Official Protocol Title:	A Phase 3 Randomized, Open-Label, Study of Pembrolizumab (MK-3475) Plus Lenvatinib (E7080/MK-7902) Versus Chemotherapy for First-line Treatment of Advanced or Recurrent Endometrial Carcinoma (LEAP-001)
NCT number:	NCT04865289
Document Date:	18-JAN-2022

TABLE OF CONTENTS

LIST OF TABLES 3
LIST OF FIGURES 4
1 INTRODUCTION....5
2 SUMMARY OF CHANGES....5
3 STATISTICAL ANALYSIS PLAN 5
3.1 Statistical Analysis Plan Summary....5
3.2 Responsibility for Analyses/In-house Blinding 7
3.3 Hypotheses/Estimation 8
3.4 Analysis Endpoints....10
3.4.1 Efficacy Endpoints....10
3.4.2 Safety Endpoints 11
3.4.3 PRO Endpoints....11
3.5 Analysis Populations....12
3.5.1 Efficacy Analysis Populations....12
3.5.2 Safety Analysis Populations 12
3.5.3 PRO Analysis Population....13
3.6 Statistical Methods....13
3.6.1 Statistical Methods for Efficacy Analyses....13
3.6.1.1 Progression-free Survival (PFS) 14
3.6.1.2 Overall Survival (OS) 16
3.6.1.3 Objective Response Rate (ORR)....16
3.6.1.4 Duration of Response (DOR) 16
3.6.1.5 Progression Free Survival 2 (PFS2) 17
3.6.1.6 Analysis Strategy for Key Efficacy Variables 18
3.6.2 Statistical Methods for Safety Analyses....19
3.6.3 Statistical Methods for PRO Analyses 21
3.6.3.1 Completion and Compliance Rate Summary for PROs....21
3.6.3.2 Analyses Methods....22
3.6.3.3 Schedule for PRO Data Collection 24
3.6.3.4 Analysis Strategy for Key PRO Endpoints....24
3.6.4 Summaries of Baseline Characteristics and Demographics....24
3.7 Interim Analyses 25
3.7.1 Safety Interim Analyses....25
3.7.2 Efficacy Interim Analyses 25
3.8 Multiplicity 27
3.8.1 Efficacy Analyses 28
3.8.1.1 Progression-free Survival....28
3.8.1.2 Overall Survival 30
3.8.2 Safety Analyses....33

3.9 Sample Size and Power Calculations 33
3.10 Subgroup Analyses....................................34
3.11 Compliance (Medication Adherence)....................................34
3.12 Extent of Exposure....................................34
4 STATISTICAL ANALYSIS PLAN FOR CHINA EXTENSION 35
4.1 Introduction....................................35
4.2 Responsibility for Analyses/In-House Blinding 36
4.3 Hypotheses/Estimation 36
4.4 The Analysis Endpoints....................................36
4.4.1 Efficacy Endpoints....................................36
4.4.2 Safety Endpoints 36
4.5 Analysis Population 36
4.5.1 Efficacy Analysis Populations....................................36
4.5.2 Safety Analysis Populations 36
4.6 Statistical Methods....................................36
4.6.1 Statistical Methods for Efficacy Analyses....................................37
4.6.1.1 Progression-Free Survival (PFS) 37
4.6.1.2 Overall Survival (OS) 37
4.6.1.3 Objective Response Rate (ORR)....................................37
4.6.2 Statistical Methods for Safety Analyses....................................37
4.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses 38
4.7 Interim Analysis and Final analysis....................................38
4.8 Multiplicity 38
4.9 Sample Size and Power Calculations 38
4.10 Subgroup Analysis....................................39
5 REFERENCES....................................40

LIST OF TABLES

Table 1 Censoring Rules for Primary and Sensitivity Analyses of Progression-free Survival 15
Table 2 Censoring Rules for DOR 17
Table 3 PFS2 – Events and Censoring Rules 18
Table 4 Efficacy Analysis Methods for Key Efficacy Endpoints 18
Table 5 Analysis Strategy for Safety Parameters 20
Table 6 PRO Data Collection Schedule and Mapping of Study Visit to Analysis Visit 24
Table 7 Analysis Strategy for Key PRO Endpoints 24
Table 8 Summary of Interim and Final Analyses 26
Table 9 Efficacy Boundaries and Properties for PFS Analyses in the pMMR Participants 29
Table 10 Efficacy Boundaries and Properties for PFS Analyses in the All-comer Participants 30
Table 11 Efficacy Boundaries and Properties for OS Analysis for pMMR Participants 31
Table 12 Efficacy Boundaries and Properties for OS Analysis for All-comer Participants 32

LIST OF FIGURES

Figure 1 Multiplicity Graph for Type I Error Control of Study Hypotheses 28

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

The changes made to the sSAP were not directly related to changes required due to a protocol amendment (MK-7902-001-05). Changes from sSAP Amend02 are summarized below:

- Update section 3.10 subgroup analysis.

3 STATISTICAL ANALYSIS PLAN

This section outlines the statistical analysis strategy and procedures for the study. If, after the study has begun, changes are made to primary and/or key secondary hypotheses, or the statistical methods related to those hypotheses, then the protocol will be amended (consistent with ICH Guideline E-9). Changes to exploratory or other non-confirmatory analyses made after the protocol has been finalized, but prior to the conduct of any analysis, will be documented in an sSAP and referenced in the CSR for the study. A separate PK analysis plan (if PK is analyzed) as well as biomarker analysis plan will be provided. Post hoc exploratory analyses will be clearly identified in the CSR. The PRO analysis plan will also be included in the sSAP.

The extension portion of the study, which includes the additional enrollment of China participants, will not be included as part of the study population used to address the protocol objectives. Additional detail regarding the analysis associated with these participants will be described in Section 4.

3.1 Statistical Analysis Plan Summary

Key elements of the SAP are summarized here. The comprehensive plan is provided in Sections 3.2 through 3.12.

Study Design Overview	A Phase 3 Randomized, Open-Label, Study of Pembrolizumab (MK-3475) Plus Lenvatinib (E7080/MK-7902) Versus Chemotherapy for First-line Treatment of Advanced or Recurrent Endometrial Carcinoma (LEAP-001)
Treatment Assignment	Approximately 875 eligible participants (612 pMMR participants and approximately 263 dMMR participants) will be randomized to one of the following 2 treatment arms in a 1:1 ratio: • Arm 1: pembrolizumab plus lenvatinib • Arm 2: paclitaxel (or docetaxel for participants who experience either a severe hypersensitivity reaction to paclitaxel or an AE requiring discontinuation of paclitaxel, see Protocol Section 4.3.3.1) and carboplatin Randomization stratification factors include: • MMR status (pMMR versus dMMR), and if pMMR: • ECOG (0 versus 1) • measurable disease (yes versus no) • prior chemotherapy and/or chemoradiation (yes versus no) Eligible participants will first be stratified by MMR status (pMMR versus dMMR), then only within the pMMR stratum, participants will be further stratified according to ECOG (0 versus 1), measurable disease (yes versus no) and prior chemotherapy and/or chemoradiation (yes versus no). A total of 9 strata will be utilized for the study (see Section 3.6).
Analysis Populations	• Efficacy: Intent to Treat (ITT) • Safety: All Participants as Treated (APaT) • ePRO: PRO Full Analysis Set (PRO FAS)
Primary Endpoints	• Progression-free survival (PFS) based on RECIST 1.1 as assessed by BICR, modified to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ • Overall survival (OS)
Secondary Endpoints	• Objective response rate (ORR) by BICR using RECIST 1.1 • Patient reported quality of life assessed by change from baseline of EORTC QLQ-C30 • Safety and tolerability of the two treatment arms
Statistical Methods for Key Efficacy Analyses	The primary hypotheses will be evaluated by comparing PFS and OS using a stratified Log-rank test. The hazard ratio will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method.
Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. The tiers differ with respect to the analyses that will be performed. There are no events of interest that warrant elevation to Tier 1 events in this study. Tier 2 parameters will be assessed via point estimates with 95% CIs provided for between-group comparison; only point estimates by treatment group are provided for Tier 3 safety parameters. The 95% CIs for the between-treatment differences in percentages will be provided using the Miettinen and Nurminen method [Miettinen, O. 1985].

