health status/QoL score. • TTD, defined as the time from baseline to the first onset of PRO deterioration in EORTC QLQ-C30 global health status/QoL score.
Tertiary/Exploratory Objectives	Tertiary/Exploratory Endpoints
To identify molecular (genomic, metabolic, or proteomic) biomarkers that may be indicative of clinical response/resistance, safety, and/or the mechanism of action of pembrolizumab and lenvatinib in all participants.	• Molecular (genomic, metabolic, or proteomic) determinants of response or resistance to treatments, using blood and/or tumor tissue.
To assess the pharmacokinetics (PK) of lenvatinib when co-administered with pembrolizumab.	Population PK parameters including clearance, volume of distribution, and absorption rate constant.
To compare pembrolizumab + lenvatinib to pembrolizumab + placebo with respect to PFS per iRECIST (RECIST 1.1 for immune-based therapeutics) by investigator.	PFS, defined as the time from randomization to the first documented progressive disease (PD) or death from any cause, whichever occurs first.
To evaluate changes in PROs from baseline using the following instruments: • EORTC QLQ-C30	Change in PROs from baseline in • Scores for the QoL of the EORTC QLQ-C30

3.4 ANALYSIS ENDPOINTS

Efficacy and safety endpoints that will be evaluated for within- and/or between-treatment differences are listed below, followed by descriptions of the derivations of selected endpoints.

3.4.1 Efficacy Endpoints

Primary

- **Progression-free Survival (PFS):** PFS is defined as the time from randomization to the first documented disease progression (PD) per RECIST 1.1 by BICR, or death due to any cause, whichever occurs first. See Section 3.6.1.1 for the definition of censoring.

- **Overall Survival (OS):** OS is defined as the time from randomization to death due to any cause.

Secondary

- **Objective Response Rate (ORR):** ORR is the percentage of participants with a confirmed complete response (CR) or partial response (PR) per RECIST 1.1 by BICR.

- **Duration of Response (DOR):** DOR is defined as the time from the first documented evidence of CR or PR to PD per RECIST 1.1 by BICR or death due to any cause, whichever occurs first.

- **Disease Control Rate (DCR):** DCR is defined as the percentage of participants with a confirmed response of CR, PR, or SD per RECIST 1.1 by BICR. Stable disease must be achieved at ≥6 weeks after randomization to be considered a best overall response.

- **Progression-free Survival (PFS):** PFS is defined as the time from randomization to the first documented disease progression (PD) per iRECIST 1.1 by investigator, or death due to any cause, whichever occurs first. See Section 3.6.1.1 for the definition of censoring.

3.4.2 Safety Endpoints

Safety and tolerability will be assessed by clinical review of all relevant parameters including AEs, SAEs, fatal AEs, laboratory tests, and vital signs. Furthermore, specific events will be collected and designated as ECIs as described in protocol Section 8.4.7.

3.4.3 Patient-reported Outcome Endpoints

The following secondary PRO endpoints will be evaluated the PRO instruments are EORTC QLQ-C30 as described protocol Section 4.2.1.3.

- Change in patient-reported outcomes from baseline in Global health status/QoL.
- TTD in EORTC QLQ-C30 Global health status/QoL.

Time to deterioration (TTD) is defined as the time from baseline to the first onset of PRO deterioration. TTD in the EORTC QLQ-C30 global health status/QoL score has been used as a key PRO endpoint in the UC study, KEYNOTE-045 [Ref. 5.4: 050DSW]. Deterioration in the global health status/quality of life is defined as a 10 points or greater worsening from baseline, with or without subsequent confirmation, under a right-censoring rule. [Ref. 5.4: 00TWVP].

All ePRO analyses will be conducted using database cutoff date of 26JUL2021 to align with CSR P011V01MK7902.

3.4.4 Derivation 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 Osoba et al. (1998) [Osoba, D., et al 1998] and King (1996) [King, M. T. 1996], a 10 points or greater worsening from baseline for each scale of the EORTC QLQ-C30 represents a clinically relevant deterioration. Changes from baseline in EORTC QLQ-C30 scores will also be interpreted according to recent subscale-specific guidelines, which indicate that clinically meaningful differences vary by scale (Cocks et al., 2012).

3.5 ANALYSIS POPULATIONS

3.5.1 Efficacy Analysis Population

The analyses of the primary efficacy endpoints are based on the intention-to-treat (ITT) population. All randomized participants will be included in this population. Participants will be analyzed in the treatment group to which they are randomized. Details of the approach to handling missing data are provided in Section 3.6.

