Official Protocol Title:	A Phase 3, randomized, double-blind trial of pembrolizumab (MK-3475) with or without lenvatinib (E7080/MK-7902) in participants with treatment-naïve, metastatic non-small cell lung cancer (NSCLC) whose tumors have a tumor proportion score (TPS) greater than or equal to 1% (LEAP-007)
NCT number:	NCT04676412
Document Date:	15-JUL-2021

Supplemental Statistical Analysis Plan (sSAP)

Table of Contents

1. INTRODUCTION 7
2. SUMMARY OF CHANGES 7
3. ANALYTICAL AND METHODOLOGICAL DETAILS FOR GLOBAL STUDY 7
3.1. Statistical Analysis Plan Summary 7
3.2. Responsibility for Analyses/In-house Blinding 8
3.3. Hypotheses/Estimation 9
3.4. Analysis Endpoints 9
3.4.1. Efficacy Endpoints 9
3.4.1.1. Primary Endpoints 9
3.4.1.2. Secondary Endpoints 9
3.4.1.3. Exploratory Endpoints 9
3.4.2. Safety Endpoints 9
3.4.3. Patient-reported Outcome Endpoints 9
3.5. Analysis Populations 11
3.5.1. Efficacy Analysis Populations 11
3.5.2. Safety Analysis Populations 11
3.5.3. Patient-reported Outcome Analysis Population 12
3.5.4. Pharmacokinetic Analysis Population 12
3.6. Statistical Methods 12
3.6.1. Statistical Methods for Efficacy Analyses 12
3.6.1.1. Progression-free Survival 13
3.6.1.2. Overall Survival 14
3.6.1.3. Objective Response Rate 15
3.6.1.4. Duration of Response 15
3.6.2. Statistical Methods for Safety Analyses 15
3.6.3. Statistical Methods for Patient-reported Outcome Analyses 17
3.6.3.1. Scoring Algorithm 17
3.6.3.2. PRO Compliance Summary 19
3.6.3.3. Mean Change from Baseline 20
3.6.3.4. Time-to-True Deterioration (TTD) Analysis 21
3.6.3.5. Overall Improvement and Overall Improvement/Stability 22
3.6.3.6. Analysis Strategy for Key PRO Endpoints 22

3.6.4. Demographic and Baseline Characteristics 23
3.7. Interim Analyses 23
3.7.1. Efficacy Interim Analysis 24
3.7.2. Safety Interim Analyses 25
3.8. Multiplicity 25
3.8.1. Objective Response Rate 26
3.8.2. Progression-free Survival 26
3.8.3. Overall Survival 27
3.8.4. Safety Analyses 29
3.9. Sample Size and Power Calculations 29
3.10. Subgroup Analyses 30
3.11. Compliance (Medication Adherence) 30
3.12. Extent of Exposure 30
4. STATISTICAL ANALYSIS PLAN FOR EXTENSION 31
4.1. Introduction 31
4.2. Responsibility for Analyses/In-House Blinding 31
4.3. Hypotheses/Estimation 31
4.4. The Analysis Endpoints 32
4.4.1. Efficacy Endpoints 32
4.4.1.1. Primary Endpoints 32
4.4.1.2. Secondary Endpoints 32
4.4.1.3. Exploratory Endpoints 32
4.4.2. Safety Endpoints 32
4.4.3. Patient-reported Outcome Endpoints 32
4.5. Analysis Populations 33
4.5.1. Efficacy Analysis Populations 33
4.5.2. Safety Analysis Populations 33
4.5.3. Patient-reported Outcome Analysis Population 33
4.5.4. Pharmacokinetic Analysis Population 34
4.6. Statistical Methods 34
4.6.1. Statistical Methods for Efficacy Analyses 34
4.6.1.1. Progression-free Survival 34
4.6.1.2. Overall Survival 35

4.6.1.3. Objective Response Rate 35
4.6.1.4. Duration of Response 35
4.6.2. Statistical Methods for Safety Analyses 36
4.6.3. Statistical Methods for PRO Analysis 36
4.6.4. Summaries of Baseline Characteristics, Demographics, and Other Analyses 36
4.7. Interim Analysis and Final analysis 36
4.8. Multiplicity 36
4.9. Sample Size and Power Calculations 36
5. APPENDICES: 37
5.1. Rules for Non-Binding Futility Analyses 37
5.2. ABBREVIATIONS 39
6. REFERENCES 42

List of Tables

Table 1 Analysis Methods for Key Efficacy Endpoints 13
Table 2 Censoring Rules for Primary and Sensitivity Analyses of Progression-free Survival 14
Table 3 Censoring Rules for Duration of Response 15
Table 4 Analysis Strategy for Safety Parameters 17
Table 5 PRO Data Collection Schedule and Mapping of Study visit to Analysis Visit 20
Table 6 Censoring Rules for Time-to-Deterioration 22
Table 7 Analysis Strategy for Key PRO Endpoints 23
Table 8 Summary of Interim and Final Analyses Strategy 25
Table 9 Efficacy Boundaries and Properties for Progression-free Survival Analyses 27
Table 10 Efficacy Boundaries and Properties for Overall Survival Analyses 28
Table 11 Futility Boundaries for OS in July 2021 and at IA 38
Table 12 Non-binding Cumulative Probability of Stopping for OS 38

List of Figures

Figure 1 Multiplicity Graph for Type I Error Control 26
Figure 2 Decision Tree for Futility Assessment at July 2021 DMC 39

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

As the observed PFS event number at the July 2021 DMC was more than the expected PFS event number at the IA (PFS final analysis), the sSAP was amended to update the non-binding futility criteria for July 2021 DMC in Appendix 5.1..