Interim Analyses	Three interim analyses (IAs) and one final analysis (FA) are planned in this study. Comparisons between two treatment arms will be conducted at the IAs and final analysis. Results of the IAs will be reviewed by the eDMC. Details are provided in Section 3.7 – Interim Analyses. • IA1: PFS and OS analysis around 27 months after the first participant is randomized. The analysis will be triggered when ~354 PFS events for pMMR participants have been observed and ~3 months after last participant randomized. • IA2: PFS and OS analysis around 36 months after the first participant is randomized. The analysis will be triggered when ~472 PFS events for pMMR participants have been observed and ~12 months after last participant randomized. • IA3: OS analysis around 42 months after the first participant is randomized. The analysis will be triggered when approximately ~316 OS events for pMMR participants have been observed and ~18 months after last participant randomized. • FA: Final OS analysis around 48 months after the first participant is randomized. The analysis will be triggered when ~359 OS events for pMMR participants have been observed and ~24 months after last participant randomized.
Multiplicity	The overall Type I error rate over the multiple endpoints will be strongly controlled at 2.5% (one-sided). A total of 0.5% Type I error rate is initially allocated to test PFS superiority between two treatment arms for pMMR participants and a total of 2.0% Type I error rate is initially allocated to test OS non-inferiority between two treatment arms for pMMR participants. The graphical approach of Maurer and Bretz [Maurer, W. 2013] will be applied to re-allocate alpha among the hypotheses of PFS and OS. The study will be considered positive if it is positive for one of two hypothesis tests (H1 and H3) outlined in Section 3.3.
Sample Size and Power	The planned sample size is approximately 875 participants (612 pMMR participants and approximately 263 dMMR participants). For pMMR participants (N = 612), with approximately 354 and 472 pooled PFS events at the planned PFS analyses, the study will have 90% power to detect a hazard ratio of 0.7 at the one-sided 0.005 significance level. With 180, 269, 316, and 359 pooled OS events at the planned OS interim and final analyses, the study will have 90% power to detect a hazard ratio of 0.7 at the one-sided 0.02 significance level. The study will have 82% power to establish the non-inferiority with the assumption of hazard ratio of 0.8 and NI margin of 1.1 for OS at one-sided 0.02 significance level for pMMR participants. Details are provided in Section 3.9 Sample Size and Power Calculations.

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 for study intervention assignment for this protocol and the randomization will be implemented in IVRS/IWRS.

Although the study is open label, analyses or summaries generated by randomized treatment assignment, or actual treatment received status will be limited and documented. In addition, the independent radiologist(s) will perform the central imaging review without knowledge of treatment group assignment.

Blinding issues related to the planned interim analyses are described in Protocol Section 9.7.

3.3 Hypotheses/Estimation

In women with Stage III, Stage IV, or recurrent endometrial carcinoma who have been treated with pembrolizumab plus lenvatinib versus chemotherapy:

Objectives	Endpoints
Primary	
• Objective: To compare progression-free survival (PFS) per Response Evaluation Criteria in Solid Tumors version 1.1 (RECIST 1.1) by blinded independent central review (BICR), modified to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ (see Protocol Section 8.2.1). Hypothesis 1 (H1): The combination of pembrolizumab plus lenvatinib is superior to chemotherapy with respect to PFS per RECIST 1.1 by BICR in mismatch repair proficient (pMMR) participants. H2: The combination of pembrolizumab plus lenvatinib is superior to chemotherapy with respect to PFS per RECIST 1.1 by BICR in all-comers.	• PFS is defined as the time from randomization to first documented disease progression or death due to any cause, whichever occurs first.
• Objective: To compare overall survival (OS) H3: The combination of pembrolizumab plus lenvatinib is non-inferior to chemotherapy with respect to OS in pMMR participants. H4: The combination of pembrolizumab plus lenvatinib is superior to chemotherapy with respect to OS in pMMR participants. H5: The combination of pembrolizumab plus lenvatinib is superior to chemotherapy with respect to OS in all-comers.	• OS is defined as the time from randomization to death due to any cause.
Secondary	
• Objective: To compare objective response rate (ORR) per RECIST 1.1 by BICR in pMMR participants and in all-comer participants who have measurable disease at study entry.	• Objective Response (OR) is defined as a confirmed complete response (CR) or partial response (PR).

Objectives	Endpoints
• Objective: To evaluate the impact of treatment on Health-Related Quality-of-Life (HRQoL) as assessed by using the global score of the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core-30 (QLQ-C30) in pMMR and in all-comer participants.	• Mean change from baseline in EORTC QLQ-C30 global health status/quality of life score.
• Objective: To compare the safety and tolerability of pembrolizumab plus lenvatinib versus chemotherapy in all-comer participants.	• AEs, SAEs and irAEs. • Study intervention discontinuations due to AEs.
Exploratory	
• Objective: To assess duration of response (DOR), disease control rate (DCR), and clinical benefit rate (CBR) per RECIST 1.1 by BICR in pMMR and in all-comer participants.	• DOR is defined as the time from the date response was first documented until the date of first documented disease progression or date of death, whichever occurs first. • Disease Control (DC) is defined as the best overall response of CR, PR, or SD. Stable disease (SD) must be achieved at ≥7 weeks after randomization to be considered best overall response. • Clinical Benefit (CB), defined as best overall response of CR, PR, or SD (duration of SD ≥23 weeks after randomization).
• Objective: To assess efficacy outcomes using RECIST by investigator in pMMR and in all-comer participants.	• PFS, OR, DOR, DC, and CB
• Objective: To assess efficacy outcomes using iRECIST by investigator in pMMR and in all-comer participants.	• PFS, OR, DOR, DC, and CB
• Objective: To evaluate the impact of treatment on HRQoL as assessed by using the EORTC QLQ-C30, EORTC QLQ-Endometrial Cancer Module (EN24) and EuroQoL 5-dimension, 5-level Questionnaire (EQ-5D-5L) instruments in pMMR participants and in all-comer participants.	• HRQoL will be assessed using the EORTC QLQ-C30 (scores other than global score), EORTC QLQ-EN24, and EuroQoL EQ-5D-5L.
• Objective: To assess PFS2 by investigator assessment in pMMR and in all-comer participants.	• PFS2 is defined as the time from randomization to disease progression, as determined by investigator assessment, on next-line of treatment or death, whichever occurs first.