3.5.2 Safety Analysis Population

Safety analyses will be conducted in the All Participants as Treated (APaT) population, which consists of all randomized participants who received at least one dose of study intervention. Participants will be included in the treatment group corresponding to the study intervention 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 receive incorrect study intervention for the entire treatment period; such participants will be included in the treatment group corresponding to the study intervention actually received. Any participant who receives the incorrect study intervention for one cycle, but receives the randomized study intervention for all other cycles, will be analyzed according to the randomized treatment group, and a narrative will be provided for any events that occur during the cycle for which the participant is incorrectly dosed.

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

3.5.3 Patient-reported Outcome Analysis Populations

The analyses of PRO endpoints will be based on a quality of life-related full analysis set (FAS) population following the ITT principle and ICH E9 guidelines. This population consists of all randomized participants who have received at least 1 dose of study intervention and have completed at least 1 PRO assessment.

3.5.4 Pharmacokinetic Analysis Population

The population PK analysis set includes all participants who have received at least one dose of study intervention with documented dosing history in the pembrolizumab + lenvatinib arm, and have measurable plasma levels of lenvatinib or serum levels of pembrolizumab.

3.6 STATISTICAL METHODS

3.6.1 Statistical Methods for Efficacy Analyses

This section describes the statistical methods that address the primary and secondary efficacy objectives. Efficacy results that are deemed to be statistically significant after consideration of the type I error control strategy are described in Section 3.8. Nominal *p*-values will be computed for other efficacy analyses but should be interpreted with caution because of potential issues of multiplicity.

The stratification factors used for randomization (see protocol Section 6.3.2) will be applied to all stratified analyses, in particular, the stratified log-rank test, stratified Cox model, and stratified Miettinen and Nurminen method [Ref. 5.4: 00VMQY]. In the event that there are small strata, for the purpose of analysis, strata will be combined to ensure sufficient number of participants, responses and events in each stratum. Details regarding the pooling strategy will be pre-specified in a future sSAP amendment prior to the database lock for the first analysis when each applicable endpoint will be analyzed, and decisions regarding the pooling will be based on a blinded review of response and event counts by stratum.

The efficacy analyses for ORR, DOR and PFS will include responses and documented progression events that occur prior to second course treatment or protocol specified treatment crossover.

3.6.1.1 Progression-free Survival

The nonparametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment arm. The hypotheses of treatment difference in PFS will be tested by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to estimate the magnitude of the treatment comparison (ie, HR) between the treatment arms. The HR and its 95% CI from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization (protocol Section 6.3.2) will be applied to both the stratified log-rank test and the stratified Cox model.

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. For the primary analysis, for the participants 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 BICR, regardless of discontinuation of study

drug. Additional analyses will be performed for comparison of PFS based on the investigator's assessment and PFS analysis for PD per RECIST 1.1 by BICR.

To evaluate the robustness of the PFS endpoint per RECIST 1.1 by BICR, one primary and two sensitivity analyses with a different set of censoring rules will be performed. For the primary analysis, if the events (PD or death) are after more than 1 missed disease assessment, the data are censored at the last disease assessment before missing visits. Also data after new anticancer therapy are censored at the last disease assessment before the initiation of new anticancer therapy. The first sensitivity analysis follows ITT principles (ie, 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 or initiation of an anticancer treatment subsequent to discontinuation of study-specified treatments, 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 primary and sensitivity analyses are summarized in Table 1.

The proportional hazards assumption on PFS may be examined using both graphical and analytical methods if warranted. The log [-log] of the survival function vs. time for PFS will be plotted for the comparison between the two arms. If the curves are not parallel, indicating that hazards are not proportional, supportive analyses may be conducted to account for the possible non-proportional hazards effect associated with immunotherapies: for example, using the Restricted Mean Survival Time method [Ref. 5.4: 045X5X].

The RMST is the population average of the amount of event-free survival time experienced during a fixed study follow-up time. This quantity can be estimated by the area under the Kaplan-Meier curve up to the follow-up time. The clinical relevance and feasibility should be taken into account in the choice of follow-up time to define RMST (e.g., near the last observed event time assuming that the period of clinical interest in the survival experience is the whole observed follow-up time for the study, but avoiding the very end of the tail where variability may be high); a description of the RMST as a function of the cutoff time may be of interest. The difference between two RMSTs for the two treatment groups will be estimated and 95% CI will be provided.

A sensitivity analysis may be performed based on the MaxCombo test with logrank(FH(0,0)), FH (0, 1), FH (1, 1) at the final analysis of PFS to account for the potential loss of power with logrank test when the proportional hazard assumption is violated.

