Official Protocol Title:	A Phase III Randomized Double-blind Study of Pembrolizumab plus Best Supportive Care vs. Placebo plus Best Supportive Care as Second-Line Therapy in Asian Subjects with Previously Systemically Treated Advanced Hepatocellular Carcinoma (KEYNOTE-394)
NCT number:	NCT03062358
Document Date:	22-Jun-2021

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

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

2 SUMMARY OF CHANGES

This sSAP is currently based on MK-3475-394-05. Protocol specific detail that is stated as deferred to the sSAP includes the following (note, not all will be addressed in the sSAP unless deemed necessary; reference page number is provided if addressed):

- IA2 will be the final analysis of ORR as all subjects in the ITT population had at least 24 weeks from randomization to data cutoff.
- The analysis methods for OS sensitivity analysis were updated at Section 3.5.1.1.
- Variable list in subgroup analysis were updated and the baseline factors of gender and Barcelona Clinic Liver Cancer (BCLC) were added to the subgroup analyses at Section 3.9.
- Description of Patient Reported Outcomes (PRO) analysis was added at Section 3.12.
- Stable disease (SD) after ≥6 weeks was changed to ≥ 5 weeks based on assessments by the blinded central imaging vendor per RECIST 1.1 at Section 3.4.1.2, according to program standard.
- Description of strata used in the analysis was simplified in Section 3.5.1.
- Adverse events of special interest (AEOSI) and events of immune-mediated hepatitis were added to Tier 3 events and determined a short list of Tier 2 events (**Table 5**) at Section 3.5.2.
- Description of ongoing responders was clarified for the analysis of duration of response (DOR) in Section 3.5.1.6.

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 Sections 3.2 through 3.12.

Study Design Overview	A Randomized Double-Blind Phase III Study of Single Agent Pembrolizumab plus BSC vs. placebo plus BSC as Second-Line Therapy in Asian Subjects with Previously Systemically Treated Advanced Hepatocellular Carcinoma after Progression on Sorafenib or Oxaliplatin-based Chemotherapy
Treatment Assignment	Subjects will be randomized in a 2:1 ratio to receive blinded treatment with pembrolizumab plus BSC or placebo plus BSC (Control Arm). Stratification factors are in protocol Section 5.4. This is a double-blinded study.
Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoints/Hypotheses	Pembrolizumab improves overall survival (OS) compared to placebo.
Statistical Methods for Key Efficacy Analyses	The primary hypothesis will be evaluated by comparing pembrolizumab to the control on OS using a stratified log-rank test. Estimation of the hazard ratio will be done using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. PFS will be analyzed using the same method for OS. Stratified Miettinen and Nurminen's method [1] with weights proportional to the stratum size will be used for comparison of the objective response rates (ORR) between the treatment arms.
Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. The tiers differ with respect to the analyses that will be performed. There are no Tier 1 events in this trial. Tier 2 parameters will be assessed via point estimates with 95% confidence intervals provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters. The 95% confidence intervals for the between-treatment differences in percentages will be provided using the Miettinen and Nurminen method [1].
Interim Analyses	Two interim analyses will be performed in this study. Results will be reviewed by an external data monitoring committee. Details are provided in Section 3.6. IA1(ORR-driven): • Timing: to be performed when 163 randomized subjects have at least 24 weeks follow-up (which is estimated to happen at approximately 17 months after study start). Approximately 174 PFS events and 100 OS events are expected to be accumulated. • Purpose: estimate treatment effect and evaluate consistency with global data, interim analysis for ORR, PFS and OS. IA2 (event-driven): • Timing: to be performed when approximately 276 OS events have been observed and enrollment is complete, estimated to be approximately 31 months after study start. Approximately 395 PFS events are expected to be accumulated at IA2. • Purpose: interim efficacy analysis for OS, final analysis for PFS and ORR. Final analysis (event-driven) • Timing: to be performed when approximately 345 OS events have been observed, estimated to be approximately 37 months after study start • Purpose: final analysis for OS.

Multiplicity	The multiplicity strategy in this study will be applied to the primary hypothesis (superiority of pembrolizumab in OS to placebo) and the secondary hypotheses of superiority of pembrolizumab to placebo in PFS and ORR. The overall Type I error across the three hypotheses above is strongly controlled at 2.5% (one- sided) by the graphical approach of Maurer and Bretz [2] as described in Section 3.7. Initially, α=2.3% will be allocated to the OS hypothesis and α=0.2% will be allocated to the PFS hypothesis. Following a group sequential approach, the Type I error rates for the two interim and final analyses will be controlled through the alpha-spending function as described in Section 3.7.
Sample Size and Power	The sample size is approximately 450. The analyses of OS endpoint are event driven (i.e., the testing of the OS hypothesis is conducted upon accumulating a preset number of events). The study is designed and will be conducted to accumulate approximately 345 OS events (unless superiority in OS is proven at the interim analysis). For primary endpoint OS, the trial has ~87% power to demonstrate that pembrolizumab is superior to the control at a one-sided 2.3% alpha-level, if the underlying hazard ratio of OS is 0.7.

3.2 Responsibility for Analyses/In-House Blinding

The statistical analysis of the data obtained from this study will be the responsibility of the Clinical Biostatistics department of the SPONSOR. The SPONSOR will generate the randomized allocation schedule(s) for study treatment assignment for this protocol, and the randomization will be implemented in IVRS.

Since the trial is double-blinded with in-house blinding, the Sponsor, the investigators, site staffs, and subjects will be blinded to the treatment assignment. In addition, the blinded central imaging vendor will perform the central imaging review without knowledge of treatment group assignment.

The eDMC will serve as the primary reviewer of the unblinded results of the interim analyses and will make recommendations for discontinuation of the study or modification to an executive oversight committee of the SPONSOR. An external unblinded statistician and statistical programmer will be responsible for generating unblinded data summaries and presenting them to the eDMC. Depending on the recommendation of the eDMC, the Sponsor may prepare a regulatory submission. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee and limited additional SPONSOR personnel may be unblinded to results at the treatment level in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of interim analyses will be documented. Additional logistical details will be provided in the eDMC Charter.

