Official Protocol Title:	A Phase 3 study of MK-4280A (coformulated favezelimab [MK-4280] plus pembrolizumab [MK-3475]) Versus Standard of Care in Previously Treated Metastatic PDL1 positive Colorectal Cancer (KEYFORM-007)
NCT number:	NCT05600309
Document Date:	02-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 study of MK-4280A (coformulated favezelimab [MK-4280] plus

pembrolizumab [MK-3475]) Versus Standard of Care in Previously Treated Metastatic PDL1 positive Colorectal Cancer (KEYFORM-007)

Protocol Number: 007-04

Compound Number: MK-4280A

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

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	04	4.1.1 China-specific Requirements	Added a clarification for PRO analysis for the China subpopulation	To clarify that separate PRO analysis for the China subpopulation is not applicable

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 study of MK-4280A Versus Standard of Care in Previously Treated Metastatic Colorectal Cancer.
Treatment Assignment	Approximately 432 participants will be randomized in a 1:1 ratio between 2 treatment groups: (1) the MK-4280A arm (Arm A) and (2) the standard of care (regorafenib or TAS-102) arm (Arm B). Stratification factors are: Geographic region (Asia Pacific; EMEA/Americas), presence of liver metastasis (Yes, No) and time from initial diagnosis of metastatic disease to randomization (≥18 months, <18 months). This is an open-label study.
Analysis Populations	Efficacy: ITT Safety: APaT
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/Immunogenicity/ Pharmacokinetic 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 secondary hypothesis testing of ORR [Miettinen, O. and Nurminen, M. 1985].
Statistical Methods for Key Safety Analyses	For analyses in which 95% CIs will be provided for between-treatment differences in the percentage of participants with events, these analyses will be performed using the M&N method [Miettinen, O. and Nurminen, M. 1985].
Interim Analyses	**Efficacy** One interim analysis is planned in this study. Results will be reviewed by an eDMC. Details are provided in Section 3.7. • CCI [redacted] **Safety** CCI [redacted] 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), CCI [redacted] By using the graphical approach of Maurer and Bretz [Maurer, W. and Bretz, F. 2013], if one hypothesis is rejected, the alpha will be shifted to other hypotheses.
Sample Size and Power	CCI [redacted]
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; CCI [redacted]; 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.

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. 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/quality of life scores (QoL), physical functioning and appetite loss and EORTC QLQ-CR29 bloating scores for MK-4280A versus standard of care

- Time to confirmed deterioration (TTD) in EORTC QLQ-C30 global health status/QoL, physical functioning and appetite loss and EORTC QLQ-CR29 bloating scores for MK-4280A versus standard of care

  Based on prior literature (Bjordal, et al., 2000; Osoba D, 1998 [Osoba, D., et al 1998]; King, 1996 [King, M. T. 1996]), 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 with confirmation at the subsequent visit of a 10 or more points deterioration from baseline.

- CCI [redacted]

- Overall improvement / stability + improvement in EORTC QLQ-C30 global health status / QoL, physical functioning and appetite loss and EORTC QLQ-CR29 bloating scores:

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 [Cocks, K., et al 2012]).

3.5 Analysis Populations

CCI [redacted]

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 after 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 efficacy objectives. Methods related to exploratory objectives will be described in the supplemental SAP.

The stratification factors used for randomization (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 M&N method [Miettinen, O. and Nurminen, M. 1985]. CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

3.6.1.1 Overall Survival

The non-parametric Kaplan-Meier method will be used to estimate the survival curve**s**. 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. Participants without documented death at the time of analysis will be censored at the date the participant was last known to be alive.

The proportional hazards assumption on OS may be examined using both graphical and analytical methods if warranted. The log [-log] of the survival function vs. time for OS will be plotted for the comparison between MK-4280A and the control arm. If the curves are not parallel, indicating that hazards are not proportional, supportive analyses may be conducted to account for the possible non-proportional hazards effect associated with immunotherapies: for example, using the Restricted Mean Survival Time method (Finkelstein DM, 1986 [Finkelstein, D. M. 1986]).

