Official Protocol Title:	Protocol Title: A Phase 3 Randomized Study of Lenvatinib in Combination with Pembrolizumab Versus Standard of Care in Participants with Metastatic Colorectal Cancer Who Have Received and Progressed on or After or Became Intolerant to Prior Treatment
NCT number:	NCT04776148
Document Date:	24-JUN-2025

Title Page

THIS SUPPLEMENTAL STATISTICAL ANALYSIS PLAN AND ALL OF THE INFORMATION RELATING TO IT ARE CONFIDENTIAL AND PROPRIETARY PROPERTY OF MERCK SHARP & DOHME LLC., A SUBSIDIARY OF MERCK & CO., INC., NJ, U.S.A (MSD).

Protocol Title: A Phase 3 Randomized Study of Lenvatinib in Combination with Pembrolizumab Versus Standard of Care in Participants with Metastatic Colorectal Cancer Who Have Received and Progressed on or After or Became Intolerant to Prior Treatment

Protocol Number: 017-03

Compound Number: MK-7902

Sponsor Name:
Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc.
(hereafter referred to as the Sponsor or MSD)

Legal Registered Address:
126 East Lincoln Ave.
P.O. Box 2000
Rahway, NJ 07065, USA

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

sSAP Version #	Protocol Amendment #	sSAP Section # and Name	Description of Change	Brief Rationale
03	03	4.1.1 China-specific Requirements	Added clarifications that PRO is not to be analyzed in China subpopulation	To clarify the PRO analysis for the China subpopulation

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 3.2-3.12.

Study Design Overview	A Phase 3 randomized study of lenvatinib in combination with pembrolizumab versus standard of care in participants with metastatic colorectal cancer who have received and progressed on or after or became intolerant to prior treatment
Treatment Assignment	Approximately 434 participants will be randomized in a 1:1 ratio between two treatment groups: (1) the lenvatinib plus pembrolizumab arm and (2) the SOC arm. Stratification factor is: Presence of liver metastasis (Yes/No). This is an open-label study.
Analysis Populations	Efficacy: ITT Safety: APaT Patient-reported outcome: FAS
Primary Endpoint(s)	Overall survival
Key Secondary Endpoints	Progression-free survival per RECIST 1.1 as assessed by BICR. Objective response rate per RECIST 1.1 as assessed by BICR.

Statistical Methods for Key Efficacy Analyses	The primary hypothesis testing for OS and secondary hypothesis testing for PFS will be evaluated by comparing the experimental group to the control group using a stratified log-rank test. The HR will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method. The stratified M&N method with strata weighted by sample size will be used for analysis of ORR [1].
Statistical Methods for Key Safety Analyses	For analyses in which 95% CIs will be provided for between-treatment differences in the percentage of participants with events, these analyses will be performed using the M&N method.
Interim Analyses	**Efficacy** One interim analysis is planned in this study. Results will be reviewed by an eDMC. Details are provided in Section 3.7. • Interim Analysis: o Timing: to be performed after both ~260 OS events have been observed and ~ 7 months after last participant randomized o Primary purpose: interim efficacy analysis for OS, final analysis for PFS and ORR • Final Analysis: o Timing: to be performed after both ~336 OS events have been observed and ~ 7 months after interim analysis o Primary purpose: final analysis for OS **Safety** An interim safety analysis will be performed and reviewed by the eDMC 6 months after first participant is randomized. Afterwards, the eDMC will review safety data periodically in the study. Details will be specified in the DMC charter.
Multiplicity	The overall type I error over the primary and secondary hypotheses is strongly controlled at 2.5% (1-sided), with 2.5% initially allocated to OS (H1), 0% to PFS (H2), and 0% to ORR (H3). By using the graphical approach of Maurer and Bretz [2], if one hypothesis is rejected, the alpha will be shifted to other hypotheses.
Sample Size and Power	The planned sample size is approximately 434 participants. It is estimated that there will be ~ 336 OS events at the final analysis. With 336 OS events, CCI [redacted] at an initially assigned 0.025 (1-sided) significance level.
Abbreviations: APaT = all participants as treated; BICR = blinded independent central review; CI = confidence interval; DMC = data monitoring committee; eDMC = external data monitoring committee; FAS = full-analysis set; HR = hazard ratio; ITT = intent to treat; M&N = Miettinen and Nurminen; ORR = objective response rate; OS = overall survival; PFS = progression free survival; SOC = standard of care	

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 as appropriate in this protocol, and the allocation will be implemented in IRT.

