Official Protocol Title:	An Open-label, Randomized Phase 3 Study of MK-6482 Versus Everolimus in Participants with Advanced Renal Cell Carcinoma That Has Progressed After Prior PD-1/L1 and VEGF-Targeted Therapies
NCT number:	NCT04195750
Document Date:	28-NOV-2022

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

1. INTRODUCTION 4
2. SUMMARY OF CHANGES 4
3. ANALYTICAL AND METHODOLOGICAL DETAILS 4
3.1 Statistical Analysis Plan Summary 4
3.2 Responsibility for Analyses/In-House Blinding 5
3.3 Hypotheses/Estimation 6
3.3.1 Primary Objectives & Hypotheses 6
3.3.2 Secondary Objectives & Hypothesis 6
3.3.3 Exploratory Objectives 6
3.4 Analysis Endpoints 7
3.4.1 Efficacy Endpoints 7
3.4.2 Safety Endpoints 7
3.4.3 Patient-Reported Outcome (PRO) Endpoints 7
3.5 Analysis Populations 9
3.5.1 Efficacy Analysis Populations 9
3.5.2 Safety Analysis Populations 9
3.5.3 PRO Analysis Populations 9
3.6 Statistical Methods 10
3.6.1 Statistical Methods for Efficacy Analyses 10
3.6.1.1 Progression-free Survival 10
3.6.1.2 Overall Survival 12
3.6.1.3 Objective Response Rate (ORR) 12
3.6.1.4 Duration of Response (DOR) 12
3.6.1.5 Analysis Strategy for Key efficacy Endpoints 13
3.6.2 Statistical Methods for Safety Analyses 14
3.6.3 Statistical Methods for Patient-Reported Outcomes (PRO) Analyses 15
3.6.3.1 PRO Scoring Algorithm 15
3.6.3.2 PRO completion and compliance summary 17
3.6.3.3 Change from baseline 20
3.6.3.4 Time to Confirmed Deterioration (TTD) 20
3.6.3.5 Overall Improvement and Overall Improvement/Stability 21
3.6.3.6 Analysis Strategy for Key PRO Endpoints 22
3.6.4 Demographic and Baseline Characteristics 22
3.6.5 Japan Safety Run-in Cohort (local Japan protocol version) 22
3.7 Interim Analyses 23
3.7.1 Efficacy Interim Analyses 23
3.7.2 Safety Interim Analyses 24
3.8 Multiplicity 24
3.8.1 Objective Response Rate 25
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 and Effect of Baseline Factors....30
3.11 Extent of Exposure....31
4. APPENDIX....31
4.1 Technical details for PRO analysis....31
4.2 Technical details for minimum spending approach....32
5. REFERENCES....33

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

- Updated details of analysis for patient-related outcomes (PRO). Summary of EQ-5D will be provided for VAS instead of the health utility scores.
- Updated schedule and plan for interim analyses and final analyses based on protocol amendment 006.
- Provided details on analyses for Japanese Safety Run -In cohort (Japan local protocol version 007).
- Provided strata collapsing strategy for analyses of objective response due to potential small strata

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 – Responsibility for Analyses/In-House Blinding through 3.12 – Extent of Exposure.

Study Design Overview	A randomized, open-label, Phase 3 controlled study of the HIF-2α inhibitor, MK-6482, versus everolimus in participants with advanced renal cell carcinoma after prior therapy
Treatment Assignment	Participants will be randomly assigned in a 1:1 ratio to receive either MK-6482 at 120 mg QD or everolimus at 10 mg QD. Stratification factors are as follows: IMDC prognostic scores: 0 vs 1-2 vs 3-6 Number of prior VEGF/VEGF receptor targeted therapies for advanced RCC: 1 vs 2-3
Analysis Populations	Efficacy: ITT Safety: APaT
Primary Endpoints	• PFS • OS
Secondary Endpoints	• ORR • DOR • PRO assessment • AEs and discontinuations due to AEs

Statistical Methods for Key Efficacy Analyses	The primary hypotheses comparing MK-6482 to everolimus with respect to PFS and OS will be evaluated using a stratified log-rank test. The hazard ratio will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. The stratified Miettinen and Nurminen method with strata weighted by sample size will be used for analysis of ORR.
Statistical Methods for Key Safety Analyses	For analyses in which 95% CIs will be provided for between-treatment differences in the percentage of participants with events, these analyses will be performed using the Miettinen and Nurminen method [1]
Interim Analyses	Two IAs are planned for the study. Results will be reviewed by an external DMC. Details are provided in Section 9.7. First IA (IA1): • Timing: to be performed after ~563 PFS events have occurred AND ~7 months after last participant randomized. • Primary purpose: efficacy analyses for PFS, OS and ORR. Second IA (IA2): • Timing: to be performed after ~410 OS events have occurred AND ~17 months after last participant randomized. • Primary purpose: efficacy analysis for OS and PFS (final analysis) FA: • To be performed after ~483 OS events have occurred AND ~27 months after last participant randomized. • Primary purpose: efficacy analysis for OS.
Multiplicity	The overall Type I error rate over the primary and secondary hypotheses is strongly controlled at 2.5% (1-sided), with 0.5% initially allocated to PFS (H1), 1.9% initially allocated to OS (H2) and 0.1% initially allocated to ORR (H3). By using the graphical approach of Maurer and Bretz, if one hypothesis is rejected, the alpha will be shifted to other hypotheses [2].
Sample Size and Power	The planned sample size is approximately 736 participants. There will be ~483 deaths at the final OS analysis. With 483 deaths, the study has ~85.4% power for detecting a HR of 0.75 at an initially assigned 0.019 (1-sided) significance level. It is estimated that there will be ~626 events at the final PFS analysis (ie, the second IA of the study). With 626 PFS events, the study has ~96.9% power for detecting a HR of 0.70 at an initially assigned 0.005 (1-sided) significance level. Based on all randomized participants, the power of the ORR testing at the allocated α=0.001 is approximately 99.9% to detect a 15-percentage point difference between an underlying 5% response rate in the control arm and a 20% response rate in the experimental arm.