Table 1 Censoring Rules for Primary and Sensitivity Analyses of Progression-free Survival

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	Progressed at treatment discontinuation due to reasons other than complete response; otherwise, censored at last disease assessment if still receiving study intervention or completed study intervention
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
PD or death documented after ≤1 missed disease assessment and before new anticancer treatment	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 immediately after ≥2 consecutive missed disease assessments or after new anticancer treatment	Censored at last disease assessment prior to the earlier date of ≥ 2 consecutive missed disease assessments and new anticancer treatment	Progressed at date of documented PD or death	Progressed at date of documented PD or death
Abbreviation: PD = progressive disease.			

3.6.1.2 Overall Survival

The nonparametric Kaplan-Meier method will be used to estimate the OS curve in each treatment arm. The hypotheses of treatment difference in OS will be tested by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to estimate the magnitude of the treatment comparison (ie, HR) between the treatment arms. The HR and its 95% CI from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization (protocol 6.3.2) will be applied to both the stratified log-rank test and the stratified Cox model. Participants without documented death at the time of analysis will be censored at the date of last known contact.

The proportional hazards assumption on OS may be examined using both graphical and analytical methods if warranted. The log [-log] of the survival function vs. time for OS will be plotted for the comparison between the two arms. If the curves are not parallel, indicating that hazards are not proportional, supportive analyses may be conducted to account for the possible non-proportional hazards effect associated with immunotherapies: for example, using the Restricted Mean Survival Time method [Ref. 5.4: 045X5X].

The RMST is the population average of the amount of event-free survival time experienced during a fixed study follow-up time. This quantity can be estimated by the area under the Kaplan-Meier curve up to the follow-up time. The clinical relevance and feasibility should be taken into account in the choice of follow-up time to define RMST (e.g., near the last observed event time assuming that the period of clinical interest in the survival experience is the whole observed follow-up time for the study, but avoiding the very end of the tail where variability may be high); a description of the RMST as a function of the cutoff time may be of interest. The difference between two RMSTs for the two treatment groups will be estimated and 95% CI will be provided.

A sensitivity analysis may be performed based on the MaxCombo test with logrank, FH (0, 1), FH (1, 1) at the final analysis of OS to account for the potential loss of power with logrank test when the proportional hazard assumption is violated.

3.6.1.3 Objective Response Rate

The stratified Miettinen and Nurminen method will be used for comparison of ORR between the treatment groups. The difference in ORR and its 95% CI from the stratified Miettinen and Nurminen method with strata weighting by sample size will be provided. The same stratification factors used for randomization (protocol Section 6.3.2) will be used as stratification factors in the analysis.

3.6.1.4 Analysis Strategy for Key Efficacy Endpoints

A summary of the primary analysis strategy for the key efficacy endpoints is provided in Table 2.

Table 2 Analysis Methods for Key Efficacy Endpoints

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Primary Analyses			
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 last known alive date
Key Secondary Analyses			
ORR per RECIST 1.1 by BICR	Testing and estimation: stratified Miettinen and Nurminen method	ITT	Participants with missing data are considered nonresponders
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).

3.6.1.5 Duration of Response

If sample size permits, DOR will be summarized descriptively using the Kaplan-Meier method. Range and median survival time will be reported. Only the subset of participants who show a confirmed CR or PR will be included in this analysis.

Censoring rules for DOR are summarized in Table 3. For the 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 3 Censoring Rules for Duration of Response

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new anticancer treatment initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anticancer treatment initiated	Last adequate disease assessment before new anticancer treatment 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 treatment, if any	Censor (non-event)
Death or progression after ≤1 missed disease assessment and before new anticancer treatment, if any	PD or death	End of response (event)
Abbreviation: PD = progressive disease. A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response.		

3.6.1.6 Disease Control Rate

The stratified Miettinen and Nurminen method will be used for comparison of DCR between the treatment arms. The difference in DCR and its 95% CI from the stratified Miettinen and Nurminen method with strata weighting by sample size will be provided. The same stratification factors used for randomization (protocol Section 6.3.2) will be used as stratification factors in the analysis.

3.6.2 Statistical Methods for Safety Analyses

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

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

Tier 1 Events
Safety parameters or AEs 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. Similarly, the combination

of pembrolizumab and lenvatinib has not been associated with any new safety signals. Therefore, there are no Tier 1 events for this protocol.

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 [Ref. 5.4: 00VMQY].

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

Tier 3 Events

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

Continuous Safety Measures

For continuous measures such as changes from baseline in laboratory, vital signs, and ECG 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 combo arm, AE incidence adjusted for treatment exposure analyses may be performed as appropriate.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Any AE (≥10% of participants in one of the treatment groups)	X	X
	Any Grade 3 to 5 AE (≥5% of participants in one of the treatment groups)	X	X
	Any serious AE (≥5% of participants in one of the treatment groups)	X	X
Tier 3	Any AE		X
	Change from baseline results (laboratory test toxicity grade)		X
Abbreviations: AE = adverse event; CI = confidence interval; X = results will be provided.			

