Official Protocol Title:	A Randomized, Double-Blind, Phase III Study of Pembrolizumab (MK-3475) plus Chemotherapy vs Placebo plus Chemotherapy for Previously Untreated Locally Recurrent Inoperable or Metastatic Triple Negative Breast Cancer – (KEYNOTE-355)
NCT number:	NCT02819518
Document Date:	14 Jan 2020

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

This sSAP aligns with the protocol amendment 05 (MK-3475-355-05) for the global study with regard to the statistical analysis plan. In addition, the following changes were made to the sSAP which were not directly related to changes required due to this protocol amendment:

• Added max-combo test as a sensitivity analysis for overall survival (OS) (Section 3.6.1.2).

• Added sensitivity analysis to evaluate efficacy endpoints by PD-L1 cutoffs (Section 3.6.1).

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 Section 3.2– Responsibility for Analyses/In-House Blinding through Section 3.12– Extent of Exposure.

Study Design Overview	A Randomized, Double-Blind, Phase III Study of Pembrolizumab (MK-3475) plus Chemotherapy vs Placebo plus Chemotherapy for Previously Untreated Locally Recurrent Inoperable or Metastatic Triple Negative Breast Cancer
Treatment Assignment	Part 1: Approximately 30 subjects will be partially-randomized (unblinded open-label) among 3 treatment arms: (1) pembrolizumab + nab-paclitaxel, (2) pembrolizumab + paclitaxel and (3) pembrolizumab + gemcitabine/carboplatin. Part 2: Approximately 828 subjects will be randomized (double-blind) in a 2:1 ratio between 2 treatment arms: (1) pembrolizumab + chemotherapy and (2) placebo + chemotherapy. Stratification factors are as follows: 1. Chemotherapy on study (taxane [i.e., paclitaxel or nab-paclitaxel] vs gemcitabine/carboplatin). 2. Tumor PD-L1 status (CPS ≥1 vs CPS <1). 3. Prior treatment with same class of chemotherapy in the (neo)adjuvant setting (yes vs no).
Analysis Populations	Part 1 and Part 2 subjects will be analyzed separately. Part 1 Efficacy: All Subjects as Treated (ASaT) Part 2 Efficacy: Intention-to-Treat Population (ITT); ORR ITT Population for ORR endpoints. Safety: All Subjects as Treated (ASaT)
Primary Endpoint(s)	Part 1: Safety and tolerability. Part 2: 1. Progression-free survival (PFS) based on Response Evaluation Criteria in Solid Tumors Version 1.1 (RECIST 1.1) as assessed by a blinded central imaging vendor (CIV) in all subjects. 2. PFS based on RECIST 1.1 as assessed by a blinded CIV in subjects with PD-L1 positive tumors. (CPS ≥1).