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in protocol Section 3.0.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

3.4.1.1 Primary

Overall Survival

OS is defined as the time from randomization to death due to any cause. Subjects without documented death at the time of the final analysis will be censored at the date of the last follow-up.

3.4.1.2 Secondary

Progression-Free Survival (PFS) – RECIST 1.1 by BICR

PFS is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded central imaging vendor review or death due to any cause, whichever occurs first. See Section 3.5.1 for the censoring rules.

Objective Response Rate (ORR) – RECIST 1.1 by BICR

Objective response rate is defined as the proportion of the subjects in the analysis population who have a complete response (CR) or partial response (PR) per RECIST 1.1. Disease Control Rate (DCR) - RECIST 1.1 by BICR

DCR is defined as the percentage of subjects who have achieved CR, PR, or have demonstrated SD for at least 5 weeks prior to any evidence of progression based on assessments by the central imaging vendor per RECIST 1.1.

Time to Progression (TTP) – RECIST 1.1 by BICR

TTP is defined as the time from randomization to the first documented disease progression per RECIST 1.1. See Section 3.5.1 for the censoring rules.

Duration of Response (DOR) - RECIST 1.1 by BICR

For subjects who demonstrate CR or PR, duration of response is defined as the time from first documented evidence of CR or PR per RECIST 1.1 until disease progression per RECIST 1.1 or death due to any cause, whichever occurs first.

3.4.1.3 Exploratory Endpoints

Quality of life and health utilities will be examined between groups. EORTC QLQ-C30 will be used for evaluating changes from baseline in health related quality of life outcomes; and EuroQol EQ-5D-3L will be used to characterize utilities between two treatment arms.

3.4.2 Safety Endpoints

Safety measurements are described in protocol Section 4.2.3.4 Safety Endpoints and protocol Section 7.

Safety and tolerability will be assessed by clinical review of all relevant parameters including adverse events (AEs), laboratory tests, and vital signs. Safety parameters to be analyzed include, but are not limited to, AEs, SAEs, fatal AEs, and laboratory changes. Furthermore, specific events will be collected and designated as events of clinical interest (ECIs) as described in protocol Section 7.2.3.

Events of clinical interest for this trial include:

1. An overdose of Sponsor's product, as defined in protocol Section 7.2.1 - Definition of an Overdose for This Protocol and Reporting of Overdose to the Sponsor, that is not associated with clinical symptoms or abnormal laboratory results.

2. Hepatic ECIs as defined in protocol Section 7.2.3.

There are no “Tier 1” events in this trial. In addition, the broad clinical and laboratory AE categories consisting of the percentage of subjects with any AE, any drug related AE, any Grade 3-5 AE, any serious AE, any AE which is both drug-related and Grade 3-5, any AE which is both serious and drug-related, dose modification due to AE, and who discontinued due to an AE, and death will be considered Tier 2 endpoints. AEs (specific terms as well as system organ class terms) will be classified as belonging to "Tier 2" or "Tier 3", based on the number of events observed. Membership in Tier 2 requires that at least 4 subjects in any treatment group exhibit the event; all other AEs and predefined limits of change will belong to Tier 3. Analysis Populations

3.4.3 Efficacy Analysis Populations

The Intention-to-Treat (ITT) population will serve as the population for primary efficacy analysis. All randomized subjects will be included in this population. Subjects will be included in the treatment group to which they are randomized.

Details on the approach to handling missing data are provided in Section 3.5 Statistical Methods.

3.4.4 Safety Analysis Populations

The All Subjects as Treated (ASaT) population will be used for the analysis of safety data in this study. The ASaT population consists of all randomized subjects who received at least one dose of study treatment. Subjects will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the ASaT population. For most subjects this will be the treatment group to which they are randomized. Subjects who take incorrect study treatment for the entire treatment period will be included in the treatment group corresponding to the study treatment actually received. Any subject who receives the incorrect study medication for one cycle but receives the correct treatment for all other cycles will be analyzed according to the randomized treatment group and a narrative will be provided for any events that occur during the cycle for which the subject is 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.

Details on the approach to handling missing data for safety analyses are provided in Section 3.5 Statistical Methods

3.5 Statistical Methods

3.5.1 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 3.7, Multiplicity. Nominal p-values may be computed for other efficacy analyses, but should be interpreted with caution due to potential issues of multiplicity.

The stratification factors applied to all stratified analyses of efficacy endpoints will be: macrovascular invasion (Yes, No), α-fetoprotein (ng/mL) (<200, ≥200) and region (China, ex-China). Due to the small number of subjects in the stratum of presence of macrovascular invasion, subjects from any region or with any level of α-fetoprotein will be combined in this stratum. Thus, the following 5 strata will be applied to all stratified analyses:

- Macrovascular invasion (No) + Region (China) + α-Fetoprotein (ng/mL) (<200)
- Macrovascular invasion (No) + Region (ex-China) + α-Fetoprotein (ng/mL) (<200)
- Macrovascular invasion (No) + Region (China) + α-Fetoprotein (ng/mL) (≥200)
- Macrovascular invasion (No) + Region (ex-China) + α-Fetoprotein (ng/mL) (≥200)
- Macrovascular invasion (Yes)

These five strata will be used in all stratified analyses of efficacy endpoints and PRO endpoints.

3.5.1.1 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The superiority hypothesis of treatment difference in survival 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 (i.e., the hazard ratio). The hazard ratio and its 95% confidence interval from the stratified Cox model with a single treatment covariate will be reported.