Objectives	Endpoints
• Objective: To identify molecular (genomic, metabolic, and/or proteomic) biomarkers that may be indicative of clinical response/resistance, safety, and/or the mechanism of action of pembrolizumab and lenvatinib in all participants.	• Molecular (genomic, metabolic, and/or proteomic) determinants of response or resistance to treatments, using blood and/or tumor tissue.
• Objective: To characterize the population pharmacokinetics (PK) of lenvatinib when co-administered with pembrolizumab in pMMR participants and in all-comer participants	• Plasma concentration of lenvatinib versus time.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Primary

PFS is defined as the time from randomization to the first documented disease progression per RECIST 1.1 as assessed by BICR, modified to follow a maximum of 10 target lesions and a maximum of 5 target lesions per organ or death due to any cause, whichever occurs first.

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

Secondary

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

Exploratory

DOR is defined as the time from the date a response was first documented until the date of the first documentation of disease progression, by BICR of objective radiographic disease assessment per RECIST 1.1, or date of death (whichever occurs first).

DCR is the proportion of participants who have best overall response of CR or PR or SD by BICR per RECIST 1.1. Stable disease must be achieved at ≥7 weeks after randomization to be considered best overall response.

CBR is the proportion of participants who have best overall response of CR, PR, or durable SD (duration of SD ≥23 weeks after randomization) by BICR per RECIST 1.1.

PFS, ORR, DOR, DCR, and CBR as determined by investigator assessment using RECIST will be used for exploratory analyses.

PFS, ORR, DOR, DCR, and CBR as determined by investigator assessment using iRECIST will be used for exploratory analyses. PFS per iRECIST is defined as specified for the respective endpoints using RECIST 1.1 above, with the exception that: 1) a confirmation assessment of PD (at least 4 weeks after the initial PD assessment) is required for participants who remain on study intervention following a documented PD per RECIST 1.1, and 2) responses of iSD, iPR, and iCR are permitted as a best overall response after iUPD. Participants who discontinue study intervention following a documented PD assessment per RECIST 1.1 will be counted as having disease progression on the date of the documented PD assessment. Responses will be based on Investigator assessment.

PFS2 is defined as the time from randomization to disease progression on next-line of treatment, or death (whichever occurs first).

3.4.2 Safety Endpoints

Safety will be assessed by the number of AEs, SAEs, and irAEs and the number of participants discontinuing study intervention due to AEs.

3.4.3 PRO Endpoints

Unless otherwise specified below, the primary timepoint for assessment for the analyses of PRO is the latest timepoint at which PRO data for both groups was collected, and the overall completion is at least 60% and Week 18 was selected based on blinded data review prior to the database lock for any PRO analysis.

The mean score change from baseline to the primary timepoint based on blinded data review as measured by the EORTC QLQ-C30 global health status/quality of life (GHS/QoL) scale.

- The mean score change from baseline to the primary timepoint based on blinded data review for the functional and symptom scales other than the EORTC QLQ-C30 GHS/QoL .
- The mean score change from baseline to the primary timepoint based on blinded data review for the functional and symptom scales of the EORTC QLQ-EN24.
- The mean score change from baseline to the primary timepoint based on blinded data review for EQ-5D VAS.
- Empirical mean change from baseline in scores over time for the following scales:
    - EORTC QLQ-C30 GHS/QoL
    - EORTC QLQ-C30 Physical Functioning
    - EORTC QLQ-C30 Role Functioning
    - EORTC EN24 Urological Symptoms
    - EQ-5D-5L VAS

- Overall improvement/stability rate for the following scales
  - EORTC QLQ-C30 GHS/QoL
  - EORTC QLQ-C30 Physical Functioning
  - EORTC QLQ-C30 Role Functioning
  - EORTC EN24 Urological Symptoms

Improvement is defined as a 10-point or more increase in score (in the positive direction) from baseline at any time during the study and confirmed by a 10-point or more improvement at the next consecutive visit. Stability is defined as, when the criteria for improvement are not met, a less than 10 points worsening in score from baseline at any time during the study and confirmed by a less than 10 points worsening at the next consecutive visit. Overall improvement/stability is defined as the composite of improvement and stability.

3.5 Analysis Populations

Extension Portion of the Study in China

After the sample size required for the Global Cohort is reached, the study will continue to randomize participants in China until the sample size for the China participants meets the target for China. The China participants randomized after the enrollment of the Global Cohort is closed will not be included in the primary analysis population which is based on the Global Cohort. The China Cohort will also be analyzed separately per local regulatory requirement.

3.5.1 Efficacy Analysis Populations

The Intention-to-Treat (ITT) population will serve as the population for the primary efficacy analyses. All randomized participants will be included in this population. Participants will be analyzed in the treatment arm to which they are randomized.

ORR will be analyzed in participants who have measurable disease at study entry. ORR may also be analyzed in ITT population as sensitivity analyses.

3.5.2 Safety Analysis Populations

The All Participants as Treated (APaT) population will be used for the analysis of safety data in this study. The APaT population consists of all randomized/allocated participants who received at least one dose of study intervention. Participants will be included in the treatment group corresponding to the study intervention they actually received for the analysis of safety data using the APaT population. This will be the treatment group to which they are randomized except for participants who take incorrect study intervention for the entire treatment period; such participants will be included in the treatment group corresponding to the study intervention actually received.

At least one laboratory, vital sign, or ECG measurement obtained subsequent to at least one dose of study intervention is required for inclusion in the analysis of the respective safety parameter. To assess change from baseline, a baseline measurement is also required.

Details on the approach to handling safety analyses are provided in Section 3.6.2.

3.5.3 PRO Analysis Population

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

3.6 Statistical Methods

Statistical methods for efficacy analyses are described in Section 3.6.1. Efficacy results that will be deemed to be statistically significant after consideration of the Type I error control strategy as 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 etc. Statistical methods for safety analyses and PRO analyses are described in Section 3.6.2 and 3.6.3, respectively.

The stratification factors used for randomization

- MMR status (pMMR versus dMMR), and if pMMR further stratify by
    - ECOG (0 versus 1)
    - Measurable disease (yes versus no)
    - Prior chemotherapy and/or chemoradiation (yes versus no)

(See protocol Section 6.3.2 for additional detail) will be applied to all stratified analyses, in particular, the stratified log-rank test, stratified Cox model for efficacy analyses, and the stratified Miettinen and Nurminen method [Miettinen, O. 1985]. Based on a blinded review of the number of randomized participants by stratum prior to the first efficacy interim analysis, the total number of randomized participants in no measurable disease groups is ~5% of the ITT population, and hence, stratification factors will be combined for analysis to ensure sufficient number of participants and/or events in each stratum, by eliminating measurable disease from analysis stratification and leaving stratification by MMR status (pMMR versus dMMR), and if pMMR further stratify by

- ECOG (0 versus 1)
- Prior chemotherapy and/or chemoradiation (yes versus no).

3.6.1 Statistical Methods for Efficacy Analyses

This section describes the statistical methods that address the primary, secondary and exploratory objectives as stated in the protocol.

3.6.1.1 Progression-free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, the HR) between the treatment arms. The HR and its 95% CI from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. The stratification factors used for randomization (See Section 3.6) will be applied to both the stratified log-rank test and the stratified Cox model.

Since disease progression is assessed periodically, progressive disease (PD) can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. 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 as a PD event.

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

In order to evaluate the robustness of the PFS endpoint per RECIST 1.1 based on BICR, two 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 primary and sensitivity analyses are summarized in
[Table 1].