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 IRT.

Although this is an open-label study, analyses or summaries generated by randomized treatment assignment, or actual treatment received will be limited and documented.

An independent radiologist(s) will perform the central imaging review without knowledge of treatment assignments.

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Protocol Section 3.0 – Objective(s) & Hypothesis(es) and are listed in this section.

3.3.1 Primary Objectives & Hypotheses

1) **Objective:** To compare MK-6482 to everolimus with respect to PFS per RECIST 1.1 as assessed by BICR.

   Hypothesis (H1): MK-6482 is superior to everolimus with respect to PFS per RECIST 1.1 by BICR.

2) **Objective**: To compare MK-6482 to everolimus with respect to OS.

   Hypothesis (H2): MK-6482 is superior to everolimus with respect to OS.

3.3.2 Secondary Objectives & Hypothesis

1) **Objective**: To compare MK-6482 to everolimus with respect to ORR based on RECIST 1.1 as assessed by BICR.

   Hypothesis (H3): MK-6482 increases ORR according to RECIST 1.1 by BICR compared to everolimus.

2) **Objective:** To evaluate the DOR as assessed by BICR according to RECIST 1.1.
3) **Objective:** To evaluate the safety and tolerability of MK-6482 compared to everolimus.
4) **Objective:** To evaluate TTD and change from baseline in HRQoL using the EORTC QLQ-C30 and the FKSI-DRS
5) **Objective:** To characterize VAS as measured using the EuroQoL EQ-5D-5L.

3.3.3 Exploratory Objectives

1) **Objective**: To evaluate the PK of MK-6482 administered orally as monotherapy.

2) **Objective:** To identify molecular (genomic, metabolic, and/or proteomic) biomarkers that may be indicative of clinical response/resistance, safety, pharmacodynamic activity, and/or the mechanism of action of MK-6482 and other treatments.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Primary Endpoints

Progression-free survival (PFS) – RECIST 1.1 assessed by BICR

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded independent central imaging vendor review or death due to any cause, whichever occurs first.

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

Secondary Endpoints

Objective Response Rate (ORR) – ORR is defined as the proportion of participants in the analysis population who have a best overall response of either confirmed complete response (CR) or partial response (PR) per RECIST 1.1 as assessed by BICR.

Duration of Response (DOR) – For participants who demonstrated confirmed CR or PR, DOR is defined as the time from the first documented evidence of confirmed CR or PR until the first documented date of disease progression or death due to any cause, whichever occurs first. Responses and progression will be assessed using RECIST 1.1 by BICR.

3.4.2 Safety Endpoints

A description of safety endpoint assessment is provided in Section 4.2.1.3 of the protocol. Assessments include, but not limited to, the incidence of, causality of, and outcome of AEs/SAEs; and changes in laboratory values.

3.4.3 Patient-Reported Outcome (PRO) Endpoints

As described in Section 4.2.1.4 of the Protocol, the following secondary PRO assessments will be evaluated:

- **Change from baseline** in EORTC QLQ-C30 global health status/quality of life scores, physical functioning score, role functioning score, FKSI-DRS score, and EQ-5D visual analogue scale (VAS).
- **Time to confirmed deterioration (TTD)** as measured by EORTC QLQ-C30 global health status/quality of life scores, physical functioning score, role functioning score and FKSI-DRS score

  Based on prior literature (Osoba et al., 1998; King, 1996), a 10 points or greater worsening from baseline for each scale represents a clinically relevant deterioration for EORTC QLQ-C30. Three points or more decrease from baseline represents a clinically relevant deterioration for FKSI-DRS. TTD is defined as the time from baseline to the first onset of a 10 or more points for EORTC QLQ-C30 and 3 or more points for FKSI-DRS

deterioration with confirmation by the subsequent visit of a 10 or more points for EORTC QLQ-C30 and 3 or more points for FKSI-DRS deterioration from baseline. If the first deterioration is at the last PRO assessment timepoint in the current database, then no confirmation is required.

- **Overall improvement / stability / stability + improvement / deterioration** in EORTC QLQ-C30 global health status / QoL, physical functioning, role functioning, and FKSI-DRS score where:

  The assessment for possible PRO response at a time point considering subsequent confirmation is defined as follows:

Assessment Category at a time point (one analysis visit)	**Change from baseline at a time point (one analysis visit)**	**Change from baseline at the subsequent time point (the next consecutive analysis visit)**
Improvement	score improved from baseline by ≥10 points for EORTC QLQ-C30 and ≥3 points for FKSI-DRS	score improved from baseline by ≥10 points for EORTC QLQ-C30 and ≥3 points for FKSI-DRS
Stability	score improved from baseline by ≥10 points for EORTC QLQ-C30 and ≥3 points for FKSI-DRS	score improved or worsened from baseline by <10 points for EORTC QLQ-C30 and <3 points for FKSI-DRS
	score improved or worsened from baseline by <10 points for EORTC QLQ-C30 and <3 points for FKSI-DRS	score improved or worsened from baseline by <10 points for EORTC QLQ-C30 and <3 points for FKSI-DRS
	score improved or worsened from baseline by <10 points for EORTC QLQ-C30 and <3 points for FKSI-DRS	score improved from baseline by ≥10 points for EORTC QLQ-C30 and ≥3 points for FKSI-DRS
Worsening	score worsened from baseline by ≥10 points for EORTC QLQ-C30 and ≥3 points for FKSI-DRS	not required
Unconfirmed	A time point assessment that doesn't meet any of the above criteria.	