Prior to verification of PD by blinded independent central review, switching to another treatment is discouraged. Following verification of PD, subjects may switch to another anti-cancer treatment. Sensitivity analyses to adjust for the effect of the next line anti-cancer therapies of HCC on OS in both treatment arms would be considered based on the recognized methods. Specifically, three methods would be performed: (1) an analysis with survival censored at the start of the next line of HCC anticancer therapy; (2) the inverse probability of censoring weighting (IPCW) model proposed by Robins and Finkelstein [3]; and (3) the simplified two-stage survival model without re-censoring [4].

3.5.1.2 Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The superiority hypothesis 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 (i.e., hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported.

Since disease progression is assessed periodically, progressive disease (PD) can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. For the primary analysis, for the subjects who have PD, the true date of disease progression will be approximated by the date of the first assessment at which PD is objectively documented per RECIST 1.1 by blinded independent central review, regardless of discontinuation of study drug. Death is always considered as a confirmed PD event. Sensitivity analyses may be performed for comparison of PFS based on investigator's assessment per RECIST 1.1 and PFS analysis for PD per irRECIST by blinded independent central review.

In order to evaluate the robustness of the PFS endpoint per RECIST 1.1 by blinded independent central review, we will perform two sensitivity analyses with a different set of censoring rules. The first sensitivity analysis is the same as the primary analysis except that (1) the date of documented PD or death will be the progression date, regardless of whether or not new anti-cancer treatment is initiated and (2) it censors at the last disease assessment, regardless of whether or not new anti-cancer treatment is initiated if no PD and no death occur. The second sensitivity analysis is the same as the first sensitivity analysis, except that it considers discontinuation of treatment due to reasons other than complete response or initiation of an anti-cancer treatment subsequent to discontinuation of study specified treatment, whichever

occurs later, to be a PD event for subjects without documented PD or death. The censoring rules for primary and sensitivity analyses are summarized in Table 1.

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
PD or death documented immediately after ≥2 consecutive missed disease assessments or after new anticancer therapy, if any	Censored at last disease assessment prior to the earlier date of ≥2 consecutive missed disease assessment and new anticancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death
No PD and no death; and new anticancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment	Progressed at treatment discontinuation due to reasons other than complete response; otherwise censored at last disease assessment if still on study treatment or completed study treatment.
No PD and no death; new anticancer treatment is initiated	Censored at last disease assessment before new anticancer treatment	Censored at last disease assessment	Progressed at date of new anticancer treatment
Abbreviations: PD = progressive disease			

The proportional hazards assumption on PFS will be examined using both graphical and analytical methods if warranted. The log[-log] of the survival function vs. time for PFS will be plotted for the comparison between pembrolizumab and the control arm. If the curves are not parallel, indicating that hazards are not proportional, supportive analyses may be conducted to account for the possible non-proportional hazards effect associated with immunotherapies: for example, using the Restricted Mean Survival Time (RMST) method [5], parametric method [6], etc.

The PFS analyses are described in Section 3.6 Interim Analyses and Section 3.7 Multiplicity. The supportive analysis of the PFS data available at the time of the final OS analysis will be also conducted.

3.5.1.3 Objective Response Rate

Stratified Miettinen and Nurminen's method [1] with weights proportional to the stratum size will be used for the comparison of the objective response rates between the treatment arms. A 95% confidence interval for the difference in response rates between the pembrolizumab arm and the control arm will be provided.

The ORR analysis will be conducted according to the hypotheses testing plan as described in Section 3.6 Interim Analyses and Section 3.7 Multiplicity.

3.5.1.4 Disease Control Rate

Stratified Miettinen and Nurminen's method [1] with weights proportional to the stratum size will be used for the comparison of the DCR between the treatment arms. A 95% confidence interval for the difference in response rates between the pembrolizumab arm and the control arm will be provided.

3.5.1.5 Time to Progression

The non-parametric Kaplan-Meier method will be used to estimate the TTP curve in each treatment group. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to estimate the magnitude of the treatment difference (i.e., hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported.

Since disease progression is assessed periodically, progressive disease (PD) can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. For the analysis, for the subjects who have PD, the true date of disease progression will be approximated by the date of the first assessment at which PD is objectively documented per RECIST 1.1 by blinded independent central review, regardless of discontinuation of study drug. Unlike in PFS analysis, death is not considered as an event.

The censoring rules for TTP are summarized in **Table 2**.

Table 2
Censoring rules for TTP

Situation	Primary Analysis
Death without a preceding disease progression	Censored at date of randomization or date of last non-PD disease assessment, whichever is later
No PD and no death; new anticancer treatment is not initiated	Censored at last non-PD disease assessment
No PD and no death; new anticancer treatment is initiated	Censored at last non-PD disease assessment before new anticancer treatment is initiated
PD documented after ≤ 1 missed disease assessment	Progressed at date of documented PD
PD documented immediately after ≥ 2 missed disease assessments	Censored at last non-PD disease assessment prior to the ≥ 2 consecutive missed disease assessments

3.5.1.6 Duration of Response

Subjects who achieved confirmed CR or PR and are alive, have not progressed, have not initiated new anti-cancer treatment, have not been determined to be lost to follow-up, and have had a disease assessment within ~5 months of the data cutoff are considered ongoing responders at the time of analysis.

The non-parametric Kaplan-Meier method will be used to estimate the DOR curve in each treatment group; estimates and 95% CIs at specific duration time points will be provided.

Censoring rules for DOR are summarized in **Table 3**.

Table 3
Censoring Rules for DOR

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new anti-cancer therapy initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anti-cancer therapy initiated	Last adequate disease assessment before new anti-cancer therapy initiated	Censor (non-event)
Death or progression immediately after ≥2 consecutive missed disease assessments or after new anti-cancer therapy	Earlier date of last adequate disease assessment prior to ≥2 missed adequate disease assessments and new anti-cancer therapy, if any	Censor (non-event)
Death or progression after ≤1 missed disease assessments and before new anti- cancer therapy, if any	PD or death	End of response (Event)
Subjects are considered to have an ongoing response if censored, alive, have not progressed, have not started a new anti-cancer therapy, have not been determined to be lost to follow-up, and have had a disease assessment within ~5 months of the data cutoff. A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response.		