Table 1 Censoring Rules for Primary and Sensitivity Analyses of Progression-free Survival

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
PD or death documented after ≤1 missed disease assessment, and before new anticancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death	Progressed at date of documented PD or death
PD or death documented immediately after ≥2 consecutive missed disease assessments or after new anticancer therapy, if any	Censored at last disease assessment prior to the earlier date of ≥2 consecutive missed disease assessment and new anticancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death
No PD and no death; and new anticancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment	Progressed at treatment discontinuation due to reasons other than complete response; otherwise censored at last disease assessment if still on study or completed study intervention.
No PD and no death; new anticancer treatment is initiated	Censored at last disease assessment before new anticancer treatment	Censored at last disease assessment	Progressed at date of new anticancer treatment

PFS by investigator using RECIST and PFS by investigator using iRECIST, will be analyzed using the approach specified for the PFS by BIRC above. Results based on the primary censoring rules for PFS summarized in [Table 1] above will be provided.

The proportional hazards assumption on PFS 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 lenvatinib plus pembrolizumab and the paclitaxel and carboplatin 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 by Uno et al. (2014) [Uno, H., et al 2014], parametric method by Anderson (1991) [Anderson, K. M. 1991], etc.

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.

3.6.1.2 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (ie, the HR). The HR and its 95% CI from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization (see Section 3.6) 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. The Restricted Mean Survival Time (RMST) method may be conducted for OS to account for the possible non-proportional hazards effect.

The non-inferiority hypothesis of OS will be evaluated using a stratified Cox regression model.

The same method as described in Section 3.6.1.1 for PFS to exam the proportional hazards assumption may be applied to OS.

Adjustment for the effect of crossover on OS may be performed based on recognized methods, e.g. the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis (1989) [Robins, J. M. and Tsiatis, A. A. 1991], two stage model proposed by Latimer et al. (2019) [Latimer, N. R., et al 2019], etc., based on an examination of the appropriateness of the data to the assumptions required by the methods.

3.6.1.3 Objective Response Rate (ORR)

Stratified Miettinen and Nurminen’s method will be used for comparison of the objective response rate (ORR) between two treatment groups [Miettinen, O. 1985]. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen’s method with strata weighting by sample size will be reported. The stratification factors used for randomization (See Section 3.6) 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) [Clopper, C. J. 1934].

ORR by investigator review, DCR and CBR by BICR per RECIST 1.1 and ORR, DCR, and CBR as determined by investigator assessment using iRECIST will be evaluated using the approach specified for the endpoint ORR above.

3.6.1.4 Duration of Response (DOR)

If sample size permits, DOR by BICR per RECIST 1.1 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. Censoring rules for DOR are summarized in [Table 2].

For each DOR analysis, a corresponding summary of the reasons for censoring subjects from the analysis 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.

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.		

DOR by investigator using iRECIST will be analyzed using the approach specified for the DOR by BIRC above.

3.6.1.5 Progression Free Survival 2 (PFS2)

An analysis of PFS2, defined as the time from randomization to subsequent disease progression after initiation of new anticancer therapy, or death from any cause, whichever first, will be carried out [Table 3]. Participants alive and for whom a disease progression following initiation of new anticancer treatment has not been observed will be censored at the last time the participant was known to be alive and without disease progression. The same stratified Cox proportional hazard model as the primary analyses of PFS will be used to estimate the HR and its 95% CI.

Table 3 PFS2 – Events and Censoring Rules

Situation	Analyses
1) First next-line therapy started, progression per investigator after first next-line start	Event on date of progression per investigator
2) If not 1), and date of progression on first next-line therapy before datacut	Event on date of progression on first next-line therapy
3) if not 1) and 2), date of progression on first next-line therapy after datacut or second next-line therapy not started, and death occurred	Event on date of death
4) If not 1) and 2), date of progression on first next-line therapy after datacut or second next-line therapy not started, and no death	Censor at uncut last known alive date
5) If not 1) to 4), and second next-line therapy started	Censor on date of start of second next-line therapy

3.6.1.6 Analysis Strategy for Key Efficacy Variables

A summary of the primary analysis strategy for the key efficacy endpoints is provided in [Table 4].

Table 4 Efficacy Analysis Methods for Key Efficacy Endpoints

Endpoint/Variable	Statistical Method	Analysis Population	Missing Data Approach
Primary Analyses:			
PFS (RECIST 1.1) by BICR	Testing: Stratified Log-rank Test Estimation: Stratified Cox model with Efron's tie handling method.	ITT	Censored according to rules in [Table 1].
OS	Testing: Stratified Log-rank Test Estimation: Stratified Cox model with Efron's tie handling method.	ITT	Censored at last known alive date.
Secondary Analyses:			
ORR (RECIST 1.1) by BICR	Testing: Stratified Miettinen and Nurminen method [Miettinen, O. 1985].	ITT with measurable disease at baseline	Participants with missing data are considered non-responders.
Sensitivity analyses will be performed for PFS, and ORR based on investigator's assessment.			

3.6.2 Statistical Methods for Safety Analyses

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

The analysis of safety results will follow a tiered approach [Table 5]. The tiers differ with respect to the analyses that will be performed. AEs (specific terms as well as system organ class terms) and events that meet predefined limits of change (PDLCs) in laboratory values, vital signs, and ECG parameters are either prespecified as “Tier 1” endpoints, or will be classified as belonging to "Tier 2" or "Tier 3" based on the observed proportions of participants with an event.

Tier 1 Events

Safety parameters or adverse events of special interest (AEOSIs) that are identified *a priori* constitute “Tier 1” safety endpoints that will be subject to inferential testing for statistical significance. AEOSIs 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. Further, the combination of pembrolizumab plus lenvatinib included in this study has not been found to impact safety. Additionally, there are no known AEs associated with participants for which determination of a p-value is expected to impact the safety assessment. Therefore, there are no Tier 1 events in this study.

Tier 2 Events

Tier 2 parameters will be assessed via point estimates with 95% confidence intervals provided for differences in the proportion of participants with events (via the Miettinen and Nurminen method [Miettinen, O. 1985]).

Membership in Tier 2 requires that at least 10% 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 differences. In addition, Grade 3 to 5 AEs (≥5% of participants in 1 of the treatment groups) and SAEs (≥1% of participants in 1 of the treatment groups) will be considered Tier 2 endpoints. Because many 95% confidence intervals for Tier 2 events may be provided without adjustment for multiplicity, the confidence intervals should be regarded as a helpful descriptive measure to be used in review, not a formal method for assessing the statistical significance of the between-group differences.

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 to 5 AE, a drug-related Grade 3 to 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, vital signs, and ECG parameters, summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group in table format.

Table 5 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 ≥1% 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
	Any AE leading to 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 event; CI=confidence interval; SOC=system organ class.			

Exposure-adjusted rate 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 person years of exposure is calculated based on the number of participants at risk for the event during the particular time period, where at risk is defined as participants who are exposed to drug at the start of indicated time interval. The numerator is the number of events occurring in the interval.

In addition, to properly account for the potential difference in follow-up time between the study arms, which is expected to be longer in the lenvatinib plus pembrolizumab arm, AE incidence adjusted for treatment exposure analyses will be performed. For exposure adjusted analyses, events count as the numerator, and person-months of exposure as the denominator.

3.6.3 Statistical Methods for PRO Analyses

This section describes the planned analyses for the PRO endpoints.

The patient-reported outcomes are secondary and exploratory objectives in the trial. No formal hypotheses were formulated. Nominal p-value to compare lenvatinib in combination with pembrolizumab versus paclitaxel in combination with carboplatin in pMMR participants and in all-comer participants may be provided as appropriate. The PRO instruments are EORTC QLQ-C30 (global health status, physical function and role function), EORTC QLQ-EN24 (Urological scale) and EuroQol-5D (EQ-5D VAS).