This study is being conducted as a randomized, open-label study, i.e., participants, investigators, and Sponsor personnel will be aware of participant treatment assignments after each participant is enrolled and treatment is assigned. Although the study is open label, analyses or summaries generated by randomized intervention assignment, or actual intervention received will be limited and documented.

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

3.3 Hypotheses/Estimation

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

3.4 Analysis Endpoints

Efficacy and safety endpoints that will be evaluated are listed below.

3.4.1 Efficacy Endpoints

Primary

- **Overall Survival**

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

Secondary

- **Progression-free survival**

  PFS is defined as the time from randomization to the first documented disease progression per RECIST 1.1 by BICR or death due to any cause, whichever occurs first.

- **Objective Response Rate**

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

- **Duration of Response**

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

3.4.2 Safety Endpoints

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

3.4.3 PRO Endpoints

- Change from baseline in EORTC QLQ-C30 global health status/QoL, physical functioning and appetite loss and EORTC QLQ-CR29 bloating scores for the combination of lenvatinib plus pembrolizumab versus SOC
- Time to first deterioration (TTD) in EORTC QLQ-C30 global health status/QoL, physical functioning and appetite loss and EORTC QLQ-CR29 bloating scores for the combination of lenvatinib plus pembrolizumab versus SOC

  Based on prior literature (Bjordal, et al., 2000; Osoba D, 1998 [3]; King, 1996 [4]), a 10 points or greater worsening from baseline for each scale represents a clinically relevant deterioration for EORTC. TTD is defined as the time from baseline to the first onset of a 10 or more points deterioration from baseline.
- Change from baseline in CCI for the combination of lenvatinib plus pembrolizumab versus SOC
- CCI

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	score improved from baseline by ≥10 points
Stability	score improved from baseline by ≥10 points	score improved or worsened from baseline by <10 points
	score improved or worsened from baseline by <10 points	score improved or worsened from baseline by <10 points
	score improved or worsened from baseline by <10 points	score improved from baseline by ≥10 points
Worsening	score worsened from baseline by ≥10 points	not required
Unconfirmed	A time point assessment that doesn't meet any of the above criteria.	

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

Changes from baseline in EORTC QLQ-C30 scores will also be interpreted according to recent subscale-specific guidelines, which indicate that clinically meaningful differences vary by scale (Cocks et al., 2012) [5].

3.5 Analysis Populations

China participants randomized after enrollment of the global portion is closed (if any) will not be included in the global analysis populations. The China subpopulation (including China participants in global portion and China extension portion) will also be analyzed separately per local regulatory requirements.

3.5.1 Efficacy Analysis Populations

The ITT population will serve as the primary population for the analysis of efficacy data in this study. The ITT population consists of all randomized participants. Participants will be analyzed in the treatment arm to which they are randomized. Details of the approach to handling missing data are provided in Section 3.6.1.4.

3.5.2 Safety Analysis Populations

Safety Analyses will be conducted in the APaT population, which consists of all randomized participants who received at least one dose of study treatment. Participants will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the APaT population. This will be the treatment group to which they are randomized except for participants who take incorrect study treatment for the entire treatment period; such participants will be included in the treatment group corresponding to the study treatment actually received.

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 the respective safety 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 FAS population, defined as all randomized participants who have at least one PRO assessment available for the specific endpoint and have received at least one dose of the study intervention. Participants will be analyzed in the treatment group to which they are randomized.