Table 4 summarizes the primary analysis approach for key efficacy endpoints. Sensitivity analysis methods are described above for each endpoint.

Analyses of the DCR, TTP, and DOR data will be performed at the time of the interim and final analyses of OS. Analyses of the imaging endpoints based on the investigator assessment per RECISIT 1.1 will also be provided.

The strategy to address multiplicity issues with regard to multiple efficacy endpoints, multiple populations, and interim analyses is described in Section 3.6 Interim Analyses and in Section 3.7 Multiplicity.

Table 4
Analysis Strategy for Key Efficacy Hypotheses

Endpoint/Variable (Description, Time Point)	Statistical†Method	Analysis Population	Missing Data Approach
Primary Hypothesis			
OS	Test: Stratified Log-rank test. Estimation: Stratified Cox model with Efron's tie handling method.	ITT	Censored at the last date the subject was known to be alive
Secondary Endpoints			
PFS per RECIST 1.1 by BICR	Test: Stratified Log- rank test. Estimation: Stratified Cox model with Efron's tie handling method	ITT	Primary censoring rule Sensitivity analysis 1 Sensitivity analysis 2 (More details are in Table 9)
ORR per RECIST 1.1 by BICR	Stratified M& N method‡	ITT	Subjects with missing data are considered non-responders

† Statistical models are described in further detail in the text. For stratified analyses, the stratification factors applied to the analysis model will be: macrovascular invasion (Yes, No), α-fetoprotein (ng/mL) (<200, ≥200) and region (China, ex- China), with all cells that correspond to macrovascular invasion=Yes combined.

‡ Miettinen and Nurminen method.

3.5.2 Statistical Methods for Safety Analyses

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

Tiered Approach

The analysis of safety results will follow a tiered approach (**Table 5**). The tiers differ with respect to the analyses that will be performed. For this protocol, there are no Tier 1 events.

Tier 2 parameters will be assessed via point estimates with 95% confidence intervals provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters.

AEs (specific terms as well as system organ class terms) will be classified as belonging to "Tier 2" or "Tier 3", based on the number of events observed. Membership in Tier 2 requires that at least 4 subjects in any treatment group exhibit the event; all other AEs and predefined limits of change will belong to Tier 3.

The threshold of at least 4 events was chosen because the 95% confidence interval for the between-group difference in percent incidence will always include zero when treatment groups of equal size each have less than 4 events and thus would add little to the interpretation of potentially meaningful differences. Because many 95% confidence intervals may be provided without adjustment for multiplicity, the confidence intervals should be regarded as a helpful descriptive measure to be used in review, not a formal method for assessing the statistical significance of the between-group differences in AEs and predefined limits of change.

Continuous measures such as changes from baseline in laboratory values and vital signs, that are not pre-specified as Tier-1 endpoints will be considered Tier 3 safety parameters. Summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group in table format.

In addition, the broad clinical and laboratory AE categories consisting of the percentage of subjects with any AE, any drug related AE, any Grade 3-5 AE, any serious AE, any AE which is both drug-related and Grade 3-5, any AE which is both serious and drug-related, dose modification due to AE, and who discontinued due to an AE, and death will be considered Tier 2 endpoints.

The 95% confidence intervals will be provided for between-treatment differences in the percentage of subjects with Tier 2 events; these analyses will be performed using the Miettinen and Nurminen method [1], an unconditional, asymptotic method. Safety analyses will not be stratified.

Table 5
Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint†	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	AEs (incidence ≥10% of participants in one of the treatment groups)	X	X
	Grade 3-5 AE (incidence ≥2% of participants in one of the treatment groups)	X	X
	Serious AE (incidence ≥1% 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 <4% of subjects in all of the treatment groups), AEOSI, Immune-mediated Hepatitis Events		X
	Change from baseline results (laboratory test toxicity grade, vital signs)		X

AE = adverse event; CI = confidence interval; X = results will be provided.
The rainforest plots including the treatment difference and its 95% CI will be applied for Tier2 AEs.

3.5.3 Summaries of Demographic and Baseline Characteristics

The comparability of the treatment groups for each relevant characteristic will be assessed by the use of tables and/or graphs. No statistical hypothesis tests will be performed on these characteristics. The number and percentage of subjects screened, randomized, the primary reasons for screening failure, and the primary reason for discontinuation will be displayed. Demographic variables (e.g., age), baseline characteristics, primary and secondary diagnoses, and prior and concomitant therapies will be summarized by treatment either by descriptive statistics or categorical tables.

3.6 Interim Analyses

An external data monitoring committee (eDMC) will be convened to review the unblinded efficacy results and accumulating safety at the planned IAs.

3.6.1 Efficacy Interim Analyses

Two interim analyses and one final analysis are planned in this trial. The timing and the purpose of each analysis are summarized in Table 6. A detailed description of the multiplicity adjustment and hypotheses testing plan is provided in Section 3.7 Multiplicity.

The first interim analysis will be performed at approximately Month 17, at which time the first 163 randomized subjects have at least 24 weeks follow-up. The main purpose of this interim analysis is to conduct the analysis for ORR on the first 163 randomized subjects, evaluate the consistency of efficacy and safety in this study to the reference global data, and to conduct an interim testing for efficacy on the PFS and OS endpoints for all the subjects randomized at that time (superiority only). Consistency of efficacy will be evaluated by comparing the treatment effect observed in this study to the reference global data (based on point estimates). Approximately 225 subjects will be enrolled at the first interim analysis cutoff, with approximately 174 PFS events and 100 OS events projected to be accumulated. The main purpose of the second interim analysis is to conduct the primary testing for efficacy on the PFS hypothesis and to conduct an interim testing for efficacy on the OS hypothesis (superiority only). It will be performed when enrollment is complete and approximately 276 OS events (~80% of the target 345 total OS events) are observed. It is projected that this event count will be accumulated at approximately 31 months after the start of the study. Approximately 395 PFS events are expected to be accumulated by then. Under assumptions specified in Section 3.8 sample size and power calculation, at the time of the secondary interim analysis: 1) a total of approximately 395 PFS events for testing the PFS hypothesis at Type I error α=0.2% provides approximately 88% power to successfully demonstrate the PFS hypothesis; and 2) a total of approximately 276 OS events for testing the OS hypothesis at Type I error α=1.10% provides approximately 70% power to successfully demonstrate the OS hypothesis.