3.6.3.1 Completion and Compliance Rate Summary for PROs

Completion and compliance of EORTC QLQ-C30, EORTC QLQ-EN24 and EQ-5D by visit and by treatment will be described. Numbers and percentages of complete and missing data at each visit will be summarize. An instrument is considered complete if at least one valid score is available according to the missing item rules outlined in the scoring manual for the instrument.

Completion rate of treated participants (CR-T) at a specific time point is defined as the percentage of the number of treated participants who complete at least one item over the number of treated participants in the PRO population.

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

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

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

The reasons of non-completion and non-compliance will be provided in supplementary table:

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

In addition, reasons for non-completion as scheduled of these measures will be collected using “miss_mode” forms filled by site personnel and will be summarized in table format. The schedule (study visits and estimated study times) and mapping of study visit to analysis visit for PRO data collection is provided in [Table 6].

3.6.3.2 Analyses Methods

Mean Score Change from Baseline

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

To assess the treatment effects on the PRO score change from baseline in the EORTC QLQ-C30 global health status scale, physical functioning scale; EORTC QLQ-EN24 Urological symptoms scale; and EQ-5D-5L VAS at prespecified time points, a constrained longitudinal data analysis (cLDA) model proposed by Liang and Zeger [Liang, Kung-Yee and Zeger, Scott L. 2000] will be applied, with the PRO score as the response variable, and treatment, time, the treatment by time interaction and stratification factors used for randomization (See Section 3.6) as covariates.

The treatment difference in terms of least square (LS) mean change from baseline will be estimated from this model together with 95% CI. Model-based LS mean with 95% CI will be provided by treatment group for PRO scores at baseline and post-baseline time point. These timepoints are consistent with the information included in the [Table 6] below.

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:

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

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).

Empirical Mean Score Change from Baseline

Line plots for the empirical mean change from baseline in EORTC QLQ-C30 global health status scale, physical functioning scale and role functioning scale; EORTC QLQ-EN24 Urological symptom scale; and EQ-5D-5L VAS will be provided across all time points (as indicated in below) up to the final assessment timepoint as a supportive analysis.

In addition, the model-based LS mean change from baseline to the specified post-baseline time point together with 95% CI will be plotted in bar charts for EORTC QLQ-C30 global health status/quality of life scores, functional and symptom scales; and EORTC QLQ-EN24 functional and symptom scales.

Overall Improvement and Overall Improvement/Stability

Overall improvement rate will be analyzed, which is defined as the proportion of participants who have achieved an improvement as defined in Section 3.4.3 PRO Endpoints. Stratified Miettinen and Nurminen’s method will be used for comparison of the overall improvement rate between the treatment groups. The difference in overall improvement rate and its 95% CI from the stratified Miettinen and Nurminen’s method with strata weighting by sample size will be provided. The stratification factors used for randomization (See Section 3.6) 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) [Clopper, C. J. 1934].

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

3.6.3.3 Schedule for PRO Data Collection

Table 6 PRO Data Collection Schedule and Mapping of Study Visit to Analysis Visit

Visit	C1	C2	C3 to C34 (Every cycle)	C35	C39 to C60 (Every 4 cycles)	EOT Discontinuation Visit	30-day Safety Follow-up Visit	90-day Safety Follow-up Visit
Calendar	W0 (Baseline)	W3	W6 to W99 (Every 3 weeks)	W102	W114 to W177 (Every 12 weeks)	X	X	X
Day	1	22	Week number*7+1	715	Week number*7+1	X	X	X
Range	≤1	2 to 32	Week number*7-9 to Week number*7+11	705 to 756	Week number*7-41 to Week number*7+42	X	X	X
C: Cycle; W: Week. Every cycle is 3 weeks.								

3.6.3.4 Analysis Strategy for Key PRO Endpoints

Table 7 Analysis Strategy for Key PRO Endpoints

Endpoints	Statistical Method	Missing Data Approach
Change from baseline at primary timepoint in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning and in EORTC QLQ-EN24 Urological symptoms scale and in EQ-5D VAS	cLDA model	Model-based
Overall improvement and overall improvement/stability in EORTC QLQ-C30 • Global health status/QoL • Physical functioning • Role functioning and in EORTC QLQ-EN24 Urological symptoms scale	Stratified Miettinen and Nurminen method	Participants with missing data are considered not achieving improvement/stability.
Abbreviations: cLDA = constrained longitudinal data analysis, FAS = full analysis set, QoL = quality of life.		

3.6.4 Summaries of Baseline Characteristics and Demographics

The comparability of the treatment groups for each relevant characteristic will be assessed by the use of tables and/or graphs. No statistical hypothesis tests will be performed on these characteristics. The number and percentage of participants randomized, and the primary reason for discontinuation will be displayed. Demographic variables (such as age) and baseline characteristics will be summarized by treatment either by descriptive statistics or categorical tables. The reasons for exclusion from the ITT population (if any) will be summarized.

3.7 Interim Analyses

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

An eDMC will serve as the primary reviewer of the results of the interim analyses of the study and will make recommendations for discontinuation of the study or protocol modifications to an 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 results at the treatment level in order to act on these recommendations. The extent to which individuals are unblinded with respect to results of interim analyses will be documented by the unblinded statistician. Additional logistical details will be provided in the eDMC Charter. Key aspects of the interim analyses are described in Section 3.7.2.

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

3.7.1 Safety Interim Analyses

The eDMC will conduct regular safety interim analyses. The timing of these safety interim analyses will be specified in the eDMC charter.

3.7.2 Efficacy Interim Analyses

Three interim analyses are planned in addition to the final analysis for this study. Results of the interim analyses will be reviewed by the eDMC. There is no expectation to stop the study before superiority hypotheses for OS have been adequately evaluated. However, earlier positive findings may form the basis for earlier regulatory submission based on the recommendation of the eDMC. Details on how the above planned analyses are incorporated into establishing statistical significance and the boundaries with regard to efficacy are discussed further in Section 3.8.

The analyses planned, endpoints evaluated, and drivers of the timing are summarized in [Table 8]. We would like to use AND to ensure enough FU time for patients' PFS for data maturity.

Table 8 Summary of Interim and Final Analyses

Analyses	Key Endpoints	Events required for the analysis	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
IA1	PFS OS	Timing of analysis will be triggered when ~354 PFS events for pMMR participants have been observed and ~3 months after last participant randomized. It is estimated that ~180 OS events for pMMR participants will be observed at the same time.	~27 months	• Interim PFS analysis • Interim OS analysis
IA2	PFS OS	Timing of analysis will be triggered when ~ 472 PFS events for pMMR participants have been observed and ~12 months after last participant randomized. It is estimated that ~269 OS events for pMMR participants will be observed at the same time.	~36 months	• Demonstrate PFS superiority • Interim OS analysis
IA3	OS	Timing of analysis will be triggered when ~316 OS events for pMMR participants have been observed and ~18 months after last participant randomized.	~42 months	• Interim OS analysis

Analyses	Key Endpoints	Events required for the analysis	Estimated Time after First Participant Randomized	Primary Purpose of Analysis
Final Analysis	OS	Timing of analysis will be triggered when ~359 OS events for pMMR participants have been observed and ~24 months after last participant randomized.	~48 months	• Demonstrate OS non-inferiority/superiority

Note: For IA1 and IA2, if the PFS events accrue slower than expected such that the targeted number of events cannot be reached in the anticipated timeframe, the Sponsor may conduct the analysis with approximately 2 additional months of follow-up, or when the specified number of events is observed, whichever occurs first.

