Official Protocol Title:	A Randomized, Double-Blind, Phase 3 Study of Pembrolizumab/Vibostolimab Coformulation (MK-7684A) in Combination with Chemotherapy Versus Pembrolizumab Plus Chemotherapy as First Line Treatment for Participants with Metastatic Non-Small Cell Lung Cancer (MK-7684A-007/KEYVIBE-007)
NCT number:	NCT05226598
Document Date:	SUPPLEMENTAL SAP AMENDMENT 02 (16-Sep-25)

Supplemental Statistical Analysis Plan (sSAP)

Confidential

TABLE OF CONTENTS

LIST OF TABLES

C Confidential

LIST OF FIGURES

C Confidential

1 INTRODUCTION

This supplemental SAP (sSAP) is a companion document to the protocol. While Section 9 of the protocol provides the principal features of confirmatory analyses for this trial, this supplemental SAP provides additional statistical analysis details/data derivations and may document modifications or additions to the protocol-specified analysis plan that are not principal in nature and/or result from information that was not available at the time of protocol finalization.

2 SUMMARY OF CHANGES

This is Amendment 02 of the sSAP and aligns with protocol Amendment 05. The following changes were made compared to the previous version of the sSAP:

- The pre-specified futility criteria were met at the Futility Analysis before efficacy IA, and the Sponsor decided to discontinue treatment with MK-7684A.
- No additional efficacy analysis will be conducted at IA or FA.
- Analyses on Secondary efficacy endpoints specifically PFS, DOR, ORR and ePRO will not be conducted (Section 3.1, Section 3.6 and Section 3.7).

3 ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

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

Study Design Overview	A Randomized, Double-Blind, Phase 3 Study of Pembrolizumab/Vibostolimab Coformulation (MK-7684A) Plus Chemotherapy Versus Pembrolizumab Plus Chemotherapy as First Line Treatment for Participants With Metastatic Non-Small Cell Lung Cancer (MK-7684A-007/KEYVIBE-007)
Treatment Assignment	Approximately 700 participants will be randomized in a 1:1 ratio between 2 treatment arms: (1) MK-7684A plus chemotherapy and (2) pembrolizumab plus chemotherapy. Stratification factors are as follows: • ECOG Performance Status (0 vs 1) • Predominant tumor histology (Squamous vs nonsquamous)
	• PD-L1 expression (TPS<50% vs ≥50%) • Geographic Region (East Asia vs North America/Western Europe vs. Rest of the World)
Analysis Populations	Efficacy: ITT Safety: APaT
Primary Endpoints	• OS in participants with PD-L1 TPS≥1%

Secondary Endpoints	• OS in all participants • PFS per RECIST 1.1 as assessed by BICR in participants with PD-L1 TPS≥1% and all participants • ORR per RECIST 1.1 as assessed by BICR in participants with PD-L1 TPS≥1% and all participants • DOR per RECIST 1.1 as assessed by BICR in participants with PD-L1 TPS≥1% and all participants • Change from baseline in global health status/QoL, cough, chest pain, dyspnea, and physical functioning scores in participants with PD-L1 TPS≥1% and all participants • TTD in global health status/QoL, cough, chest pain, dyspnea, and physical functioning in participants with PD-L1 TPS≥1% and all participants • Safety and tolerability
Statistical Methods for Key Efficacy Analyses	The primary and key secondary hypotheses testing of OS in participants with PD-L1 TPS≥1% and all participants, and PFS in all participants will be evaluated by comparing the MK-7684A plus chemotherapy treatment arm with the pembrolizumab plus chemotherapy treatment arm using a stratified log-rank test. The 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. **Prior to IA, the Sponsor decided to discontinue MK-7684A due to futility. No additional efficacy analysis will be conducted. Analyses on secondary efficacy endpoints, specifically PFS, ORR, DOR and PRO, will not be conducted.**
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 M&N method [Miettinen, O. and Nurminen, M. 1985].
Interim Analyses	One interim analysis and 1 final analysis are planned in this study: IA: • Timing: to be performed after both ~237 OS events in participants with PD-L1 TPS≥1% have been observed and ~22 months after last participant randomized • Primary purpose: Interim OS analysis in participants with PD-L1 TPS≥ 1%; OS in all participants will be tested if OS null hypothesis for participants with PD-L1 TPS≥1% is rejected; PFS in all participants will be tested if OS null hypotheses for both participants with PD-L1 TPS≥1% and all participants are rejected. FA: • Timing: to be performed after both ~302 OS events in participants with TPS≥1% have been observed and ~36 months after last participant randomized

	• Primary purpose: Final OS analysis in participants with PD-L1 TPS≥1%; OS in all participants will be tested if OS null hypothesis for participants with PD-L1 TPS≥1% is rejected. At IA, non-binding futility analysis based on OS in participants with PD-L1 TPS≥1% will be performed. Additionally, non-binding futility analysis based on OS in participants with PD-L1 TPS ≥1% and in all percipients will be performed before the efficacy interim analysis when ~198 OS events have been observed in participants with PD-L1 TPS≥1% and ~16 months after last participant randomized, whichever comes later. The overall safety and efficacy data will be reviewed for the decision. Results will be reviewed by an eDMC. Details are provided in Section 3.7. **Prior to IA, the Sponsor decided to discontinue MK-7684A due to futility. No additional efficacy analysis will be conducted at IA or FA; information regarding planned IAs in Section 3.7.1 is being retained for historical purposes.**
Multiplicity	The overall Type I error over the primary and key secondary hypotheses is strongly controlled at 2.5% (1-sided), with 2.5% initially allocated to OS (H1). The graphical approach of Maurer and Bretz [Maurer, W., et al 2011] will be applied to reallocate α among the hypotheses for OS in participants with PD-L1 TPS≥1%, OS in all participants and PFS in all participants. Lan-DeMets and O'Brien-Fleming group sequential methods will be used to allocate α among the interim and final analyses for the PFS and OS endpoints [Lan, K. K. G. and DeMets, D. L. 1983] [O'Brien, P. C. and Fleming, T. R. 1979].
Sample Size and Power	Overall, 739 participants were enrolled into the study. A target sample size of approximately 370 participants per treatment arm will be used for study planning purposes. It is estimated there will be ~302 deaths at the OS final analysis in participants with PD-L1 TPS≥1%. With 302 deaths, the study has ~81% power for detecting an HR of 0.72 (MK-7684A plus chemotherapy vs pembrolizumab plus chemotherapy) at the assigned 0.025 (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-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 Sponsor will generate the randomized allocation schedule(s) for study treatment assignment for this protocol, and the randomization will be implemented in IRT.