If superiority of PFS or OS (pembrolizumab vs. placebo) is demonstrated at one of the interim analyses, the ORR hypothesis will be tested according to the group-sequential boundaries for ORR analysis (see Table 9 in Section 3.7 Multiplicity).

The final analysis (FA) will be performed when approximately 345 OS events are observed which is expected at Month 37.

Table 6
Summary of Interim and Final Analyses Strategy

Analysis	Endpoint(s)	Criteria for Conduct of Analysis	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA 1	ORR, DOR, PFS, OS	First 163 randomized subjects have at least 24 weeks follow-up	~17 months	Descriptive ORR analysis, consistency evaluation
IA 2	ORR, PFS, OS	~276 OS events observed	~31 months	PFS FA ORR FA OS IA
FA	OS, ORR	~345 OS events observed	~37 months	OS FA

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

The eDMC will serve as the primary reviewer of the unblinded results of the interim analyses and will make recommendations. Depending on the recommendation of the eDMC, the Sponsor may prepare a regulatory submission. NOTE: no futility test is planned in the interim analysis.

3.6.2 Safety Interim Analyses

As noted in protocol Section 7.3.2 – Data Monitoring Committee, the eDMC will be responsible for periodic interim safety reviews as specified in the DMC charter.

3.7 Multiplicity

The multiplicity strategy specified in this section will be applied to the primary hypothesis (superiority of pembrolizumab to placebo in OS) and the secondary hypotheses of superiority of pembrolizumab in PFS or ORR.

The overall Type-I error across the testing of the OS, PFS, and ORR hypotheses is strongly controlled at α=2.5% (one-sided). The multiplicity strategy will follow the graphical approach of Mauer and Bretz [2]. Figure 1 provides the multiplicity strategy diagram of the study.

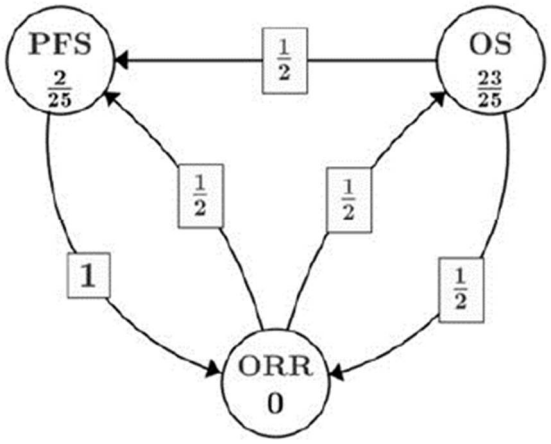

Figure 1 Multiplicity Strategy

In the diagram shown in Figure 1, when a particular null hypothesis is rejected, the arrows leading to it are removed, and the Type I error allocated to the null hypothesis that was rejected are re-distributed to other hypotheses. The arrows on the diagram show how the Type I error allocated to a hypothesis that was successfully tested will be re-distributed. Initially, α=2.3% (23/25 of the overall total α=2.5% for testing the OS, PFS and ORR) is allocated to the OS hypothesis and α=0.2% is allocated to the PFS hypothesis.

In detail, if PFS and OS hypotheses are both rejected on the initial alpha, then ORR hypothesis will be tested on the 2.5% alpha level. If PFS hypothesis is rejected on the initial alpha level 0.2% while OS is NOT rejected on the initial alpha 2.3%, then ORR hypothesis will be tested on the 0.2% alpha level. If ORR hypothesis can be rejected on the 0.2% alpha level, then OS hypothesis will be retested on the 2.5% alpha level. If PFS hypothesis is NOT rejected on the

initial alpha level 0.2% but OS hypothesis is rejected on the initial alpha 2.3%, then PFS hypothesis can be re-tested on the 1.35% alpha level and ORR hypothesis can be tested on the 1.15% level. After that, we will follow the same procedure of hypothesis testing and type-I error re-distribution as illustrated in Figure 1.

OS hypothesis: The OS hypothesis will be tested following a group sequential approach. The testing of the OS hypothesis will be based on an OS test statistic calculated from study data at the first interim analysis. If unsuccessful at the first IA, the OS hypothesis will be tested at the second interim analysis time and, if unsuccessful at the second IA, the OS hypothesis will be tested at the final analysis time. No futility test is planned at the interim analyses. The nominal Type I error rates for the two interim analyses and final analysis that will allow tight control of the overall Type I error for testing the OS hypothesis will be derived using the alpha-spending function approach based on the overall Type I error allocated to the OS hypothesis. The group sequential testing of the OS hypothesis will be conducted with efficacy boundaries (superiority test only). The OS efficacy boundary will be set using the Lan-DeMets spending function that approximates an O'Brien-Fleming boundary. The OS hypothesis will initially be tested at the overall Type I error α=2.3%. If both the PFS and ORR null hypotheses have been rejected at the α=0.2% level, the OS hypothesis will be tested at Type I error level of α=2.5%.

Table 7 summarizes the timing, sample size and decision guidance of the interim analyses and final OS analysis. Interim analysis spending for OS analyses will be based on the expectation of 345 OS events at the final analysis and the final analysis spending will be updated based on spending alpha using the actual number of OS events at times of analysis using spending functions as noted above.