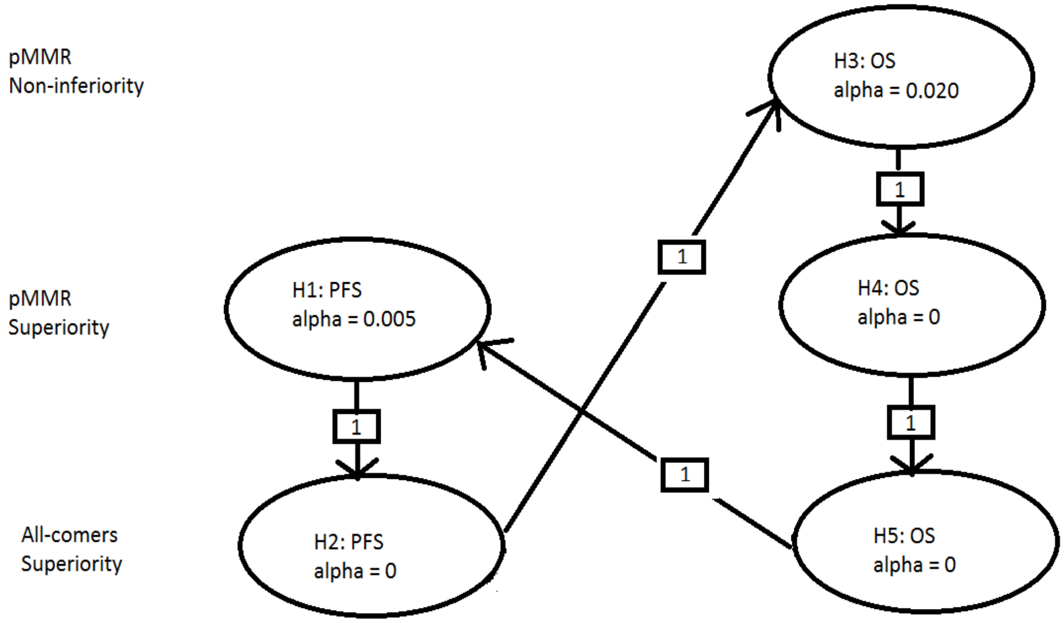
IA3 may be skipped if the actual OS event number at IA2 is close to the prespecified event number for IA3(if more than ~83% info fraction, ie. ~298 OS events).

- For IA3 and FA, if the OS events accrue slower than expected such that the targeted number of events cannot be reached in the anticipated timeframe, the Sponsor may conduct the analysis with approximately 4 additional months of follow-up.
- For IA3 (if not skipped) and FA, if planed analysis is too close to actual preceding IA, sponsor may conduct the analysis ~6 months after preceding IA for operational consideration.
- IA2 is PFS final analysis. All alpha allocated/passed to PFS would be used up at IA2 (ie. the second efficacy analysis), regardless how many PFS events per BICR are observed in IA2.
- All alpha allocated/passed to OS would be used up at the planned FA (ie. the fourth efficacy analysis if the third IA is not skipped), regardless how many OS events observed.

3.8 Multiplicity

The study uses the graphical method of Maurer and Bretz [Maurer, W. 2013] to control multiplicity for multiple hypotheses as well as interim analyses. [Figure 1] shows the initial 1-sided α allocation for each hypothesis in the ellipse representing the hypothesis. The weights for reallocation from each hypothesis to the others are represented in the boxes on the lines connecting hypotheses.

Figure 1 Multiplicity Graph for Type I Error Control of Study Hypotheses



3.8.1 Efficacy Analyses

3.8.1.1 Progression-free Survival

The study initially allocates one-sided α = 0.005 to test PFS between two treatment groups for pMMR participants. If the null hypothesis of PFS for pMMR participants is rejected, its α = 0.005 will be reallocated to the PFS analysis for all-comers. [Table 9] below shows the bounds and boundary properties for PFS testing which were derived using a Lan-DeMets O'Brien-Fleming α-spending function. If the actual number of events at the PFS analyses differ from those specified in the table, the bounds will be adjusted using the Lan-DeMets O'Brien-Fleming spending function accordingly.

[Table 10] below shows the bounds and boundary properties for PFS testing in all-comers with α = 0.005 (if H1 is rejected), assuming 612 pMMR participants and 192 dMMR participants.

Table 9 Efficacy Boundaries and Properties for PFS Analyses in the pMMR Participants

Analysis[†]	Value	α = 0.005	α = 0.025
IA1: 75%	Z	3.0382	2.3397
N: 612	p (1-sided)	0.0012	0.0096
Events: 354	HR at bound	0.7239	0.7796
Month: 27	P(Cross) if HR=1	0.0012	0.0096
	P(Cross) if HR=0.7	0.6268	0.8452
IA2: 100%	Z	2.6025	2.0118
N: 612	p (1-sided)	0.0046	0.0221
Events: 472	HR at bound	0.7869	0.8308
Month: 36	P(Cross) if HR=1	0.0050	0.0250
	P(Cross) if HR=0.7	0.9015	0.9703

Abbreviations: HR=hazard ratio; IA=interim analysis.

[†] This column displays the number (Events) and percentage (%) of needed PFS events, the expected sample size (N) and the estimated months (Month) after first participant is randomized for each analysis.

P(1-sided): the nominal α for testing.

~HR at bound: the approximate hazard ratio required to reach an efficacy bound.

P(Cross if HR=1): the probability of crossing a bound at or before each analysis under the null hypothesis.

P(Cross if HR=0.7): the probability of crossing a bound at or before each analysis under the alternative hypothesis.

Table 10 Efficacy Boundaries and Properties for PFS Analyses in the All-comer Participants

Analysis†	Value	α = 0.005	α = 0.025
IA1: 75%	Z	3.0382	2.3397
N: 804	p (1-sided)	0.0012	0.0096
Events: 466	HR at bound	0.7546	0.8049
Month: 27	P(Cross) if HR=1	0.0012	0.0096
	P(Cross) if HR=0.7	0.7930	0.9343
IA2: 100%	Z	2.6025	2.0118
N: 804	p (1-sided)	0.0046	0.0221
Events: 621	HR at bound	0.8115	0.8508
Month: 36	P(Cross) if HR=1	0.0050	0.0250
	P(Cross) if HR=0.7	0.9685	0.9929

Abbreviations: HR=hazard ratio; IA=interim analysis.

† This column displays the number (Events) and percentage (%) of needed PFS events, the expected sample size (N) and the estimated months (Month) after first participant is randomized for each analysis.

P(1-sided): the nominal α for testing.

~HR at bound: the approximate hazard ratio required to reach an efficacy bound.

P(Cross if HR=1): the probability of crossing a bound at or before each analysis under the null hypothesis.

P(Cross if HR=0.7): the probability of crossing a bound at or before each analysis under the alternative hypothesis.

3.8.1.2 Overall Survival

The study initially allocates one-sided α = 0.02 for non-inferiority test of OS between two treatment groups for the pMMR participants. The OS non-inferiority test for pMMR may be tested at one-sided α = 0.02 (initially allocated) or higher if the PFS test for all-comers is rejected. If the non-inferiority test of OS for pMMR is rejected, as shown in [Figure 1], its α will be allocated to the OS superiority test for pMMR participants, and if the OS superiority test for pMMR participants is rejected, its α will be reallocated to the OS superiority test for all-comers.