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

Based on prior literature, Osoba et al. (1998) and King (1996), a 10-point or greater worsening from baseline for each scale of the EORTC QLQ-C30 and a 3-point or greater worsening from baseline for FKSI-DRS represent a clinically relevant deterioration.

Changes from baseline in EORTC QLQ-C30 scores and FKSI-DRS will also be interpreted according to recent subscale-specific guidelines, which indicate that clinically meaningful differences vary by scale (Cocks et al., 2012, Cella et.al. 2007, Cella et.al. 2008).

3.5 Analysis Populations

The study includes a main Phase 3 global cohort and Japan safety run-in cohort. The analysis population for the primary efficacy, safety and PRO analyses will only include participants from the global cohort. The participants in Japan safety run-in cohort who received at least 1 dose of study treatment will be analyzed separately.

3.5.1 Efficacy Analysis Populations

The ITT population will serve as the population for primary efficacy analyses. All randomized subjects will be included in this population. Subjects will be included in the treatment group to which they are randomized.

3.5.2 Safety Analysis Populations

The All Participants as Treated (APaT) population will be used for the analysis of safety data in this study. The APaT population consists of all randomized participants who received at least 1 dose of study treatment. Participants will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the APaT population. For most participants this will be the treatment group to which they are randomized. Participants who take incorrect study treatment for the entire treatment period will be included in the treatment group corresponding to the study treatment actually received. Any participant who receives the incorrect study treatment for a short amount of time (e.g., no greater than 4 weeks), but receives the correct treatment for the rest of the time, will be analyzed according to the correct treatment group and a narrative will be provided for any events that occur during the time when the participant is incorrectly dosed.

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

3.5.3 PRO Analysis Populations

The PRO analyses are based on the PRO Full Analysis Set (FAS) population, defined as 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.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

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

The stratification factors used for randomization (see Section 6.3.2 of 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 [1]. In the event that there are small strata, for the purpose of analysis, strata will be combined to ensure sufficient number of participants, responses, and events in each stratum.

Based on a blinded review of response counts by stratum prior to the first efficacy interim analysis, if there are ≤ 5 responses in one or more strata, stratification factors will be combined for analysis of objective response to ensure sufficient number of events in each stratum. That is, VEGF receptor therapies stratum of 1 and 2-3 will be collapsed withing the stratum of IMDC score equals 3 to 6 for analysis of objective response.

No strata collapsing is planned for PFS and OS analyses.

3.6.1.1 Progression-free Survival

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, the 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 of the protocol) will be applied to both the stratified log-rank test and the stratified Cox model.

Since disease progression is assessed periodically, disease progression can occur any time in the time interval between the last assessment where disease progression was not documented and the assessment when disease progression is documented. The true date of disease progression will be approximated by the earlier of the date of the first assessment at which disease progression is objectively documented per RECIST 1.1 by BICR and the date of death.

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

Sensitivity analyses will be performed for comparison of PFS based on investigator's assessment.

In order to evaluate the robustness of the PFS endpoint per RECIST 1.1 by BICR, 2 sensitivity analyses with different sets of censoring rules will be performed. The first sensitivity analysis follows the intention-to-treat principle. That is, disease progressions/deaths are counted as events regardless of missed study visits or initiation of new anti-cancer therapy. The second sensitivity analysis considers discontinuation of treatment due to reasons other than CR or initiation of new anticancer treatment, whichever occurs later, to be a disease progression event for participants without documented disease progression or death. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied. The censoring rules for primary and sensitivity analyses are summarized in Table 1.

Table 1 Censoring Rules for Primary and Sensitivity Analyses of PFS

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

3.6.1.2 Overall Survival

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, the 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 of 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 the participant was last known to be alive.

Additional supportive unstratified analyses may also be provided.

3.6.1.3 Objective Response Rate (ORR)

The stratified Miettinen and Nurminen's method will be used for comparison of the ORR between 2 treatment groups. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be reported. The stratification factors used for randomization (See Section 6.3.2 of the protocol) will be applied to the analysis. A sensitivity analysis will be performed for the comparison of ORR based on investigator's assessment.

The point estimate of ORR will be provided by treatment group, together with 95% CI using exact binomial method proposed by Clopper and Pearson (1934).

3.6.1.4 Duration of Response (DOR)

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

Sensitivity analyses will be performed to assess DOR based on investigator's assessment.

Censoring rules for DOR are summarized in Table 2.

For each DOR analysis, a corresponding summary of the reasons responding subjects are censored will also be provided. Responding participants who are alive, have not progressed, have not initiated new anticancer treatment, have not been determined to be lost to follow-up, and have had a disease assessment within ~5 months of the data cutoff date are considered ongoing responders at the time of analysis. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied.

Table 2 Censoring Rules for DOR

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new 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, if any	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	Disease progression or death	End of response (Event)
A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response.		

3.6.1.5 Analysis Strategy for Key efficacy Endpoints

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

Table 3 Analysis Strategy for Key Efficacy Endpoints

Endpoint/Variable	Statistical Method†	Analysis Population	Missing Data Approach
Primary Hypothesis 1			
PFS as assessed by BICR according to RECIST 1.1	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 1)
Primary Hypothesis 2			
OS	Test: stratified log-rank test Estimation: stratified Cox model with Efron's tie handling method	ITT	Censored at last known alive date

Endpoint/Variable	Statistical Method†	Analysis Population	Missing Data Approach
Key Secondary Hypothesis 3			
ORR as assessed by BICR according to RECIST 1.1	Testing and estimation: stratified Miettinen and Nurminen method	ITT	Participants with missing data are considered non-responders
Abbreviations: BICR = blinded independent central review; ITT = intent-to-treat; ORR = objective response rate; OS = overall survival; PFS = progression-free survival; RECIST 1.1 = Response Evaluation Criteria in Solid Tumors Version 1.1 Note: Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization will be used as stratification factors for analysis.			