3.6.3 Analysis Methods for Patient-reported Outcome Endpoints

This section describes the planned analyses for the PRO endpoints.

EORTC QLQ-C30 Scoring: Each scale or item is scored between 0 and 100, according to the EORTC QLQ-C30 standard scoring algorithm {04HN7P}. For global health status/quality of life, a higher value indicates a better level of function.

PRO Scoring Algorithm

The QLQ-C30 is composed of both multi-item scales and single-item measures. These include five functional scales, three symptom scales, a global health status / QoL scale, and six single items. Each of the multi-item scales includes a different set of items - no item occurs in more than one scale.
EORTC QLQ-C30 Scoring: All the scales and single-item measures range in score from 0 to 100. A high scale score represents a higher response level.
Thus, a high score for a functional scale represents a *high / healthy level of functioning*, a high score for the global health status / QoL represents a *high QoL*, but a high score for a symptom scale / item represents a *high level of symptomatology / problems*., according to the EORTC QLQ-C30 standard scoring algorithm [Scott, N. W., et al 2008]. For global health status/QoL and all functional scales, a higher value indicates a better level of function; for symptom scales and single items, a higher value indicates increased severity of symptoms.
According to the QLQ-C30 Manuals, if items $I_1, I_2, \ldots, I_n$ are included in a scale, the linear transformation procedure is as follows:

1. Compute the raw score: $RS = (I_1 + I_2 + \ldots + I_n)/n$
2. Linear transformation to obtain the score S:

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

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

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

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

3.6.3.1 PRO Compliance Summary

Completion and compliance of EORTC QLQ-C30/QoL by visit and by treatment will be described. Numbers and percentages of complete and missing data at each visit will be summarized. An instrument is considered complete if at least one valid is available according to the missing item rules outlined in the scoring manual for the instrument.

Completion rate of treated participants (CR-T) at a specific time point is defined as the number of treated participants who complete at least one item over the number of treated participants in the PRO analysis population.

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

The completion rate is expected to shrink in the later visit during study period due to the participants who discontinued early. Therefore, another measurement, compliance rate of eligible participants (CR-E) will also be employed as the support for completion rate. CR-E is defined as the number of treated participants who complete at least one item over number of eligible participants who are expected to complete the PRO assessment, not including the participants missing by design such as death, discontinuation, translation not available.

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

The reasons of non-completion and non-compliance will be provided in supplementary table:

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

In addition, reasons for non-completion as scheduled of these measures will be collected using "miss_mode" forms filled by site personnel and will be summarized in 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

Data Collection Visit	Scheduled Day	Window of Day Range
Week 1	1	[-28,1]
Week 5	20	[2,43]
Week 8	55	[44,65]
Week 11	77	[66,87]
Week 14	99	[88,109]
Week 17	121	[110,131]
Week 20	143	[132,153]
Week 25	175	[154,219]
Week 36	252	[220,285]
Week 45	318	[286,351]
Week 56	395	[352,439]
Week 69	483	[440,527]
Week 82	571	[528,615]
Week 94	659	[616,703]
Week 105	736	[704,769]
EoT or Safety FU *	796	[770,822]

** If the D/C visit takes place ≥30 days from the last dose of study intervention, a safety FU visit is not required. In that event, all procedures required for both the D/C visit and the safety FU visit will be performed at the D/C visit. The D/C date is the date when the participant discontinues all study intervention. SoA as detailed in Protocol 7902-011-00.*

3.6.3.2 Mean change from baseline

Change from baseline in the following secondary PRO endpoint from EORTC QLQ C30 will be assessed:

- Global health status/QoL score (EORTC QLQ-C30)

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% based on blinded data review prior to the database lock for any PRO analysis and will be documented in a future sSAP amendment.

To assess the treatment effects on the PRO score change from baseline outcomes, a

constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [Ref. 5.4: 00QJ0M] will be applied, with the PRO score as the response variable, and treatment, time, the treatment by time interaction, and stratification factors used for randomization (see protocol Section 6.3.2) 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 PROscores at baseline and post-baseline time point.

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 Yijt is the PRO score for participant i, with treatment assignment j at visit t; γ0 is the baseline mean for all treatment groups, γjt is the mean change from baseline for treatment group j at time t; Xi 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).

Line plots for the empirical mean change from baseline will be provided across all time points 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.

3.6.3.3 Time to Traditional 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. 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.