3.6 Statistical Methods

Statistical testing and inference for safety analyses are described in Section 3.6.2. 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 may be computed for other efficacy analyses, but should be interpreted with caution due to potential issues of multiplicity, sample size, etc.

3.6.1 Statistical Methods for Efficacy Analyses

This section describes the statistical methods that address the primary and secondary objectives.

The stratification factors used for randomization (see Section 6.3.2 of protocol) will be applied to all stratified analyses, in particular, the stratified log-rank test, stratified Cox model, and stratified M&N method [1].

3.6.1.1 Overall Survival

The nonparametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. 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 (see Section 6.3.2 of 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.

3.6.1.2 Progression-Free Survival

The nonparametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test. 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, HR) between the treatment arms. The HR and its 95% CI from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization (see Section 6.3.2 of protocol) will be applied to both the stratified log-rank test and the stratified Cox model.

Since disease progression is assessed periodically, PD can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. The true date of disease progression will be approximated by the earlier of the date of the first assessment at which PD is objectively documented per RECIST 1.1 by BICR and the date of death. Death is always considered a PD event. Surgical participants (i.e., those who undergo oncologic surgeries with curative intent) will be followed to the disease recurrence after the surgery for PFS analysis.

For the primary analysis, any participant who experiences an event (PD or death) immediately after 2 or more missed disease assessments will be censored at the last disease assessment prior to the missed visits. In addition, any participant who initiates new anti-cancer therapy prior to documented progression will be censored at the last disease assessment prior to the initiation of new anti-cancer therapy. Participants who do not start new anti-cancer 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.

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, PDs/deaths are counted as events regardless of missed study visits or initiation of new anti-cancer therapy. The second sensitivity analysis considers initiation of new anticancer treatment or discontinuation of treatment due to reasons other than complete response, whichever occurs later, to be a PD event for participants without documented PD or death. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied. The censoring rules for the 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 anti-cancer 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 anti-cancer therapy	Censored at last disease assessment prior to the earlier date of ≥2 consecutive missed disease assessment and new anti-cancer 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			

3.6.1.3 Objective Response Rate

The stratified M&N method will be used for the comparison of ORR between 2 treatment groups. The difference in ORR and its 95% CI from the stratified M&N method with strata weighting by sample size will be reported. The stratification factors used for randomization (see Section 6.3.2 of protocol) will be applied to the analysis.

The point estimate of ORR will be provided by treatment group, together with 95% CI using exact binomial method proposed by Clopper and Pearson (1934) [6].

3.6.1.4 Duration of Response

If sample size permits, DOR will be summarized descriptively using Kaplan-Meier method. Only the subset of participants who show a confirmed complete response or partial response will be included in this analysis. Censoring rules for DOR are summarized in Table 2.

For each DOR analysis, a corresponding summary of the reasons responding participants are censored will also be provided. Responding participants who are alive, have not progressed, have not initiated new 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 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. For surgical participants, DOR will be defined using time from first documented evidence of response to the disease recurrence or death after the surgery.

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	PD 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 Variables

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

Table 3 Analysis Strategy for Key Efficacy Variables

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Primary Analyses			
OS	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at the date participant last known to be alive
Key Secondary Analyses			
PFS per RECIST 1.1 by BICR	Testing: stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored according to rules in Table 1
ORR per RECIST 1.1 by BICR	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.			

3.6.2 Statistical Methods for Safety Analyses

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

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

Tier 1 Events

Safety parameters or AEs 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 Tier 1 events for this protocol. Adverse events that are immune-mediated or potentially immune-

mediated are well documented and will be evaluated separately; however, these events have been characterized consistently throughout the pembrolizumab clinical development program, and determination of statistical significance is not expected to add value to the safety evaluation. The combination of lenvatinib plus pembrolizumab has not been found to be associated with any new safety signals. Finally, there are no known AEs associated with participants with CRC for which determination of a p-value is expected to impact the safety assessment.