3.6.2 Statistical Methods for Safety Analyses

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

The safety analysis 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 and vital signs parameters are either pre-specified as Tier-1 endpoints, or will be classified as belong to "Tier 2" or "Tier 3", based on the number of events observed.

Tier 1 Events

Safety parameters or adverse events of special interest that are identified a priori constitute Tier 1 safety endpoints that will be subject to inferential testing for statistical significance.

There are no known AEs associated with participants with RCC 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 [1].

Membership in Tier 2 requires that at least 10% of participants in any treatment group exhibit the event; all other AEs and predefined limits of change will belong to Tier 3. The threshold of at least 10% of participants was chosen for Tier 2 events because the population enrolled in this study is in critical condition and usually experiences various AEs of similar types regardless of treatment; events reported less frequently than 10% of participants would obscure the assessment of the overall safety profile and add little to the interpretation of potentially meaningful treatment differences. In addition, Grade 3 to 5 AEs (≥5% of participants in 1 of the treatment groups) and SAEs (≥5% of participants in 1 of the treatment

groups) will be considered Tier 2 endpoints. Because many 95% CIs may be provided without adjustment for multiplicity, the CIs should be regarded as a helpful descriptive measure to be used in safety review, not as a formal method for assessing the statistical significance of the between-group differences.

Tier 3 Events

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

Continuous Safety Measures

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

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Any AE (≥10% of participants in one of the treatment groups)	X	X
	Any serious AE (≥5% of participants in one of the treatment groups)	X	X
	Any Grade 3 to 5 AE (≥5% of participants in one of the treatment groups)	X	X
Tier 3	AEs, Specific AEs, SOCs		X
	Discontinuation due to AE		X
	Dose interruption due to AE		X
	Change from baseline results (laboratory, ECGs, Vital Signs)		X
Abbreviations: SOC = system organ class; X = results will be provided			

3.6.3 Statistical Methods for Patient-Reported Outcomes (PRO) Analyses

This section describes the planned analyses for the PRO endpoints.

3.6.3.1 PRO Scoring Algorithm

QLQ-C30 Scoring

The QLQ-C30 is composed of both multi-item scales and single-item measures. These include a global health status / QoL scale, five functional scales, three symptom scales, and six single

items. Each of the multi-item scales includes a different set of items - no item occurs in more than one scale.

All of the scales and single-item measures will follow a standardization procedure prior to analysis so that scores range from 0 to 100. A high scale score represents a higher response level. Thus a high score for a functional scale represents a high / healthy level of functioning; a high score for the global health status / QoL represents a high QoL; but a high score for a symptom scale / item represents a high level of symptomatology / problems.

According to the EORTC QLQ-C30 Scoring Manual [9], the principle for scoring these scales is the same in all cases:

1. Estimate the average of the items that contribute to the scale; this is the raw score.

2. Use a linear transformation to standardize the raw score, so that scores range from 0 to 100; a higher score represents a higher ("better") level of functioning, or a higher ("worse") level of symptoms.

Specifically, if items $I_1, I_2, ..., I_n$ are included in a scale, the scoring 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*:

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

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

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. The QLQ-C30 has been designed so that all items in any scale take the same range of values. Therefore, the range of *RS* equals the range of the item values. 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.

EQ-5D-5L Scoring

The EQ-5D-5L utility score will be calculated based on the European algorithm [10] based on responses on the five health state dimensions, including mobility, self-care, usual activities, pain / discomfort, and anxiety / depression. The EQ-5D also includes a graded (0 to 100) vertical visual analog scale (VAS) on which the participant rates his or her general state of health at the time of the assessment.

FKSI-DRS scoring

The FKSI-DRS includes 9 questions which are negatively stated items [6]. So the score for each item must be reversed by subtracting the response from "4". After that, all the item scores are summed to a total, which is the subscale score. The final FKSI-DRS score is defined as below.

Higher scores correspond with better QoL.

FKSI-DRS score = Subscale score × 9 ÷ Number of items answered

Such FKSI-DRS score is a prorated score if there are missing items. It is acceptable as long as more than 50% of the items were answered (e.g., a minimum of 5 of 9 items). Otherwise, the FKSI-DRS score is considered as missing.

3.6.3.2 PRO completion and compliance summary

Completion and compliance of EORTC QLQ-C30, FKSI-DRS, and EQ-5D VAS by visit and by treatment will be described. Numbers and percentages of complete and missing data at each visit will be summarized.

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

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

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

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

The completion and compliance status will be summarized as below:

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

The reasons for non-completion as scheduled of these measures are collected using "miss_mode" forms filled by site personnel and will be summarized in a table format. The

schedule (study visits and estimated study times) and mapping of study visit to analysis visit for PRO data collection is provided in Table 5 and Table 6.

