Official Protocol Title:	A Phase II Study of Navarixin (MK-7123) in Combination with Pembrolizumab (MK-3475) in Participants with Selected Advanced/Metastatic Solid Tumors
NCT number:	NCT02054806
Document Date:	23-Apr-2019

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

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

2 SUMMARY OF CHANGES

The following changes made to the sSAP compared to the previous version:

- Added max-combo test as a sensitivity analysis for overall survival (OS) (Section 3.6.1.1)
- Updated details of analysis of patient reported outcome (PRO) data (Section 3.6.2).

All other changes are to ensure consistency with changes made to the protocol (MK-3475-119-05).

3 ANALYTICAL AND METHODOLOGICAL DETAILS

3.1 Statistical Analysis Plan Summary

Key elements of the statistical analysis plan are summarized below; the comprehensive plan is provided in Sections 3.2 – Responsibility for Analyses/In-House Blinding through 3.12 –Extent of Exposure.

Study Design Overview	A Phase III Randomized Trial of Single Agent Pembrolizumab versus Single Agent Chemotherapy per Physician's Choice for Metastatic Triple Negative Breast Cancer (mTNBC)
Treatment Assignment	Approximately 600 subjects will be randomized in a 1:1 ratio between two treatment groups: (1) pembrolizumab arm and (2) Treatment of Physician's Choice (TPC) arm. Stratification factors are: 1) PD-L1 tumor status (positive [CPS ≥1] vs negative [CPS <1]), and 2) history of prior (neo)adjuvant treatment vs de novo metastatic disease at initial diagnosis.
Analysis Populations	Efficacy: Intention-to-treat Population (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoint(s)	1 Overall survival (OS) in subjects with CPS ≥10 2 OS in subjects with CPS ≥1 3 OS in all subjects
Statistical Methods for Key Efficacy Analyses	The primary hypotheses will be evaluated by comparing pembrolizumab to TPC on OS using a stratified Log-rank test. Estimation of the hazard ratio will be performed 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. Safety parameters or adverse experiences of special interest that are identified *a priori* constitute "Tier 1" safety endpoints that will be subject to inferential testing for statistical significance with p-values and 95% confidence intervals (CIs) provided for between-group comparisons. Other safety parameters will be considered Tier 2 (with 95% confidence intervals provided) or Tier 3 (with point estimates by treatment group provided) The between-treatment difference will be analyzed using the Miettinen and Nurminen method.
Interim Analyses	One interim efficacy analysis is planned in this study. Details are provided in Section 8.7. • Interim Analysis o Timing: approximately 14 months after enrollment is completed. It is estimated that approximately 130, 284 and 445 OS events will be accrued in subjects with CPS ≥10, CPS ≥1 and all subjects. o Purpose: interim efficacy analysis for OS • Final analysis o Triggered by duration of follow-up and OS events in subjects with CPS ≥1: Approximately 24 months after enrollment is completed or 334 OS events accrue in subjects with CPS ≥1, whichever occurs later. Approximately 154, 334 and 520 OS events will be accrued in subjects with CPS ≥10, CPS ≥1 and all subjects. o If OS events in subjects with CPS ≥1 accrue slower than expected and fewer than 334 events are observed 26 months after enrollment is completed, then the Sponsor will conduct the final analysis at that time o Purpose: final efficacy analysis for OS
Multiplicity	The overall Type I error over the multiple endpoints will be strongly controlled at 2.5% (one-sided) for primary hypotheses of OS endpoints (initial α of 1.7% in subjects with CPS ≥10 and 0.8% in subjects with CPS ≥1) and secondary hypotheses of PFS and ORR in all subjects. Using an extension of the graphical approach of Maurer and Bretz, alpha can be re-allocated between hypotheses. Group sequential methods will be used to allocate alpha between the interim and final analyses for OS endpoints.
Sample Size and Power	The planned sample size is approximately 600 subjects. Details are provided in Section 3.9. • For the primary OS endpoint in subjects with CPS ≥10, with approximately 154 OS events the trial has ~85% power at a one-sided 1.7% alpha-level, if the underlying HR is 0.60, with a HR at boundary for success of ~0.70 (~4.2 month improvement). • For the primary OS endpoint in subjects with CPS ≥1, with approximately 334 OS events the trial has ~80% (90%) power at a one-sided 0.8% (2.5%) alpha-level, if the underlying HR is 0.70, with a HR at boundary for success of ~0.76 (0.80) (~3.1 [2.5] month improvement). • For the primary OS endpoint in all subjects, with approximately 520 OS events the trial has ~66% (80%) power at a one-sided 0.8% (2.5%) alpha-level, if the underlying HR is 0.78, with a HR at boundary for success of ~0.80 (0.84) (~2.4 [1.9] month improvement).

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 IVRS vendor will generate the randomized allocation schedule(s) for study treatment assignment for this protocol, and the randomization will be implemented in IVRS.

Although the trial is open label, analyses or summaries generated by randomized treatment assignment, actual treatment received, and/or PD-L1 biomarker status will be limited and documented. In addition, the independent radiologist(s) will perform the central imaging review without knowledge of treatment group assignment. The study team at the Sponsor consisting of clinical, statistical, statistical programming and data management personnel, will be blinded to subject-level PD-L1 biomarker results. An unblinded Sponsor clinical scientist, unblinded Sponsor statistician and unblinded Sponsor statistical programmer will have access to the subject-level PD-L1 results for the purpose of data review and event monitoring, and will have no other responsibilities associated with the study. A summary of PD-L1 biomarker prevalence may be provided to the study team at the Sponsor by the IVRS vendor or the unblinded Sponsor statistician.

Access to the allocation schedule and the subject-level PD-L1 results for summaries or analyses will be restricted to an unblinded external statistician, and, as needed, an external scientific programmer performing the analysis, who will have no other responsibilities associated with the study.

Treatment-level results at the interim efficacy will be provided by the external unblinded statistician to the eDMC. The external unblinded statistician will have access to allocation schedule, treatment group, PD-L1 status for any analysis required for eDMC review. Limited additional SPONSOR personnel may be unblinded to the treatment level and/or PD-L1 biomarker results of the efficacy interim analyses, if required, in order to act on the recommendations of the eDMC or facilitate regulatory filing after the interim efficacy analysis. The extent to which individuals are unblinded with respect to results of interim analyses will be documented by the unblinded statistician.

The eDMC will serve as the primary reviewer of the unblinded results of the interim efficacy analyses and will make recommendations for discontinuation of the study or modification to an executive oversight committee of the SPONSOR. Depending on the recommendation of the eDMC, the Sponsor may prepare a regulatory submission. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee may be unblinded to results at the treatment level in order to act on these recommendations. Additional logistical details, revisions to the above plan and data monitoring guidance will be provided in the eDMC Charter.

3.3 Hypotheses/Estimation

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

3.3.1 Primary Objective(s) & Hypothesis(es)

Pembrolizumab will be compared to TPC as 2L or 3L monotherapy in subjects with centrally confirmed mTNBC:

PD-L1 Positive Populations: subjects with PD-L1 positive expression defined by ≥10 Combined Positive Score (CPS) and by ≥1 CPS; henceforth abbreviated as CPS ≥10 and CPS ≥1 respectively.