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 M&N 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. The broad AE categories consisting of the proportion of participants with any AE, a drug related AE, a serious AE, an AE which is both drug-related and serious, a Grade 3-5 AE, a drug-related Grade 3-5 AE, and discontinuation due to an AE will be considered Tier 3 endpoints. Only point estimates by treatment group are provided for Tier 3 safety parameters.

Continuous Safety Measures

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

Table 4 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Grade 3-5 AE (incidence ≥5% of participants in one of the treatment groups)	X	X
	Serious AE (incidence ≥5% of participants in one of the treatment groups)	X	X
	AEs (incidence ≥10% of participants in one of the treatment groups)	X	X
Tier 3	Any AE		X
	Any Grade 3-5 AE		X
	Any Serious AE		X
	Any Drug-Related AE		X
	Any Serious and Drug-Related AE		X
	Any Grade 3-5 and Drug-Related AE		X
	Discontinuation due to AE		X
	Death		X
	Specific AEs, SOCs (incidence <10% of participants in all of the treatment groups)		X
	Change from Baseline Results (lab toxicity shift, vital signs)		X
Abbreviations: AE = adverse events; SOC = standard of care			

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

Time to Grade 3-5 AE

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

3.6.3 Statistical Methods for Patient-Reported Outcome Analyses

This section describes the planned analyses for the PRO endpoints.

3.6.3.1 PRO Scoring Algorithm

EORTC QLQ-C30 Scoring: Each scale or item is scored between 0 and 100, according to the EORTC QLQ-C30 standard scoring algorithm [7]. For global health status/quality of life and all functional scales, a higher value indicates a better level of function; for symptom scales and items, a higher value indicates increased severity of symptoms.

EORTC QLQ-CR29 Scoring: All of the scales and single-item measures range in score from 0 to 100. A high score for the functional scale and functional single-items represents a high level of functioning, whereas a high score for the symptom scales and symptom single-items represents a high level of symptomatology or problems.

3.6.3.2 PRO Completion and Compliance Summary

Completion and compliance of EORTC QLQ-C30, EORTC QLQ-CR29, and EQ-5D by visit and by treatment will be described. Numbers and percentages of complete and missing data at each visit will be summarized.

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.

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

Treatment Week	**Week 0**	**Week 2**	**Week 4**	**Week 6**	**Week 8**	**Week 10**	**Week 12**	**Week 16**	**Week 20**	**Week 24**	**Week 28**
Day	1	15	29	43	57	71	85	113	141	169	197
Range (relative day to first dose date)	[-5, 8]	[9, 22]	[23, 36]	[37, 50]	[51,64]	[65,78]	[79, 99]	[100, 127]	[128, 155]	[156, 183]	[184, 211]
Treatment Week	**Week 32**	**Week 36**	**Week 40**	**Week 44**	**Week 52**	**Week 60**	**Week 68**	**Week 76**	**Week 84**	**Week 92**	**Week 100**
Day	225	253	281	309	365	421	477	533	589	645	701
Range (relative day to first dose date)	[212, 239]	[240, 267]	[268, 295]	[296, 337]	[338, 393]	[394, 449]	[450, 505]	[506, 561]	[562, 617]	[618, 673]	[674, 729]

3.6.3.3 Change from Baseline

The time point for the mean change from baseline analysis is defined as the latest time point at which CR-T ≥ 60% and CR-E ≥ 80%, and week 8 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 EORTC QLQ-C30 global health status/QoL, physical functioning and appetite loss, EORTC QLQ-CR29 bloating scores, and EQ-5D-5L VAS, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [8] will be applied, with the PRO score as the response variable, and treatment, time, the treatment by time interaction, and the stratification factors used for randomization (see Section 6.3.2 of protocol) as covariates. The treatment difference in terms of least square mean change from baseline will be estimated from this model together with 95% CI. Model-based least square 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 and appetite loss, EORTC QLQ-CR29 bloating scores, and EQ-5D-5L VAS 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/QoL scores, all functioning and symptom scores, and for EORTC QLQ-CR29 all functioning and symptom scores.