Table 7
Summary of Timing, Sample Size and Decision Guidance for the Interim Analyses and Final Analysis of Overall Survival

Type I Error (Overall α)	Analysis	Study Calendar Time	N[§]	Events	Information Fraction	Efficacy Boundary Crossing[†]		
						Nominal α	Hazard Ratio	Power
2.3%	IA1	Month 17	225	100	0.30	<0.01%	~0.42	0.9%
	IA2	Month 31	450	276	0.80	1.10%	~0.75	70%
	Final	Month 37	450	345	1.00	1.98%	~0.79	87%
2.5%	IA1	Month 17	225	100	0.30	<0.01%	~0.43	1.1%
	IA2	Month 31	450	276	0.80	1.22%	~0.75	72%
	Final	Month 37	450	345	1.00	2.14%	~0.79	88%

[†] Based on Lan-DeMets spending function that approximates an O'Brien-Fleming boundary.
[§] Expected number of subjects randomized into the study at the time of analysis.

PFS hypothesis: The group sequential testing of the PFS hypothesis will be conducted with efficacy boundaries (superiority test only). The PFS efficacy boundary will be set using the Lan-DeMets spending function that approximates an O'Brien-Fleming boundary. The interim analysis of PFS will be conducted at the time of the first interim OS analysis; the final PFS

analysis will be conducted at the time of the second OS analysis. The testing of the PFS hypothesis will initially be tested at one-sided Type I error α=0.2%. Depending on the results of testing of the OS and ORR hypothesis, the PFS hypothesis can be tested at a one-sided Type I error level of α=1.35% or α=2.5%. At the second interim analysis, it is expected that approximately 395 PFS events would have been accumulated assuming 1) a hazard ratio of 0.65; 2) Progression-free survival follows an exponential distribution with a median of 1.5 months in the control arm; 3) an enrollment time of 31 months with a recruitment rate of 16 subjects per month and a 6-month ramp up time, and 4) a monthly drop-out rate of ~1%.

Table 8 summarizes the timing, sample size and decision guidance of the interim and final PFS analyses. Interim analysis spending for PFS analysis will be based on the expectation of 395 PFS events at the final analysis and the final analysis spending will be updated on spending alpha using the actual number of PFS events at times of analysis when the criteria for triggering the final PFS analysis is complete.

Table 8
Summary of Timing, Sample Size and Decision Guidance for the Interim Analysis and Final Analysis of Progression Free Survival

Type I Error (Overall α)	Analysis	Study Calendar Time	N[§]	Events	Information Fraction	Efficacy Boundary Crossing[†]		
						Nominal α	Hazard Ratio	Power
0.2%	IA	Month 17	225	174	0.44	<0.01%	~0.48	3.4%
	FA	Month 31	450	395	1.00	0.2%	~0.74	88%
1.35%	IA	Month 17	225	174	0.44	0.02%	~0.57	20%
	FA	Month 31	450	395	1.00	1.34%	~0.79	97%
2.5%	IA	Month 17	225	174	0.44	0.07%	~0.60	31%
	FA	Month 31	450	395	1.00	2.5%	~0.81	98%

[†] Based on Lan-DeMets spending function that approximates an O'Brien-Fleming boundary.

[§] Expected number of subjects randomized into the study at the time of analysis.

ORR hypothesis: ORR by treatment group will be estimated at each interim analysis and the final analysis. The initial alpha allocated to ORR is zero. If PFS or OS null hypothesis is rejected, depending on the results of testing of the OS and PFS hypothesis, the ORR hypothesis can be tested at a one-sided Type I error level of α=0.2%, α=1.15%, or α=2.5%. The ORR hypothesis will be tested following a group sequential approach. The testing of the ORR hypotheses will be at the first interim analysis time. If unsuccessful at the first IA, the ORR hypotheses will be tested at the second interim analysis time which will be the final analysis for ORR. Subjects who will be included in the ORR analysis are those who have "mature ORR information", defined as subjects who were enrolled at least 24 weeks prior to the interim data cutoff dates and thus had an opportunity to have at least 4 scheduled scans if not discontinued. For the first interim analysis, 163 subjects with at least 24 weeks follow up will be included for ORR analysis and information fraction is expected to be about 0.36. At the time of the second interim analysis, all randomized subjects will have at least 24 weeks follow-up and thus, 100% of information will be reached. At this time, the final ORR analysis will be

conducted. The nominal Type I error rates for the interim analyses and final analysis that will allow tight control of the overall Type I error for testing the ORR hypothesis will be derived using the alpha-spending function approach. The group sequential testing of the ORR hypothesis will be conducted with an efficacy boundary only. The efficacy boundary for the ORR will be set using an Exponential spending function $f(t) = \alpha^{t^{-\nu}}$ [7] with parameter ν=0.25, which yields a Pocock-like boundary. The ORR hypothesis is initially allocated a Type I error α=0% and thus, cannot be tested unless one or both of the PFS and OS null hypotheses have been rejected. Depending on the results of the OS and PFS hypotheses testing, the ORR hypothesis can be tested at Type I error levels of α=0.2%, 1.15%, or 2.5%.

Table 9 shows the boundary thresholds corresponding to a group sequential testing of the ORR hypothesis at each of these Type I error levels. The p-value at the boundary for the final analysis will be adjusted according to the exact number of subjects enrolled in the study.

Table 9
Efficacy Boundaries for Testing the ORR Hypothesis

Type I Error (Overall α)	Analysis	N‡	Information Fraction	Efficacy Boundary Crossing†		
				Nominal α	ORR Δ§	Power
0.2%	IA1	163	0.36	0.03%	~16.77%	17%
	Final	450	1.00	0.17%	~8.61%	88%
1.15%	IA1	163	0.36	0.31%	~12.95%	42%
	Final	450	1.00	0.92%	~6.70%	97%
2.50%	IA1	163	0.36	0.86%	~11.09%	58%
	Final	450	1.00	1.89%	~5.79%	99%

† Based on Exponential (ν=0.25) spending function. ORR is tested for superiority only.