(1) **Objective:** To compare OS in subjects with PD-L1 positive tumors (CPS ≥10).
Hypothesis (H1): Pembrolizumab prolongs OS compared to TPC in subjects with PD-L1 positive tumors (CPS ≥10).

(2) **Objective:** To compare OS in subjects with PD-L1 positive tumors (CPS ≥1).
Hypothesis (H2): Pembrolizumab prolongs OS compared to TPC in subjects with PD-L1 positive tumors (CPS ≥1).

(3) **Objective:** To compare OS in all subjects.
Hypothesis (H3): Pembrolizumab prolongs OS compared to TPC in all subjects.

The study is considered to have met its primary objective if pembrolizumab is superior to TPC in OS in either subjects with PD-L1 positive tumors (CPS ≥10 [H1] or CPS ≥1 [H2]) or in all subjects (H3), at either the interim analysis or the final analysis.

3.3.2 Secondary Objective(s) & Hypothesis(es)

Pembrolizumab will be compared to TPC as 2L or 3L monotherapy in subjects with centrally confirmed mTNBC:

(1) **Objective:** To compare progression-free survival (PFS) based on RECIST 1.1 as assessed by blinded central imaging vendor in all subjects.

Hypothesis (H4): Pembrolizumab prolongs PFS based on RECIST 1.1 as assessed by blinded central imaging vendor compared to TPC in all subjects.

(2) **Objective:** To compare overall response rate (ORR) per RECIST 1.1 by blinded central imaging vendor in all subjects.

Hypothesis (H5): Pembrolizumab increases ORR per RECIST 1.1 by blinded central imaging vendor compared to TPC in all subjects.

(3) **Objective:** To evaluate PFS and ORR based on RECIST 1.1 as assessed by blinded central imaging vendor in subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1).

(4) **Objective:** To evaluate duration of response (DOR), and disease control rate (DCR) based on RECIST 1.1 as assessed by blinded central imaging vendor in subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1) and in all subjects.

(5) **Objective:** To determine the safety and tolerability of pembrolizumab.

The secondary hypotheses regarding PFS in all subjects (H4) and ORR in all subjects (H5) can only be formally tested once the null hypothesis regarding the primary endpoint OS in all subjects (H3) has been rejected.

3.3.3 Exploratory Objectives

Pembrolizumab will be compared to TPC as 2L or 3L monotherapy in subjects with centrally confirmed mTNBC:

(1) **Objective:** To evaluate PFS, ORR, DOR, and DCR based on irRECIST as assessed by blinded central imaging review in subjects with PD-L1 positive tumors (CPS ≥10 and CPS≥1) and all subjects.

(2) **Objective:** To evaluate changes in health-related quality-of-life assessments from baseline in subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1) and all subjects using the European Organisation for Research and Treatment of Cancer (EORTC) QLQ-C30 and QLQ-BR23.

(3) **Objective:** To characterize utilities in subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1) and all subjects using EuroQol (EQ)-5D.

(4) **Objective:** To investigate the association between antitumor activity of pembrolizumab in mTNBC and efficacy/resistance biomarkers, utilizing tumor and blood specimens obtained before/during treatment and at disease progression.

(5) **Objective:** To explore the relationship between genetic variation and response to the treatment(s) administered. Variation across the human genome will be analyzed for association with clinical data collected in this study.

3.4 Analysis Endpoints

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

3.4.1 Efficacy Endpoints

Primary

Overall Survival (OS)

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

Secondary/Exploratory

Progression-free survival (PFS) – based on RECIST 1.1 as assessed by blinded central imaging vendor (RECIST 1.1 by site Investigator/local radiology review as a supportive analysis).

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression based on RECIST 1.1 as assessed by blinded central imaging vendor radiology review or death due to any cause, whichever occurs first. See Section 3.6.1 for the definition of censoring.

Overall Response Rate (ORR) – based on RECIST 1.1 (Secondary) and irRECIST (Exploratory) as assessed by blinded central imaging vendor (RECIST 1.1 by site Investigator/local radiology review as the corresponding supportive analyses; irRECIST by site Investigator/local radiology review may also be evaluated).

Overall response rate is defined as the proportion of the subjects in the analysis population who have a complete response (CR) or partial response (PR).

Duration of Overall Response (DOR) – based on RECIST 1.1 (Secondary) and irRECIST (Exploratory) as assessed by blinded central imaging vendor (RECIST 1.1 by site Investigator/local radiology review as the corresponding supportive analyses; irRECIST by site Investigator/local radiology review may also be evaluated).

For subjects who demonstrated CR or PR, response duration is defined as the time from first documented evidence of CR or PR until disease progression or death. See Section 3.6.1 for the definition of censoring.

Disease Control Rate (DCR) – based on RECIST 1.1 (Secondary) and irRECIST (Exploratory) as assessed by blinded central imaging vendor (RECIST 1.1 by site Investigator/local radiology review as the corresponding supportive analyses; irRECIST by site Investigator/local radiology review may also be evaluated).

Disease control rate (DCR) is defined as the proportion of the subjects in the analysis population who have a CR, partial response (PR) and stable disease (SD), the latter for at least 8 cycles.

PFS–based on irRECIST (Exploratory) as assessed by blinded central imaging vendor (irRECIST by site Investigator/local radiology review may be performed as supportive analysis).

3.4.2 Safety Endpoints

Safety measurements are described in Protocol Section 4.2.3 – Rationale for Endpoints and Protocol Section 7.0 – Trial Procedures.

3.5 Analysis Populations

3.5.1 Efficacy Analysis Populations

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

Details on the approach to handling missing data are provided in Section 3.6 – Statistical Methods.

3.5.2 Safety Analysis Populations

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

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

Details on the approach to handling missing data for safety analyses are provided in Section 3.6 – Statistical Methods.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

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

both stratification factors [PD-L1 tumor status (positive [CPS ≥1] vs negative [CPS <1]), and prior (neo)adjuvant therapy vs de novo metastatic disease at initial diagnosis] will be included. For efficacy analyses performed in subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1), only the stratification factor of prior (neo)adjuvant therapy vs de novo metastatic disease at initial diagnosis will be considered in the analysis models, i.e., not considering the stratification factor of PD-L1 tumor status. If there is too small number (<10) of subjects in specific stratum for analyses in all subjects or subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1) based on blinded data review, then strata will be combined with the neighboring strata. If the mis-stratification rate at randomization is greater than 10% based on review of blinded data, additional sensitivity analyses may be performed for primary efficacy and key secondary efficacy endpoints according to each subject's stratum based on actual data observed/collected.

3.6.1.1 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 hazard ratio). The hazard ratio and its 95% confidence interval from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization [PD-L1 tumor status (positive [CPS ≥1] vs negative [CPS <1]), and prior (neo)adjuvant therapy vs de novo metastatic disease at initial diagnosis] will be applied, as stratification factors used for analysis, to both the stratified log-rank test and the stratified Cox model.