3.6.3.4 Time-to-First-Deterioration

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 (ie, HR). The HR and its 95% CI will be reported. The stratification factors used for randomization (see Section 6.3.2 of 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 6 provides censoring rule for TTD analysis.

Table 6 Censoring Rules for Time to First 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 / Overall Improvement and Stability

Overall improvement rate will be analyzed, which is defined as the proportion of participants who have achieved an overall 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 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) [6].

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

3.6.3.6 Analysis Strategy for Key PRO Endpoints

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

Table 7 Analysis Strategy for Key PRO Variables

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Mean change from baseline in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Appetite loss EORTC QLQ-CR29 • Bloating And EQ-5D VAS	cLDA model	FAS	Model-based.
TTD in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Appetite loss And EORTC QLQ-CR29 • Bloating	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.
CCI [redacted]			

Abbreviations: cLDA = constrained longitudinal data analysis, FAS = full analysis set, TTD = time to deterioration, HR = hazard ratio, QoL = quality of life.

3.6.4 Demographic and Baseline Characteristics

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

3.7 Interim Analyses

The eDMC will serve as the primary reviewer of the results of the IAs and will make recommendations for discontinuation of the study or modification to the executive oversight committee of the Sponsor. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee and potentially

other limited Sponsor personnel may be unblinded to the treatment level results in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of IAs will be documented by the unblinded statistician. Additional logistic details will be provided in the eDMC Charter.

Treatment-level results of the interim analysis will be provided by the unblinded statistician to the eDMC. Prior to final study unblinding, the unblinded statistician will not be involved in any discussions regarding modifications to the protocol or statistical methods, identification of protocol deviations, or data validation efforts after the IAs.

Efficacy Interim Analysis

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

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

Table 8 Summary of Interim and Final Analyses Strategy

Analyses	Key Endpoints	Timing	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA	OS (PFS and ORR if OS is rejected)	Both ~260 OS events have been observed and ~ 7 months after last participant randomized	~ 18 months	• Interim OS analysis • Final PFS and ORR analysis
FA	OS	both ~336 OS events have been observed and ~ 7 months after interim analysis	~ 25 months	• Final OS analysis
Abbreviations: FA = final analysis; IA = interim analysis; ORR = objective response rate; OS = overall survival; PFS = progression free survival.				

3.7.1 Safety Interim Analysis

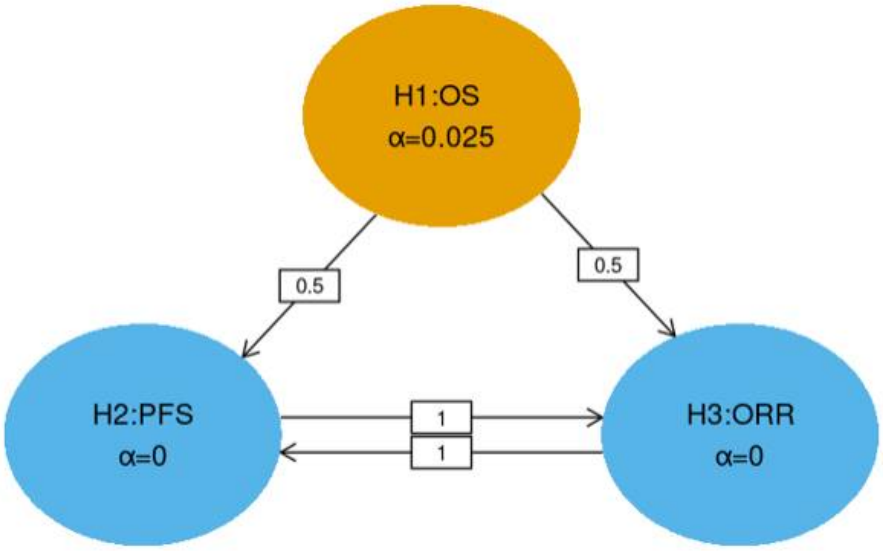
The eDMC will be responsible for periodic interim safety reviews as specified in the eDMC charter. An interim safety analysis will be performed 6 months since first participant is randomized. Afterwards, the eDMC will review safety data periodically in the study. Interim safety analyses will also be performed at the time of interim efficacy analyses. Details will be specified in the eDMC charter.