Table 5 PRO Data Collection Schedule

Study Period	Screening	Treatment Period				EOT	Posttreatment	
Visit:	**Screening[a]**	**Wk 1 Day 1[a]**	**Wk 3 Day 1**	**Wk 5 Day 1**	**Wk 9 Day 1**	**Wk 13+ Day 1[b]**	**DC**	**Safety Follow-up**
Scheduling Window (Days):	**–28 to –1**	**+3**	**±3**	**±3**	**±3**	**±5**	**At time of DC**	**30 Days After Last Dose (+7 days)**
Patient-reported Outcomes								
FKSI-DRS		X	X	X	X	Q4W	X	X
EORTC QLQ-C30		X	X	X	X	Q4W	X	X
EuroQoL EQ-5D-5L		X	X	X	X	Q4W	X	X

a. Week 1 Day 1 denotes the first dose of study treatment, which should be on the date of randomization, but can be within 3 days following randomization. Every effort should be made to ensure the participants receive the first dose of study intervention on the day of randomization.
b. Clinic visits after Week 13 are Q4W.

Table 6 Mapping of Study visit to Analysis Visit

Treatment Week	1	3	5	9	13	17	21	25	29	33
Target Day [a]	1	15	29	57	85	113	141	169	197	225
Range [a]	-28 - 1	2 - 22	23 - 43	44 - 71	72 - 99	100 - 127	128 - 155	156 - 183	184 - 211	212 - 239
Treatment Week	37	41	45	49	53	57	61	65	69	73
Target Day [a]	253	281	309	337	365	393	421	449	477	505
Range [a]	240 - 267	268 - 295	296 - 323	324 - 351	352 - 379	380 - 407	408 - 435	436 - 463	464 - 491	492 - 519
Treatment Week	77	81	85	89	93	97	101	105	109	113
Target Day [a]	533	561	589	617	645	673	701	729	757	785
Range [a]	520 - 547	548 - 575	576 - 603	604 - 631	632 - 659	660 - 687	688 - 715	716 - 743	744 - 771	772 - 799
Treatment Week	117	121	125	129	133	137	141	145	149	153
Target Day [a]	813	841	869	897	925	953	981	1009	1037	1065
Range [a]	800 - 827	828 - 855	856 - 883	884 - 911	912 - 939	940 - 967	968 - 995	996 - 1023	1024 - 1051	1052 - 1079
Treatment Week	157	161	165	169	173	177	181	185	189	193
Target Day [a]	1093	1121	1149	1177	1205	1233	1261	1289	1317	1345
Range [a]	1080 - 1107	1108 - 1135	1136 - 1163	1164 - 1191	1192 - 1219	1220 - 1247	1248 - 1275	1276 - 1303	1304 - 1331	1332 - 1359

a. Day = Date of the assessment – date of the first dose +1

3.6.3.3 Change from baseline

The time point for the mean change from baseline is defined as the latest time point at which CR-T ⩾ 60% and CR-E⩾ 80%, and week 17 was selected 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, role functioning, FKSI-DRS score, and EQ-5D VAS, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [11] 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 of 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. Model-based LS mean with 95% CI will be provided by treatment group for PRO scores at baseline and post-baseline time point.

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

Line plots for the empirical mean change from baseline in EORTC QLQ-C30 global health status/QoL, physical functioning, role functioning, and FKSI-DRS score 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 EORTC QLQ-C30 global health status/quality of life scores, all functioning, symptom scores, FKSI-DRS score, and EQ-5D VAS.

3.6.3.4 Time to Confirmed Deterioration (TTD)

The Kaplan-Meier method will be used to estimate the TTD curve for each treatment group. The estimate of median time to deterioration and its 95% confidence interval will be obtained from the Kaplan-Meier estimates. The treatment difference in TTD will be assessed by the stratified log-rank test, and two-sided nominal p-value will be reported. A stratified Cox proportional hazard model with Efron's method of tie handling and with a single treatment covariate will be used to assess the magnitude of the treatment difference (i.e, HR). The HR and its 95% CI will be reported. The same stratification factors used for randomization (See Section 6.3.2 of the protocol) will be used as the stratification factors in both the stratified log-rank test and the stratified Cox model.

The approach for the TTD analysis will be based on the assumption of non-informative censoring. The participants who do not have deterioration on the last date of evaluation will be censored. Table 7 provides censoring rule for TTD analysis.

Table 7 Censoring Rules for Time to Confirmed Deterioration

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

3.6.3.5 Overall Improvement and Overall Improvement/Stability

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

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

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

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Mean change from baseline in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning FKSI-DRS And EQ-5D VAS	cLDA model	FAS	Model-based.
TTD in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning And FKSI-DRS	stratified log-rank test and HR estimation using stratified Cox model with Efron's tie handling method	FAS	Censored according to rules in Table 6.
Overall improvement and overall improvement/stability in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning And FKSI-DRS	Stratified Miettinen and Nurminen method	FAS	Participants with missing data are considered not achieving improvement/stability.
Abbreviations: cLDA = constrained longitudinal data analysis, FAS = full analysis set, QoL = quality of life. TTD=time to confirmed deterioration, HR = hazard ratio.			

3.6.4 Demographic and Baseline Characteristics

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

3.6.5 Japan Safety Run-in Cohort (local Japan protocol version)

For participants from Japan safety run-in cohort, listings of participant demographics, discontinuations, death, adverse events, drug exposure, and efficacy endpoints of OS, PFS, and best response will be provided. An AE summary table will also be provided.

3.7 Interim Analyses

The results of IAs will not be shared with the investigators prior to the completion of the study. Participant-level unblinding will be restricted to an internal unblinded statistician and scientific programmer performing the IA, who will have no other responsibilities associated with the study.

An external DMC will serve as the primary reviewer of the results of the IAs of the study and will make recommendations for discontinuation of the study or protocol modifications to the EOC of the Sponsor (Appendix 1 of the protocol). If the DMC recommends modifications to the design of the protocol or discontinuation of the study, this executive committee (and potentially other limited Sponsor personnel) may be unblinded to results at the treatment level in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of interim analyses will be documented. Additional logistical details will be provided in the DMC Charter.