Subjects in the TPC arm are expected to discontinue treatment earlier compared to subjects in the pembrolizumab arm and are not allowed to crossover to the pembrolizumab arm; however, they may be treated with another anti PD-1 drug following the verification of progressive disease by blinded central imaging vendor. As an exploratory analysis, the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis (1991) [1] and two stage-model [2] may be used to adjust for the effect of crossover to other PD-1 therapies on OS, based on an examination of the appropriateness of the data to the assumptions required by the methods.

The RPSFT model provides a randomization-based estimate of the treatment effect corrected for bias introduced by crossover from the control arm to the experimental treatment. This method is rank-preserving in the sense that it assumes that given two subjects *i* and *j*, if subject *i* failed before subject *j* when both were on one treatment, then subject *i* would also fail before subject *j* if both subjects took any other alternative treatment. The method is structural in the sense that it assumes a defined relationship between the observed survival time and the survival time that would have been observed if crossover had not occurred. It is also assumed that the treatment effect is the same before and after progression. More specifically, the RPSFT method first relates the observed survival time to a latent survival time (experimental treatment-free survival time if the patient was never to receive the experimental treatment) through an accelerated failure time model. The treatment effect will then be estimated under the assumption that the latent survival curves are identical between

the control and experimental treatment arms. Re-censoring of the latent survival time using the treatment effect will be applied in order to preserve the independent censoring assumption. The OS analysis will then be applied to the "corrected" survival dataset, which includes the adjusted survival time for subjects in the control arm so that it reflects the OS had they not received the experimental treatment as well as the observed survival time for subjects in the experimental treatment arm. The HR and the associated 95% CI for OS after adjustment of the crossover effect using the RPSFT method will be provided.

Under the assumptions of no unmeasured confounders at the secondary baseline time-point (disease progression), treatment switching only happens after progression, and happens soon after progression, the "two-stage" approach may be appropriate. At Stage 1, the date of disease progression is used as a secondary baseline for subjects who have a documented progression in the control arm and data from these subjects beyond this time-point are considered as an observational dataset. An accelerated failure time model including covariates for crossover and other prognostic covariates measured at the secondary baseline will be applied to this observational dataset to estimate an acceleration factor. At Stage 2, a counterfactual survival dataset will be constructed such that survival time of subjects with treatment switching will be shrunk by the inverse of the acceleration factor, while no shrinkage is performed for the survival time of subjects in the control group without treatment switching or subjects in the experimental arm. The OS analysis will then be applied to this counterfactual survival dataset to estimate the HR from this two-stage method.

It is very important to assess trial data, crossover mechanism, and treatment effect to determine which method is likely to be most appropriate to evaluate the crossover effect. Additional supportive unstratified analyses may also be provided.

Due to multiple occurrences of delayed separation phenomena observed in monotherapy Immuno-Oncology (I/O) RCTs, another sensitivity analysis, which evaluates the treatment difference for OS using the stratified max-combo test, will be conducted at the final analysis. The max-combo test statistic is the maximum of the log-rank test statistic and weighted log-rank variation of the Fleming-Harrington test statistics; $Z_m = \max(Z_1, Z_2, Z_3)$, where Z_1, Z_2 and Z_3 are the test statistics from the FH (0, 0), FH (1, 1) and FH (0, 1) family of test statistics, respectively. FH (0, 0) corresponds to the log-rank test, while FH (1, 1) and FH (0, 1) are more sensitive to middle and late-difference alternatives, respectively. The adjusted nominal p-value, which can be derived by integrating under the multivariate normal density [3] will be reported. No formal hypothesis test will be conducted.

3.6.1.2 Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The 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 (i.e. hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. The stratification factors used for randomization (see Protocol Section 5.4 –

Stratification) will be applied, as stratification factors used for analysis, 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. For the subjects who have PD, the true date of disease progression will be approximated by the date of the first assessment at which PD is objectively documented based on RECIST 1.1 as assessed by blinded central imaging vendor. Death is always considered as a confirmed PD event. Subjects who do not experience a PFS event will be censored at the last disease assessment (with exceptions described below). Sensitivity analyses will be performed for comparison of PFS based on Investigator's assessment.

In order to evaluate the robustness of the PFS endpoint based on RECIST 1.1 as assessed by blinded central imaging vendor, we will perform one primary and two sensitivity analyses with a different set of censoring rules. For the primary analysis, if the events (PD or death) are immediately after more than one missed disease assessment, the data are censored at the last disease assessment prior to missing visits. Also data after new anti-cancer therapy are censored at the last disease assessment prior to the initiation of new anti-cancer therapy. The first sensitivity analysis follows the complete follow up intention-to-treat principle. That is, PDs/deaths are counted as events regardless of missed study visits or initiation of new anti-cancer therapy. The second sensitivity analysis considers discontinuation of treatment or initiation of an anticancer treatment subsequent to discontinuation of study-specified treatments, whichever occurs later, to be a PD event for subjects without documented PD or death. If a subject 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
No PD and no death; new anticancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment	Progressed at treatment discontinuation due to reasons other than 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 or death documented after ≤ 1 missed disease assessment, and before new anti-cancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death	Progressed at date of documented PD or death
PD or death documented immediately after ≥ 2 missed disease assessments or after new anti-cancer therapy, if any	Censored at last disease assessment prior to the earlier date of ≥ 2 consecutive missed disease assessment and new anti-cancer therapy, if any	Progressed at date of documented PD or death	Progressed at date of documented PD or death

The proportional hazards assumption on PFS will be examined using both graphical and analytical methods if warranted. The log[-log] of the survival function vs time for PFS may be plotted for the comparison between pembrolizumab and the TPC 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 Restricted Mean Survival Time (RMST) method [4] and parametric method [5]. 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 trial, 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.

One assumption for stratified Cox proportional hazard model is that the treatment HR is constant across the strata. If strong departures from the assumption of the HR being the same

for all the strata observed (which can result in a notably biased and/or less powerful analysis), a sensitivity analysis may be performed based on a two-step weighted Cox model approach by Mehrotra et al., 2012 [6], in which the treatment effect is first estimated for each stratum, and then the stratum specific estimates are combined for overall inference using sample size weights. The RMST method and two-step weighted Cox model may also be applied to OS endpoints as appropriate as sensitivity analyses.

In case there is an imbalance between the treatment groups on disease assessment schedules or censoring patterns, we may also perform one additional PFS supportive analysis using Finkelstein (1986)'s likelihood-based score test [7] for interval-censored data, which modifies the Cox proportional hazard model for interval-censored data. The interval will be constructed so that the left endpoint is the date of the last disease assessment without documented PD and the right endpoint is the date of documented PD or death, whichever occurs earlier.

Additional PFS supportive analyses may be performed as appropriate, including a PFS analysis using time to scheduled tumor assessment visit from randomization as opposed to the actual tumor assessment time. Additional supportive unstratified analyses may also be provided.

3.6.1.3 Objective Response Rate (ORR)

Stratified Miettinen and Nurminen's method will be used for the comparison of ORR between 2 treatment arms. The difference in ORR and its 95% CI from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be reported. The stratification factors used for randomization (see Protocol Section 5.4 – Stratification) will be applied to the analysis.

The ORR hypotheses will be tested according to the hypotheses testing plan as described in Section 3.8 – Multiplicity.