3. ANALYTICAL AND METHODOLOGICAL DETAILS FOR GLOBAL STUDY

3.1. Statistical Analysis Plan Summary

Study Design Overview	Phase 3 study of pembrolizumab plus lenvatinib vs pembrolizumab plus placebo for 1L treatment of metastatic NSCLC in participants whose tumors express PD-L1 (TPS ≥1%)
Treatment Assignment	Approximately 620 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: • Geographic region (East Asia vs non-East Asia) • ECOG PS (0 vs 1) • TPS (1%-49% vs ≥50%)
Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Participants as Treated (APaT) PRO: Full Analysis Set (FAS)
Primary Endpoints	• PFS per RECIST 1.1 based on BICR • OS
Secondary Endpoints	• OR per RECIST 1.1 based on BICR • Safety and tolerability • Mean change from baseline in Global health status/QoL, cough, chest pain, dyspnea and physical functioning scores • Time to True Deterioration (TTD) in global health status/QoL, cough, chest pain, dyspnea, physical functioning, and composite endpoint (combination of cough, chest pain, or dyspnea items)
Statistical Methods for Key Efficacy Analyses	The primary hypotheses will be evaluated by comparing pembrolizumab + lenvatinib to pembrolizumab + placebo in PFS and OS using a stratified log-rank test and in ORR using the stratified Miettinen and Nurminen method [Miettinen, O. 1985]. 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. The difference in ORR will be estimated using the stratified Miettinen and Nurminen method with strata weighting by sample size.

Statistical Methods for Key Safety Analyses	The primary hypotheses will be evaluated by comparing pembrolizumab + lenvatinib to pembrolizumab + placebo in PFS and OS using a stratified log-rank test and in ORR using the stratified Miettinen and Nurminen method [Miettinen, O. 1985]. 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. The difference in ORR will be estimated using the stratified Miettinen and Nurminen method with strata weighting by sample size.

Interim Analyses	One interim analysis and one final analysis are planned for this study. Results will be reviewed by an external Data Monitoring Committee (DMC). Details are provided in Section 9.7.
	• Interim analysis (IA): to be performed after ~416 PFS events are observed and ~8.8 months after last participant randomized
	  ○ Primary purpose: final PFS analysis and interim OS analysis
	• Final analysis: to be performed after ~388 deaths are observed and ~18.5 months after last participant randomized
	  ○ Primary purpose: final OS analysis
	Note that for the IA and the FA, if the events accrue slower than expected such that the targeted number of events cannot be reached in the anticipated timeframe, the Sponsor may conduct the analysis with an additional 2 months and 7 months of follow-up for the IA and FA respectively, or the specified number of events is observed, whichever occurs first.

Multiplicity	The overall Type I error rate over the multiple endpoints will be strongly controlled at 2.5% (1-sided). A 0.55% (1-sided) will be initially allocated to test PFS and 1.95% (1-sided) allocated to test OS. The graphical approach of Maurer and Bretz [Maurer, W., et al 2011] will be applied to reallocate α among the hypotheses for ORR, PFS, and OS. Lan-DeMets [Lan, K. K. G. and DeMets, D. L. 1983] and O'Brien-Fleming [O'Brien, P. C. and Fleming, T. R. 1979] group sequential methods will be used to allocate α among the interim and final analysis for the OS endpoint.
Sample Size and Power	The planned sample size is approximately 620 participants. For PFS, based on 416 events, the study has 86.5% power to detect a HR of 0.7 (pembrolizumab + lenvatinib vs pembrolizumab + placebo) at α=0.55% (1-sided). For OS, based on 388 events, the study has 90% power to detect a HR of 0.71 (pembrolizumab + lenvatinib vs pembrolizumab + placebo) at α=1.95% (1-sided).
China Extension Study	China participants randomized during the global study phase will be included in all global study analyses (efficacy and safety). China participants randomized during the China extension phase will be excluded from all global study analyses. China participants randomized during global and extension phases will both be included in the China-specific analyses.

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 intervention assignment.

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

3.3. Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Section 3 of the study protocol.

3.4. Analysis Endpoints

Efficacy and safety endpoints that will be evaluated for within- and between-treatment differences are listed below.

3.4.1. Efficacy Endpoints

3.4.1.1. Primary Endpoints

***PFS*:** the time from randomization to the first documented PD per RECIST 1.1, adjusted to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ, based on BICR or death due to any cause, whichever occurs first. See Section 9.6.1 for the definition of censoring.

***OS*:** the time from randomization to death due to any cause.

3.4.1.2. Secondary Endpoints

***OR*:** a confirmed CR or PR per RECIST 1.1, adjusted to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ, based on BICR.

3.4.1.3. Exploratory Endpoints

***DOR*:** the time from the earliest date of qualifying response until earliest date of PD or death from any cause, whichever comes first.

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 Section 8.4.7 of the study protocol.

3.4.3. Patient-reported Outcome Endpoints

Secondary

- Mean change from baseline in EORTC QLQ-C30 global health status/QoL (items 29-30), physical functioning (items 1-5), and dyspnea scores (item 8), and EORTC QLQ-LC13 cough (item 31) and chest pain (item 40) scores.

- TTD as measured by each of the EORTC QLQ-C30 global health status/QoL, physical functioning, and dyspnea scores, and EORTC QLQ-LC13 cough and chest pain scores.

- TTD as measured by a composite symptom endpoint: cough (QLQ-LC13 item 31), chest pain (QLQ-LC13 item 40), or dyspnea (QLQ-C30 item 8)

Based on prior literature (Osoba D et al., 1998; King et al., 1996, Maringwa et al., 2011), a ≥10 point or greater worsening from baseline for each scale represents a minimally important difference (MID) that represents a clinically relevant deterioration. TTD in the individual scores specified above is defined as the time to first onset of ≥10 (out of 100) point deterioration from baseline in a given scale/subscale/item and confirmed by a second adjacent ≥10 point deterioration from baseline. TTD in the composite endpoint is defined as the time to first onset of a ≥10 point deterioration (i.e., increase in score) from baseline in any one of the 3 lung cancer symptom scale items (EORTC-QLQ-C30 [dyspnea] and EORTC-QLQ-LC13 [cough and chest pain]) with confirmation by the subsequent visit of a ≥10 point deterioration (i.e., increase in score) from baseline in the same scale as the first onset. 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)

Exploratory

- Change from baseline in the additional scales of the EORTC QLQ-C30, EORTC QLQ-LC13, and EuroQoL EQ-5D visual analogue scale (VAS) scores. Specifically:

    - Each EORTC QLQ-C30 functional scales: role functioning, emotional functioning, cognitive functioning, and social functioning

    - Each of three EORTC QLQ-C30 symptom scales (fatigue, nausea/vomiting, and pain), and single item measures (sleep disturbance, appetite loss, constipation, diarrhea, and financial difficulties)

    - Each EORTC QLQ-LC13 item: site-specific pain (pain in arm or shoulder, and pain in other parts), sore mouth, dysphagia, peripheral neuropathy, alopecia, and hemoptysis

    - The EQ-5D VAS

- Proportions of improvement / stability / stability + improvement / deterioration in EORTC QLQ-C30 global health status / QoL, physical functioning, and dyspnea scores, and EORTC QLQ-LC13 cough and chest pain scores where

    - Improvement

        - Improvement for the global health status/QoL and physical functioning scores is defined by an increase of 10 points or more in score from baseline at any time during the study and confirmed by an increase of 10 points or more in score at the next consecutive visit.