[Table 11] below shows the bounds and boundary properties for OS testing which were derived using a Lan-DeMets O'Brien-Fleming α-spending function. If the actual number of events at the OS analyses differ from those specified in the table, the bounds will be adjusted using the Lan-DeMets O'Brien-Fleming spending function accordingly.

[Table 12] below shows the bounds and boundary properties for OS testing in all-comers with α = 0.02 (if H3 and H4 are rejected), assuming 612 pMMR participants and 192 dMMR participants.

Table 11 Efficacy Boundaries and Properties for OS Analysis for pMMR Participants

Analysis†	**Value**	**Superiority α = 0.02**	**Non-inferiority α = 0.02**	**Superiority α = 0.025**	**Non-inferiority α = 0.025**
IA1: 50%	Z	3.0862	3.0862	2.9592	2.9592
N: 612	p (1-sided)	0.0010	0.0010	0.0015	0.0015
Events:180	HR at bound	0.6308	0.6939	0.6429	0.7072
Month: 27	P(Cross) if HR=1 (or 1.1) for superiority (or non-inferiority)	0.0010	0.0010	0.0015	0.0015
	P(Cross) if HR=0.7 (or 0.8) for superiority (or non-inferiority)	0.2431	0.1702	0.2844	0.2044
IA2: 75%	Z	2.4625	2.4625	2.3592	2.3592
N: 612	p (1-sided)	0.0069	0.0069	0.0092	0.0092
Events: 269	HR at bound	0.7405	0.8145	0.7499	0.8249
Month: 36	P(Cross) if HR=1 (or 1.1) for superiority (or non-inferiority)	0.0072	0.0072	0.0096	0.0096
	P(Cross) if HR=0.7 (or 0.8) for superiority (or non-inferiority)	0.6804	0.5615	0.7167	0.6026
IA3: 88%	Z	2.2904	2.2904	2.1970	2.1970
N: 612	p (1-sided)	0.0110	0.0110	0.0140	0.0140
Events: 316	HR at bound	0.7727	0.8499	0.7809	0.8590
Month: 42	P(Cross) if HR=1 (or 1.1) for superiority (or non-inferiority)	0.0132	0.0132	0.0169	0.0169
	P(Cross) if HR=0.7 (or 0.8) for superiority (or non-inferiority)	0.8203	0.7182	0.8443	0.7500
Final: 100%	Z	2.1525	2.1525	2.0654	2.0654
N: 612	p (1-sided)	0.0157	0.0157	0.0194	0.0194
Events: 359	HR at bound	0.7966	0.8762	0.8040	0.8844
Month: 48	P(Cross) if HR=1 (or 1.1) for superiority (or non-inferiority)	0.0200	0.0200	0.0250	0.0250
	P(Cross) if HR=0.7 (or 0.8) for superiority (or non-inferiority)	0.8990	0.8200	0.9140	0.8430

Abbreviations: HR=hazard ratio; IA=interim analysis.

† This column displays the number (Events) and percentage (%) of needed OS events, the expected sample size (N) and the estimated months (Month) after first participant is randomized for each analysis.

p (1-sided): the nominal α for testing.

~HR at bound: the approximate hazard ratio required to reach an efficacy bound.

P(Cross if HR=1 or 1.1): the probability of crossing a bound at or before each analysis under the null hypothesis.

P(Cross if HR=0.7 or 0.8): the probability of crossing a bound at or before each analysis under the alternative hypothesis.

Table 12 Efficacy Boundaries and Properties for OS Analysis for All-comer Participants

Analysis†	Value	Superiority α = 0.02	Superiority α = 0.025
IA1: 50%	Z	3.0862	2.9592
N: 804	p (1-sided)	0.0010	0.0015
Events:236	HR at bound	0.6690	0.6801
Month: 27	P(Cross) if HR=1 for superiority	0.0010	0.0015
	P(Cross) if HR=0.7 for superiority	0.3634	0.4115
IA2: 75%	Z	2.4625	2.3592
N: 804	p (1-sided)	0.0069	0.0092
Events: 353	HR at bound	0.7694	0.7779
Month: 36	P(Cross) if HR=1 for superiority	0.0072	0.0096
	P(Cross) if HR=0.7 for superiority	0.8140	0.8403
IA3: 88%	Z	2.2904	2.1970
N: 804	p (1-sided)	0.0110	0.0140
Events: 415	HR at bound	0.7985	0.8059
Month: 42	P(Cross) if HR=1 for superiority	0.0132	0.0169
	P(Cross) if HR=0.7 for superiority	0.9153	0.9290
Final: 100%	Z	2.1525	2.0654
N: 804	p (1-sided)	0.0157	0.0194
Events: 471	HR at bound	0.8200	0.8266
Month: 48	P(Cross) if HR=1 for superiority	0.0200	0.0250
	P(Cross) if HR=0.7 for superiority	0.9610	0.9679

Abbreviations: HR=hazard ratio; IA=interim analysis.

† This column displays the number (Events) and percentage (%) of needed OS events, the expected sample size (N) and the estimated months (Month) after first participant is randomized for each analysis.

p (1-sided): the nominal α for testing.

~HR at bound: the approximate hazard ratio required to reach an efficacy bound.

P(Cross if HR=1 or 1.1): the probability of crossing a bound at or before each analysis under the null hypothesis.

P(Cross if HR=0.7 or 0.8): the probability of crossing a bound at or before each analysis under the alternative hypothesis.

3.8.2 Safety Analyses

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. However, to account for any multiplicity concerns raised by the eDMC review of unplanned efficacy data when prompted by safety concerns, a sensitivity analysis for OS adopting a conservative multiplicity adjustment will be prespecified in the sSAP.

3.9 Sample Size and Power Calculations

The sample size is estimated based on the primary endpoints PFS and OS. Approximately 875 participants (including approximately 612 pMMR participants and approximately 263 dMMR participants) will be randomized in a 1:1 ratio for the global study.

The study will be considered to be fully enrolled when 612 pMMR participants have enrolled. (By the time all global participants were enrolled (642 pMMR participants and 200 dMMR participants), actual dMMR prevalence is ~24%.)

Sample size and power calculations are based on pMMR participants:

This study is well-powered for both primary endpoints. Enrollment of ~875 participants is expected to take 24 months.

The PFS hypothesis testing was designed for one-sided $\alpha = 0.005$ and power of 90% to detect a HR of 0.7 with approximately 354 and 472 pooled PFS events at the planned PFS analyses. The duration of PFS in the control group is assumed to follow an exponential distribution with a median of 8.8 months based on historical data.

The OS hypothesis testing was designed for one-sided $\alpha = 0.02$ and power of 90% to detect a HR of 0.7 with approximately 180, 269, 316, and 359 pooled OS events at the planned OS interim and final analyses. The study has power of 82% with a hazard ratio of 0.8 to establish non-inferiority test (NI margin = 1.1) at one-sided $\alpha = 0.02$ level for pMMR participants. The duration of OS in the control group is assumed to follow an exponential distribution with a median of 23 months based on historical data. Further justification of the non-inferiority margin can be found in Protocol Appendix 10.12.

Non-inferiority will be declared if the upper limit of the 2-sided 96% CI for HR is less than the HR boundary calculated based on actual number of OS events at the time of analysis.