An eDMC will serve as the primary reviewer of the results of the interim analyses of the study and will make recommendations for discontinuation of the study or protocol modifications to the study EOC. Treatment-level results of the interim analyses will be provided by the unblinded statistician to the eDMC. If the eDMC 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 team. Additional logistical details will be provided in the eDMC charter. Key aspects of the interim analyses are described in Section 3.7.

Before 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.

3.3 Hypotheses/Estimation

Hypotheses are aligned with objectives in the Objectives and Endpoints table.

In males and females with treatment-naïve metastatic NSCLC (squamous or nonsquamous):

Primary Objective	**Primary Endpoint**
Objective: To compare MK-7684A in combination with chemotherapy to pembrolizumab in combination with chemotherapy with respect to Overall Survival (OS) in participants with PD-L1 TPS ≥1% Hypothesis (H1): MK-7684A in combination with chemotherapy is superior to pembrolizumab combination with chemotherapy with respect to OS in participants with PD-L1 TPS ≥1%	OS, defined as the time from randomization to the date of death due to any cause
Secondary Objectives	**Secondary Endpoints**
Objective: To compare MK-7684A in combination with chemotherapy to pembrolizumab in combination with chemotherapy with respect to OS in all participants Hypothesis (H2): MK-7684A in combination with chemotherapy is superior to pembrolizumab combination with chemotherapy with respect to OS in all participants	OS, defined as the time from randomization to the date of death due to any cause

Confidential

Objective: To compare MK-7684A in combination with chemotherapy to pembrolizumab in combination with chemotherapy with respect to progression-free survival (PFS) in participants with PD-L1 TPS≥1% and in all participants per Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 as assessed by blinded independent central review (BICR) Hypothesis (H3): MK-7684A in combination with chemotherapy is superior to pembrolizumab in combination with chemotherapy with respect to PFS in all participants per RECIST 1.1 as assessed by BICR	PFS, defined as the time from randomization to the first documented disease progression or death due to any cause, whichever occurs first
Objective: To evaluate MK-7684A in combination with chemotherapy to pembrolizumab in combination with chemotherapy with respect to Objective Response Rate (ORR) in participants with PD-L1 TPS≥1% and in all participants per RECIST 1.1 as assessed by BICR	Objective response, defined as a confirmed Complete Response (CR) or Partial Response (PR)
Objective: To evaluate the mean change from baseline in global health status/quality of life (QoL), physical functioning, role functioning, dyspnea, cough, and chest pain for MK-7684A in combination with chemotherapy compared to pembrolizumab in combination with chemotherapy in participants with PD-L1 TPS≥1% and in all participants	Change from baseline in the following patient-reported outcomes (PROs) scales/items: - Global health status/QoL score (EORTC QLQ-C30 items 29 and 30) - Physical functioning score (EORTC QLQ-C30 items 1-5) - Role functioning score (EORTC QLQ-C30 items 6-7) - Dyspnea score (EORTC QLQ-C30 item 8) - Cough (EORTC QLQ-LC13 item 31) - Chest pain (EORTC QLQ-LC13 item 40)

Secondary Objectives	Secondary Endpoints
Objective: To evaluate the time to deterioration in global health status/QoL, physical functioning, role functioning, dyspnea, cough, and chest pain for MK-7684A in combination with chemotherapy compared to pembrolizumab in combination with chemotherapy in participants with PD-L1 TPS≥1% and in all participants	Time to deterioration: time from baseline to the first onset of a ≥10-point (out of 100 points) deterioration from baseline in a given scale/subscale/item with confirmation at a subsequent visit of a ≥10-point deterioration from baseline. If the first deterioration is at the last patient-reported outcomes assessment timepoint, then no confirmation is required. Time to deterioration in the following scales/items: - Global health status/QoL score (EORTC QLQ-C30 items 29 and 30) - Physical functioning score (EORTC QLQ-C30 items 1-5) - Role functioning score (EORTC QLQ-C30 item 6-7) - Dyspnea score (EORTC QLQ-C30 item 8) - Cough (EORTC QLQ-LC13 item 31) - Chest pain (EORTC QLQ-LC13 item 40)
Objective: To evaluate the safety and tolerability of MK-7684A in combination with chemotherapy compared to pembrolizumab in combination with chemotherapy	• Adverse Events (AEs) • Discontinuations of study intervention due to an AE
Objective: To evaluate DOR per RECIST 1.1 as assessed by BICR for MK-7684A plus chemotherapy compared to pembrolizumab plus chemotherapy in participants with PD-L1 TPS≥1% and in all participants	DOR: for participants who demonstrate confirmed CR or PR, DOR is defined as the time from first documented evidence of CR or PR until disease progression or death due to any cause, whichever occurs first

C Confidential

Tertiary/Exploratory Objectives	Tertiary/Exploratory Endpoints
CCI	

Note: This study will be considered to have met its success criteria if MK-7684A plus chemotherapy is superior to pembrolizumab plus chemotherapy with respect to OS in participants with PD-L1 TPS≥1%.

Throughout this protocol, the term RECIST 1.1 refers to the modification of RECIST 1.1 to include a maximum of 10 target lesions and a maximum of 5 target lesions per organ. Refer to Section 4.2.1.1 in protocol for further details.