- Improvement for the dyspnea, cough, and chest pain scores is defined by a decrease of 10 points or more in score from baseline at any time during the study and confirmed by a decrease of 10 points or more in score at the next consecutive visit.

- Stability is defined as follows, when the criteria for improvement are not met:
  - an improvement confirmed by a less than 10 points change from baseline at the next consecutive visit
  - less than 10 points change in score confirmed by a less than 10 points change in score at the next consecutive visit
  - less than 10 points change in score and confirmed by an improvement at the next consecutive visit
- Stability + improvement is defined as the sum of proportions with stability and improvement as specified above.
- Deterioration is defined as a 10 points or greater worsening from baseline at any time during the study when the criteria for improvement or stability are not met.

3.5. Analysis Populations

3.5.1. Efficacy Analysis Populations

The analyses of the primary efficacy endpoints are based on the ITT population. All randomized participants will be included in this population. Participants will be analyzed in the treatment group to which they were randomized. Based on actual enrollment, all participants who were randomized on or prior to January 28, 2021 will be included in the ITT population of the global study. All participants who failed screening in addition to all randomized participants on or prior to January 28, 2021will be included in the screening population. Details on the approach to handling missing data are provided in Section 3.6.

3.5.2. Safety Analysis Populations

Safety analyses will be conducted in the APaT population, which consists of all randomized participants who received at least 1 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 take 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 1 cycle, but receives the randomized treatment 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 1 laboratory, vital sign, or ECG measurement obtained subsequent to at least 1 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 Population

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

3.5.4. Pharmacokinetic Analysis Population

The population of PK analysis set includes all the participants who have received at least 1 dose of study intervention with documented dosing history in the lenvatinib + pembrolizumab 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 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 because of potential issues of multiplicity.

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

The efficacy analyses for OR, DOR and PFS will include responses and documented progression events that occur prior to second course treatment.

A summary of the primary analysis strategy for the key efficacy endpoints is provided in Table 1.

Table 1 Analysis Methods for Key Efficacy Endpoints

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Primary Analyses:			
PFS (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 2
OS	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie-handling method	ITT	Censored at last known alive date
Secondary Analyses:			
ORR (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 = intention-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 in Section 3.8, Multiplicity.

3.6.1.1. Progression-free Survival

The nonparametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. 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 difference (ie, HR) between the treatment arms. The HR and its 95% CI 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 (Section 6.3.2 in the protocol) will be applied to both the stratified log-rank test and the stratified Cox model.

Since PD 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 PD will be approximated by the date of the first assessment at which PD is objectively documented per RECIST 1.1 by the BICR vendor, regardless of discontinuation of study drug.

To evaluate the robustness of the PFS endpoint per RECIST 1.1 via BICR by the imaging vendor, 1 primary and 2 sensitivity analyses with a different set of censoring rules will be performed. For the primary analysis, if the events (PD or death) are immediately after more than 1 consecutive 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 2.

Table 2 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 for reasons other than CR; otherwise, censored at last disease assessment if still on 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 therapy	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 therapy	Censored at last disease assessment before the earlier date of ≥2 consecutive missed disease assessment and new anticancer therapy	Progressed at date of documented PD or death	Progressed at date of documented PD or death
Abbreviation: CR = complete response; PD = progressive disease.			

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

3.6.1.2. Overall Survival

The nonparametric 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 the 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 HR). 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 (Section 6.3.2 in the 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 of last known contact.

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

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 (Section 6.3.2 in the protocol) will be used as stratification factors in the analysis.

3.6.1.4. Duration of Response

For subjects who demonstrate CR or PR, DOR is defined as the time from first documented evidence of CR or PR until PD or death due to any cause, whichever occurs first.

The nonparametric Kaplan-Meier method will be used to summarize the DOR. The median and range of DOR will be provided.

Censoring rules for DOR are summarized in Table 3.

Table 3 Censoring Rules for Duration of Response

Situation	Date of Progression or Censoring	Outcome
No PD or death, no new anticancer therapy initiated	Last adequate disease assessment	Censor (nonevent)
No PD or death, new anticancer therapy initiated	Last adequate disease assessment before new anticancer therapy initiated	Censor (nonevent)
Death or PD immediately after ≥2 consecutive missed disease assessments or after new anticancer therapy, if any	Earlier date of last adequate disease assessment before ≥2 missed adequate disease assessments and new anticancer therapy, if any	Censor (nonevent)
Death or PD after ≤1 missed disease assessments and before new anticancer therapy, if any	PD or death	End of response (event)
Abbreviation: PD = progressive disease. Note: A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response.		

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. Finally, there are no known AEs associated with participants with NSCLC for which determination of a *p* value is expected to impact the safety assessment. 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 [Miettinen, O. 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 randomized in this study are in critical conditions and usually experience 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 one of the treatment groups) and SAEs (≥5% of participants in one of the treatment groups) will be considered Tier 2 endpoints. Because many 95% CIs may be provided without adjustment for multiplicity, the CIs should be regarded as a helpful descriptive measure to be used in safety review, not 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.

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	p-Value	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
	Serious AE (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 Serious AE			X
	Any Drug-Related AE			X
	Any Serious and Drug-Related AE			X
	Any Grade 3-5 and Drug-Related AE			X
	Discontinuation due to AE			X
	Death			X
	Specific AEs, SOCs (incidence <10% of participants in all of the treatment groups)			X
	Change from Baseline Results (lab toxicity shift)			X
Abbreviations: AE = adverse event; CI = confidence interval; SOC = system organ class.				