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [2] to provide strong multiplicity control for multiple hypotheses as well as interim analyses. According to this approach, study hypotheses may be tested more than once, and when a particular null hypothesis is rejected, the α allocated to that hypothesis can be reallocated to other hypothesis tests. Note that if the OS null hypothesis is rejected at FA of the study, the previously computed PFS and ORR test statistics at IA may be used for inferential testing with its updated bounds considering the α reallocation from the OS hypothesis. 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.

CCI [redacted]

Figure 1 Multiplicity Diagram for Type I Error Control



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

Note: If OS null hypothesis is rejected, the allocation strategy allows testing of PFS and ORR at α = 0.0125, separately.

3.8.1 Overall Survival

The study will test OS at IA and FA. Following the multiplicity strategy CCI [redacted].

Table 9 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	**Value**	**α=0.025**
IA: 77%* N: 434 Events: 260 Month: 18	Z	2.2976
	p (1-sided)[a]	0.0108
	HR at bound[b]	0.7517
	P(Cross) if HR=1[c]	0.0108
	P(Cross) if HR=0.7[d]	0.7185
FA N: 434 Events: 336 Month: 25	Z	2.0177
	p (1-sided)[a]	0.0218
	HR at bound[b]	0.8022
	P(Cross) if HR=1[c]	0.0250
	P(Cross) if HR=0.7[d]	0.9000

Abbreviations: HR = hazard ratio; IA = interim analysis, FA = final analysis.
The number of events and timings are estimated.
*Percentage of total planned events at the interim analysis.
[a]*p* (1-sided) is the nominal α for group sequential testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P(Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P(Cross if HR=0.7) is the probability of crossing a bound under the alternative hypothesis.

The bounds provided in the table above are based on the assumptions that the expected number of events at IA and FA are 260 and 336, respectively. At the time of an analysis, the observed number of events may differ substantially from the expected. To avoid overspending at an IA and leave reasonable α for the FA, the minimum α 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 IA 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 IA 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 observed 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.

3.8.2 Progression-free Survival

The study will test PFS at IA only if the OS null hypothesis is rejected. Following the multiplicity strategy as outlined in Figure 1, the PFS hypothesis may be tested at α=0.0125 (if the OS null hypothesis is rejected, but not the ORR hypothesis) or at α= 0.025 (if both the OS and ORR null hypothesis is 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.

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

Analysis	Value	α=0.0125	α=0.025
IA N = 434 Events*: 404 Month: 18	Z	2.2414	1.9600
	p (1-sided) [a]	0.0125	0.025
	HR at bound[b]	0.8000	0.8227
	P(Cross) if HR=1[c]	0.0125	0.025
	P(Cross) if HR=0.65[d]	0.9820	0.9912

Abbreviations: HR = hazard ratio; IA = interim analysis.
*The number of events and timing is estimated.
[a]*p* (1-sided) is the nominal α for group sequential testing.
[b]HR at bound is the approximate HR required to reach an efficacy bound.
[c]P (Cross if HR=1) is the probability of crossing a bound under the null hypothesis.
[d]P(Cross if HR=0.65) is the probability of crossing a bound under the alternative hypothesis.

3.8.3 Objective Response Rate

The study will test ORR only once at the IA if the OS null hypothesis is rejected. Following the multiplicity strategy as outlined in Figure 1, the ORR hypothesis may be tested at α=0.0125 (if the OS null hypothesis is rejected, but not the PFS hypothesis) or at α= 0.025 (if both the OS and PFS null hypothesis is rejected). Power at the possible α-levels as well as the approximate treatment difference required to reach the bound (ΔORR) are shown in Table 11, assuming underlying CCI response rates in the control and experimental groups, CCI

Table 11 Possible α Levels and Approximate ORR Difference Required to Demonstrate Efficacy for Objective Response at IA

α	~Δ Objective Response Rate (ORR)	Power (ΔORR=0.1)
0.0125	0.0549	0.970
0.025	0.0480	0.985

3.8.4 Safety Analysis

The eDMC has responsibility for assessment of overall risk/benefit. When prompted by safety concerns, the eDMC can request corresponding efficacy data. eDMC 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 434 participants in a 1:1 ratio into the lenvatinib plus pembrolizumab arm the SOC arms. OS are primary endpoint for the study, with PFS and ORR as the key secondary endpoints.