3.4 Analysis Endpoints.

Efficacy and safety endpoints that will be evaluated are listed below, followed by the descriptions of the derivations of selected endpoints.

3.4.1 Efficacy Endpoints

Primary

- **Overall Survival**

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

Secondary

- **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.

- **Objective Response Rate**

ORR is defined as the percentage of participants who achieve a confirmed CR or PR per RECIST 1.1 as assessed by BICR.

- **Duration of Response**

For participants who demonstrate confirmed CR or PR, DOR 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 and tolerability of study treatment will be assessed by clinical review of all relevant parameters including AEs, laboratory tests, and vital signs. AEs will be assessed as defined by NCI CTCAE 5.0. A description of safety measures is provided in Section 8.3 in protocol.

3.4.3 Patient-reported Outcomes

Secondary

- ***Change from baseline in***

1. Global health status/QoL score (QLQ-C30 items 29-30)
2. Single-item symptom scores: cough (QLQ-LC13 item 31), chest pain (QLQ-LC13 item 40), and dyspnea (QLQ-C30 item 8)
3. Functioning scores: physical functioning (QLQ-C30 items 1-5) and role functioning (QLQ-C30 items 6-7)

- ***Time to deterioration in***

4. Global health status/QoL score (QLQ-C30 items 29-30)

5. Single-item symptom scores: cough (QLQ-LC13 item 31), chest pain (QLQ-LC13 item 40), and dyspnea (QLQ-C30 item 8)
6. Functioning scores: physical functioning (QLQ-C30 items 1-5) and role functioning (QLQ-C30 items 6-7)

Based on prior literature [Maringwa, J. T., et al 2011] [Osoba, D., et al 1998] [King, M. T. 1996], a 10-point or greater worsening from baseline for each scale represents a clinically relevant deterioration. TTD is defined as the time to first onset of 10-point or more (out of 100) deterioration from baseline in a given scale/subscale/item and confirmed by a second adjacent 10-point or more deterioration from baseline. If the first deterioration is at the last PRO assessment timepoint (in the current database lock), then no confirmation is required. 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, K., et al 2011].

Exploratory

CCI

CCI

[redacted]

3.5 Analysis Populations

3.5.1 Efficacy Analysis Populations

The ITT population will serve as the population for primary efficacy analysis. All randomized participants will be included in this population. Participants will be included in the treatment group to which they are randomized.

The analysis population for DOR consists of participants in the analysis population of OR who demonstrate confirmed CR or PR.

3.5.2 Safety Analysis Populations

The APaT population will be used for the analysis of safety data in this study. The APaT population 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. For most participants this will be the treatment group to which they are randomized. Participants who take incorrect study treatment for the entire treatment period will be included in the treatment group corresponding to the study treatment actually received. Any participants who receive the incorrect study medication for one cycle but receives the correct treatment for all other cycles will be analyzed according to the participant's randomized treatment group and a narrative will be provided for any events that occur during the cycle for which the participant was incorrectly dosed.

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

3.5.3 Patient-reported Outcome Analysis Population

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

3.6 Statistical Methods

Prior to IA, the Sponsor decided to discontinue MK-7684A due to futility. No additional efficacy analysis will be conducted. Analyses on secondary efficacy endpoints, specifically PFS, ORR, DOR and PRO, will not be conducted.

3.6.1 Statistical Methods for Efficacy Analyses

This section describes the statistical methods that address the primary and secondary 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. Nominal p-values will be computed for other efficacy analyses, but should be interpreted with caution due to potential issues of multiplicity.

The stratification factors used for randomization (see Section 6.3.2 in protocol) will be applied to all stratified analyses, in particular, the stratified log-rank test, stratified Cox model, and stratified Miettinen and Nurminen method. 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. No pooling has been conducted.

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

3.6.1.1 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 (based on the stratification factors defined in Section 6.3.2 in protocol). 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 (Section 6.3.2 in protocol) 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 the participant was last known to be alive.

In case the proportional hazards assumption does not hold, RMST method may be conducted for OS to account for the possible non-proportional hazards effect as a sensitivity analysis.

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, based on the stratification factors defined in Section 6.3.2 in protocol. 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 Section 6.3.2 in protocol) 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 a BICR, regardless of discontinuation of study drug. Death is always considered as a confirmed PD event. Sensitivity analyses will be performed for comparison of PFS based on investigator's assessment.

For the primary analysis, any participant who experiences an event (PD or death) immediately after 2 or more 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 prior to documented progression 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. Sensitivity analyses will be performed for comparison of PFS based on investigator’s assessment.

In order to evaluate the robustness of the PFS endpoint per RECIST 1.1 by a BICR, two sensitivity analyses with a different set of censoring rules will be performed. The first sensitivity analysis is the same as the primary analysis except that it censors at the last disease assessment without PD when PD or death is documented after more than one missed disease assessment. The second sensitivity analysis is the same as the primary analysis except that it considers discontinuation of treatment or initiation of an anticancer treatment subsequent to discontinuation of study-specified treatments, whichever occurs later, to be a PD event for 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.

In case the proportional hazards assumption is not valid, RMST method may be conducted for PFS to account for the possible non-proportional hazards effect as a sensitivity analysis.

C Confidential

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
Death or progression immediately after ≥ 2 consecutive missed disease assessments, or after new anticancer therapy	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 CR; 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
Abbreviations: CR=complete response; PD=progressive disease.			

3.6.1.3 Objective Response Rate

The stratified Miettinen and Nurminen method will be used for comparison of the ORR between the 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 (Section 6.3.2 in protocol) will be applied to the 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 [Clopper, C. J. and Pearson, E. S. 1934].