Sensitivity analyses will be performed for ORR based on site investigator/local radiology review. Additional supportive unstratified analyses may also be provided.

3.6.1.4 Disease Control Rate (DCR)

Stratified Miettinen and Nurminen's method will be used for the comparison of DCR between 2 treatment arms. The difference in DCR and its 95% CI from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be reported. The stratification factors used for randomization (see Protocol Section 5.4 – Stratification) will be applied to the analysis.

Sensitivity analyses will be performed for DCR based on site investigator/local radiology review.

3.6.1.5 Duration of Response (DOR)

If sample size permits, DOR will be summarized descriptively using the non-parametric Kaplan-Meier method. Only the subset of subjects who achieved CR or PR will be included in this analysis.

Censoring rules for DOR are summarized in Table 2. If a subject meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied.

For each DOR analysis, a corresponding summary of the reasons responding subjects are censored will also be provided. Responding subjects 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 subject 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.		

3.6.1.6 Summary of Statistical Methods for Efficacy

Table 3 summarizes the primary analysis approach for primary and secondary efficacy endpoints of Part 2. Sensitivity analysis methods are described above for each endpoint as applicable.

The strategy to address multiplicity issues with regard to multiple efficacy endpoints, multiple populations, and interim analyses is described in Section 3.7 – Interim Analyses and in Section 3.8 – Multiplicity.

Table 3 Analysis Strategy for Key Efficacy Endpoints

Endpoint/Variable (Description, Time Point)	Statistical Method†	Analysis Population	Missing Data Approach
Primary Endpoints/Hypothesis			
Primary Hypothesis #1			
OS in subjects with CPS ≥10	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at last known alive date
Primary Hypothesis #2			
OS in subjects with CPS ≥1	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at last known alive date
Primary Hypothesis #3			
OS in all subjects	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at last known alive date
Secondary Endpoints/Hypothesis			
Secondary Hypothesis #4			
PFS based on RECIST 1.1 by blinded central imaging vendor in all subjects	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	• Primary censoring rule • Sensitivity analysis 1 • Sensitivity analysis 2 (More details are in Table 1)
Secondary Hypothesis #5			
ORR based on RECIST 1.1 by blinded central imaging vendor in all subjects	Stratified M & N method‡	ITT	Subjects with missing data are considered non-responders
Secondary Efficacy Objectives			
PFS based on RECIST 1.1 by blinded central imaging vendor in subjects with CPS ≥10 and subjects with CPS ≥1	Test: Stratified Log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	• Primary censoring rule • Sensitivity analysis 1 • Sensitivity analysis 2 (More details are in Table 1)
ORR based on RECIST 1.1 by blinded central imaging vendor in subjects with CPS ≥10 and subjects with CPS ≥1	Stratified M & N method‡	ITT	Subjects with missing data are considered non-responders

Endpoint/Variable (Description, Time Point)	Statistical Method†	Analysis Population	Missing Data Approach
DCR based on RECIST 1.1 by blinded central imaging vendor in subjects with CPS ≥10, subjects with CPS ≥1 and all subjects	Stratified M & N method‡	ITT	Subjects with missing data are considered non-responders
DOR based on RECIST 1.1 by blinded central imaging vendor in subjects with CPS ≥10, subjects with CPS ≥1 and all subjects	Summary statistics using Kaplan-Meier method	All responders in ITT	See Table 2

† Statistical models are described in further detail in the text. For stratified analyses, the stratification factors used for randomization will be as stratification factors for analysis.

‡ Miettinen and Nurminen method.

3.6.2 Statistical Considerations for Patient-Reported Outcomes (PRO)

3.6.2.1 Patient Reported Outcome (PRO) Endpoints

The PRO endpoints include results from the EORTC QLQ-C30, EORTC QLQ-BR23, and EQ-5D™ questionnaires.

The EORTC QLQ C30 is a self-reported 30-item cancer specific instrument that assesses 15 domains: 5 functional scales (physical, role, emotional, cognitive and social functioning), 9 symptom scales or single items (fatigue, nausea and vomiting, pain, dyspnea, insomnia, appetite loss, constipation, diarrhea and financial difficulties), and a global health status / QoL.

The EORTC QLQ-BR23 is a breast-specific module of the EORTC QLQ. It includes 23 items composed of 4 functional scales (i.e., body image, sexual functioning, sexual enjoyment and future perspective) and 4 symptom scales (systemic therapy side effects, breast symptoms, arm symptoms and upset by hair loss).

EQ-5D is a standardized measure of health status developed by the EuroQol Group in order to provide a simple, generic measure of health for clinical and economic appraisal [8]. EQ-5D comprises two separate elements. Utility score (or descriptive system), the first of these captures health state across five dimensions: mobility, self-care, usual activities, pain / discomfort, anxiety / depression. Unique health states are defined by combining response levels from each of the five dimensions. The second EQ-5D element is based on a vertical visual analogue scale (VAS). The VAS records the respondent's self-rated health on a vertical, visual analogue scale ranging from 0 to 100 where the end points are labelled 'Best imaginable health state' (100) and 'Worst imaginable health state' (0). This information can be used as a quantitative measure of health outcome as judged by the individual respondents.

In this trial, the PROs will be assessed according to the following schedule (Table 4). For the analysis, PROs assessed at visits of "End of Treatment" and "Safety Follow-up" will be

mapped into different time points according to the actual visit time. If there are multiple PRO collections within the time window of a specific visit, the collection closest to the target day will be used in the analysis.

Table 4 PRO Assessment Schedule

	Week[1]								End of Treatment	Safety Follow-up[3]
	Baseline (0)	3	6	15	24	33	42	51[2]		
Treatment Cycle (C)	C1	C2	C3	C6	C9	C12	C15	C18[2]	X	X

[1]PRO collections are scheduled on the 1st day of a cycle, and Week is counted as number of weeks elapsed since the start of treatment. E.g, the first day of C2 is mapped to Week 3.

[2]After the 3rd cycle and until the end of Year 1, PROs will be collected every 3rd cycle (every 9 weeks) until PD, while the subject is receiving study treatment. During Year 2, they will occur every 4th cycle (every 12 weeks) until PD, while the subject is receiving study treatment.

[3]If the End of Treatment Visit occurs 30 days from the last dose of study treatment, at the time of the mandatory Safety Follow up Visit, PROs do not need to be repeated.

Key PRO Endpoint

Primary analysis time point: the primary ePRO analysis time point is defined as the latest time point where the completion and compliance rates are still high enough based on blinded data review (~60% completion rate and ~80% compliance rate). The key PRO endpoint is:

- The mean score changes from baseline to the primary analysis time point in EORTC QLQ-C30 global health status/QoL score.

Supportive PRO Endpoints

The following are supportive PRO endpoints and may be analyzed as appropriate.

1. The mean score changes from baseline to the primary analysis time point in VAS as measured by EQ-5D.
2. The mean score changes from baseline to the primary analysis time point for:
    - The QLQ-C30 symptom scale Nausea and Vomiting.
    - The QLQ-C30 symptom scale Diarrhea.
    - The QLQ-C30 symptom scale Physical Functioning.
    - The QLQ-BR23 symptom scale Systemic Therapy Side Effects.