	3. PFS based on RECIST 1.1 as assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥10). 4. Overall survival (OS) in all subjects. 5. OS in subjects with PD-L1 positive tumors (CPS ≥1). 6. OS in subjects with PD-L1 positive tumors (CPS ≥10).
Key Secondary Endpoint(s)	Part 2: Objective response rate (ORR) based on RECIST 1.1 as assessed by a blinded CIV in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1).
Statistical Methods for Key Efficacy Analyses	Part 2: The primary hypotheses will be evaluated by comparing pembrolizumab + chemotherapy vs placebo + chemotherapy in PFS and OS using a stratified log-rank test. The hazard ratio (HR) will be estimated using a stratified Cox model. The key secondary hypotheses of ORR will be evaluated by comparing pembrolizumab + chemotherapy vs placebo + chemotherapy in ORR using a stratified Miettinen and Nurminen method.
Statistical Methods for Key Safety Analyses	Part 1: Descriptive summary statistics will be provided for safety endpoints by treatment as appropriate. Part 2: The analysis of safety will follow a tiered approach. There is no Tier 1 safety endpoint for this trial. Point estimates and 95% confidence intervals (CIs) for between-treatment comparisons via the Miettinen and Nurminen method will be provided for Tier 2 safety endpoints; only point estimates by treatment group will be provided for Tier 3 safety endpoints.
Interim Analyses	One safety interim analysis for Part 1 and 3 efficacy interim analyses for Part 2 will be performed. Results will be reviewed by an external DMC. Details are provided in Section 3.7– Interim Analyses. Part 1 – Safety Interim Analysis: ~ 3 months after first subject randomized. o Timing: after all Part 1 subjects have completed the first 21 or 28 days (depending on chemotherapy treatment) of study treatment. o Primary purpose: interim safety evaluation. Part 2 – Efficacy Interim and Final Analyses • Interim analysis 1 (IA1): ~ 9 months after first 640 Part 2 subjects are randomized. o Primary purpose: final ORR analysis, interim PFS and interim OS analysis. • Interim analysis 2 (IA2): after ~ 185 OS events among subjects with CPS ≥10 have been observed. o Primary purpose: interim OS analysis and final PFS analysis. • Interim analysis 3 (IA3): after ~ 210 OS events among subjects with CPS ≥10 have been observed. o Primary purpose: interim OS analysis. • Final analysis (FA): after ~ 664 OS events among all subjects, ~ 482 OS events among subjects with CPS ≥1, and ~ 240 OS events among subjects with CPS ≥10 have been observed. o Primary purpose: final OS analysis.
Multiplicity	Part 1: Multiplicity adjustment not applicable. Part 2: The family-wise type-I error rate over the 6 primary hypotheses and the 2 secondary hypotheses will be strongly controlled at 2.5% (one-sided) with 0.5% allocated to PFS, 1.8% allocated to OS, and 0.2% allocated to ORR hypotheses. An extension [1] of the graphical approach of Maurer and Bretz [2] will be applied to re-allocate alpha between PFS, OS and ORR hypotheses. The Spiessens and Debois method [3] will be used to adjust the nominal alphas in ORR between all subjects and subjects with CPS ≥1. Group sequential methods will be used to allocate alpha between the interim and final analyses for OS endpoints.

Sample Size and Power	Part 1: Approximately **30** subjects will be enrolled. Part 2: It is expected that ~ 664 OS events among all subjects, ~ 482 OS events among subjects with CPS ≥1, and ~ 240 OS events among subjects with CPS ≥10 have been observed at the FA. The planned sample size is approximately **828** subjects. (1) PFS in all subjects: at IA2 the analysis has ~ 89% power at a one-sided 0.111% alpha level, if the true HR is 0.70. At IA2, with ~ 634 events the HR at boundary for success is ~ 0.77 (~ 1.6 months improvement over control median PFS of 5.5 months). At IA2, PFS in all subjects can only be tested if both hypotheses of PFS in subjects with CPS ≥10 and PFS in subjects with CPS ≥1 are supported. (2) PFS in subjects with CPS ≥1: at IA2 the analysis has ~ 97% power at a one-sided 0.111% alpha level, if the true HR is 0.62. At IA2, with ~ 463 events the HR at boundary for success is ~ 0.74 (~ 1.9 months improvement over control median PFS of 5.5 months). At IA2, PFS in all subjects with CPS ≥1 can only be tested if the hypothesis of PFS in subjects with CPS ≥10 is supported. (3) PFS in subjects with CPS ≥10: at IA2 the analysis has ~ 86% power at a one-sided 0.411% alpha level, if the true HR is 0.60. At IA2, with ~ 235 events the HR at boundary for success is ~ 0.69 (~ 2.4 months improvement over control median PFS of 5.5 months). (4) OS in all subjects: the trial has ~ 60% power at a one-sided 0.75% alpha level, if the true HR is 0.80. With ~ 664 events, the HR at boundary for success at FA is ~ 0.81 (~ 4.0 months improvement over control median OS of 17.5 months). After IA1, OS in all subjects can be tested if hypothesis of OS in subjects with CPS ≥1 is supported. (5) OS in subjects with CPS ≥1: the trial has ~ 87% power at a one-sided 0.75% alpha level, if the true HR is 0.71. With ~ 482 events, the HR at boundary for success at FA is ~ 0.78 (~ 4.8 months improvement over control median OS of 17.5 months). (6) OS in subjects with CPS ≥10: the trial has ~ 79% power at a one-sided 1.011% alpha level, if the true HR is 0.65. With ~ 240 events, the HR at boundary for success at FA is ~ 0.72 (~ 6.8 months improvement over control median OS of 17.5 months).