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 CR or PR 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 or death, no new anticancer therapy initiated	Last adequate disease assessment	Censor (nonevent)
No progression or death, new anticancer therapy initiated	Last adequate disease assessment before new anticancer therapy initiated	Censor (nonevent)
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 (nonevent)
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=progressive disease. 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 Efficacy Variables

A summary of the primary analysis strategy for the efficacy endpoints is provided in Table 3. The strategy to address multiplicity issues with regard to multiple endpoints is described in Section 3.8.

Table 3 Analysis Strategy for Efficacy Variables

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Primary Analyses			
OS	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	Participants with PD-L1 TPS ⩾ 1% in ITT	Censored at the date participant last known to be alive
Secondary Analyses			
OS	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at the date participant last known to be alive

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
PFS per RECIST 1.1 by BICR	Testing: stratified log-rank test	ITT	Censored according to rules in Table 1
	Estimation: Stratified Cox model with Efron's tie handling method		
Abbreviations: BICR=blinded independent central review; ITT=intent-to-treat; TPS=tumor proportion score; OS=overall survival; PFS=progression-free survival; RECIST=Response Evaluation Criteria in Solid Tumors.			

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

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) are either prespecified as "Tier 1" endpoints or will be classified as belonging to "Tier 2" or "Tier 3" based on the number of events observed.

Tier 1 Events

Safety parameters or adverse events of special interest that are identified a priori constitute Tier 1 safety events that will be subject to inferential testing for statistical significance. There are no Tier 1 events for this protocol. AEs that are immune-related or potentially immune-related 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. Based on a review of historic data from MK-7684A clinical studies, there are no AEs that warrant inferential testing between-treatment groups.

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 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.

For continuous measures such as changes from baseline in laboratory parameters, summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group in table format.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Grade 3-5 AE (incidence ≥5% of participants in one of the treatment groups)	X	X
	SAE (incidence ≥5% of participants in one of the treatment groups)	X	X
	AEs (incidence ≥10% of participants in one of the treatment groups)	X	X
Tier 3	Any AE		X
	Any Grade 3-5 AE		X
	Any SAE		X
	Any Drug-Related AE		X
	Any Drug-Related SAE		X
	Any Grade 3-5 and Drug-Related AE		X
	Discontinuation due to AE		X
	Death		X
	Specific AEs, SOCs (incidence <10% of participants in all of the treatment groups)		X
	Change from baseline results (laboratory toxicity shifts)		X
Abbreviations: AE=adverse event; CI=confidence interval; SAE=serious adverse event; SOC=system organ class.			

C Confidential

3.6.3 Statistical Methods for Patient-reported Outcome Analyses

This section describes the planned analyses for the PRO endpoints.

3.6.3.1 Scoring Algorithm

EORTC QLQ-C30 Scoring: For each scale or item, a linear transformation will be applied to standardize the score as between 0 and 100, according to the corresponding scoring standard. For 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.

According to the EORTC 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:

Functional scales: $$S = \left\{1 - \frac{RS - 1}{Range}\right\} \times 100$$

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

Global health status/quality of life scale: $$S = \frac{RS - 1}{Range} \times 100$$

Confidential

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 [Scott, N. W., et al 2008].

EORTC QLQ-LC13 scoring: The lung cancer questionnaire module comprises both multi-item and single-item measures of lung cancer-associated symptoms (i.e., coughing, hemoptysis, dyspnea, and pain) and treatment related symptoms (i.e., sore mouth, dysphagia, peripheral neuropathy, and alopecia). A linear transformation will be applied to standardize the scores between 0 and 100 as described above for the EORTC QLQ-C30 symptom scales/items scoring.

NSCLC-SAQ scoring: The *NSCLC-SAQ* consists of seven items covering five domains: Cough, Pain, Dyspnea, Fatigue, Appetite (see table below). All five of these domains must be non-missing to compute a total score. Two of the domains contain 2 items: Pain and Fatigue.

Domain	**Item**		**Response**
Cough	1. How would you rate your coughing at its worst…?		0, 1, 2, 3, 4
Pain	2. How would you rate the worst pain in your chest…? 3. How would you rate the worst pain in areas other than your chest…?	*Create a single score by selecting the highest severity (i.e., value on either item)*	0, 1, 2, 3, 4
Dyspnea	4. How often did you feel short of breath during usual activities…?		0, 1, 2, 3, 4
Fatigue	5. How often did you have low energy…? 6. How often did you tire easily…?	*Create a single score by calculating the mean of these 2 items.*	0, 1, 2, 3, 4
Appetite	7. How often did you have a poor appetite over the last 7 days?		0, 1, 2, 3, 4
NSCLC-SAQ Total Score (Sum the 5 domains)			**Range 0 to 20**

PAIN: The two pain items [2. "How would you rate the worst pain in your chest over the last 7 days?" and 3. "How would you rate the worst pain in areas other than your chest over the past 7 days?"] are combined into a score by selecting the most severe response from the two items (or the single response if both items have the same score). The goal of the *NSCLC-SAQ* is to assess worst pain, wherever it manifests, hence a score will be derived by taking the most severe answer to either of the items, becoming a single "Pain" score. If one of these two items is missing, the included response (from the remaining item) is used as the "Pain" score.

FATIGUE: The two fatigue items [5. "How often did you have low energy over the last 7 days?" and 6. "How often did you tire easily over the last 7 days?"] are also combined. Given the high correlation between the two items (0.84), indicating considerable conceptual redundancy, a score will be derived by taking the mean of the two items, thus becoming a single "Fatigue" score. If one of these two items is missing, the included response (from the remaining item) is used as the "Fatigue" score.

For both "Pain" and "Fatigue" domains, if both items are missing responses, then the score would not be computed, it would remain missing.

C Confidential

The provisional scoring algorithm of the *NSCLC-SAQ* total score is as follows:

- **Cough Domain Score**: score of the cough item, or missing if skipped
- **Fatigue Domain Score**: if both items present, compute mean; or use score from 1 item if the other is missing; or set to missing if both are skipped
- **Pain Domain Score**: if both items present, use most severe of both; or use score from 1 item if the other is missing; or set to missing if both are skipped
- **Dyspnea Domain Score**: score of the shortness of breath item, or missing if skipped
- **Appetite Domain Score**: score of the poor appetite item, or missing if skipped
- **NSCLC-SAQ Total Score**: sum all five domain scores; if any are missing, a total score is not computed. This creates a total score ranging between 0 and 20 with higher scores indicating more severe symptomatology.