Treatment-level results from the IA will be provided to the DMC by the unblinded statistician. Prior to final study unblinding, 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 IA.

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

3.7.1 Efficacy Interim Analyses

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

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

Table 8 Summary of Interim and Final Analyses Strategy

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	PFS OS ORR	Enrollment is complete with approximately ~7 months of follow-up and ~563 PFS events have been observed.	~29 months	• Interim PFS analysis • Interim OS analysis • Final ORR analysis
IA2	OS PFS	~410 deaths have occurred and ~17 months after last participant randomized.	~39 months	• Interim OS analysis • Final PFS analysis
FA	OS	~483 deaths have occurred and ~27 months after last participant randomized.	~49 months	• Final OS analysis

Abbreviations: FA = final analysis; IA1 = interim analysis 1; IA2 = interim analysis 2; ORR = objective response rate; OS = overall survival; PFS = progression-free survival

Note: If the PFS events accrue slower than expected when participants have been followed up for approximately 7 months, IA1 can take place when at least ~525 PFS events have occurred.

Note: If the OS events accrue slower than expected for IA2 and/or FA, the Sponsor may conduct the analysis with up to additional 3 months of follow-up, or the specified number of events is observed, whichever occurs first.

If an efficacy boundary is crossed at any interim analysis or FA for either PFS or OS, the study will be declared to have met its primary objective.

3.7.2 Safety Interim Analyses

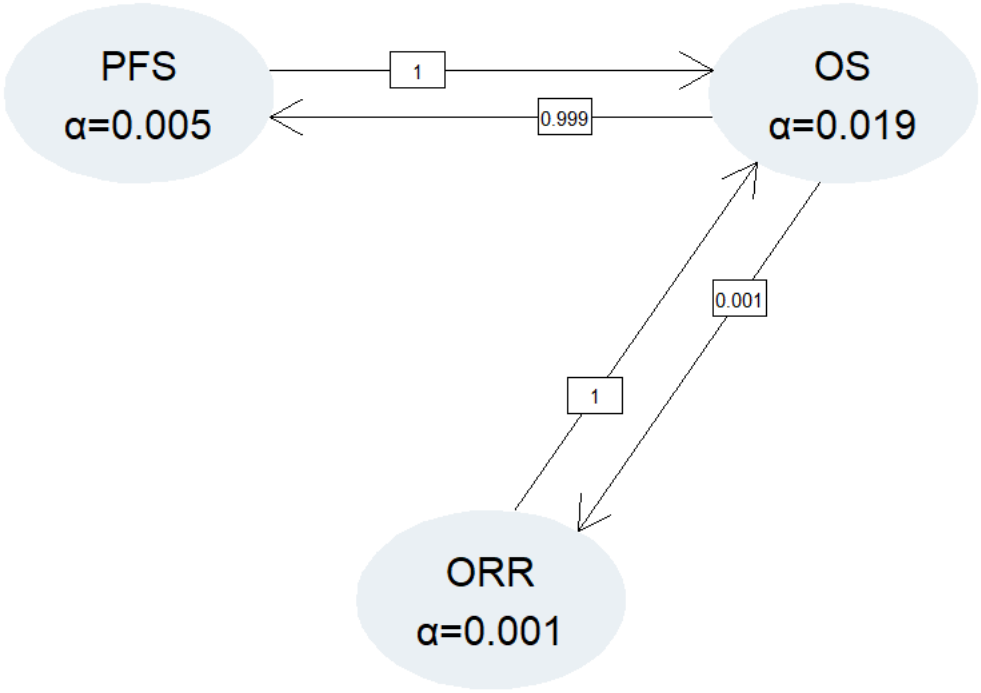
The DMC will be responsible for periodic interim safety reviews, as specified in the DMC charter. Interim safety analyses will also be performed at the time of interim efficacy analyses. Details will be specified in the DMC charter.

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [2] to control multiplicity for multiple hypotheses as well as IAs. According to this approach, study hypotheses may be tested more than once, and when a particular null hypothesis is rejected, the α allocated to that hypothesis can be reallocated to other hypothesis tests. Figure 1 shows the initial 1-sided α allocation for each hypothesis in the ellipse representing the hypothesis. The weights for reallocation from each hypothesis to the others are shown in the boxes on the lines connecting hypotheses.

Initial α assigned to OS, PFS and ORR will be 0.019, 0.005 and 0.001, respectively. If any hypothesis is rejected, α will be reallocated to the other hypotheses.

Figure 1 Multiplicity Diagram for Type I Error Control (One-Sided)



3.8.1 Objective Response Rate

The primary test of the ORR hypothesis will occur at IA1, at an initial one-sided α level of 0.001. The testing of ORR will be based on all randomized participants in the study. If ORR null hypothesis is not rejected at IA1 at the initially allocated alpha level and if the null hypotheses for OS and PFS are rejected at any analysis time, the p-value from the ORR test at IA1 analysis will be compared to an updated α-level of 0.025.

Based on all randomized participants, the power at α=0.001 as well as the approximate treatment difference required to reach the bound (ΔORR) are shown in Table 9, assuming underlying 5% and 20% response rates in the control and experimental groups, respectively.

Table 9 Possible α Levels and Approximate Objective Response Rate Difference Required to Demonstrate Efficacy for Objective Response at Interim Analysis

α	~Δ Objective Response Rate (ORR)	Power (ΔORR=0.15)
0.001	0.073	99.9%
0.025	0.047	>99.9%

3.8.2 Progression-free Survival

The initial α-level for testing the PFS hypothesis is 0.005. If the null hypothesis for OS is rejected, Figure 1 shows that α=0.019 from the OS hypothesis is almost fully reallocated to PFS hypothesis testing. Thus, the PFS null hypothesis may be tested at α=0.005 (initially allocated α), α=0.024 if the OS hypothesis is rejected and the ORR hypothesis is not rejected, or at α=0.025 if both the OS and ORR null hypotheses are rejected. Table 10 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. A Lan-DeMets O'Brien-Fleming spending function was used to derive the bounds and boundary properties of the PFS hypothesis at each analysis based on the estimated number of events.