Exposure-Adjusted Approach

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

Time to Grade 3-5 AE

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

3.6.3. Statistical Methods for Patient-reported Outcome Analyses

3.6.3.1. Scoring Algorithm

EORTC QLQ-C30:

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.

The scoring method for each subscale of the EORTC-QLQ-C30 will be performed according to the scoring manual (Scott, et al., 2008). According to the EORTC QLQ-C30 manual, if items $I_1, I_2, \cdots, I_n$ are included in a scale, the linear transformation procedure is as follows:

1. Compute the raw score:

$$RS = (I_1 + I_1 + \cdots + I_n)/n$$

2. Linear transformation to obtain the score *S*:

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

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

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

The range is defined as the difference between the maximum possible value of *RS* and the minimum possible value for RS. 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.

EORTC QLQ-LC13:

The lung cancer questionnaire module is a validated self-reported PRO questionnaire that is intended to be used in conjunction with the QLQ-C30 and includes 13 questions that share the four level ordinal response with the QLQ-C30 and can be summarized and scored as described above (1= “not at all”; 2= “a little”; 3= “quite a bit”; 4= “very much”). The LC13 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 (least severe symptom) and 100 (most severe symptom) as described above for the EORTC QLQ-C30 symptom scales/items scoring.

EQ-5D:

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 (Packard et al., 2007).

3.6.3.2. PRO Compliance Summary

Completion and compliance of EORTC QLQ-C30, EORTC-QLQ-LC13, 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 each respective 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.

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

The completion rate is expected to 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.

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

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. If there are multiple PRO collections within any of the stated time windows, the assessment completed closest to the target collection day will be used in the analyses.

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

Treatment Week	Target Day	Range
0	1	-28 to 1
3	22	2 to 32
6	43	33 to 53
9	64	54 to 74
12	85	75 to 95
15	106	96 to 116
18	127	117 to 137
21	148	138 to 158
24	169	159 to 179
27	190	180 to 200
30	211	201 to 221
33	232	222 to 242
36	253	243 to 263
39	274	264 to 284
42	295	285 to 305
45	316	306 to 326
48	337	327 to 358
54	379	359 to 400
60	421	401 to 442
66	463	443 to 484
72	505	485 to 526
78	547	527 to 568
84	589	569 to 610
90	631	611 to 652
96	673	653 to 694
102	715	695 to 736

3.6.3.3. Mean Change from Baseline

The time point for the mean change from baseline analysis will be 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 global health status/QoL, physical functioning and dyspnea scores of the EORTC QLQ-C30 and the cough and chest pain scores of the EORTC QLQ-LC13 scales 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 the 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 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 Y_{ijt} is the PRO score for participant i, with treatment assignment j at visit t; γ_0 is the baseline mean for all treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t; X_i is the stratification factor (binary) vector for this participant, and β is the coefficient vector for stratification factors. An unstructured covariance matrix will be used to model the correlation among repeated measurements. If the unstructured covariance model fails to converge with the default algorithm, then Fisher scoring algorithm or other appropriate methods can be used to provide initial values of the covariance parameters. In the rare event that none of the above methods yield convergence, a structured covariance such as Toeplitz can be used to model the correlation among repeated measurements. In this case, the asymptotically unbiased sandwich variance estimator will be used. The cLDA model implicitly treats missing data as missing at random (MAR).

Line plots for the empirical mean change from baseline in the EORTC QLQ-C30 global health status/QoL, physical functioning and dyspnea scores and EORTC QLQ-LC13 cough and chest pain scores (as specified in section 3.4.3) 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 for the EORTC QLQ-C30 global health status/QoL, physical functioning and symptom scales, and EORTC QLQ-LC13 symptom scales (as specified in section 3.4.3).

3.6.3.4. Time-to-True Deterioration (TTD) Analysis

The non-parametric Kaplan-Meier method will be used to estimate the deterioration curve in each group. The estimate of median time to deterioration and its 95% confidence interval will be obtained from the Kaplan-Meier survival estimates. The treatment difference in TTD will be assessed by the stratified log-rank test, and nominal two-sided p-values will be reported. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (hazard ratio) between treatment arms. Stratification factors used for randomization (See Section 6.3.2 in the protocol) will be applied to the analysis will be used in the stratified Cox PH 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, death 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.

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 • Global health status/QoL • Physical functioning • dyspnea – EORTC QLQ-LC13 • Cough • Chest pain	cLDA model	FAS	Model-based.= cLDA
TTD in – EORTC QLQ-C30 • Global health status/QoL • Physical 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 • 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.			

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 reason 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

Blinding to treatment assignment will be maintained at all investigational sites. The results of interim analyses will not be shared with the investigators before the 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 the interim analyses of the study and will make recommendations for discontinuation of the study or protocol modifications to the study Executive Oversight Committee (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.

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

3.7.1. Efficacy Interim Analysis

One interim analysis is planned in addition to the final analysis for this study. For the interim and final analyses, all the randomized participants will be included. Results of the interim analysis will be reviewed by the DMC. Details on 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 Analyses Strategy

Analyses	Key Endpoints	Timing	Estimated Time After First Participant Randomized	Primary Purpose of Analysis
IA	PFS OS	Both ~416 PFS events have been observed and ~8.8 months after last participant randomized	~30.5 months	• Final PFS analysis • Interim OS analysis
Final analysis	OS	Both ~388 deaths have occurred and ~18.5 months after last participant randomized	~40.2 months	• Final OS analysis
Note that for the IA and the FA, if the events accrue slower than expected such that the targeted number of events cannot be reached in the anticipated timeframe, the Sponsor may conduct the analysis with an additional 2 months and 7 months of follow-up for the IA and FA respectively, or the specified number of events is observed, whichever occurs first. Abbreviations: IA = interim analysis; OS = overall survival; PFS = progression-free survival.				