3.2 Responsibility for Analyses/In-House Blinding

The statistical analysis of the data obtained from this study will be the responsibility of the Clinical Biostatistics department of the Sponsor.

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

The Investigators, other study site staff, and subjects will be blinded to subject-level PD-L1 biomarker results. Analysis or summaries generated by PD-L1 status will be limited and documented.

Part 1:

Part 1 of this study is being conducted as a partially randomized, open-label study, i.e., subjects, investigators, and Sponsor personnel will be aware of subject treatment assignments after each subject is enrolled and treatment is assigned.

Planned safety interim analysis is described in Section 3.7 – Interim Analyses. Study enrollment is likely to be ongoing at the time of the safety interim analysis (i.e., no planned enrollment pause between Part 1 and Part 2).

The external DMC will serve as the primary reviewer of the results of the safety interim analysis and will make recommendations for discontinuation of the study or modification to an EOC of the Sponsor. Additional logistical details, revisions to the above plan and data monitoring guidance will be provided in the DMC Charter. Key aspects of the interim analysis are described in Section 3.7– Interim Analyses.

Part 2:

Part 2 of this study will be conducted as a double-blind, Phase III study part under in-house blinding procedures. The official, final database of Part 2 will not be unblinded until medical/scientific review has been performed, protocol deviations have been identified, and data have been declared final and complete. In addition, the independent radiologist(s) will perform the central imaging review without knowledge of treatment group assignment. Additional details regarding trial blinding/unblinding including unblinding required for operational purposes (e.g., unblinded pharmacist) are described in Protocol Section 5.2.3 – Trial Blinding.

Planned efficacy interim analyses are described in Section 3.7 – Interim Analyses. Study enrollment is likely to be ongoing at the time of any efficacy interim analyses. Blinding to treatment assignment will be maintained at all investigational sites.

Treatment-level results of the efficacy interim analyses will be provided by an external unblinded statistician to the external DMC. The external DMC will serve as the primary reviewer of the results of the interim analyses and will make recommendations for discontinuation of the study or modification to an EOC of the Sponsor. Depending on the recommendation of the DMC, the Sponsor may prepare a regulatory submission. If the DMC recommends modifications to the design of the protocol or discontinuation of the study, this EOC may be unblinded to results at the treatment level in order to act on these recommendations or facilitate regulatory filing. Limited additional Sponsor personnel may be unblinded to the treatment level results of the interim analysis (analyses), if required, in order to act on the recommendations of the DMC or facilitate regulatory filing. The extent to which individuals are unblinded with respect to results of interim analyses will be documented. Additional logistical details, revisions to the above plan and data monitoring guidance will be provided in the DMC Charter. Key aspects of the interim analyses are described in Section 3.7 – Interim Analyses.

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.

3.3 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)

Part 1 (Safety Run-In):

(1) **Objective**: To evaluate the safety and tolerability of 3 pembrolizumab + chemotherapy combinations, namely, pembrolizumab + paclitaxel, pembrolizumab + nab-paclitaxel, and pembrolizumab + gemcitabine/carboplatin.

Part 2 (Phase III study):

The combination of pembrolizumab and chemotherapy will be compared to placebo and chemotherapy for the treatment of previously untreated locally recurrent inoperable or metastatic centrally confirmed triple negative breast cancer (TNBC):

(1) **Objective:** To compare progression-free survival (PFS) based on Response Evaluation Criteria in Solid Tumors Version 1.1 (RECIST 1.1) as assessed by a blinded central imaging vendor (CIV) in all subjects.

Hypothesis: The combination of pembrolizumab and chemotherapy prolongs PFS compared to placebo and chemotherapy in all subjects.

(2) **Objective:** To compare PFS based on RECIST 1.1 as assessed by a blinded CIV in subjects with programmed cell death ligand 1 (PD-L1) positive tumors (combined positive score [CPS] ≥1).

Hypothesis: The combination of pembrolizumab and chemotherapy prolongs PFS compared to placebo and chemotherapy in subjects with PD-L1 positive tumors (CPS ≥1).

(3) **Objective:** To compare PFS based on RECIST 1.1 as assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥10).