3. Time to deterioration (TTD), defined as time from start of treatment to first onset of 10 points or more worsening from baseline, for
    - The QLQ-C30 symptom scale Nausea and Vomiting.
    - The QLQ-C30 symptom scale Diarrhea.
    - The QLQ-C30 symptom scale Physical Functioning.
    - The QLQ-BR23 symptom scale Systemic Therapy Side Effects.
4. Overall improvement rate, where improvement is defined as a 10 points or more improvement from baseline at any time during the trial, for
    - The QLQ-C30 symptom scale Nausea and Vomiting.
    - The QLQ-C30 symptom scale Diarrhea.
    - The QLQ-C30 symptom scale Physical Functioning.
    - The QLQ-BR23 symptom scale Systemic Therapy Side Effects.

3.6.2.2 Patient Reported Outcome (PRO) Analysis Population

The PRO Full Analysis Set (FAS) population will be used for PRO analyses. The PRO FAS population consists of all randomized subjects who received at least one dose of study medication and completed at least one PRO assessment.

The analysis will be conducted in all subjects and in subjects with CPS ≥10, and subjects with CPS ≥1.

3.6.2.3 Analysis Approaches

The PROs are exploratory objectives in this study, thus no formal hypotheses are formulated. Nominal p-values without multiplicity adjustment will be provided and should be interpreted with caution.

3.6.2.3.1 Scoring Algorithm

QLQ-C30 Scoring

The QLQ-C30 is composed of both multi-item scales and single-item measures. These include a global health status / QoL scale, five functional scales, three symptom scales, and six single items. Each of the multi-item scales includes a different set of items - no item occurs in more than one scale.

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

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

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

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

Specifically, if items $I_1, I_2, \ldots, I_n$ are included in a scale, the scoring procedure is as follows:

1. Compute the raw score: $RS = (I_1 + I_2 + \ldots + I_n)/n$

2. Linear transformation to obtain the score S:

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

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

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

Range is the difference between the maximum possible value of *RS* and the minimum possible value. The QLQ-C30 has been designed so that all items in any scale take the same range of values. Therefore, the range of *RS* equals the range of the item values. If more than half of the items within one scale are missing, then the scale is considered missing, otherwise, the score will be calculated as the average score of those available items.

QLQ-BR23 Scoring

The scoring approach for the QLQ-BR23 is identical in principle to that for the function and symptom scales / single items of the QLQ-C30. A linear transformation will be applied to standardize the scores between 0 and 100 as described above for the EORTC QLQ-C30 scoring.

EQ-5D Scoring

The EQ-5D utility score will be calculated based on the European algorithm [8] based on responses on the five health state dimensions, including mobility, self-care, usual activities, pain / discomfort, and anxiety / depression.

3.6.2.3.2 Patient Reported Outcome (PRO) Score Analysis

To assess the treatment effect on the PROs, for each continuous PRO endpoint defined, a constrained longitudinal data analysis (cLDA) model will be used as the primary analysis method, with the PRO score as the response variable [10]. Only PRO data up to the primary analysis time point will be included in this analysis model.

The cLDA model is specified as follows:

$$E(Y_{ijt}) = \gamma_0 + \gamma_{jt} I(t > 0) + \beta X_i, \qquad j = 1,2, \;\; t = 0,1, \ldots T$$

Where Y_{ijt} is the PRO score for subject i, with treatment 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 vector of stratification values (Protocol Section 5.4) for subject i, and β is the corresponding coefficient for strata.

The treatment effect on PRO score change from baseline will be evaluated at the primary analysis time point. Between-group comparison will be performed and the differences in the least-squares mean change from baseline at the primary analysis time point will be reported, together with 95% CI and nominal p-value. In addition, model-based least-squares mean score with corresponding 95% CI will be provided by treatment group at the primary analysis time point.

Patients with disease progression confirmed or feeling worse due to drug-related AE may have missing PRO assessments. The missing data must be handled accordingly to obtain valid statistical analysis results. The cLDA model implicitly treats missing data as missing at random (MAR). Sensitivity analyses may be conducted in case the robustness of MAR assumption is questionable.

Descriptive statistics (mean, SE) of change from baseline no imputation for missing data on the following score/scales will be plotted for QLQ-C30 global health status/QoL, symptom scales nausea and vomiting, diarrhea, the QLQ-BR23 symptom scale systemic therapy side effects, and the QLQ-C30 physical functioning.

3.6.2.3.3 Analysis of the Time to Deterioration (TTD)

The non-parametric Kaplan-Meier method will be used to estimate the deterioration curve in each group. The treatment difference in time-to-deterioration 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 (hazard ratio) between treatment arms. The stratification factors used for randomization (see Protocol Section 5.4 – Stratification) will be applied to the analysis. This analysis will be conducted for QLQ-C30 global health status/QoL, symptom scales nausea and vomiting, diarrhea, physical functioning, and the QLQ-BR23 symptom scale systemic therapy side effects.

3.6.2.3.4 Analysis of Overall Improvement

Stratified Miettinen and Nurminen's method will be used for the comparison of overall improvement rates between 2 treatment arms. The difference in overall improvement rates and its 95% CI from the stratified Miettinen and Nurminen's method with strata weighting by sample size will be reported. The stratification factors used for randomization (see Protocol Section 5.4 – Stratification) will be applied to the analysis. This analysis will be conducted for QLQ-C30 global health status/QoL, symptom scales nausea and vomiting, diarrhea, physical functioning, and the QLQ-BR23 symptom scale systemic therapy side effects.

Of note, for all above ePRO analysis (Section 3.6.2.3.2 to Section 3.6.2.3.5) in subjects with PD-L1 positive tumors (CPS ≥10 and CPS ≥1), only the stratification factor of prior (neo)adjuvant therapy vs de novo metastatic disease at initial diagnosis will be considered in the analysis models, i.e., not considering the stratification factor of PD-L1 tumor status.

3.6.2.3.5 Summary of Completion and Compliance

Completion and compliance of QLQ-C30, QLQ-BR23 and EQ-5D by treatment and visit will be described based on the PRO FAS population.

Completion Rate is defined as the percentage of subjects who completes at least one score/item over the number of subjects in the PRO FAS population at each time point.

The completion rate is expected to shrink in the later visits during due to early discontinuations. Therefore, another measurement, Compliance Rate, defined as the percentage of subjects who completes at least one score/item over the number of eligible subjects who are expected to complete the PRO assessment (not including the subjects missing by design (such as death, discontinuation, translation not available, etc.), will be employed as a supportive measure.

The reasons of non-completion and non-compliance will also be summarized.

3.6.3 Statistical Methods for Safety Analyses

Safety and tolerability will be assessed by clinical review of all relevant parameters including adverse experiences, laboratory tests, vital signs, etc. Of note, safety analyses will be conducted in all subjects.

The analysis of safety results will follow a tiered approach (Table 5). The tiers differ with respect to the analyses that will be performed. For this protocol, there are no Tier 1 safety endpoints. Tier 2 parameters will be assessed via point estimates with 95% CIs provided for between-group comparisons; only point estimates by treatment group will be provided for Tier 3 safety parameters.