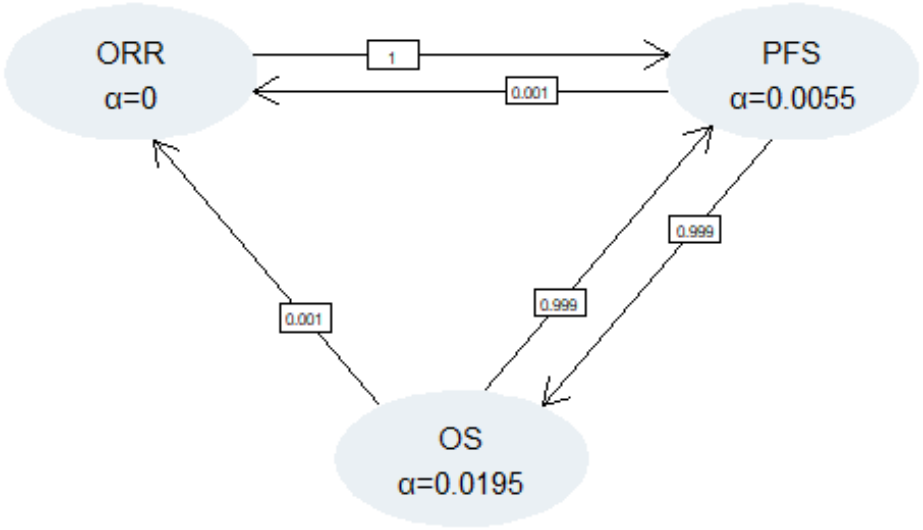
Non-binding futility analyses will be conducted in July 2021 (data cutoff in May 2021) to coincide with an anticipated safety DMC meeting and at the IA as per DMC recommendation. Further details of any evaluation are specified in the Appendix 5.1.

3.7.2. Safety Interim Analyses

The DMC will conduct regular safety monitoring. The timing of the safety monitoring will be specified in the DMC charter.

3.8. Multiplicity

The study uses the graphical method of Maurer and Bretz [Maurer, W., et al 2011] to control multiplicity for multiple hypotheses as well as interim analysis. 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 represented in the boxes on the lines connecting the hypotheses.



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

Note: If both PFS and OS null hypotheses are rejected, the reallocation strategy allows retesting of ORR at α = 0.025.

Figure 1 Multiplicity Graph for Type I Error Control

3.8.1. Objective Response Rate

No initial alpha is allocated to test ORR. If superiorities for both the PFS and OS hypotheses are declared, α = 0.025 will be rolled over to the hypothesis for ORR and the test statistics computed at the IA for the ORR hypothesis will be used for inferential testing with an updated alpha level of 0.025. The power at the α-level of 0.025 is 99.9%, with an approximate treatment difference (ΔORR) required for reaching the efficacy bound being 7.5%, assuming underlying 28% and 48% ORR in the control and experimental groups, respectively.

3.8.2. Progression-free Survival

The study will test PFS only at the interim analysis, at an initial α level of 0.0055. Table 9 shows the boundary properties for each of these α-levels for the PFS analysis. Note that the final row indicates the total power to reject the null hypothesis for PFS at each α level. Also note that if the OS null hypothesis is rejected at the FA, the previously computed PFS test statistics may be used for inferential testing with its updated bounds considering the α re-allocation from the OS hypothesis. Following the multiplicity strategy as outlined in Figure 1, if superiority for both the ORR and OS hypotheses are declared, α = 0.025 will be rolled over to the hypothesis for PFS.

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

Analysis	Value	α=0.0055	α=0.025
IA: 100%[a] N: 620 Events: 416 Month: 30.5[f]	Z	2.5427	1.9600
	p (1-sided)[b]	0.0055	0.0250
	~HR at bound[c]	0.7793	0.8252
	P(Cross) if HR=1[d]	0.0055	0.0250
	P(Cross) if HR=0.7[e]	0.8650	0.9536

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated.
[a] Percentage of total planned events at the interim analysis.
[b] The nominal α for testing.
[c] The approximate HR required to reach an efficacy bound
[d] The probability of crossing a bound under the null hypothesis
[e] The probability of crossing a bound under the alternative hypothesis
[f] The approximate number of months since first participant randomized

3.8.3. Overall Survival

The study will test OS at IA and FA. Following the multiplicity strategy as outlined in Figure *1*, the OS hypothesis may be tested at α=0.0195 (initially allocated α) or α=0.025 (if the null hypotheses for both PFS and ORR are rejected). Table 10 shows the boundary properties for each of these α-levels for OS at IA and FA, which were derived using a Lan-DeMets and O'Brien-Fleming spending function. Note that if the PFS null hypothesis is rejected at IA, OS interim and final analysis test may be compared with its updated bounds considering the α reallocation from the PFS hypothesis.

Table 10 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	**Value**	**α=0.0195**	**α=0.025**
IA: 78%[a] N: 620 Events: 302 Month:30.5 [f]	Z	2.4029	2.2869
	p (1-sided)[b]	0.0081	0.0111
	~HR at bound[c]	0.7584	0.7686
	P(Cross) if HR=1[d]	0.0081	0.0111
	P(Cross) if HR=0.71[e]	0.7173	0.7548
Final N: 620 Events: 388 Month: 40.2 [f]	Z	2.1181	2.0193
	p (1-sided)[b]	0.0171	0.0217
	~HR at bound[c]	0.8064	0.8145
	P(Cross) if HR=1[d]	0.0195	0.0250
	P(Cross) if HR=0.71[e]	0.9000	0.9165

Abbreviations: HR = hazard ratio; IA = interim analysis.
The number of events and timings are estimated.

[a] Percentage of total planned events at each interim analysis.
[b] The nominal α for group sequential testing.
[c] The approximate HR required to reach an efficacy bound
[d] The probability of crossing a bound under the null hypothesis
[e] The probability of crossing a bound under the alternative hypothesis
[f] The approximate number of months since first participant randomized

The bounds provided in Table *10* are based on the assumption that the expected number of events at the IA and FA for OS are 302 and 388, 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 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.

3.8.4. Safety Analyses

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

3.9. Sample Size and Power Calculations

The study will randomize 620 participants in a 1:1 ratio into the pembrolizumab + lenvatinib and pembrolizumab + placebo arms. PFS and OS are primary endpoints for the study, with OR as the secondary endpoint.