Hypothesis: The combination of pembrolizumab and chemotherapy prolongs PFS compared to placebo and chemotherapy in subjects with PD-L1 positive tumors (CPS ≥10).

(4) **Objective:** To compare overall survival (OS) in all subjects.

Hypothesis: The combination of pembrolizumab and chemotherapy prolongs OS compared to placebo and chemotherapy in all subjects.

(5) **Objective:** To compare OS in subjects with PD-L1 positive tumors (CPS ≥1).

Hypothesis: The combination of pembrolizumab and chemotherapy prolongs OS compared to placebo and chemotherapy in subjects with PD-L1 positive tumors (CPS ≥1).

(6) **Objective:** To compare OS in subjects with PD-L1 positive tumors (CPS ≥10).

Hypothesis: The combination of pembrolizumab and chemotherapy prolongs OS compared to placebo and chemotherapy in subjects with PD-L1 positive tumors (CPS ≥10).

The study is considered to have met its primary objective if the combination of pembrolizumab and chemotherapy is superior to placebo and chemotherapy in either PFS or OS in either all subjects or in subjects with PD-L1 positive tumors (CPS ≥1 or CPS ≥10) at either an interim analysis or the final analysis (OS only).

3.3.2 Secondary Objective(s) & Hypothesis(es)

Part 2 (Phase III study):

For comparisons, the combination of pembrolizumab and chemotherapy will be compared to placebo and chemotherapy for the treatment of previously untreated locally recurrent inoperable or metastatic centrally confirmed TNBC:

(1) **Objective**: To compare objective response rate (ORR) based on RECIST 1.1 as assessed by a blinded CIV in all subjects.

Hypothesis: The combination of pembrolizumab and chemotherapy increases ORR compared to placebo and chemotherapy in all subjects.

(2) **Objective**: To compare ORR based on RECIST 1.1 as assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥1).

Hypothesis: The combination of pembrolizumab and chemotherapy increases ORR compared to placebo and chemotherapy in subjects with PD-L1 positive tumors (CPS ≥1).

(3) **Objective**: To compare ORR based on RECIST 1.1 as assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥10).

(4) **Objective**: To evaluate duration of response (DOR) based on RECIST 1.1 as assessed by a blinded CIV in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1 and CPS ≥10).

(5) **Objective**: To compare disease control rate (DCR) based on RECIST 1.1 as assessed by a blinded CIV in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1 and CPS ≥10).

(6) **Objective**: To evaluate the safety and tolerability of 3 pembrolizumab + chemotherapy combinations.

(7) **Objective:** To evaluate changes in health-related quality-of-life (QoL) assessments from baseline in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1 and CPS ≥10) using the European Organization for Research and Treatment of Cancer (EORTC) Quality of Life Questionnaire Core 30 (EORTC QLQ-C30) and EORTC Breast Cancer–Specific Quality of Life Questionnaire (EORTC QLQ-BR23).

3.3.3 Exploratory Objectives

Part 2 (Phase III study):

For comparisons, the combination of pembrolizumab and chemotherapy will be compared to placebo and chemotherapy for the treatment of previously untreated locally recurrent inoperable or metastatic centrally confirmed TNBC:

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

(2) **Objective:** To investigate association(s) between anti-tumor activity of study treatments and efficacy/resistance biomarkers, utilizing tumor and blood specimens obtained before randomization, during treatment, and at disease progression.

(3) **Objective:** To identify molecular (genomic, metabolic, and/or proteomic) determinants of response or resistance to pembrolizumab and other treatments in this study, so as to define novel predictive and pharmacodynamic biomarkers and understand the mechanism of action of pembrolizumab.

3.4 Analysis Endpoints

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

3.4.1 Efficacy Endpoints

Primary

Progression-free survival (PFS) – based on RECIST 1.1 as assessed by a CIV

Progression-free survival is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on assessments by a CIV or death due to any cause, whichever occurs first. See Section 3.6.1 – Statistical Methods for Efficacy Analyses for the definition of censoring.

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

Objective Response Rate (ORR) – based on RECIST 1.1 as assessed by a CIV

Objective response rate is defined as the proportion of the subjects in the analysis population who have a CR or PR. Responses are based on assessments by a CIV per RECIST 1.1.