‡ Expected number of subjects who were enrolled at least 24 weeks prior to the data cutoff date and thus had an opportunity to have at least 4 scheduled scans at the time of ORR analysis.

§ Δ = ORR in pembrolizumab group – ORR in control group. The assumed expected ORR in pembrolizumab and control groups are 15% and 3%, respectively.

The spending function planned to be used for the testing of the OS and ORR hypotheses satisfy the requirements laid out in Maurer and Bretz [2].

3.8 Sample Size and Power Calculations

The study will randomize subjects in a 2:1 ratio into the pembrolizumab plus BSC arm and the control arm (placebo plus BSC).

The final analysis is event driven (i.e., the testing of the OS and PFS hypotheses will be conducted upon accumulation of a preset number of events). The study is designed and will be conducted to accumulate approximately 345 OS events unless superiority in OS is proven at the interim analyses.

OS Analysis: A total of approximately 345 OS events are required to test the OS hypothesis at Type I error rate of α=2.3% with ~87% power (see Table 7) if the underlying OS hazard ratio (pembrolizumab/control) is 0.7. A total of approximately 450 subjects are needed to be enrolled into the study in order to accumulate approximately 345 OS events at approximately Month 37 after study start. The sample size and power calculation is based on the following assumptions: 1) a hazard ratio of 0.7; 2) overall survival follows an exponential distribution with a median of 6.0 months in the control arm; 3) an enrollment rate of 16 subjects per month with a 6-month ramp up time; and 4) a monthly dropout rate of ~0.2%.

PFS Analysis: As described in Section 3.6 Interim Analysis, the final PFS analysis will be conducted at the same time as the OS interim analysis 2 at approximately Month 31 after study start. It is projected that approximately 395 PFS events will be accumulated at this time. With 395 PFS events, the testing of the PFS hypothesis at Type I error α=0.2% has approximately 88% power to demonstrate that pembrolizumab is superior to the control with respect to PFS if the underlying PFS hazard ratio (pembrolizumab/control) is 0.65.

The sample size and power calculation is based on the following assumptions: 1) a hazard ratio of 0.65; 2) progression-free survival follows an exponential distribution with a median of 1.5 months in the control arm; 3) an enrollment rate of 16 subjects per month with a 6- month ramp up time; and 4) a monthly dropout rate of ~1%.

The sample size and power calculation were performed in the software EAST and R (package "gsDesign").

3.9 Subgroup Analyses and Effect of Baseline Factors

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

- Prior treatment (Sorafenib, chemotherapy)
- Macrovascular invasion (Yes, No)
- HBV (Active Positive, Negative)
- α-Fetoprotein (ng/mL) (<200, ≥200)

- α-Fetoprotein (ng/mL) (<400, ≥400)
- Prior Locoregional Therapy (Yes, No)
- Prior Treatment Surgery (Yes, No)
- ECOG performance status (0, 1)
- Age (<65 years, ≥65 years)
- Extrahepatic spread (Yes, No)
- Region (China, ex-China)
- Gender (Male, Female)
- Current disease overall BCLC stage (B, C)

3.10 Compliance (Medication Adherence)

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

3.11 Extent of Exposure

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

3.12 Statistical considerations for Patient-Reported Outcomes (PRO):

The patient-reported outcomes endpoints are exploratory objectives in KN394. No formal hypotheses were formulated. Since multiplicity adjustments are only applied to primary and secondary hypotheses, nominal p-value to compare the pembrolizumab arm to the control arm will be provided without multiplicity adjustment

The PRO instruments included in the study are the EORTC QLQ-C30 and EuroQol-5D-3L (EQ-5D3L).

3.12.1 PRO Endpoints

Exploratory PRO endpoints include mean score changes from baseline to the latest time point, where the completion rate and compliance rates are still high enough (e.g. close to 60 and 80%, respectively) based on blinded data review, as measured by:

- EORTC QLQ-C30 global health status/quality of life scale
- All EORTC QLQ-C30 sub-scales/items
- EQ-5D VAS

The other PRO endpoints are:

- The time to deterioration (TTD) for the QLQ-C30 global health status/quality of life scale.
- The number and proportions of deterioration/stable/improvement from baseline to the latest time point, where the completion rate and compliance rates are still high enough (e.g. close to 60 and 80%, respectively) for all QLQ-C30 sub-scales/items.

3.12.2 Scoring Algorithm

The QLQ-C30 includes five functional dimensions (physical, role, emotional, cognitive, and social), three symptom scales (fatigue, nausea/vomiting, and pain), and six single item measures (dyspnea, sleep disturbance, appetite loss, constipation, diarrhea, and financial difficulties).

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 functioning and global health status/quality-of-life 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 QLQ-C30 Manuals [12], if items I_1, I_2, …, I_n are included in a scale, the linear transformation procedure is as follows:

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

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

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

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

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

3.12.3 The Schedule for PRO Data Collection

Table 10 provides the schedule for PRO data collection.

Table 10
PRO Data Collection Schedule

Week	Cycle	MK-3475	Placebo
Week 0	C1	✓	✓
Week 3	C2	✓	✓
Week 6	C3	✓	✓
Week 9	C4	✓	✓
Week 12	C5	✓	✓
Week 18	C7	✓	✓
Week 27	>=C8	✓	✓
Week 36	>=C8	✓	✓
Week 45	>=C8	✓	✓
End of Treatment		✓	✓
30-day Safety follow-up		✓	✓
All PROs are to be performed at Cycle 1, Cycle 2, Cycle 3, Cycle 4, Cycle 5, and Cycle 7 before dosing. After Cycle 7 (Week 18), PROs are to be performed every 9 weeks (e.g., Week 27, Week 36, Week 45). PROs are to be performed up to a year or End of Treatment, whichever comes first, at treatment discontinuation, and at the 30-day post-treatment discontinuation follow-up visit.			