The RMST is simply the population average of the amount of event-free survival time experienced during a fixed study follow-up time. This quantity can be estimated by the area under the Kaplan-Meier curve up to the follow-up time. The clinical relevance and feasibility should be taken into account in the choice of follow-up time to define RMST (e.g., near the last observed event time assuming that the period of clinical interest in the survival experience is the whole observed follow-up time for the study, but avoiding the very end of the tail where variability may be high); a description of the RMST as a function of the cutoff time may be of interest. The difference between two RMSTs for the two treatment groups will be estimated and 95% CI will be provided.

A sensitivity analysis may be performed based on the MaxCombo test with logrank FH (0, 1), FH (1, 1) at the final analysis of OS to account for the potential loss of power with logrank test when the proportional hazard assumption is violated.

3.6.1.2 Progression-Free Survival

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test. 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 (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, 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 subjects (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 anticancer therapy prior to documented progression will be censored at the last disease assessment prior to the initiation of new anticancer therapy. Participants who do not start new anticancer therapy and who do not experience an event will be censored at the last disease assessment. If a participant meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied.

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

The proportional hazards assumption on PFS may be examined using the same approach as OS, and supportive analyses may be conducted when the proportional hazard assumption is violated: for example, using the Restricted Mean Survival Time method (Finkelstein DM, 1986 [Finkelstein, D. M. 1986]).

3.6.1.3 Objective Response Rate

The stratified M&N method will be used for the comparison of ORR between 2 treatment groups [Miettinen, O. and Nurminen, M. 1985]. 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 the protocol) will be applied to the analysis.

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

3.6.1.4 Duration of Response

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

CCI [redacted]

Table 2 Censoring Rules for DOR

CCI [redacted]

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 Analysis			
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 M&N method	ITT	Participants with missing data are considered non-responders
Abbreviations: BICR=blinded independent central review; ITT=intent-to-treat; M&N=Miettinen & Nurminen; 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 and laboratory tests.

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 observed proportion of participants with events.

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 coformulated MK-4280A 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 [Miettinen, O. and Nurminen, M. 1985].

Membership in Tier 2 requires that at least 10% of participants in any treatment group exhibit the event; all other AEs and predefined limits of change will belong to Tier 3. The threshold of at least 10% of participants was chosen for Tier 2 events because the population enrolled in this study is in critical condition and usually experiences various AEs of similar types regardless of treatment; events reported less frequently than 10% of participants would obscure the assessment of the overall safety profile and add little to the interpretation of potentially meaningful treatment differences. In addition, Grade 3 to 5 AEs (≥5% of participants in 1 of the treatment groups) and SAEs (≥5% of participants in 1 of the treatment groups) will be considered Tier 2 endpoints. Because many 95% CIs may be provided without adjustment for multiplicity, the CIs should be regarded as a helpful descriptive measure to be used in safety review, not as a formal method for assessing the statistical significance of the between-group differences.

Tier 3 Events

Safety endpoints that are not Tier 1 or 2 events are considered Tier 3 events. 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 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, system organ class (incidence <10% of participants in all of the treatment groups)		X
	Change from Baseline Results (laboratory toxicity shift)		X

AE=adverse events; CI=confidence interval.

Frequency of AE by time period from first dose (e.g., 0-3, 3-6, 6-12 months) may also be provided. In each time interval, the denominator is the number of participants at risk for the event during the particular time period, defined as participants who are event-free up until the start of the interval.

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 [Scott, N. W., et al 2008]. 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 and [redacted] 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 4**	**Week 8**	**Week 12**	**Week 24**	**Week 36**
Day	1	29	57	85	169	253
Range (relative day to first dose date)	[-7, 7]	[8, 42]	[43, 70]	[71, 126]	[127, 210]	[211, 294]
Treatment Week	**Week 48**	**Week 60**	**Week 72**	**Week 84**	**Week 96**	
Day	337	421	505	589	673	
Range (relative day to first dose date)	[295, 378]	[379, 462]	[463, 546]	[547, 630]	[631, 714]	

3.6.3.3 Change from Baseline

The time point for the change from baseline analysis is defined as the latest time point at which CR-T $\geqslant$ 60% and CR-E $\geqslant$ 80%, and week 8 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, appetite loss, bloating, and CCI [redacted], a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [Liang, K.-Y. and Zeger, S. L. 2000] will be applied, with the PRO score as the response variable, and treatment, time, the treatment by time interaction, and the stratification factors used for randomization (Section 6.3.2 of the 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, appetite loss and bloating 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 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. 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 (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 6 provides censoring rule for TTD analysis.