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

group exhibit the event; all other adverse experiences and predefined limits of change will belong to Tier 3.

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

Continuous measures such as changes from baseline in laboratory and vital signs will be considered Tier 3 safety parameters. Summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group.

The broad clinical and laboratory AE categories consisting of the percentage of subjects with any AE, any drug related AE, any Grade 3-5 AE, any serious AE, any AE which is both drug-related and Grade 3-5, any AE which is both serious and drug-related, dose modification due to AE, and who discontinued due to an AE, and death will be considered Tier 2 endpoints. For Tier 2 endpoints, point estimates and 95% CIs will be provided for between-treatment differences in the percentage of subjects with events; these analyses will be performed using the Miettinen and Nurminen method [11].

Table 5 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint†	p-Value	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Any AE		X	X
	Any Serious AE		X	X
	Any Grade 3-5 AE		X	X
	Any Drug-Related AE		X	X
	Any Serious and Drug-Related AE		X	X
	Any Grade 3-5 and Drug-Related AE		X	X
	Dose Modification due to AE		X	X
	Discontinuation due to AE		X	X
	Death		X	X
	Specific AEs, System Organ Class, or Pre-Defined Limit of Change ‡ (incidence ≥4 of subjects in one of the treatment groups)		X	X
Tier 3	Specific AEs, System Organ Class or Pre-Defined Limit of Change ‡ (incidence <4 of subjects in all of the treatment groups)			X
	Change from Baseline Results (Labs, ECGs, Vital Signs)			X

† Adverse Experience references refer to both Clinical and Laboratory AEs.

‡ Includes only those endpoints not pre-specified as Tier 1 or not already pre-specified as Tier-2 endpoints.

Note: X = results will be provided.

To properly account for the potential difference in follow-up time between treatment arms, which is expected to be longer in the pembrolizumab arm, AE incidence density adjusted for treatment exposure analyses may be performed as appropriate.

In addition to the tiered approach, exploratory analysis may be performed on time to first Grade 3-5 AE. Time to first Grade 3-5 AE is defined as the time from the first day of study medication to the first event of Grade 3-5 AE. The Kaplan-Meier method will be used to estimate the curve of time to first Grade 3-5 AE. The treatment difference in time to first Grade 3-5 AE will be assessed by the log-rank test. A Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (i.e., the HR). The HR and its 95% CI from the Cox model with a single treatment covariate will be reported.

3.6.4 Summaries of Demographic and Baseline Characteristics

The comparability of the treatment groups for each relevant characteristic will be assessed by the use of tables and/or graphs. No statistical hypothesis testing will be performed on these

characteristics. The number and percentage of subjects screened, randomized, the primary reasons for screening failure, and the primary reason for discontinuation will be displayed. Demographic variables (e.g., age) and baseline characteristics will be summarized by treatment either by descriptive statistics or categorical tables.

3.6.5 Statistical Methods for Exploratory Analyses

Exploratory analyses related to PFS, ORR, DOR, and DCR based on irRECIST will be conducted using the same primary statistical methods as those for RECIST 1.1 based PFS, ORR, DOR and DCR, respectively.

The analyses plan for the exploratory objectives regarding biomarkers will be provided in separate SAP(s) as appropriate.

An exploratory analysis of PFS2, defined as the time from randomization to subsequent disease progression after initiation of new anti-cancer therapy, or death from any cause, whichever occurs first, may be carried out. Subjects alive and for whom a PFS event has not been observed should be censored at the last time the subject was known alive and without disease progression.

3.7 Interim Analyses

There is one planned interim efficacy analysis in this trial. Results will be reviewed by an external data monitoring committee (eDMC). Results will be reviewed by the external DMC.

Interim Analysis

The purpose of the interim analysis (IA) is to evaluate the superiority of pembrolizumab compared to TPC in OS. The IA is triggered by the specified duration of follow-up after enrollment is completed rather than the number of OS events. It will be performed approximately 14 months after enrollment is completed. It is estimated at this time approximately 130 OS events will be accrued in subjects with CPS ≥10, approximately 284 OS events in subjects with CPS ≥1, and approximately 445 OS events in all subjects.

If the pembrolizumab arm demonstrates a superior OS to TPC in all subjects in this interim analysis, the secondary hypotheses for PFS and ORR in all subjects will be tested at the same time. It is estimated that approximately 575 PFS events will have accumulated in all subjects by that time.

The estimated boundaries for OS endpoints at IA are provided in Table 6, assuming the planned numbers of events are analyzed.

Final Analysis

The purpose of the final analysis (FA) is to evaluate the superiority of pembrolizumab compared to TPC in OS. The final analysis is triggered by duration of follow-up and OS events in subjects with CPS ≥1: approximately 24 months after enrollment is completed or

334 OS events accrue in subjects with CPS ≥1, whichever occurs later. It is estimated at this time approximately 154 OS events will be accrued in subjects with CPS ≥10 and approximately 520 OS events in all subjects. If OS events in subjects with CPS ≥1 accrue slower than expected and fewer than 334 events are observed 26 months after enrollment is completed, then the Sponsor will conduct the final analysis at that time.

Table 6 summarizes the timing, sample size and decision guidance of the interim analysis and FA, assuming there is no alpha re-allocation between hypotheses. Boundaries are based on estimated number of events and may be updated at times of the analyses using the actual observed number of events and spending functions as noted above.

Table 6 Summary of Timing, Sample Size, and Decision Guidance of Interim Efficacy Analyses and Final Analysis under Initial Alpha Allocation

Analysis	Criteria for Conduct of Analysis	Endpoint	Parameter	Value
Interim Analysis: Interim OS Analysis	Triggered by duration of follow-up: Approximately 14 months after enrollment is completed. Projected OS events in subjects with CPS ≥10: approximately 130	OS in subjects with CPS ≥10	p value (1-sided) at boundary	0.0089
			~ HR at boundary	0.66
	Projected OS events in subjects with CPS ≥1: approximately 284	OS in subjects with CPS ≥1	p value (1-sided) at boundary	0.0043
			~ HR at boundary	0.73
	Projected OS events in all subjects: approximately 445	OS in all subjects	p value (1-sided) at boundary	n/a
			~ HR at boundary	n/a
Final Analysis: Final OS Analysis	Triggered by duration of follow-up and OS events in subjects with CPS ≥1: Approximately 24 months after enrollment is completed or 334 OS events accrue in subjects with CPS ≥1, whichever occurs later. If OS events in subjects with CPS ≥1 accrue slower than expected and fewer than 334 events are observed 26 months after enrollment is completed, then the Sponsor will conduct the final analysis at that time. Projected OS events in subjects with CPS ≥10: approximately 154	OS in subjects with CPS ≥10	p value (1-sided) at boundary	0.0146
			~ HR at boundary	0.70
	Projected OS events in subjects with CPS ≥1: approximately 334	OS in subjects with CPS ≥1	p value (1-sided) at boundary	0.0066
			~ HR at boundary	0.76
	Projected OS events in all subjects: approximately 520	OS in all subjects	p value (1-sided) at boundary	n/a
			~ HR at boundary	n/a

If a hypothesis is supported, the alpha can be re-allocated to another hypothesis following the pre-specified rules in Section 3.8 – Multiplicity. The alternative efficacy decision guidance for OS with respect to the re-allocated alpha from the support of other hypothesis(es) is summarized in Table 7 below, assuming the same numbers of events specified in Table 7 available for analyses at each time point. Please note that the actual boundaries will be adjusted based on the actual number of events observed at the time of the corresponding analysis based on the spending functions specified in this section.