Since the timing of the PFS interim analysis at IA1 will be dependent on both the number of PFS events and a minimum follow-up time, for the PFS hypothesis, alpha will be spent as a function of the minimum of the actual event information fraction and the expected event information fraction at IA1. If events accrue faster than expected, this approach ensures that the actual spending will be no more aggressive than the planned, while at the same time ensuring that not all alpha is spent prior to the accrual of the final planned event counts.

The expected number of events at the final PFS analysis is ~626. The final PFS analysis at IA2 will use the remaining Type I error not spent at the IA1, regardless of the actual number of PFS events observed. The p-value bound at the final PFS analysis will be calculated by considering the correlation between the test statistics as determined by the actual number of PFS events at IA1 and the final PFS analysis at IA2.

Table 10 summarizes the boundary properties of 3 possible scenarios.

Table 10 Efficacy Boundaries and Properties for Progression-Free Survival Analyses

Analysis	Value	α=0.005	α=0.024	α=0.025
IA1: 90%* N = 736 Events: 563 Month: 29	Z	2.7383	2.1119	2.0937
	p (1-sided) [a]	0.0031	0.0173	0.0181
	HR at bound[b]	0.7938	0.8369	0.8382
	P(Cross) if HR=1[c]	0.0031	0.0173	0.0181
	P(Cross) if HR=0.7[d]	0.9330	0.9831	0.9839
IA2:100% N = 736 Events: 626 Month: 39	Z	2.6421	2.0694	2.0529
	p (1-sided) [a]	0.0041	0.0193	0.0200
	HR at bound[b]	0.8095	0.8474	0.8486
	P(Cross) if HR=1[c]	0.0050	0.0240	0.0250
	P(Cross) if HR=0.7[d]	0.9685	0.9928	0.9931

Abbreviations: HR = hazard ratio; IA = interim analysis; OS = overall survival; PFS = progression-free survival

The number of events and timings are estimated approximately.

*Percentage of the target number of events at PFS final analysis anticipated at IA.

[a]*p* (1-sided) is the nominal α for testing.

[b]HR at bound is the approximate HR required to reach an efficacy bound.

[c]P (Cross if HR=1) is the probability of crossing a bound under the null hypothesis.

[d]P(Cross if HR=0.7) is the probability of crossing a bound under the alternative hypothesis.

Note that if the α-reallocation from OS hypothesis testing occurs at an analysis after hypothesis testing for PFS has been completed, the previously computed PFS test statistic for the PFS final analysis may be re-evaluated based on the updated bounds.

3.8.3 Overall Survival

The OS hypothesis may be tested at α=0.019 (initially allocated α), α=0.02 (if only the ORR null hypothesis is rejected), α=0.024 (if the PFS null hypothesis is rejected only), or α=0.025 (if both the PFS and ORR null hypotheses are rejected). Table 11 summarizes the boundary properties of 4 possible scenarios.

Table 11 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	Value	α=0.019	α=0.020	α=0.024	α=0.025
IA1: 62%* N = 736 Events: 300 Month: 29	Z	2.7596	2.7336	2.6396	2.6158
	p (1-sided)[a]	0.0029	0.0031	0.0041	0.0045
	HR at bound[b]	0.7269	0.7289	0.7370	0.7392
	P(Cross) if HR=1[c]	0.0029	0.0031	0.0041	0.0045
	P(Cross) if HR=0.75[d]	0.3943	0.4037	0.4408	0.4508
IA2: 85%* N: 736 Events: 410 Month: 39	Z	2.3275	2.3058	2.2271	2.2095
	p (1-sided)[a]	0.0100	0.0106	0.0130	0.0136
	HR at bound[b]	0.7945	0.7961	0.8024	0.8038
	P(Cross) if HR=1[c]	0.0109	0.0115	0.0142	0.0149
	P(Cross) if HR=0.75[d]	0.7264	0.7331	0.7592	0.7646
FA: 100% N: 736 Events: 483 Month: 49	Z	2.1541	2.1343	2.0629	2.0467
	p (1-sided)[a]	0.0156	0.0164	0.0196	0.0203
	HR at bound[b]	0.8220	0.8234	0.8288	0.8300
	P(Cross) if HR=1[c]	0.0190	0.0200	0.0240	0.0250
	P(Cross) if HR=0.75[d]	0.8536	0.8578	0.8740	0.8773

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

The bounds provided in the table above are based on the assumption that the expected number of events at IA1, IA2 and FA are 300, 410 and 483, 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 IA.

3.9 Sample Size and Power Calculations

The study will randomize approximately 736 participants in a 1:1 ratio to the MK-6482 and everolimus arms. PFS and OS are primary endpoints for the study, with ORR as the key secondary endpoint.

Based on all randomized participants, the power of the ORR testing at the allocated α=0.001 is approximately 99.9% to detect a 15-percentage point difference between an underlying 5% response rate in the control arm and a 20% response rate in the experimental arm.

For the PFS endpoint, based on an expected number of 626 events at the final analysis and one interim analysis at 90% of the final target number of events, the study has approximately 96.9% power to demonstrate an HR of 0.7 at an overall α level of 0.005 (1 sided), 99.3% power at an α level of 0.024 (1-sided), and 99.3% power at an α level of 0.025 (1-sided).