Table 6 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 / 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 [Miettinen, O. and Nurminen, M. 1985] 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 (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 [Clopper, C. J. and Pearson, E. S. 1934].

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

CCI

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 (including safety reviews) 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 are 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.

3.7.1 Efficacy Interim Analyses

One IA is planned in addition to the FA for this study. For the IA and FA, all randomized participants will be included. Results of the IA 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 ~309 OS events have been observed and ~ 10 months after last participant randomized	~ 21 months	• Interim OS analysis • Final PFS and ORR analysis
FA	OS	Both ~386 OS events have been observed and ~ 12 months after IA	~ 33 months	• Final OS analysis

Abbreviations: FA=final analysis; IA=interim analysis; ORR=objective response rate; OS=overall survival; PFS=progression-free survival.
Note that for FA, if the OS events accrual is slower than expected and the final targeted OS events cannot be reached by ~33 months after the first participant randomized, the Sponsor may conduct the analysis with up to additional ~4 months of follow-up, or the specified number of events is observed, whichever occurs first.

3.7.2 Safety Interim Analyses

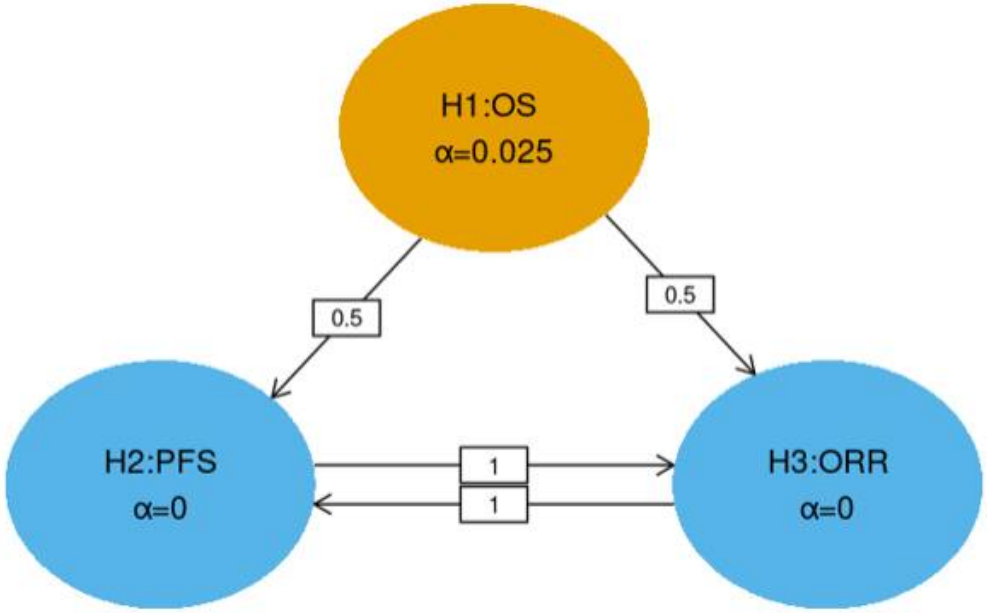
The eDMC will be responsible for periodic interim safety reviews as specified in the eDMC charter. An interim safety analysis will be performed 6 months since first participant is randomized or 2 months after 60th participant is randomized, whichever comes first. Afterwards, the eDMC will review safety data periodically in the study. Interim safety analyses will also be performed at the time of interim efficacy analyses. Details are specified in the eDMC charter.

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [Maurer, W. and Bretz, F. 2013] 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.

The initial α assigned to OS, PFS and ORR will be 0.025, 0 and 0, respectively. If OS hypothesis is rejected, the corresponding alpha can be reallocated equally to PFS and ORR. If the PFS hypothesis is rejected, the corresponding alpha can be reallocated to ORR. If the ORR hypothesis is rejected, the corresponding α can be reallocated to PFS.