For the OS endpoint, based on a target number of 336 events and 1 IA at approximately 77% of the target number of events, CCI at the initially allocated α=0.025 (1-sided).

For the PFS endpoint, based on a target number of 404 events at the IA (final PFS analysis), the study has CCI at the reallocated α=0.0125 (1-sided) if OS hypothesis is rejected.

Based on the 434 participants with at least 7 months of follow-up, the power of the ORR testing at the reallocated α=0.0125 (1-sided) if OS hypothesis rejected is approximately 97.0% to detect a 10-percentage point difference between CCI .

Note that the above power calculations are based on a constant HR assumption.

Based on CORRECT and RECOURSE studies, the above sample size and power calculations for OS and PFS assume the following:

CCI [redacted]

- Enrollment period of 11 months with enrollment ramp-up over first 2 months.
- CCI [redacted]
- A follow-up period of 14 and 7 months and for OS and PFS, respectively, after the last participant is randomized.

CCI [redacted]

3.10 Subgroup Analyses

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

- Geographic region (Asia vs. Western Europe/North America vs. Rest of World)
- ECOG performance status (0, 1)
- Age category (<65 years, ≥65 years)
- Sex (female, male)
- Race (white, all others)
- Presence of liver metastasis (Yes, No)
- PD-L1 expression level (CPS≥1, CPS<1)
- BRAF (wild type, mutant)
- RAS (wild type, mutant)
- Investigators' choice of standard of care chemotherapy prior to randomization (Regorafenib versus TAS102)

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 OS and PFS will be conducted using an unstratified Cox model, and the subgroup analyses for ORR will be conducted using the unstratified Miettinen and Nurminen method.

3.11 Compliance (Medication Adherence)

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

3.12 Extent of Exposure

Extent of Exposure for a participant is defined as the number of cycles and number of days for which the participant receives the study intervention. Summary statistics will be provided on the extent of exposure for the overall study intervention, and for lenvatinib and pembrolizumab separately, for the APaT population.

4 APPENDIX

4.1 Region/Country-specific Requirements

4.1.1 China-specific Requirements

This section outlines the statistical analysis strategy and procedures for China subpopulation (including China participants randomized in the global portion and China extension portion). China refers to China mainland in this section.

CCI

[redacted]

After the enrollment for the global portion is completed, participants in China will continue to be enrolled in a 1:1 ratio into the lenvatinib plus pembrolizumab arm and the SOC arm until the sample size for the China subpopulation reaches approximately 100 in total (global and China extension portion combined).

CCI

[redacted]

Key elements of the statistical analysis plan for China subpopulation are summarized below. More details are provided in following sections.

Analysis Populations	Efficacy: ITT China subpopulation (including China participants randomized in the global portion and the China extension portion) Safety: APaT China subpopulation (including China participants randomized in the global portion and the China extension portion who received at least 1 dose of study intervention) PRO: PRO will not be analyzed in the China subpopulation
Efficacy Endpoint(s)	Efficacy endpoints are the same as described in Section 3.4.1
Safety Endpoint (s)	Safety endpoints are the same as described in Section 3.4.2
PRO Endpoint(s)	PRO will not be analyzed in the China subpopulation
Statistical Methods for Efficacy Analyses	No formal hypothesis testing is planned, and no multiplicity adjustment will be applied to the analysis for China subpopulation. Unstratified methods will be used for China subpopulation analyses.
Statistical Methods for Safety Analyses	Safety analyses for China subpopulation are the same as those for the global portion as described in Section 3.6.2 if applicable.