The *NSCLC-SAQ* total score ranges between 0 and 20. Higher scores indicate more severe NSCLC-related symptomatology.

EQ-5D scoring: The EQ-5D-5L is primarily designed for self-completion and consists of 2 parts: a descriptive system and the VAS. The EQ-5D-5L descriptive system includes 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The 5 dimensions each have 5 levels: no problems, slight, moderate, severe problems, and extreme problems. The responses patterns on the 5 dimensions are scored using country-specific population weights to provide an aggregate index score anchored at 0 (death) and 1 (perfect health); depending on the algorithm used some states may be considered worse than death.

The EQ-5D VAS records the respondent's self-rated health on a vertical, visual analogue scale (0-100), with endpoints labeled 'the best health you can imagine' and 'the worst health you can imagine'. The recall period is current health today (the day of completion). For the EQ-5D VAS scale, A ≥7-point change from baseline in VAS is considered to be a MID [Pickard, A. S., et al 2007].

3.6.3.2 PRO Compliance Summary

Completion and compliance of EORTC QLQ-C30, EORTC QLQ-LC13, NSCLC-SAQ, and EQ-5D 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 score is available according to the missing item rules outlined in the scoring manual for the instrument.

C Confidential

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.

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

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.

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

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

Study Week	Week 0 (Baseline)	Week 3	Week 6 to Week 45 (Every 3 weeks)	Week 48	Week 54 to Week 102 (every 6 weeks)
Start Day	1	22	Week number *7+1	337	Week number *7+1
Day Range (relative day to first dose)	[-28, 1]	[2, 32]	[Week number*7-9, week number*7+11	[327,357]	[Week number*7-20, week number*7+21]

3.6.3.3 Mean Change from Baseline

The time point for the mean change from baseline analysis is defined as the latest time point at which approximately CR-T ≥ 60% and CR-E ≥ 80% based on blinded data review prior to the database lock for any PRO analysis.

To assess the treatment effects on the PRO score change from baseline in the PRO endpoint (global health status/QoL, physical functioning, dyspnea, cough, chest pain) defined in Section 3.4.3, a constrained longitudinal data analysis (cLDA) model proposed by Liang and

Zeger [Liang, K-Y. and Zeger, S. L. 2000] 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 Section 6.3.2 in protocol) 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 and nominal two-sided p-value. Model-based LS mean with 95% CI will be provided by treatment group for PRO scores at baseline and postbaseline 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:

$$E(Y_{ijt}) = y_0 + y_{jt}I(t > 0) + fJX_i, j = 1,2,,3,..,n;\ t = 0,1,2,3,..k$$

where Y_{ijt} is the PRO score for participant i, with treatment assignment j at visit t; y_0 is the baseline mean for all treatment groups, y_{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 fJ 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 in EORTC QLQ-C30 global health status/QoL, its physical functioning, dyspnea score, and EORTC QLQ-LC13 cough and chest pain scores will be provided across all time points as a supportive analysis.

In addition, the model-based LS mean change from baseline to the specified postbaseline time point together with 95% CI will be plotted in bar charts for EORTC QLQ-C30 global health status/QoL scale, functioning scales, and symptoms scales and EORTC QLQ-LC13 symptom scales.

3.6.3.4 Time to Deterioration (TTD)

For the TTD endpoint defined in Section 3.4.3, the Kaplan-Meier method will be used to estimate the TTD curve for each treatment group. The estimate of median TTD and its 95% CI 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 (ie, HR). The HR and its 95% CI will be reported. The same stratification factors used for randomization (see Section 6.3.2 in protocol) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model.

The approach for the time-to-deterioration 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

3.6.3.5 Overall Improvement and Overall Improvement/Stability

Overall improvement rate will be analyzed, which is defined as the proportion of participants who have achieved an improvement as defined in Section 3.4.3 PRO Endpoints. The point estimate of overall proportions of participants who have achieved an improvement, stability and deterioration will be provided by treatment group together with 95% CI using exact binomial method by Clopper and Pearson (1934). 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, along with nominal two-sided p-values, from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be provided. The stratification factors used for randomization (see Section 6.3.2 in protocol) will be applied to the analysis.

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

3.6.3.6 Analysis Strategy for Key PRO Endpoints

A summary of the analysis strategy for the key PRO endpoints is provided in Table 7 below.

C Confidential

Table 7 Summary of 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 • Dyspnea – EORTC QLQ-LC13 • Cough • Chest pain	cLDA model	FAS	Model-based.
TTD in – EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning • Dyspnea – EORTC QLQ-LC13 • Cough • Chest pain	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
Overall improvement and overall improvement/stability in – EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning • Dyspnea – EORTC QLQ-LC13 • Cough • Chest pain	Stratified Miettinen and Nurminen method	FAS	Participants with missing data are considered not achieving improvement/stability.
Abbreviations: cLDA = constrained longitudinal data analysis, FAS = full analysis set, QoL = quality of life, TTD = Time to deterioration.			

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

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 EOC of the Sponsor. If the eDMC recommends modifications to the design of the protocol or

C Confidential

discontinuation of the study, this EOC and potentially other limited Sponsor personnel may be unblinded to the treatment-level results 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 analyses 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.

The pre-specified futility criteria (Table 11 and Table 12) were met at the futility analysis before efficacy IA, and the Sponsor decided to discontinue treatment with MK 7684A. No additional efficacy analysis will be conducted at IA or FA; information regarding planned IAs in Section 3.7.1 are being retained for historical purposes.