Figure 1 Multiplicity Diagram for Type I Error Control



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 $\alpha = 0.0125$, separately.

3.8.1 Overall Survival

The study will test OS at IA and FA. Following the multiplicity strategy as outlined in Figure 1, the OS hypothesis will be tested at α=0.025. Table 9 shows the bounds and boundary properties for OS hypothesis testing derived using a Lan-DeMets spending function approximating O'Brien-Fleming bounds.

Table 9 Efficacy Boundaries and Properties for Overall Survival Analyses

Analysis	**Value**	**α=0.025**
IA: 80%* N: 432 Events: 309 Month: 21	Z	2.2504
	p (1-sided)[a]	0.0122
	HR at bound[b]	0.7738
	P(Cross) if HR=1[c]	0.0122
	P(Cross) if HR=0.7[d]	0.8123
FA N: 432 Events: 386 Month: 33	Z	2.0249
	p (1-sided)[a]	0.0214
	HR at bound[b]	0.8135
	P(Cross) if HR=1[c]	0.0250
	P(Cross) if HR=0.7[d]	0.9350

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 309 and 386, 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 analysis. 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.

3.8.2 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 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.

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

Analysis	**Value**	**α=0.0125**	**α=0.025**
IA N = 432 Events*: 419 Month: 21	Z	2.2414	1.9600
	p (1-sided) [a]	0.0125	0.025
	HR at bound[b]	0.8032	0.8256
	P(Cross) if HR=1[c]	0.0125	0.025
	P(Cross) if HR=0.65[d]	0.9850	0.9929

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

Note that if the OS null hypothesis is rejected, the PFS test statistics computed at IA will be used for inferential testing with its corresponding alpha levels.

3.8.3 Objective Response Rate

The study will test ORR at IA only 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 2% and 12% response rates in the control and experimental groups, respectively.

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

α	**~Δ ORR**	**Power (ΔORR=0.1)**
0.0125	0.0550	0.970
0.025	0.0481	0.984

IA=interim analysis; ORR=Objective Response Rate.

Note that if the OS null hypothesis is rejected, the ORR test statistics computed at IA will be used for inferential testing with its corresponding alpha levels.

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 432 participants in a 1:1 ratio into the MK-4280A arm (Arm A) and the standard of care arm (Arm B). OS is the primary endpoint for the study, with PFS and ORR as the key secondary endpoints.

For the OS endpoint, based on a target number of 386 events and 1 IA at approximately 80% of the target number of events, the study has approximately 93.5% power to detect a HR of 0.7 at the initially allocated α=0.025 (1-sided).

For the PFS endpoint, based on a target number of 419 events at the IA (final PFS analysis), the study has approximately 99% power to detect a HR of 0.65 at the reallocated α=0.0125 (1-sided) if only OS hypothesis is rejected.

Based on the 432 participants with at least approximately 10 months of follow-up, the power of the ORR testing at the reallocated α=0.0125 (1-sided) if only OS hypothesis rejected is approximately 97% to detect a 10-percentage point difference between an underlying 2% ORR in the control arm (Arm B) and a 12% ORR in the experimental arm (Arm A).

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

Based on CORRECT, RECOURSE, and SUNLIGHT studies (Section 4.2.1.1 of protocol), the above sample size and power calculations for OS and PFS assume the following:

- OS follows an exponential distribution with a median of 7.5 months for the control group.
- PFS follows an exponential distribution with a median of 2 months for the control group.
- Enrollment period of 12 months with enrollment ramp-up over first 2 months.
- An annual dropout rate of 2% and 5% for OS and PFS, respectively
- A follow-up period of 22 and 10 months and for OS and PFS, respectively, after the last participant is randomized.

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

3.10 Subgroup Analyses and Effect of Baseline Factors

CCI

[redacted]

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 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.

The purpose of the China extension portion is to ensure an adequate sample size to evaluate the consistency of effect for efficacy and safety between the China subpopulation and the global population. Country-specific analyses may also be conducted per local regulatory requirement.