For the OS endpoint, based on a target number of 483 events at the final analysis and 2 interim analyses at approximately 62% and 85% of the final target number of events, the study has approximately 85.4% power to detect an HR of 0.75 at an overall α level of 0.019 (1-sided), 85.8% power at an α level of 0.020 (1-sided), 87.4% power at an α level of 0.024 (1-sided), and 87.7% power at an α level of 0.025 (1-sided).

The above sample size and power calculations for PFS and OS assume the following:

- PFS follows an exponential distribution with a median of 4.4 months for the control group NCT01668784 study (nivolumab versus everolimus in advanced RCC).

- OS follows an exponential distribution with a median of 20.0 months for the control group NCT01668784 study (nivolumab versus everolimus in advanced RCC).

- Enrollment period of 22 months

- An annual dropout rate of 20% and ~1% for PFS and OS, respectively

- A follow-up period of 17 and 27 months for PFS and OS, respectively, after the last participant enrolls.

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

3.10 Subgroup Analyses and Effect of Baseline Factors

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

- IMDC risk category (favorable vs intermediate vs poor; favorable vs intermediate plus poor)
- Geographic region (North America vs Western Europe vs Rest of the World)
- Age category (<65 vs ≥65 years)
- Sex (male vs female)
- Race (white vs non-white)
- Number of prior VEGF/VEGF receptor targeted therapies for advanced RCC (1 vs 2-3)

- Number of prior lines of therapy (1 vs 2 vs 3)

The consistency of the treatment effect will be assessed using descriptive statistics for each category of the subgroup variables listed above. If the number of participants in a category of a subgroup variable is less than 10 % of the ITT population, the subgroup analysis will not be performed for this category of the subgroup variable, and this subgroup variable will not be displayed in the forest plot. The subgroup analyses for PFS and OS will be conducted using an unstratified Cox model, and the subgroup analyses for ORR will be conducted using the unstratified Miettinen and Nurminen method. Compliance (Medication Adherence)

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

For each participant, percent compliance will be calculated using the following formula:

$$\text{Percent Compliance} = \frac{\text{Number of Days on Therapy}}{\text{Number of Days Should Be on Therapy}} \times 100\,\%$$

For participants who are still on study treatment (ongoing at the cutoff date), the "Number of Days Should be on Therapy" is the total number of days from the first scheduled intervention day to the last scheduled intervention day. For participants who discontinued from the study treatment permanently, the "Number of Days Should Be on Therapy" is the total number of days from the first scheduled intervention day to the last dose day.

Summary statistics will be provided on percent compliance by treatment group for the APaT population

3.11 Extent of Exposure

Extent of exposure for a participant is defined as number of days in which the participant receives the study intervention. Summary statistics will be provided on Extent of Exposure for the APaT population.

4. APPENDIX

4.1 Technical details for PRO analysis

The cLDA model assumes a common mean across treatment groups at baseline and a different mean for each treatment at each of the post-baseline time points. In this model, the response vector consists of baseline and the values observed at each post-baseline time point. Time is treated as a categorical variable so that no restriction is imposed on the trajectory of the means over time. The cLDA model is specified as follows:

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

where Y_{ijt} is the PRO score for participant i, with treatment assignment j at visit t; γ_0 is the baseline mean for all treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t; X_i is the stratification factor (binary) vector for this participant, and β is the

coefficient vector for stratification factors. An unstructured covariance matrix will be used to model the correlation among repeated measurements. If the unstructured covariance model fails to converge with the default algorithm, then Fisher scoring algorithm or other appropriate methods can be used to provide initial values of the covariance parameters. In the rare event that none of the above methods yield convergence, a structured covariance such as Toeplitz can be used to model the correlation among repeated measurements. In this case, the asymptotically unbiased sandwich variance estimator will be used. The cLDA model implicitly treats missing data as missing at random (MAR).

4.2 Technical details for minimum spending approach

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

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

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

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

To illustrate how the minimun spending approach is implemented, examples with 2 hypothetical scenarios where events accrue slower and faster than expected are given below for the OS first interim analysis with the total alpha of 1.9% (initially allocated).

IA1 boundary calculation:

For the first OS interim analysis at IA1, the p-value boundary is the same as alpha spending determined from the Lan-DeMets spending function. At the time of the analysis, 300 events are expected over the target 483 events at the FA.

- Hypothetical scenario 1 (events accrue slower than expected): 290 events are observed. The spending time is calculated as *t* = 290/483 = 60.0% and p-value boundary = 0.0025.
- Hypothetical scenario 2 (events accrue faster than expected): 320 events are observed. The spending time is calculated as *t* = 300/483 = 62.1%, p-value boundary = 0.0029.

5. REFERENCES

[1] Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med. 1985 Apr-Jun;4(2):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] 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.

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

[5] 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.

[6] Cella D, Yount S, Brucker PS, Du H, Bukowski R, Vogelzang N, et al. Development and validation of a scale to measure disease-related symptoms of kidney cancer. Value Health. 2007 Jul-Aug;10(4):285-93.

[7] Cella D, Li JZ, Cappelleri JC, Bushmakin A, Charbonneau C, Kim ST, et al. Quality of life in patients with metastatic renal cell carcinoma treated with sunitinib or interferon alfa: results from a phase III randomized trial. J Clin Oncol. 2008 Aug 1;26(22):3763-9

[8] Clopper C; Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26 (4): 404-413.

[9] The EORTC QLQ-C30 Manuals, Reference Values and Bibliography.

[10] EuroQol Research Foundation. EQ-5D-5L User Guide, 2019.

[11] 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.

Revision History

Date	Summary of Change
03 Jan 2020	Original Document
28-Nov-2022	First Amendment: Update for efficacy IA1