Duration of Overall Response (DOR) – based on RECIST 1.1 as assessed by a CIV

For subjects who demonstrate CR or PR, duration of response is defined as the time from first documented evidence of CR or PR until disease progression or death due to any cause, whichever occurs first, based on assessments by a CIV per RECIST 1.1. See Section 3.6.1 – Statistical Methods for Efficacy Analyses for the definition of censoring.

Disease Control Rate (DCR) – based on RECIST 1.1 as assessed by a CIV

Disease control rate is defined as the percentage of subjects who have achieved CR or PR or have demonstrated SD for at least 24 weeks, based on assessments by a CIV per RECIST 1.1.

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 of Part 2. All randomized subjects will be included in this population. Subjects will be included in the treatment group to which they are randomized. For the primary analysis of ORR endpoints, all subjects randomized on or prior to the date the 640th Part 2 subject is randomized will be included in the analysis, which is considered the ORR ITT population. The ORR endpoints will also be analyzed in the ITT population for estimation only. The analysis of DOR and DCR will be conducted in both the ORR ITT and the ITT population.

Part 1 subjects will be excluded from all Part 2 efficacy analyses and therefore will not contribute to the analyses to address the primary/secondary objectives of Part 2. The All Subjects as Treated (ASaT) population will be used for the analysis of Part 1 efficacy 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 Part 1 efficacy 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 treatment for one cycle, but receives the correct treatment for all other cycles, will be analyzed according to the correct treatment group.

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 treatment 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.

Part 1 and Part 2 subjects will be analyzed separately.

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

For Part 1, descriptive summaries will be provided for efficacy endpoints (e.g., PFS, OS, ORR, DOR, DCR) as appropriate.

The efficacy analysis methods specified in this section apply to Part 2. Part 1 subjects will be excluded from all Part 2 efficacy analyses and therefore will not contribute to the analyses to address the primary/secondary objectives of Part 2.

Efficacy results that will be deemed to be statistically significant after consideration of the Type I error control strategy are described in Section 3.8 – Multiplicity. Nominal p-values 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, all stratification factors [chemotherapy on study (taxane vs gemcitabine/carboplatin); tumor PD-L1 status (positive vs negative); prior treatment with same class of chemotherapy in the (neo)adjuvant setting (yes vs no)] will be included. For efficacy analyses performed in subjects with PD-L1 positive tumors, only the stratification factors of chemotherapy on study (taxane vs gemcitabine/carboplatin) and prior treatment with same class of chemotherapy in the (neo)adjuvant setting (yes vs no) will be considered in the analysis models, i.e., not considering the stratification factor of PD-L1 tumor status. If based on blinded data review, there is too small number (<10) of subjects in a specific stratum for analyses in all subjects or subjects with PD-L1 positive tumors, in the ITT or the ORR ITT population, that stratum will be combined with the neighboring stratum for the analysis. 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 collected.

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 (i.e., HR) between the treatment arms. The HR and its 95% confidence interval (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 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, PD can occur any time in the time interval between the last assessment in which PD was not documented and the assessment when PD is

documented. For 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 a CIV. 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.

In order to evaluate the robustness of the PFS endpoint based on RECIST 1.1 as assessed by a CIV, one primary and two sensitivity analyses with a different set of censoring rules will be performed. 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 intention-to-treat principle. That is, PDs/deaths are counted as events regardless of missed study visits or initiation of new anti-cancer therapy. The second sensitivity analysis considers initiation of new anticancer treatment or discontinuation of treatment due to reasons other than complete response to be a PD event for subjects without documented PD or death. The censoring rules for primary and sensitivity analyses are summarized in Table 1. If a subject meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied.

Table 1 Censoring Rules for Primary and Sensitivity Analyses of PFS

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

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.

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

Kaplan-Meier plots among different PD-L1 cutoff points may be provided for each treatment arm separately, in order to estimate the influence of patients' CPS scores on PFS within each arm.

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 (i.e., 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 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. The Kaplan-Meier estimates of the OS rate at selected time points of interest (e.g., 6 months, 12 months, 18 months etc.) will also be estimated.