After the enrollment for the global portion is completed, participants in China will continue to be enrolled in a 1:1 ratio into the MK-4280A arm and the standard of care arm until the sample size for the China subpopulation reaches approximately 94 in total (global and China extension portion combined).

The China extension portion will be completed at about 14 months after enrollment of China extension is completed.

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 one dose of study treatment)
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)	Separate PRO analysis for the China subpopulation is not applicable

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]
Hypotheses and Multiplicity	
Sample Size Calculations	

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]

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 94. 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.

The China extension portion will complete at about 14 months after the enrollment of China extension is completed. CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

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:

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

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).

4.3 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 scenarios where events accrue slower and 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, 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, 309 events are expected over the target 386 events at the FA.

- Hypothetical scenario 1 (events accrue slower than expected): 296 events are observed. The spending time is calculated as t = 296/386 = 77% and p-value boundary = 0.0122
- Hypothetical scenario 2 (events accrue faster than expected): 320 events are observed. The spending time is calculated as t = 309/386 = 80% and p-value boundary = 0.0122

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.

The table below summarizes the boundary properties of the hypothetical scenarios, together with the planned scenario.

Value	Planned scenario	Hypothetical scenario 1 (events accrue slower)	Hypothetical scenario 2 (events accrue faster)
IA: Month: 21			
Events (I.F.)	309 (80%)	296 (77%*)	320 (80%*)
Z	2.2504	2.2504	2.2504
p (1-sided)	0.0122	0.0122	0.0122
HR at bound	0.7738	0.7698	0.7776
P(Cross) if HR=1	0.0122	0.0122	0.0122
P(Cross) if HR=0.7	0.8123	0.7933	0.8263
FA: Month: 33			
Events	386	375	395
Z	2.0249	2.0289	2.0212
p (1-sided)	0.0214	0.0212	0.0216
HR at bound	0.8135	0.8110	0.8160
P(Cross) if HR=1	0.0250	0.0250	0.0250
P(Cross) if HR=0.7	0.9350	0.9276	0.9339
Information fraction is the minimum of the expected information fraction and the actual information fraction used for the alpha spending calculation. The actual information fraction based on the observed number of events is used for computing correlations needed to derive the group sequential bounds.			

5. REFERENCES

[Clopper, C. J. and Pearson, E. S. 1934]	Clopper CJ and Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404-13.	[03Y75Y]
[Cocks, K., et al 2012]	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.	[04SL8J]
[Finkelstein, D. M. 1986]	Finkelstein DM. A proportional hazards model for interval-censored failure time data. Biometrics 1986;42:845-54.	[03QGDL]
[King, M. T. 1996]	King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res. 1996;5:555-67.	[052KJK]
[Liang, K.-Y. and Zeger, S. L. 2000]	Liang K-Y and Zeger SL. Longitudinal data analysis of continuous and discrete responses for pre-post designs. Sankhya: The Indian Journal of Statistics 2000;62(Series B, Pt. 1):134-48.	[00QJ0M]

[Maurer, W. and Bretz, F. 2013]	Maurer W and Bretz F. Multiple testing in group sequential trials using graphical approaches. Stat Biopharm Res 2013;5(4):311-20.	[03XQVB]
[Miettinen, O. and Nurminen, M. 1985]	Miettinen O and Nurminen M. Comparative analysis of two rates. Stat Med 1985;4:213-26.	[00VMQY]
[Osoba, D., et al 1998]	Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998 Jan;16(1):139-44.	[03Q3WF]
[Scott, N. W., et al 2008]	Scott NW, Fayers PM, Aaronson NK, Bottomley A, de Graeff A, Groenvold M, et al. This manual presents reference data for the QLQ-C30 based upon data provided by EORTC Quality of Life Group Members and other users of the QLQ-C30 [Internet]. Belgium: EORTC Quality of Life Group; 2008. Available from: http://groups.eortc.be/qol/sites/default/files/img/newsletter/reference_values_manual2008.pdf.	[04HN7P]

6. SUPPORTING DOCUMENTATION

6.1 Appendix 1: Approval Information

The sSAP Amendment 03 of Protocol MK-4280A-007-04 was approved by the BARDS TA head (or designee).

Name: PPD Date: 02-JUN-2025