For the PFS endpoint, based on a target number of 416 events at the final PFS analysis , the study has approximately 86.5% power to detect an HR of 0.7 at the initially allocated α=0.0055 (1-sided).

For the OS endpoint, based on a target number of 388 events and one interim analysis at approximately 78% of the target number of events, the study has approximately 90% power to detect an HR of 0.71 at the initially allocated α=0.0195 (1-sided).

Based on KEYNOTE-042 data, the above sample size and power calculations for PFS and OS assume the following:

- PFS follows a piecewise exponential distribution, with a median of 5 months up to 6.5 months and then a median of 12 months thereafter for the control group
- OS follows an exponential distribution, with a median of 16.7 months for the control group
- Enrollment period is approximately 21.7 months
- Annual drop-out rate is 20% and 2% for PFS and OS, respectively

- Follow-up period is approximately 8.8 and 18.5 months for PFS and OS, respectively, after the last participant is randomized

The sample size and power calculations were performed in R (package "gsDesign") 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 OR (with a nominal 95% CI) will be estimated and plotted by treatment group within each category of the following subgroup variables:

- Geographic region (East Asia, non-East Asia)
- ECOG performance status (0, 1)
- Predominant tumor histology (squamous, nonsquamous)
- TPS (1%-49%, ≥50%)
- Age category (<65, ≥65 years)
- Sex (female, male)
- Race (white, nonwhite)
- Smoking status (never, former/current smoker)
- Brain metastasis at baseline (presence, absence)
- Liver metastasis at baseline (presence, 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 unstratified Cox model, and the subgroup analyses for OR will be conducted using 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 number of cycles in which the participant receives the study intervention. Summary statistics will be provided on extent of exposure for the APaT population.

4. STATISTICAL ANALYSIS PLAN FOR EXTENSION

This section outlines the statistical analysis strategy and procedures for China specific subgroup analysis, which is required by the Chinese regulatory authority.

4.1. Introduction

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

After the enrollment for the global study is completed, subjects in China will continue to be enrolled in a 1:1 ratio into the pembrolizumab + lenvatinib arm and pembrolizumab + placebo arm until the sample size for the Chinese subpopulation reaches approximately 120 in total between the global and extension portion.

After the cut-off date for the primary analyses of the global study (including interim and final analyses), all Chinese subjects, including subjects enrolled in the global study and the extension portion, will continue their randomized treatment and continue to be followed up for PFS and OS events for China registration purpose. The extension portion will be completed after target number of PFS or OS events (Section 4.7) has been observed between the two arms in the Chinese subpopulation.

However, if the target number of PFS or OS events in the Chinese subpopulation is reached before an IA for the global study, the corresponding analysis for Chinese subpopulation will occur at the same time as the global IA or the final analysis (global study).

4.2. Responsibility for Analyses/In-House Blinding

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

4.3. Hypotheses/Estimation

No hypothesis testing is planned for the China extension portion. Nominal p values will be reported for efficacy analyses if applicable but should be interpreted with caution due to potential issues of multiplicity.

After succeeding in the global study, the consistency of efficacy and safety in the Chinese subpopulation to the global population will be evaluated. Consistency of efficacy will be evaluated using the percentage of risk reduction preserved in the Chinese subpopulation from the empirical

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

4.4. The Analysis Endpoints

4.4.1. Efficacy Endpoints

4.4.1.1. Primary Endpoints

***PFS*:** the time from randomization to the first documented PD per RECIST 1.1, adjusted to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ, based on BICR or death due to any cause, whichever occurs first. See Section 3.6.1 for the definition of censoring.

***OS*:** the time from randomization to death due to any cause.

4.4.1.2. Secondary Endpoints

***OR*:** a confirmed CR or PR per RECIST 1.1, adjusted to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ, based on BICR.

4.4.1.3. Exploratory Endpoints

***DOR*:** the time from the earliest date of qualifying response until earliest date of PD or death from any cause, whichever comes first.

4.4.2. 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 Section 8.4.7 of the study protocol.

4.4.3. Patient-reported Outcome Endpoints

Secondary

- Mean change from baseline in EORTC QLQ-C30 global health status/QoL (items 29-30), physical functioning (items 1-5), and dyspnea scores (item 8), and EORTC QLQ-LC13 cough (item 31) and chest pain (item 40) scores.
- TTD as measured by each of the EORTC QLQ-C30 global health status/QoL, physical functioning, and dyspnea scores, and EORTC QLQ-LC13 cough and chest pain scores.
- TTD as measured by a composite symptom endpoint: cough (QLQ-LC13 item 31), chest pain (QLQ-LC13 item 40), or dyspnea (QLQ-C30 item 8)

Based on prior literature (Osoba D et al., 1998; King et al., 1996, Marringwa et al., 2011), a ≥10 point or greater worsening from baseline for each scale represents a minimally important difference (MID) that represents a clinically relevant deterioration. TTD in the individual scores specified above is defined as the time to first onset of ≥10 (out of 100) point deterioration from baseline in a given scale/subscale/item and confirmed by a second adjacent ≥10 point deterioration from baseline. TTD in the composite endpoint is defined as the time to first onset of a ≥10 point deterioration (i.e., increase in score) from baseline in any one of the 3 lung cancer symptom scale items (EORTC-QLQ-C30 [dyspnea] and EORTC-QLQ-LC13 [cough and chest pain]) with confirmation by the subsequent visit of a ≥10 point deterioration (i.e., increase in score) from baseline in the same scale as the first onset. 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)

4.5. Analysis Populations

4.5.1. Efficacy Analysis Populations

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

Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary efficacy analysis population for the global study.

4.5.2. Safety Analysis Populations

Safety analysis will be carried out in the All Patients as Treated (APaT) China subpopulation, i.e., all randomized Chinese participants (in the global study and extension portion) who received at least 1 dose of study treatment.

Participants from China randomized and treated in the extension portion of this study after completion of the global enrollment will not be included in the primary safety analysis population for the global study.

4.5.3. Patient-reported Outcome Analysis Population

The PRO analyses are based on the PRO Full Analysis Set (PRO FAS) China subpopulation, defined as all randomized participants who have at least one 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. The PRO FAS China subpopulation will include all Chinese participants (in the global study and extension portion) in this population.

Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary PRO analysis population for the global study.

4.5.4. Pharmacokinetic Analysis Population

The analysis of PK data will be carried out in the population PK analysis set China subpopulation. The Population PK Analysis set will include all participants who have received at least 1 dose of study intervention with documented dosing history in the lenvatinib + pembrolizumab arm and have measurable plasma levels of lenvatinib or serum levels of pembrolizumab. Population PK Analysis Set China subpopulation will include all Chinese participants (in the global study and extension portion) in this population.

Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary PK analysis population for the global study.

4.6. Statistical Methods

Regarding the analysis for extension, no formal hypothesis testing is planned. No multiplicity adjustment will be applied to the analysis for extension.

4.6.1. Statistical Methods for Efficacy Analyses

Analyses regarding objective response and disease progression will include all events in all participants excluding events that occur in the Second Course Treatment.

4.6.1.1. Progression-free Survival

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

The nonparametric 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 the protocol). . A stratified Cox proportional hazard model with Efron's method of tie handling will be used to estimate the magnitude of the treatment difference (ie, HR) between the treatment arms. The HR and its 95% CI from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. The same stratification factors used in the global study will be used (see section 6.3.2 in the protocol). For the Chinese subgroup analyses, the stratified method will only be used if applicable. The factor of geographic region (East Asia vs. non-East Asia) will not be included in the Chinese subgroup stratified analyses. An analysis using the restricted mean survival time method may be conducted for PFS to account for the possible nonproportional hazards effect.

Consistency of efficacy will be evaluated using the percentage of risk reduction preserved in the Chinese subpopulation from the empirical risk reduction from the global primary efficacy analyses (based on point estimates). Sample size is designed to provide about 80% chance of observing the point estimate of Chinese subpopulation preserves ≥ approximately 50% of empirical risk

reduction from the global primary efficacy analysis assuming the same hazard ratio used in the sample size and power calculation for the global study. The primary analysis for PFS will be conducted in the Chinese subpopulation when approximately 75 PFS events have been collected.

4.6.1.2. Overall Survival

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

The nonparametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in OS will be assessed by the stratified log-rank test (based on the stratification factors defined in Section 6.3.2 in the 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 HR). 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 (Section 6.3.2 in the protocol) will be applied to both the stratified log-rank test and the stratified Cox model. The same stratification factors used in the global study will be used. For the Chinese subgroup analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for the Chinese subgroup analysis. Participants without documented death at the time of analysis will be censored at the date of last known contact. An analysis using the restricted mean survival time method may be conducted for OS to account for the possible nonproportional hazards effect.

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

4.6.1.3. Objective Response Rate

Analysis of ORR or DOR for extension is the same to that for the global study if applicable.

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 (Section 6.3.2 of the study protocol) will be used as stratification factors in the analysis. The same stratification factors used in the global study will be used. For the Chinese subpopulation analysis, the stratified method will only be used if applicable. The factor of Geography (East Asia vs. non-East Asia) will not be included in the stratified analysis for Chinese subgroup analysis.

4.6.1.4. Duration of Response

For subjects who demonstrate CR or PR, DOR is defined as the time from first documented evidence of CR or PR until PD or death due to any cause, whichever occurs first.

The nonparametric Kaplan-Meier method will be used to summarize the DOR. The median and range of DOR will be provided.

4.6.2. Statistical Methods for Safety Analyses

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

4.6.3. Statistical Methods for PRO Analysis

PRO Analysis are the same for extension to that for the global study as described in Section 3.6.3.

4.6.4. Summaries of Baseline Characteristics, Demographics, and Other Analyses

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

4.7. Interim Analysis and Final analysis

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

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

4.8. Multiplicity

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

4.9. Sample Size and Power Calculations

After the completion of global study enrollment, the extension portion will continue to enroll participants and randomize eligible participants until the sample size for the overall randomized Chinese subpopulation reaches approximately 120. Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary efficacy analysis population for the global study.

The extension portion will complete after approximately 75 OS events have been observed between the two arms in the Chinese subpopulation. With 75 OS events and a true hazard ratio of 0.7, the extension portion has ~80% chance to observe a point estimate of OS that preserves ≥ approximately 50% of the empirical risk reduction from the global analysis in the Chinese subpopulation. With 75 PFS events and a true hazard ratio of 0.7, the extension portion has ~80% chance to observe a point estimate of PFS that preserves ≥ approximately 50% of the empirical risk reduction from the global analysis in the Chinese subpopulation.

The above calculations for the consistency evaluation in OS and PFS are based on the same assumptions in the global study for sample size and power evaluation as specified in Section 3.9.