3.7.1 Efficacy Interim Analyses

One interim analysis is planned in addition to the FA for this study. Results of the interim analyses will be reviewed by the eDMC. 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.

C Confidential

Table 8 Summary of Interim and Final Analyses Strategy

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA	PFS OS	Both ~237 OS events have been observed in in participants with PD-L1 TPS≥1% and ~22 months after last participant randomized	~36 months	• Interim OS analysis in participants with PD-L1 TPS≥1% and all participants • Final PFS analysis in all participants
FA	OS	Both ~302 OS events have been observed in participants with PD-L1 TPS⩾1% and ~36 months after last participant randomized	~50 months	• Final OS analysis in participants with PD- L1 TPS≥1% and all participants

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

Note that for IA, if events accrue slower than expected, the Sponsor may conduct the analysis with up to an additional 3 months of follow-up beyond the planned calendar time, or the specified number of events is observed, whichever occurs first.

For FA, if events accrue slower than expected, the Sponsor may conduct the analysis with up to an additional 8 months of follow-up beyond the planned calendar time, or the specified number of events is observed, whichever occurs first.

3.7.2 Safety Interim Analyses

The eDMC will conduct regular safety monitoring. The timing of the safety monitoring will be specified in the eDMC charter. eDMC monitoring for safety will be conducted approximately every 6 months until such time that the eDMC determines that monitoring at a different frequency is appropriate.

3.7.3 Futility Analyses

At IA, non-binding futility analysis based on OS in participants with PD-L1 TPS≥1% will be performed. The overall safety and efficacy data will be reviewed for the decision. Details of the boundaries for establishing statistical significance with regard to futility are discussed further in Section 3.8.

One additional non-binding futility analysis will be conducted prior to the efficacy interim analysis, and will be triggered when ~198 OS events have been observed in participants with PD-L1 TPS≥1% and ~16 months after last participant randomized, whichever comes later.

C Confidential

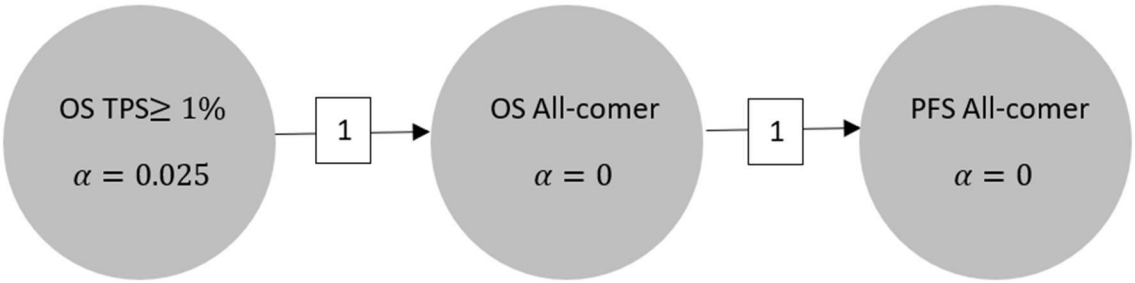
The results of the non-binding futility analysis along with safety data will be reviewed by eDMC to assess the overall risk/benefit to study participants Details of the boundaries for establishing statistical significance with regard to futility are discussed further in Section 3.8.

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [Maurer, W., et al 2011] to provide strong multiplicity control 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 reallocation from each hypothesis to the others are shown in the boxes on the lines connecting hypotheses.

The initial α of 2.5% will be allocated to OS in participants with PD-L1 TPS≥1%, and no alpha will be allocated to OS in all participants or PFS in all participants initially. If the OS null hypothesis for participants with PD-L1 TPS≥1% is rejected, then all alpha will be reallocated to the OS hypothesis for all participants. If the OS null hypotheses for both participants with PD-L1 TPS≥1% and all participants are rejected, then all 0.025 alpha will be reallocated to the PFS hypothesis for all participants.

Figure 1 Multiplicity Diagram for Type I Error Control



Abbreviations: OS=overall survival; PFS=progression-free survival; TPS=tumor proportion score.

The study will be considered a success if OS in participants with PD-L1 TPS≥1% is demonstrated to be statistically significant at either the interim analysis or the final analysis under multiplicity control.

3.8.1 Progression-free Survival

The study will test PFS at IA only. Following the multiplicity strategy as outlined in Figure 1, the PFS hypothesis may be tested at α=0.025 only if the OS null hypotheses for both participants with PD-L1 TPS≥1% and all participants are rejected. Table 9 shows the boundary properties for the PFS analysis. Note that if the OS null hypothesis for all

C Confidential

participants is also rejected at either IA or FA after the OS null hypothesis for participants with PD-L1 TPS≥1% is rejected at either IA or FA, then the test statistic computed at IA for the PFS hypothesis will be used for inferential testing with an alpha level of 0.025.

Table 9 Efficacy Boundaries and Properties for PFS Analyses

Analysis	**Value**	**α=0.025**
IA: 100%* Events: 517 Month: 36	Z	1.9600
	p (1-sided)[a]	0.0250
	HR at bound[b]	0.8416
	P(Cross) if HR=1[c]	0.0250
	P(Cross) if HR=0.7[d]	0.9820

Abbreviations: HR=hazard ratio; IA=interim analysis.
The number of events and timings are estimated.
*Percentage of total planned events at the interim analysis.
[a] *p* (1-sided) is the nominal α for group sequential 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.

3.8.2 Overall Survival

The study will test OS at IA and FA. The OS hypothesis for all participants will be tested only if OS hypothesis for participants with PD-L1 TPS≥1% is rejected. A futility analysis on OS in participants with PD-L1 TPS≥1% will also be conducted at IA. For the non-binding futility analysis at IA, the study will pass futility for OS in participants with PD-L1 TPS≥1% if the observed 1-sided nominal p-value <0.38. Table 10 shows the futility bounds and boundary properties for IA OS futility analysis in participants with PD-L1 TPS≥1%.