- Time to Traditional Deterioration (TTD) as measured by EORTC QLQ-C30 global health status/quality of life scores. TTD is defined as the time from baseline to the first onset of a 10 or more points deterioration from baseline with a subsequent visit of 10 or more points deterioration from baseline.

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-Deterioration

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

3.6.3.4 Analysis Strategy for Key PRO Endpoints

Table 7 Analysis Strategy for Key PRO Endpoints

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Mean change from baseline in EORTC QLQ-C30/QoL	cLDA model	FAS	Model-based.Table 1
TTD in EORTC QLQ-C30/QoL	stratified log-rank test and HR estimation using stratified Cox model with Efron's tie handling method	FAS	Censored according to rules in Table 6.

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 AND FINAL ANALYSES

Blinding to treatment assignments will be maintained at all investigational sites. The results of interim analyses will not be shared with investigators prior to completion of the study. Participant-level unblinding will be restricted to an external unblinded statistician and scientific programmer performing the interim analysis, who will have no other responsibilities associated with the study.

An external DMC will serve as the primary reviewer of the results of interim efficacy and futility analyses and safety analyses, and will make recommendations for discontinuation of the study or protocol modifications to the study EOC. If the DMC recommends modifications to the design of the protocol or discontinuation of the study, the EOC (and potentially other limited Sponsor personnel) may be unblinded to results at the treatment level 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 statistician. Additional logistical details will be provided in the DMC charter. Key aspects of the interim analyses are described below.

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

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

If the study is positive at an interim analysis for either primary endpoint, additional analyses, including but not limited to the protocol-specified final analysis, may be carried out for exploratory purposes or upon regulatory request.

3.7.1 Efficacy and Futility Interim Analyses

Two interim efficacy analyses and two interim futility analysis are planned in addition to the final analysis for this study. For the interim and final analyses, all randomized participants will be included. Results of the interim analyses will be reviewed by the DMC. Details of the boundaries for establishing statistical significance with regard to efficacy are discussed further in Section 3.8.

The analyses planned, endpoints evaluated, and drivers of timing are summarized in Table 8.

Table 8 Summary of Interim and Final Analysis Strategy

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
Futility Analysis 1	PFS ORR	~297 participants enrolled ⩾151 PFS events	~16 months	• Non-binding futility IA for PFS and ORR
Futility Analysis 2	PFS ORR	~370 participants enrolled ⩾191 PFS events	~20 months	• Non-binding futility IA for PFS and ORR
IA1	PFS OS ORR if both PFS and OS are rejected	≥530 PFS events ~386 OS events expected at this time	~24 months	• Interim PFS and OS analyses • Final ORR analysis
IA2	PFS OS	≥558 PFS events ~469 OS events	~30 months	• Final PFS analysis • Interim OS analysis
FA	OS	~552 OS events	~39 months	• Final OS analysis