If an efficacy boundary is crossed at IA or FA for OS in subjects with CPS ≥10, with CPS ≥1 or all subjects, the study will be declared to have met its primary objective. The study may continue till completion regardless of the results of the interim analyses in order to obtain mature OS data.

Of note, these calculations are based on the assumptions that the prevalence of PD-L1 positivity in mTNBC is approximately 65% for CPS ≥1 and approximately 31% for CPS ≥10.

Table 7 Summary of Alternative Efficacy Decision Guidance

Endpoint	Scenario	Total Alpha Allocated	Analysis	Updated Efficacy Boundary (After alpha Re-Allocation)	
				p-value (1-sided) at Boundary	Approx. HR at Boundary
OS in Subjects with CPS ≥ 1	H1 supported	0.025	IA	0.0135	0.77
			FA	0.0217	0.80
OS in All Subjects	H1 and H2 supported	0.025	IA	0.0138	0.81
			FA	0.0217	0.84
OS in All Subjects	H1 is not supported and H2 supported	0.008	IA	0.0044	0.78
			FA	0.0066	0.80

3.8 Multiplicity

The multiplicity strategy specified in this section will be applied to the 3 primary hypotheses (superiority of pembrolizumab on OS in subjects with CPS ≥10, with CPS ≥1, and in all subjects), and the 2 secondary hypotheses (superiority of pembrolizumab on PFS and ORR in all subjects).

For OS endpoints, a Hwang-Shih-DeCani alpha-spending function with the gamma parameter (-4) will be used to construct group sequential boundaries to control the Type I error. Spending time will be plugged into the pre-specified alpha spending function to calculate alpha-spending. At the time of IA, for all 3 OS endpoints the spending time will be the minimum of the actual observed information fractions among these 3 OS endpoints, i.e., number of observed events at IA/number of planned events at FA. At the time of FA, the spending time will be 1 for all 3 OS endpoints. Of note, while the spending time used for alpha-spending calculation will be the same for these 3 OS endpoints at FA, the correlations

used for computing bounds for each endpoint will still be from that endpoint depending on the actual event counts. The rationale for the above strategy is to ensure both that full Type I error is spent at the final analysis without overspending at the interim. Justification for the spending time approach can be found in Anderson et.al [12].

The trial uses an extension of the graphical method of Maurer and Bretz [13] to provide strong multiplicity control for the multiple primary and secondary hypotheses.

Figure 1 shows the initial one-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. If H1 (OS in subjects with CPS ≥10) is successful, then the corresponding alpha can be reallocated to H2 (OS in subjects with CPS ≥1). If H2 is successful, the alpha for that hypothesis can be reallocated to H3 (OS in all subjects). If H3 is successful, the alpha for that hypothesis can be reallocated ½ to H4 (PFS in all subjects) and ½ to H5 (ORR in all subjects). The alpha for H4 and H5 can also be reallocated to each other, if the hypothesis was successful. Of note, H3 can only be formally tested once the null hypothesis for H2 has been rejected; H4 and H5 can only be formally tested once the null hypothesis for H3 has been rejected. See Figure 1 for the multiplicity strategy diagram of the study.

H4 (PFS in all subjects) will be tested at IA if H3 (OS in all subjects) is successful. If at FA additional alpha can be reallocated to H4 (e.g., H3 is not successful at IA but successful at FA, or H1 is not successful at IA but successful at FA and H2/H3 are successful at both IA and FA), then the p-value of PFS in all subjects from the IA will be compared to the updated alpha threshold at the time of FA. Similarly, H5 (ORR in all subjects) will be tested at IA if H3 (OS in all subjects) is successful. If at FA additional alpha can be reallocated to H5, then the p-value of ORR in all subjects from the IA will be compared to the updated alpha threshold at the time of FA.

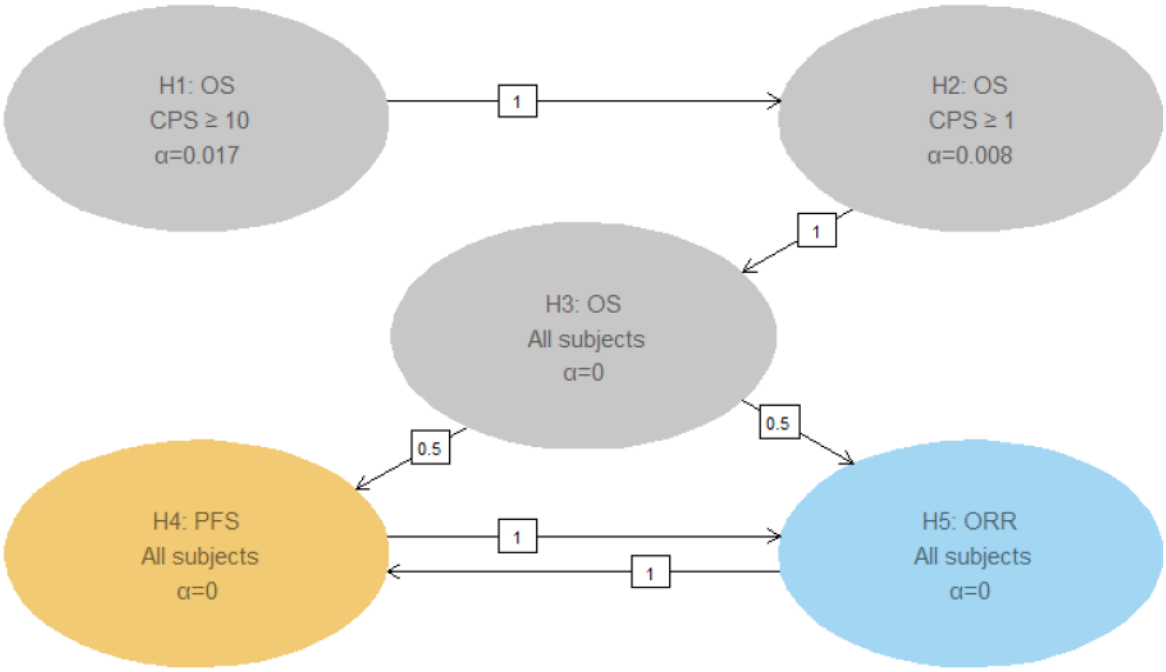

Figure 1 Multiplicity Strategy

3.9 Sample Size and Power Calculations

The study will randomize subjects in a 1:1 ratio between the pembrolizumab arm and the TPC arm. The overall sample size will be up to ~600.