Summaries of Baseline Characteristics and Demographics	They are the same for China subpopulation as those for the global portion as described in Section 3.6.4
Analyses Timing	CCI [redacted]
	At the time of global analyses, China subpopulation data including the extension portion may be provided for supportive purpose to fulfill local regulatory needs.
Hypotheses and Multiplicity	No hypothesis testing is planned for the China subpopulation analyses. No multiplicity adjustment will be applied to the analysis of China subpopulation.
Sample Size Calculations	After the completion of global portion enrollment, the China extension portion will continue to enroll participants and randomize eligible participants until the sample size for the overall China subpopulation reaches approximately 100.
	CCI [redacted]

4.1.1.1 Responsibility for Analyses/In-House Blinding

For all China participants, including participants randomized in the global portion and the China extension portion, patient level treatment randomization information will be blinded for a designated team for China analysis within the Sponsor until the China extension portion database lock is achieved.

4.1.1.2 Analyses Timing

CCI [redacted]

However, to increase operational efficiency of the global and China extension portions, if the criteria for conduct of the analyses in the China subpopulation are met before the IA or FA for the global portion, the analysis for China subpopulation may occur at the same time as the IA or FA for the global portion. If the statistical significance for the global portion has been demonstrated at the IA or FA and it is projected that the criteria for conduct of the analysis in China subpopulation including the China extension portion will be met within ~3 months after the IA or FA for the global portion, then the analysis for China subpopulation including the China extension portion may be based on the same database lock as the IA or FA for the global portion.

At the time of global analyses, China subpopulation data including the extension portion may be provided for supportive purpose to fulfill local regulatory needs.

4.1.1.3 Sample Size Calculations

After the completion of global portion enrollment, the China extension portion will continue to enroll participants and randomize eligible participants until the sample size for the overall randomized China subpopulation reaches approximately 100. Participants from China enrolled in the China extension portion of this study after completion of the global enrollment will not be included in the primary analysis population for the global portion.

CCI

4.1.1.4 Subgroup Analyses

Analyses may be considered for the China subgroup (i.e., China participants randomized in the global portion only) based on Sponsor's discretion and/or consultation with health authorities if the global interim/final analysis shows positive results and leads to filing and the China subpopulation enrollment has been completed.

4.2 Technical Details for cLDA Model

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:

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

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).

4.3 Technical Details for Minimal Spending Approach

Below are the technical details for the 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 one hypothetical scenario where events accrue faster than expected are given below for the OS analyses with the total alpha of 2.5% (initially allocated). There are 2 planned analyses for OS at IA and FA, respectively.

IA boundary calculation:

For the OS interim analysis at IA, the p-value boundary is the same as alpha spending α_1 determined from the Lan-DeMets spending function. At the time of the analysis, 260 events are expected over the target 336 events at the FA.

- Hypothetical scenario (events accrue faster than expected): 270 events are observed. The spending time is calculated as *t* = 260/336= 77%, p-value boundary = 0.0108.

FA boundary calculation:

The alpha spending at the FA is $\alpha - \alpha_1$. FA boundary (C_2) is solved from $P(Z_2 \geq C_2, Z_1 < C_1 | H_0) = \alpha - \alpha_1$, with test statistics Z_1 and Z_2 being multivariate normal and correlations based on observed event numbers.

CCI [redacted]


CCI

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. Clopper C.J., Pearson E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404-13.

7. EORTC QLQ-C30 Scoring Manual (3rd edition). Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. Brussels: EORTC, 2001. ISBN: 2-9300 64-22-6. https://www.eortc.org/app/uploads/sites/2/2018/02/SCmanual.pdf

8. 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.

6 SUPPORTING DOCUMENTATION

6.1 Appendix 1: Approval Information

The sSAP Amendment 02 of Protocol MK-7902-017-03 was approved by the BARDS TA head (or designee).

Name: PPD Date: 24-JUN-2025