5. APPENDICES:

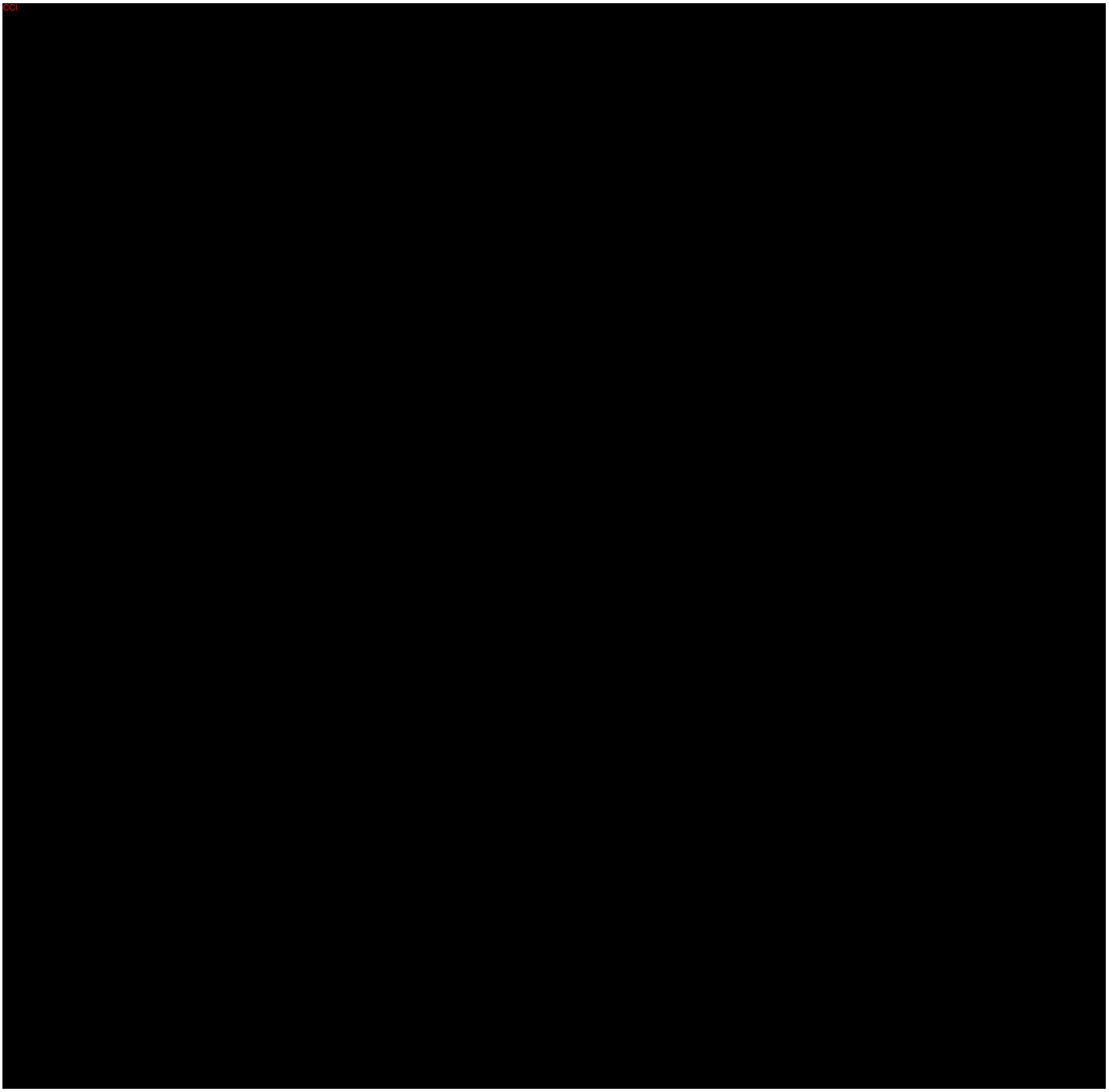
5.1. Rules for Non-Binding Futility Analyses

Non-binding futility analyses will be conducted to coincide with the safety DMC meeting in July 2021 (data cutoff in May 2021) and at the IA as per DMC recommendation. For each futility analysis, the DMC committee will evaluate the totality of the data and discuss with the EOC. The futility boundary for OS is based upon the Hwang-Shih-DeCani (HSD) beta-spending function with parameter of -15. The futility criteria are non-binding and only intended to provide guidance.



CCI

[redacted]



5.2. ABBREVIATIONS

Abbreviation	Expanded Term
1L	First-line
APaT	All Participants as Treated
BICR	blinded independent central review
CI	confidence interval

Abbreviation	Expanded Term
cLDA	constrained longitudinal data analysis
CR	complete response
CRF	case report form
CTCAE	Common Terminology Criteria for Adverse Events
DMC	data monitoring committee
DOR	duration of response
ECG	electrocardiogram
ECI	event of clinical interest
EOC	Executive Oversight Committee
EORTC	European Organisation for Research and Treatment of Cancer
ePRO	electronic patient-reported outcome
EQ-5D	EuroQoL-5D-5L
FAS	Full Analysis Set
HR	hazard ratio
HRQoL	health-related quality of life
HSD	Hwang-Shih-DeCani
IA	interim analysis
ICH	International Conference on Harmonization
iRECIST	Modified RECIST 1.1 for Immune-based Therapeutics
ITT	Intention to Treat
NPH	Non-proportional Hazards
PH	Proportional Hazards
NSCLC	non-small cell lung cancer
OR	objective response
ORR	objective response rate
OS	overall survival
PD	progressive disease
PD-L1	programmed cell death ligand 1
PFS	progression-free survival
PH	proportional hazards

Abbreviation	Expanded Term
PK	pharmacokinetic
PR	partial response
PRO	patient-reported outcomes
QLQ-LC13	Quality of Life Questionnaire and Lung Cancer Module 13
QoL	quality of life
RECIST 1.1	Response Evaluation Criteria in Solid Tumors 1.1
SAE	serious adverse event
SAP	statistical analysis plan
sSAP	supplemental statistical analysis plan
TPS	tumor proportion score
TTD	time to true deterioration

6. REFERENCES

Cocks K, King MT, Velikova G, de Castro G , Martyn St-James M, Fayers PM, et al. Evidence-based guidelines for interpreting change scores for the European organisation for the research and treatment of cancer quality of life questionnaire core 30. Eur J Cancer. 2012 Jul;48(11):1713-21.

Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation.* 2011;20(10):1727-1736.

King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Quality of Life Research 1996;5:555-67.

Lan, K. K., & DeMetes, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3): 659-63.

Liang K, Zeger, S (2000). Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhyā: The Indian Journal of Statistics, 62 (Series B), 134-148.

Maringwa JT, Quinten C, King M, 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. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer.* 2011;19(11):1753-1760.

Maurer, W., & Bretz, F. (2013). Multiple Testing in Group Sequential Trials using Graphical Approaches. *Statistics in Biopharmaceutical Research*, 5(4): 311-320.

Miettinen, O., & Nurminen, M. (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4:213-226.

O'Brien, P. C., & Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*, 35(3): 549-556.

Osoba, D., Rodrigues, G., Myles, J., Zee, B., & Pater, J. (1998). Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol*, 16:139-44.

Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. *Health and Quality of Life Outcomes.* 2007;5:70.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., Graeff , A. d., Groenvold, M., . . . Sprangers, M. A. (2008). *EORTC QLQ-C30 Reference Values.* Retrieved from https://www.eortc.org/app/uploads/sites/2/2018/02/reference_values_manual2008.pdf