Randomization will occur centrally using an interactive response system / integrated web response system (IVRS/IWRS) and will be monitored on a regular basis. When IVRS alerts that the study is approaching the desired enrollment, screening should be stopped in time. However, subjects already in screening phase may be enrolled even after we have reached the maximum sample size. The sample size is driven by the OS events in subjects with CPS ≥10, subjects with CPS ≥1, and all subjects.

Overall survival analysis in subjects with CPS ≥10: For the primary endpoint OS in subjects with CPS ≥10, with 154 OS events at a one-sided 1.7% alpha-level the trial has approximately 85% power to demonstrate that pembrolizumab is superior to TPC, if the underlying hazard ratio of OS is 0.60. Success boundary for OS at the final analysis approximately corresponds to an observed hazard ratio of ~0.70 (~4.2 month improvement over OS of 10 months in TPC).

Overall survival analysis in subjects with CPS ≥1: For the primary endpoint OS in subjects with CPS ≥1, with 334 OS events at a one-sided 0.8% (2.5%) alpha-level the trial has approximately 80% (90%) power to demonstrate that pembrolizumab is superior to TPC, if the underlying hazard ratio of OS is 0.70. With a one-sided 0.8% (2.5%) alpha-level, success boundary for OS at the final analysis approximately corresponds to an observed hazard ratio of ~0.76 (0.80) (~3.1 [2.5] month improvement over OS of 10 months in TPC).

Overall survival analysis in all subjects: For the primary endpoint OS in all subjects, with 520 OS events and a one-sided 0.8% (2.5%) alpha-level the trial has approximately 66% (80%) power to demonstrate that pembrolizumab is superior to TPC if the underlying hazard ratio of OS is 0.78. With a one-sided 0.8% (2.5%) alpha-level, success boundary for OS at the final analysis corresponds approximately to an observed hazard ratio of ~0.80 (0.84) (~2.4 [1.9] month improvement over OS of 10 months in TPC).

The sample size and power calculation for OS endpoints is based on the following assumptions for all 3 populations: 1) OS follows an exponential distribution with a median of 10 months in the control arm; 2) An enrollment period of 16.6 months and a minimum of 24 months follow-up after enrollment completion; 3) A yearly OS dropout rate of 1%.

PFS analysis in all subjects: This PFS power is calculated based on the following assumptions with approximately 575 PFS events: 1) the hypothesis of OS in all subject is supported and a total of 0.4% alpha is allocated to PFS hypothesis (H4); 2) PFS follows an exponential distribution with a median of 3 months in the TPC arm; 3) An enrollment period of 16.6 months; 4) A yearly PFS drop-out rate of 8%. The trial has ~99% power to demonstrate that pembrolizumab is superior to TPC on PFS at a one-sided 0.4% alpha-level, if the underlying hazard ratio of PFS is 0.6. If the underlying hazard ratio of PFS is 0.7, then the trial has ~95% power.

ORR analysis in all subjects: The ORR power calculation is based on the following assumptions: 1) the hypothesis of OS in subjects with CPS ≥1 is supported and a total of 0.4% alpha is allocated to ORR hypothesis (H5); 2) the underlying ORR is 10% in the TPC arm, and there is 10% increase in pembrolizumab arm (ORR of 20%) in all subjects. The trial has approximately 81% power to demonstrate that pembrolizumab is superior to TPC on ORR in all subjects at a one-sided 0.4% alpha-level.

The above statistical assumptions are based on our current understanding of the prevalence of the PD-L1 biomarker in TNBC (approximately 65% for CPS ≥10 and 31% for CPS ≥1), and it is subject to modification as needed based on emerging external data on PD-L1 prevalence in TNBC, or the correlation between PD-L1 expression and treatment effect. Any modification if occurs will be described in the supplemental SAP.

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

3.10 Subgroup Analyses and Effect of Baseline Factors

To determine whether the treatment effect is consistent across various subgroups, the estimate of the between-group treatment effect (with a nominal 95% CI) for the primary endpoint will be estimated and plotted within each category of the following classification variables in subjects with CPS ≥10 and CPS ≥1 and in all subjects:

- Age Category(<65 vs. ≥65 years)
- Menopausal status (for females only; pre- vs. post-menopausal)

- Geographic region (Europe/Israel/North America/Australia vs. Asia vs. Rest of World)
- Ethnic origin (Hispanic vs. Non-Hispanic)
- Number of prior lines of treatments in the metastatic setting (ie one vs. two)
- Time to progression on first-line (1L) therapy (<6 months vs. ≥6 months)
- TPC
- PD-L1 status by CPS cutoff (CPS ≥1 vs CPS <1; CPS ≥10 vs CPS <10). Note subgroup analysis by PD-L1 status will only be conducted in all subjects.
- Prior (neo)adjuvant therapy vs. de novo metastatic disease at initial diagnosis.

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

The extent of exposure will be summarized as duration of treatment in number of cycles or administrations as appropriate.

4 REFERENCES

[1] J. R. Tsiatis and A. Tsiatis, "Correcting for non-compliance in randomized trials using rank preserving structural failure time models," *Communications in Statistics-Theory and Methods,* vol. 20, no. 8, pp. 2609-0631, 1991.

[2] N. R. Latimer and K. R. Abrams, "NICE DSU Technical Support Document 16: Adjusting Survival Time Estimates in the Presence of Treatment Switching," 2014.

[3] T. G. Karrison, "Versatile tests for comparing survival curves based on weighted log-rank statistics," *The Stata Journal,* vol. 16, no. 3, p. 678–690, 2016.

[4] H. Uno, B. Claggett, L. Tian, E. Inoue, P. Gallo and T. Miyata, "Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis," Journal *of Clinical Oncology, vol.* 32, no. 22, pp. 2380-2385, 2014.

[5] P. Odell, K. Anderson and W. Kannel, "New models for predicting cardiovascular events," Journal *of Clinical Epidemiology, vol.* 47, no. 6, pp. 583-592, 1994.

[6] D. Mehrotra, S. Su and X. Li, "An efficient alternative to the stratified Cox model analysis," Statistics *in Medicine, vol.* 31, no. 17, pp. 1849-1856, 2012.

[7] D. Finkelstein, "A proportional hazards model for interval-censored failure time data," *Biometrics,* vol. 42, pp. 845-854, 1986.

[8] M. van Reenen and M. Oppe, EQ-5D-3L User Guide V5.1, 2015.

[9] P. Fayers, N. Aaronson, K. Bjordal, M. Groenvold, D. Curran and A. Bottomley, EORTC QLQ-C30 Scoring Manual (3rd edition), Brussels: EORTC, 2001.

[10] K. Liang and S. Zeger, "Longitudinal data analysis of continuous and discrete responses for pre-post designs," *Sankhyā: The Indian Journal of Statistics,* no. 62 (Series B), pp. 134-148, 2000.

[11] O. Miettinen and M. Nurminen, "Comparative analysis of two rates," Statistics in Medicine, no. 4, pp. 213-226, 1985.

[12] M. Anderson, "Application of graphical multiplicity methods in complex group sequential trials*," 2017 Submitted.*

[13] W. Maurer and F. Bretz, "Multiple Testing in Group Sequential Trials using Graphical Approaches," Statistics *in Biopharmaceutical Research, vol.* 5, no. 4, pp. 311-320, 2013.