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

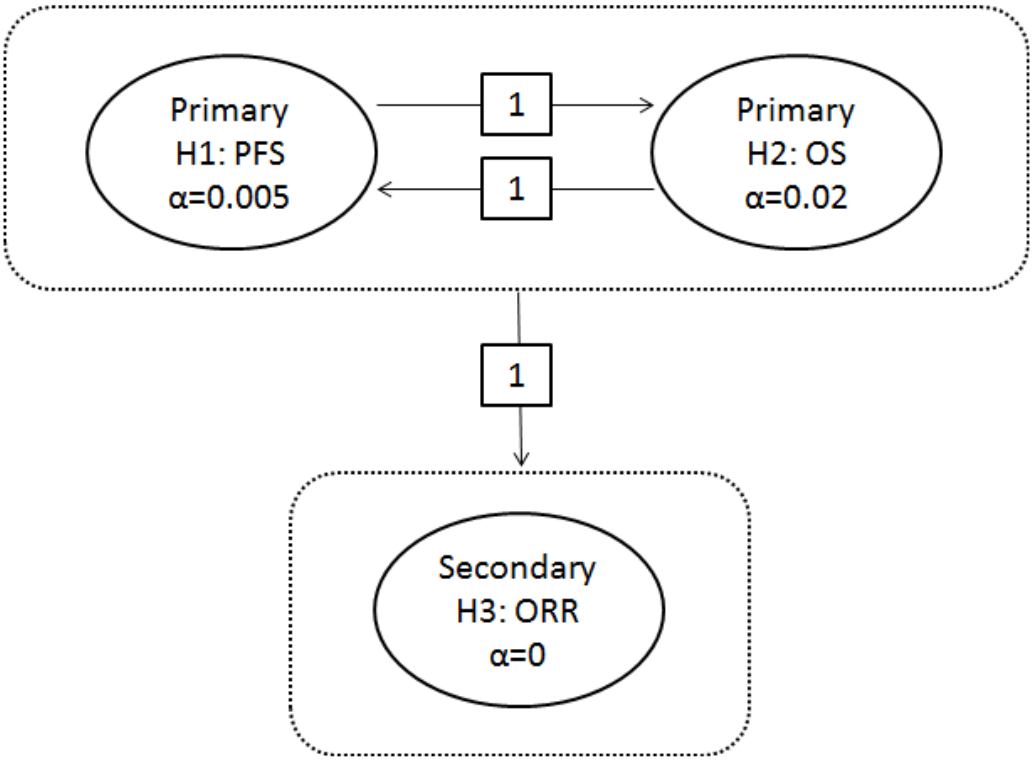
3.7.2 Safety Interim Analyses

The DMC will conduct regular safety monitoring. The safety data will be reviewed during a safety run-in after 40 participants have been enrolled and have received 2 Cycles of treatment. If the DMC notes any safety concerns, the protocol will be amended. It is estimated that this review will include approximately 15 participants and no less than 10 patients will have ECOG PS 2. If the initial review includes fewer than 10 participants with ECOG PS 2, an additional DMC review would be conducted after 15 participants with ECOG PS 2 have received 2 Cycles of treatment. The timing and interval of safety monitoring will be specified in the DMC charter.

3.8 MULTIPLICITY

The study uses the graphical method of Maurer and Bretz [Ref. 5.4: 045MYM] 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.

Initial α assigned to OS and PFS will be 0.02 and 0.005, respectively. If either hypothesis is rejected, α will be reallocated to the other hypothesis. If both are rejected, the ORR hypothesis will be tested at 0.025 (Figure 1).



Abbreviations: ORR = objective response rate; OS = overall survival; PFS = progression-free survival.
Note: If both PFS and OS null hypotheses are rejected, the allocation strategy allows testing of ORR at α = 0.025.

Figure 1 Multiplicity Diagram for Type I Error Control

3.8.1 Futility Analysis

After reviewing the safety data at the 3rd DMC safety meeting on August 7, 2020, the DMC requested that efficacy data be provided along with safety data at the 4th DMC safety review. The DMC made a similar request after the 4th DMC meeting. In order to provide a statistical framework for the DMC to evaluate benefit-risk, non-binding futility boundaries have been provided for ORR and PFS. The non-binding futility analyses on ORR and PFS will be performed at approximately 16 and 20 months from FPI, when approximately 297 and 370 participants have been enrolled and approximately 151 and 191 PFS events have been observed.

All participants randomized as of the last patient last visit for the futility analyses will be included in the comparisons of PFS between the 2 randomized treatment groups. An observed hazard ratio (combination / pembrolizumab monotherapy) greater than approximatey 1.3 and 1.2 would be sufficient for the eDMC to consider futility at each analysis. The actual PFS boundaries will be determined by the actual number of PFS events

at the time of the analysis using the pre-specified Hwang-Shih-DeCani beta-spending function with gamma parameter (-9). Table 9 shows the probability of stopping the study for futility under different scenarios.

Table 9 Futility Scenarios for PFS (Superiority in Progression Free Survival – Combo vs Pembro in All Subjects)

PFS Events	Approx. Observed HR at Boundary	Cumulative Probability (%) of Stopping for Futility Under:			
		HR = 1.3	HR = 1.2	Null HR = 1	Alternative HR = 0.7
151 (27% IF)	1.3	49.1	30.4	5.2	<1
191 (34% IF)	1.2	71.1	50.3	10.6	<1
HSD (-9) Spending function is used					

Approximately 252 and 335 participants who have available imaging responses from BICR per RECIST 1.1 will be used to compare the ORRs between the 2 randomized treatment groups in the two futility anaylses. Approximate observed differences in the ORRs (combination – pembrolizumab monotherapy) less than -10% at the first futility analysis and less than -5% at the second futility analysis would be sufficient for the eDMC to consider futility. The actual ORR futility boundaries will be determined by the actual number of participants included in the ORR analysis using a Hwang-Shih-DeCani beta-spending function with gamma parameter (-10).

Table 10 shows the probability of stopping the study for futility under different scenarios.

Table 10 Futility Scenarios for ORR (Superiority in Objective Responsive Rate – Combo vs Pembro in All Subjects with Imagine by BICR per RECIST 1.1)

Sample Size	Approx. Observed ΔORR at Boundary	Cumulative Probability (%) of Stopping for Futility Under:			
		ΔORR = -10%	ΔORR = -5%	Null ΔORR = 0	Alternative ΔORR= 17%
252 (36% IF)	-10%	60.9	24.3	5.9	<1
335 (48% IF)	-5%	88.7	54.5	17.8	<1
HSD (-10) Spending function is used					

All of the available data (ORR, PFS, and safety) will be considered by the eDMC in their deliberations in considering whether to recommend stopping the study for futility.