One additional non-binding futility analysis will be conducted prior to efficacy interim analyses per described in 3.7.3. For this additional non-binding futility analysis, the study will pass futility for OS in participants with PD-L1 TPS≥1% if the observed 1-sided nominal p-value < 0.42, and will pass futility for OS in all participants if the observed 1-sided nominal p-value < 0.50. Table 11 shows the non-binding futility bounds and boundary properties for OS in participants with PD-L1 TPS≥1% at the additional futility analysis. Table 12 shows the non-binding futility bounds and boundary properties for OS in all participants at the additional futility analysis.

Following the multiplicity strategy as outlined in Figure 1, the OS hypothesis in participants with PD-L1 TPS≥1% will be tested at α=0.025 (initially allocated α). OS hypothesis for all participants will be tested at α=0.025 only if OS hypothesis for participants with PD-L1 TPS≥1% is rejected. A Lan-DeMets O'Brien-Fleming approximation alpha-spending function is constructed to implement group sequential boundaries that control the Type I error [Lan, K. K. G. and DeMets, D. L. 1983] [O'Brien, P. C. and Fleming, T. R. 1979].

Table 10 Non-Binding Futility Bounds and Properties for OS Futility Analysis in participants with PD-L1 TPS≥1% at Interim Analysis

Analysis	**Value**	
IA: 78%* Events: 237 Month: 36	Z	0.3055
	p (1-sided)[a]	0.3800
	HR at bound[b]	0.9609
	P(Futility) if HR=0.72[c]	0.0136
	P(Futility) if HR=1 [c]	0.6200
Abbreviations: HR = hazard ratio; IA = interim analysis. The number of events and timings are estimated. * Percentage of total planned events at the interim analysis. [a] *p* (1-sided) is the value (calculated under the null hypothesis of HR=1) boundary for the futility analysis. [b] HR at bound is the approximate HR required to reach a futility bound. [c] P (Futility) is the probability of stopping for futility under the null hypothesis and different HR scenarios.		

Table 11 Non-Binding Futility Bounds and Properties for OS in participants with PD- L1 TPS≥1% at Futility Analysis

Analysis	**Value**	
Futility Analysis: 65%* Events: 198 Month: 30	Z	0.2019
	p (1-sided)[a]	0.4200
	HR at bound[b]	0.9715
	P(Futility) if HR=0.72[c]	0.0181
	P(Futility) if HR=1 [c]	0.5800
Abbreviations: HR = hazard ratio. The number of events and timings are estimated. * Percentage of total planned events at the interim analysis. [a] *p* (1-sided) is the value (calculated under the null hypothesis of HR=1) boundary for the futility analysis. [b] HR at bound is the approximate HR required to reach a futility bound. [c] P (Futility) is the probability of stopping for futility under the null hypothesis and different HR scenarios.		

C Confidential

Table 12 Non-Binding Futility Bounds and Properties for OS in all participants at Futility Analysis

Analysis	**Value**	
Futility Analysis: 66%* Events: 337 Month: 30	Z	0.0000
	p (1-sided)[a]	0.5000
	HR at bound[b]	1.0000
	P(Futility) if HR=0.72[c]	0.0014
	P(Futility) if HR=1 [c]	0.5000

Abbreviations: HR = hazard ratio.

The number of events and timings are estimated.

* Percentage of total planned events at the interim analysis.

[a] *p* (1-sided) is the value (calculated under the null hypothesis of HR=1) boundary for the futility analysis.

[b] HR at bound is the approximate HR required to reach a futility bound.

[c] P (Futility) is the probability of stopping for futility under the null hypothesis and different HR scenarios.

Table 13 shows the efficacy boundary properties for OS analyses for in participants with PD-L1 TPS≥1%

Table 13 Efficacy Boundaries and Properties for OS Analyses in participants with PD- L1 TPS≥1%

		α=0.025
Analysis	**Value**	Efficacy
IA: 78%* Events: 237 Month: 36	Z	2.2767
	p (1-sided)[a]	0.0114
	HR at bound[b]	0.7426
	P(Cross) if HR=1[c]	0.0114
	P(Cross) if HR=0.72[d]	0.5988
FA Events: 302 Month: 50	Z	2.0208
	p (1-sided)[a]	0.0216
	HR at bound[b]	0.7917
	P(Cross) if HR=1[c]	0.0250
	P(Cross) if HR=0.72[d]	0.8090

Abbreviations: HR=hazard ratio, IA=interim analysis, FA=final analysis.
The number of events and timings are estimated.
*Percentage of total planned events at the interim analysis.
[a]*p* (1-sided) is the nominal α for group sequential 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=0.72 is the probability of crossing a bound under the alternative hypothesis.

Table 14 shows the efficacy boundary properties for OS analyses for all participants.

Table 14 Efficacy Boundaries and Properties for OS Analyses for all participants

		α=0.025
Analysis	**Value**	Efficacy
IA: 79%* Events: 402 Month: 36	Z	2.2653
	p (1-sided)[a]	0.0117
	HR at bound[b]	0.7967
	P(Cross) if HR=1[c]	0.0117
	P(Cross) if HR=0.72[d]	0.8466
FA Events: 508 Month: 50	Z	2.0226
	p (1-sided)[a]	0.0216
	HR at bound[b]	0.8351
	P(Cross) if HR=1[c]	0.0250
	P(Cross) if HR=0.72[d]	0.9581

Abbreviations: HR=hazard ratio, IA=interim analysis, FA=final analysis.
The number of events and timings are estimated.
*Percentage of total planned events at the interim analysis.
[a]*p* (1-sided) is the nominal α for group sequential 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=0.72 is the probability of crossing a bound under the alternative hypothesis.

The bounds provided in Table 10 and Table 13 above are based on the assumption that the expected number of OS events in participants with PD-L1 TPS≥1% at IA and FA are ~237 and ~302, respectively. The bounds provided in Table 14 are based on the assumption that the expected numbers of OS events for all participants at IA and FA are 402 and 508, respectively. At the time of an analysis, the observed number of events may differ substantially from the expected. To avoid overspending alpha 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.