The general rule of mapping relative day to analysis visit is provided in Table 11. The relative days are counted from the date the first dose.

Table 11
Mapping Relative Day to Analysis Visit

Week	Day	Day Range
Week 0	1	-7 - 1
Week 3	21	2 - 31
Week 6	42	32 - 52
Week 9	63	53 - 73
Week 12	84	74 - 105
Week 18	126	106 - 157
Week 27	189	158 - 220
Week 36	252	221 - 283
Week 45	315	284 - 346

At each scheduled visit, two instruments, EORTC QLQ-C30 and EQ-5D-3L, will be collected. If a patient does not complete the PRO instruments, the site staff will record the reason for missingness from pre-defined choices by using a miss-mode form. If there are multiple PRO collections within any of the stated time windows, we use the closest collection to the target day.

3.12.4 Analysis Populations

The PRO analyses are based on the PRO Full Analysis Set (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. This population consists of all randomized patients who have received at least one dose of study medication and have completed at least one PRO assessment.

3.12.5 Statistical Methods

3.12.5.1 PRO Compliance Summary

Completion and compliance of EORTC QLQ-C30 and EQ-5D3L by visit and by treatment will be described based on PRO FAS population. Numbers and percentages of complete and missing data at each visit will be summarized for each of the treatment groups. 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.

Completion rate in the FAS population is defined as the percentage of number of subjects who complete at least one item over the number of subjects in the PRO FAS population at each time points.

$$Completion\ Rate = \frac{Number\ \ of\ subjects\ \ who\ complete\ \ at\ least\ \ one\ \ item}{Number\ \ of\ \ subject\ \ s\ in\ PRO\ FAS\ popula\ \ tion}$$

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

$$Compliance\ Rate = \frac{Number\ \ of\ subjects\ \ who\ complete\ \ at\ least\ \ one\ \ item}{Number\ \ of\ \ eligible\ \ subjects\ \ who\ are\ expected\ \ to\ complete}$$

The reasons for non-completion and non-compliance will be summarized.

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 the Table 10 and Table 11.

3.12.5.2 Mean Change from Baseline

The time point for the mean change from baseline is defined as the latest time point at which completion rate ≥ 60% and compliance rate ≥ 80% based on blinded data review prior to the database lock for any PRO analysis. The primary time point for the analyses of PRO endpoint is Week 12.

To assess the treatment effects based on the PRO score change from baseline, for each continuous endpoint defined, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [8] will be used. This 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 the baseline value 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 analysis model will include the PRO score as the response variable, and treatment by study visit interaction, and stratification factors used for analyses of the primary endpoints as covariates.

The cLDA model is specified as follows:

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

Where Y_{ijt} is the PRO score for subject i, with treatment assignment j at visit t, γ_0 is the baseline mean for both treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t, X_i is the stratification factor vector for this patient, and β is the coefficient vector for stratification factors. An unstructured covariance matrix will be used to model the correlation among repeated measurements. The cLDA model implicitly treats missing data as missing at random (MAR).

The treatment difference in terms of least square (LS) mean score change from baseline to the time point as specified at the beginning of this section will be estimated from this model, together with 95% CI and nominal p-value. In addition, model-based LS mean score with 95% CI will be provided by treatment group and study visit.

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 empirical option for PROC MIXED in SAS 9.4 will be used because the sandwich variance estimator is asymptotically unbiased.

3.12.5.3 Time to Deterioration

Time to deterioration (TTD): For the EORTC QLQ-C30, a 10 points or greater worsening from baseline for each scale represents a clinically relevant deterioration based on prior literature [9] [10] [11] TTD is defined as the time to first onset of 10 or more (out of 100) deterioration from baseline in a given scale/sub-scale/item and confirmed by a second adjacent 10 or more deterioration from baseline.

The Kaplan-Meier method will be used to estimate the TTD curve for each treatment group. The estimate of median time to deterioration and its 95% confidence interval will be obtained from the Kaplan-Meier estimates. The treatment difference in TTD will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling and with a single treatment covariate will be used to assess the magnitude of the treatment difference (i.e., hazard ratio). The hazard ratio and its 95% CI will be reported. The stratification factors used for analyses of the primary endpoints 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 12** provides censoring rule for TTD analysis.

Table 12
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

3.12.5.4 Proportions of Deterioration/Stable/Improvement

Patients' post-baseline PRO score will be classified as "improved", "stable" or "deteriorated" according to a 10 points or greater change for each of the instrument/scale, as this magnitude of change is perceived by patients as being clinically significant.

Since missing data cannot be ignored at the time point as specified at the beginning of Section 3.12.5.2, the number and proportion of patients who "improved", "stable", or "deteriorated", from baseline will be summarized by treatment group and at the prior analysis visit based on MAR imputation of missing data.

4 REFERENCES

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

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

3. Robins J.M., Finkelstein D.M. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000 Sep;56(3):779-88.

4. Latimer N.R., Abrams K.R., Siebert U. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. BMC Medical Research Methodology. 2019; 19:69.

5. Uno, H., et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol. 2014;32 (22):2380-2385.

6. Odell P.M., Anderson K.M., Kannel B.W. New models for predicting cardiovascular events. Journal of Clinical Epidemiology 1994;47(6):583-592.

7. Anderson K.M., Clark J. B. Fitting spending functions. Statist. Med,2010;29:321-327.

8. Liang K.Y., Zeger S.L. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhyā: The Indian Journal of Statistics, Series B. 2000 Apr 1:134-48.

9. Haes J, Curran D, Young T, Bottomley A, Flechtner H, Aaronson N, Blazeby J, Bjordal K, Brandberg Y, Greimel E, Maher J. Quality of life evaluation in oncological clinical trials: the EORTC model. Eur J Cancer. 2000;36(7):821-5.

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

11. King M.T. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res 5:555-567, 1996

12. The EORTC QLQ-C30 Manuals. Reference Values and Bibliography.