3.8.2 Objective Response Rate

The study will test the ORR hypothesis at the interim analysis 1 (IA1), at an α level of 0.025 if both OS and PFS hypotheses are rejected.

Based on the 694 randomized participants with at least 3 months of follow-up, power at α=0.025, as well as the approximate treatment difference required to reach the bound (ΔORR), are shown in Table 11, assuming underlying 30% and 47% response rates in the control and experimental groups, respectively.

If, at IA2, the OS test does not achieve statistical significance but the PFS test achieves statistical significance, the p-value of the ORR test from IA2 will be compared to 0.025 if the null hypothesis for OS is later rejected.

Table 11 Possible α Levels and Approximate Objective Response Rate Difference Required to Demonstrate Efficacy for Objective Response at Interim Analysis

α	**~Δ Objective Response Rate (ORR)**	**Power (ΔORR=0.17)**
0.025	0.073	0.996

3.8.3 Progression-free Survival

The PFS hypothesis may be tested at α=0.005 (initially allocated α) or α=0.025 (if the OS null hypothesis is rejected). Table 12 shows the bounds and boundary properties for PFS hypothesis testing derived using a Lan-DeMets O'Brien-Fleming spending function. If the actual number of PFS 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 12 Efficacy Boundaries and Properties for Progression-free Survival Analyses

Analysis	Value	α=0.005	α=0.025
IA1:95%* N = 694 Events: 530 Month: 24	Z	2.654	2.025
	p (1-sided) [a]	0.004	0.021
	HR at bound[b]	0.794	0.839
	P(Cross) if HR=1[c]	0.004	0.021
	P(Cross) if HR=0.7[d]	0.927	0.981
Final: N = 694 Events: 558 Month: 30	Z	2.657	2.062
	p (1-sided) [a]	0.004	0.02
	HR at bound[b]	0.799	0.840
	P(Cross) if HR=1[c]	0.005	0.025
	P(Cross) if HR=0.7[d]	0.947	0.986

Abbreviations: HR = hazard ratio; IA = interim OS analysis (PFS final 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.

3.8.4 Overall Survival

The OS hypothesis may be tested at α=0.02 (initially allocated α) or α=0.025 (if the PFS null hypothesis is rejected). Table 13 shows the bounds and boundary properties for OS hypothesis testing derived using a Lan-DeMets O'Brien-Fleming spending function.

Table 13 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	**Value**	**α=0.02**	**α=0.025**
IA1: 70%* N = 694 Events: 386 Month: 24	Z	2.549	2.439
	p (1-sided) [a]	0.005	0.007
	HR at bound[b]	0.771	0.780
	P(Cross) if HR=1[c]	0.005	0.007
	P(Cross) if HR=0.75[d]	0.609	0.650
IA2: 85%* N: 694 Events: 469 Month: 30	Z	2.326	2.230
	p (1-sided) [a]	0.010	0.013
	HR at bound[b]	0.807	0.814
	P(Cross) if HR=1[c]	0.012	0.015
	P(Cross) if HR=0.75[d]	0.794	0.820
Final N: 694 Events: 552 Month: 39	Z	2.138	2.050
	p (1-sided) [a]	0.016	0.020
	HR at bound[b]	0.834	0.840
	P(Cross) if HR=1[c]	0.020	0.025
	P(Cross) if HR=0.75[d]	0.901	0.915

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated approximately.
*Percentage of the target number of events at final analysis anticipated at interim analysis.
[a]*p* (1-sided) is the nominal α for testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P (Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P (Cross if HR=0.75) is the probability of crossing a bound under the alternative hypothesis.

If the actual number of OS events at the interim and final analyses differs from those specified in the table, the bounds will be updated using this spending function evaluated at the observed information fraction (fraction of observed over expected final events) at each analysis.

The bounds provided in the table above are based on the assumption that the expected number of events at IA1, IA2 and FA are 386, 469 and 552, 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 leave 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. Specifically,

- In the scenario that the events accrue slower than expected and the observed number of events is less than the expected number of events at a given analysis, the information fraction will be calculated as the observed number of events at the interim analysis over the target number of events at FA.

- In the scenario that the events accrue faster than expected and the observed number of events exceeds the expected number of events at a given 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.

The final 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 α 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.

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 α 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.