3.10 Subgroup Analyses

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

- MMR status (pMMR versus dMMR)
- ECOG (0 versus 1)
- Prior chemotherapy and/or chemoradiation (yes versus no)
- Prior neo/adjuvant chemotherapy (yes versus no)
- Age (<65 years versus ≥65 years)
- Age (<65, ≥65 to <75, ≥75 years)
- Race (White, Non-white)
- Region (North America, EU, Asia, Rest of World)
- Histology (Endometrioid, Non-endometrioid)

Efficacy subgroup analyses will be performed in both the all comer subjects and pMMR subjects , except MMR status will be only performed in the all comer subjects.

The consistency of the treatment effect will be assessed descriptively via summary statistics by category for the classification 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.

3.11 Compliance (Medication Adherence)

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

Lenvatinib compliance will be calculated by the Sponsor based on the drug accountability documented by the site staff and monitored by the Sponsor/designee. The objective is 100% compliance, and investigators and their staff should evaluate compliance at each visit and take appropriate steps to optimize compliance.

3.12 Extent of Exposure

The extent of exposure for lenvatinib will be summarized as duration of treatment in days. The extent of exposure for pembrolizumab will be summarized as duration of treatment in cycles. Dose interruption for each drug, dose reduction for lenvatinib will be summarized. Summary statistics will be provided on Extent of Exposure for the APaT population.

4 STATISTICAL ANALYSIS PLAN FOR CHINA EXTENSION

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

4.1 Introduction

After the global portion enrollment is closed, participants from China will continue to be enrolled in an extension portion designed to meet China local registration needs. The extension portion will be identical to the global portion (e.g., inclusion and exclusion criteria, primary and secondary endpoints, study procedures) in general, with the additional statistical analysis plan for the China subpopulation. The purpose of this extension portion is to evaluate the consistency of efficacy and safety in the China subpopulation to the global population. Country-specific analysis 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 pembrolizumab plus lenvatinib arm and paclitaxel/carboplatin arm. Approximately 131 participants (including approximately 92 pMMR participants and 39 dMMR participants) will be enrolled for the China subpopulation. The China subpopulation will be considered as fully enrolled when approximately 92 pMMR participants have enrolled.

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

However, if the target number of PFS or OS events in the China subpopulation is reached before an IA or final analysis for the global portion, the corresponding analysis for China subpopulation may occur at the same time as the global IA or the final analysis (global portion). If the statistical significance for the global portion has been demonstrated at an IA or final analysis 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 final analysis 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 the final analysis for the global portion.

Additional analyses may be considered for China subpopulation based on Sponsor's discretion and/or consultation with regulatory if global interim/final analysis shows positive results and leads to filing and China subpopulation enrollment has been completed.

4.2 Responsibility for Analyses/In-House Blinding

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

4.3 Hypotheses/Estimation

No hypothesis testing is planned for the China extension portion.

4.4 The Analysis Endpoints

4.4.1 Efficacy Endpoints

Efficacy endpoints are the same as described in Section 3.6.1

4.4.2 Safety Endpoints

Safety endpoints are the same as described in Section 3.6.2.

4.5 Analysis Population

4.5.1 Efficacy Analysis Populations

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

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

4.5.2 Safety Analysis Populations

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

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

4.6 Statistical Methods

No formal hypothesis testing is planned. No multiplicity adjustment will be applied to the analysis for China subpopulation. Other than formal hypothesis testing, analysis methods for the China subpopulation are the same as described in Section 3.6 for the global portion if applicable.

4.6.1 Statistical Methods for Efficacy Analyses

Degree of consistency of efficacy will be evaluated using the percentage of risk reduction preserved in the China subpopulation from the empirical risk reduction from the global primary efficacy analyses (based on point estimate).

4.6.1.1 Progression-Free Survival (PFS)

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

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 unstratified log-rank test. An unstratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (i.e., HR) between the treatment arms. The HR and its 95% confidence interval from the unstratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported.

4.6.1.2 Overall Survival (OS)

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

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the unstratified log-rank test. An unstratified 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 unstratified Cox model with a single treatment covariate will be reported.

4.6.1.3 Objective Response Rate (ORR)

Analysis of ORR for China extension portion is the same to that for the global study if applicable.

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) [Clopper, C. J. 1934]. The unstratified Miettinen and Nurminen’s method will be used for comparison of the objective response rate (ORR) between two treatment groups. The difference in ORR and its 95% confidence interval from the unstratified Miettinen and Nurminen’s method will be reported.

4.6.2 Statistical Methods for Safety Analyses

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

4.6.3 Summaries of Baseline Characteristics, Demographics, and Other Analyses

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

4.7 Interim Analysis and Final analysis

The primary analysis for PFS will be conducted in the China subpopulation when approximately 71 PFS events for pMMR participants have been collected. OS will also be analyzed.

The primary analysis for OS will be conducted in the China subpopulation when approximately 54 OS events for pMMR participants have been collected.

If the timing criteria for an analysis in the China subpopulation are met before the primary analysis of the global portion (including IA or the FA), the corresponding analysis for China subpopulation may occur at the same time as the IA or the FA for the global portion.

If the statistical significance for the global portion has been demonstrated at an IA or final analysis 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 final analysis 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 the final analysis for the global portion.

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

4.8 Multiplicity

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

4.9 Sample Size and Power Calculations

After the completion of global portion enrollment, the extension portion will continue to enroll participants and randomize eligible participants. Approximately 131 participants (including approximately 92 pMMR participants and 39 dMMR participants) will be enrolled for the China subpopulation. The China subpopulation will be considered to be fully enrolled when approximately 92 pMMR participants are enrolled. Participants from China enrolled in the extension portion of this study after completion of the global enrollment will not be included in the primary efficacy analysis population for the global portion.

The extension portion will complete after approximately 71 PFS events for pMMR participants have been observed between the two arms in the China subpopulation. With 71 PFS events for pMMR participants and a true hazard ratio of 0.7, the extension portion has ~80% chance to observe a point estimate of PFS that preserves ≥ approximately 50% of the empirical risk reduction from the global analysis in the China subpopulation. With 54 OS events for pMMR participants and a true hazard ratio of 0.7, the extension portion has ~78%

chance to observe a point estimate of OS that preserves ≥ approximately 50% of the empirical risk reduction from the global analysis in the China subpopulation.

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

4.10 Subgroup Analysis

Analyses may be considered for China subgroup (including 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 China subpopulation enrollment has been completed.

5 REFERENCES

[Anderson, K. M. 1991]	Anderson KM. A nonproportional hazards weibull accelerated failure time regression model. Biometrics 1991;47:281-8.	[00TH4Z]
[Clopper, C. J. 1934]	Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404-13.	[03Y75Y]
[Latimer, N. R., et al 2019]	Latimer NR, Abrams KR, Siebert U. Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. BMC Med Res Methodol. 2019;19:69.	[07XLD5]
[Liang, Kung-Yee and Zeger, Scott L. 2000]	Liang K-Y, 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. 2013]	Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. Stat Biopharm Res 2013;5(4):311-20.	[03XQVB]
[Miettinen, O. 1985]	Miettinen O, Nurminen M. Comparative analysis of two rates. Stat Med 1985;4:213-26.	[00VMQY]
[Robins, J. M. and Tsiatis, A. A. 1991]	Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Commun Stat-Theor M 1991;20(8):2609-31.	[023VWF]
[Uno, H., et al 2014]	Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol. 2014 Aug 1;32(22):2380-5.	[045X5X]