C Confidential

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 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.

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.

3.8.3 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 PFS and OS taking a nominal alpha penalty of 0.000001 for each of such incidence will be conducted.

3.9 Sample Size and Power Calculations

The actual sample size is 739 participants in a 1:1 ratio into the MK-7684A plus chemotherapy or pembrolizumab plus chemotherapy arms. Approximately 453 participants are participants with PD-L1 TPS≥1%. OS in participants with PD-L1 TPS≥1% is the primary endpoint for the study, with OS and PFS in all participants as key secondary endpoints.

For the primary OS endpoint, based on a target number of ~302 events in participants with PD-L1 TPS≥1% and 1 interim analysis at approximately 78% of the target number of events, the study has approximately 81% power to detect a hazard ratio of 0.72 at the initially allocated α=0.025 (1-sided).

For the secondary OS endpoint, based on a target number of ~508 events in all participants and 1 interim analysis at approximately 79% of the target number of events, the study has approximately 96% power to detect a hazard ratio of 0.72 at the initially allocated α=0.025 (1-sided) if the OS null hypothesis for in participants with PD-L1 TPS≥1% is rejected.

For the secondary PFS endpoint, based on a target number of ~517 events in all participants at the final PFS, the study has approximately 98% power to detect a hazard ratio of 0.7 at α=0.025 (1-sided) if the OS null hypotheses for both participants with PD-L1 TPS≥1% and all participants are rejected.

Note that the above power calculations are based on a constant hazard ratio assumption. The interim analysis timing and spending have been designed to ensure the final alpha available is maximized to make testing most sensitive when follow-up is available across both early and late parts of the survival and PFS distributions.

Based on KEYNOTE-189 and KEYNOTE-407 data, the above sample size and power calculations for PFS and OS assume the following:

- KEYNOTE-189 (nonsquamous)
  - PFS follows a piecewise exponential distribution with a median of 8.8 months for the first 15 months and a median of 17 months afterwards for the pembrolizumab plus chemotherapy group.
  - OS follows an exponential distribution with a median of 22.0 months for the pembrolizumab plus chemotherapy group.
  - OS follows an exponential distribution with a median of 23.0 months for the pembrolizumab plus chemotherapy group in participants with PD-L1 TPS≥1%.
- KEYNOTE-407 (squamous)
  - PFS follows a piecewise exponential distribution with a median of 7.6 months for the first 14 months and a median of 25 months afterwards for the pembrolizumab plus chemotherapy group.
  - OS follows an exponential distribution with a median of 17.2 months for the pembrolizumab plus chemotherapy group.
  - OS follows an exponential distribution with a median of 18.9 months for the pembrolizumab plus chemotherapy group in participants with PD-L1 TPS≥1%.
- Approximately 30% of the population being squamous
- Enrollment period of 13.5 months
- Annual dropout rates of 10% and 1% for PFS and OS, respectively
- Follow-up periods of 22 and 36 months for PFS and OS, respectively, after the last participant is randomized.

The sample size and power calculations were performed using R (“gsDesign2” package).

C Confidential

3.10 Subgroup Analyses

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

- Stratification factors based on eCRF collected information
  - Baseline ECOG PS (0 vs 1)
  - Predominant tumor histology (Squamous vs nonsquamous)
  - PD-L1 expression (TPS<50% vs ≥50%)
  - Geographic region (East Asia vs North America/Western Europe vs Rest of the World)
- PD-L1 TPS (<1% vs 1% to 49% vs ≥50%)
- PD-L1 TPS (<1% vs ≥1%)
- Age category (<65 years vs ≥65 years)
- Sex (female vs male)
- Race (white vs nonwhite)
- Smoking status (never vs former/current smoker)
- Baseline brain metastasis status (presence vs absence)
- Baseline liver metastasis status (presence vs absence)

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 will be conducted using the unstratified Miettinen and Nurminen method.

C Confidential

3.11 Compliance (Medication Adherence)

Drug accountability data for study 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 in which the participant receives the study intervention. Summary statistics will be provided on the extent of exposure for the APaT population.

C Confidential

4 REFERENCES

[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 2011]	Cocks K, King MT, Velikova G, Martyn St-James M, Fayers PM, Brown JM. Evidence-based guidelines for determination of sample size and interpretation of the European organisation for the research and treatment of cancer quality of life questionnaire core 30. J Clin Oncol. 2011 Jan 1;29(1):89-96.	[04SL8Q]
[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]
[Lan, K. K. G. and DeMets, D. L. 1983]	Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70(3):659-63.	[03P3QC]
[Liang, K-Y. and Zeger, S. L. 2000]	Liang K-Y and Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhya: the Indian Journal of Statistics 2000;62:134-48.	[03RSBZ]

C Confidential

[Maringwa, J. T., et al 2011]	Maringwa JT, Quinten C, King M, Ringash J, Osoba D, Coens C, et al. Minimal important differences for interpreting health-related quality of life scores from the EORTC QLQ-C30 in lung cancer patients participating in randomized controlled trials. Support Care Cancer. 2011 Nov;19(11):1753-60.	[04GM02]
[Maurer, W., et al 2011]	Maurer W, Glimm E, Bretz F. Multiple and repeated testing of primary, coprimary, and secondary hypotheses. Stat Biopharm Res. 2011;3(2):336-52.	[045MYM]
[Miettinen, O. and Nurminen, M. 1985]	Miettinen O and Nurminen M. Comparative analysis of two rates. Stat Med 1985;4:213-26.	[00VMQY]
[O'Brien, P. C. and Fleming, T. R. 1979]	O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials [Abstract]. Biometrics 1979;35(3):549-56.	[00VPPS]
[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]
[Pickard, A. S., et al 2007]	Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes 2007;5:1-8.	[00W0FM]

[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]