3.8.5 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. However, to account for any multiplicity concerns raised by the DMC review of unplanned efficacy data prompted by safety concerns, a sensitivity analysis for OR, PFS, and OS adopting a conservative multiplicity adjustment will be conducted. For each safety analysis at which an unplanned efficacy analysis is performed, an arbitrary small alpha will be assigned for a nominal bound of 0.00005. Based on this nominal bound, the final bound for OR, PFS, and OS will be adjusted so that final analysis will be controlled within the allocated Type I error.

3.9 SAMPLE SIZE AND POWER CALCULATIONS

The study will randomize 694 participants in a 1:1 ratio into the pembrolizumab + lenvatinib and pembrolizumab + placebo arms, unless a protocol-specified non-binding futility analysis leads to suspension of randomized treatment. PFS and OS are primary endpoints for the study, with ORR as the key secondary endpoint.

Based on the 694 participants with at least 7 months of follow-up, the power of the ORR testing at the allocated α=0.025 is approximately 99.6% to detect a 17-percentage point difference between an underlying 30% response rate in the control arm and a 47% response rate in the experimental arm.

For the PFS endpoint, based on a target number of 558 events and 1 efficacy interim analyses at approximately 95% of the target number of events, the study has approximately 95% power to detect an HR of 0.7 at an overall α level of 0.005 (1-sided), and 99% power at an α level of 0.025 (1-sided).

For the OS endpoint, based on a target number of 552 events and 2 efficacy interim analyses at approximately 70% and 85% of the target number of events, the study has approximately 90% power to detect an HR of 0.75 at an overall α level of 0.02 (1-sided), and 91.5% power at an α level of 0.025 (1-sided).

Based on KEYNOTE-052 and KEYNOTE-361 data, the above sample size and power calculations for PFS and OS assume the following:

- PFS follows an exponential distribution with a median of 3.2 months for the control group.
- OS follows an exponential distribution with a median of 10.0 months for the control group.
- Enrollment period of 17 months
- An annual dropout rate of 12% and 5% for PFS and OS, respectively
- A follow-up period of 13 and 22 months for PFS and OS, respectively, after the last participant enrolls.

The sample size and power calculations were performed using R (“gsDesign” package) and EAST 6.4.

3.10 SUBGROUP ANALYSES

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

- Geographic region (US, non-US)
- ECOG performance status (0, 1, 2)
- Ineligibility for cisplatin versus ineligibility for any platinum-based chemotherapy
- PD-L1 CPS (<10, ≥10)

- **Note**: Participants in the any platinum-ineligible subgroup with PD-L1 below detectable levels will be included in the CPS <10 category. Cisplatin-ineligible participants in the CPS <10 category will be ineligible for this study.

- Age category (<65 years, 65 to 74 years, ≥75 years)
- Sex (female, male)
- Race (white, nonwhite)
- Baseline hemoglobin (<10g/dL, ≥10g/dL)
- Metastatic location at baseline (visceral disease, lymph node only)
- Liver metastasis at baseline (presence, absence)
- Number of Bajorin risk factors (0, 1, 2)

The consistency of the treatment effect will be assessed using descriptive statistics for each category of the subgroup variables listed above. If the number of participants in a category of a subgroup variable is less than 10% of the ITT population, the subgroup analysis will not be performed for this category of the subgroup variable, and this subgroup variable will not be displayed in the forest plot. 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 study intervention 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 in which the participant receives the study intervention. Summary statistics will be provided for the extent of exposure in the APaT population.

4 REFERENCES

[Ref. 5.4: 00QJ0M] Liang K-Y, Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhya: The Indian Journal of Statistics 2000;62(Series B, Pt. 1):134-48.

[Ref. 5.4: 00TWVP] Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16:139-44.

[Ref. 5.4: 00VMQY] Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med 1985;4:213-26.

[Ref. 5.4: 00VPPS] O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials [Abstract]. Biometrics 1979;35(3):549-56.

[Ref. 5.4: 03P3QC] Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70(3):659-63.

[Ref. 5.4: 045MYM] Maurer W, Glimm E, Bretz F. Multiple and repeated testing of primary, coprimary, and secondary hypotheses. Stat Biopharm Res. 2011;3(2):336-52.

[Ref. 5.4: 045X5X] Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol. 2014 Aug 1;32(22):2380-5.

[Ref. 5.4: 050DSW] Vaughn DJ, Bellmunt J, Fradet Y, Lee JL, Fong L, Vogelzang NJ, et al. Health-related quality-of-life analysis from KEYNOTE-045: a phase III study of pembrolizumab versus chemotherapy for previously treated advanced urothelial cancer. J Clin Oncol. 2018 Jun 1;36(16):1579-87. Authors' disclosures of potential conflicts of interest, Acknowledgment; 3 p.