Subjects in the placebo + chemotherapy arm are expected to discontinue treatment earlier compared to subjects in the pembrolizumab + chemotherapy arm and are not allowed to crossover to the pembrolizumab + chemotherapy arm; however, they may be treated with another anti–PD-1 drug following the verification of PD by a blinded CIV. As an exploratory analysis, 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 [8], two-stage model [9]), 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.

Due to multiple occurrences of delayed separation phenomena observed in Immuno-Oncology randomized clinical trials, another sensitivity analysis, which evaluates the treatment difference for OS using the stratified max-combo test, may be conducted. 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 [10] will be reported. No formal hypothesis testing will be conducted.

Kaplan-Meier plots among different PD-L1 cutoff points may be provided for each treatment arm separately, in order to estimate the influence of patients' CPS scores on OS within each arm.

Additional supportive unstratified analyses may also be provided. Other sensitivity analyses described for the PFS endpoint may also be applied to the OS endpoint as appropriate.

3.6.1.3 Objective Response Rate (ORR)

The stratified Miettinen and Nurminen 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 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. Sensitivity analyses using the Cochran–Mantel–Haenszel test (CMH) may also be performed. Additional supportive unstratified analyses may also be provided.

3.6.1.4 Disease Control Rate (DCR)

The stratified Miettinen and Nurminen 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 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.

A corresponding summary of the reasons for which the responding subjects are censored for the DOR analysis will also be provided. Sensitivity analyses will be performed for DOR based on site investigator/local radiology review.

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 (Part 2)

Endpoint/Variable (Description, Time Point)	Statistical Method[a]	Analysis Population	Missing Data Approach
Primary Hypothesis 1			
PFS based on RECIST 1.1 assessed by a blinded CIV 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)
Primary Hypothesis 2			
PFS based on RECIST 1.1 assessed by a blinded CIV in subjects with PD-L1 positive tumors (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)
Primary Hypothesis 3			
PFS based on RECIST 1.1 assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥10)	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)

Endpoint/Variable (Description, Time Point)	Statistical Method[a]	Analysis Population	Missing Data Approach
Primary Hypothesis 4			
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
Primary Hypothesis 5			
OS in subjects with PD-L1 positive tumors (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 6			
OS in subjects with PD-L1 positive tumors (CPS ≥10)	Test: Stratified log-rank test Estimation: Stratified Cox model with Efron's tie handling method	ITT	Censored at last known alive date
Key Secondary Hypothesis 1 (Hypothesis 7)			
ORR based on RECIST 1.1 assessed by a blinded CIV in all subjects	Stratified M & N method[b]	The first ~ 640 subjects randomized in Part 2 (a subset of ITT)	Subjects with relevant data missing are considered non-responders
Key Secondary Hypothesis 2 (Hypothesis 8)			
ORR based on RECIST 1.1 assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥1)	Stratified M & N method[b]	The first ~ 640 subjects randomized in Part 2 (a subset of ITT)	Subjects with relevant data missing are considered non-responders
Other Secondary Endpoints			
ORR based on RECIST 1.1 assessed by a blinded CIV in subjects with PD-L1 positive tumors (CPS ≥10)	Stratified M & N method[b]	ITT	Subjects with relevant data missing are considered non-responders
DCR based on RECIST 1.1 assessed by a blinded CIV in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1 and CPS ≥10)	Stratified M & N method[b]	ITT	Subjects with relevant data missing are considered non-responders
DOR based on RECIST 1.1 assessed by a blinded CIV in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1 and CPS ≥10)	Summary statistics using Kaplan-Meier method	All responders in ITT	See Table 2

CIV=central imaging vendor; CPS=combined positive score; DCR=disease control rate; DOR=duration of response; ITT=intention-to-treat; M & N=Miettinen and Nurminen; ORR=objective response rate; OS=overall survival; PD-L1=programmed cell death ligand 1; PFS=progression-free survival; RECIST 1.1= Response Evaluation Criteria in Solid Tumors version 1.1.

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

[b] 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 [11]. 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:

   o The QLQ-C30 functional scale Physical Functioning.

   o The QLQ-C30 functional scale Emotional Functioning.

3. Time to deterioration (TTD), defined as time from start of treatment to first onset of 10 points or more worsening from baseline, for

   o The QLQ-C30 global health status/QoL score.

   o The QLQ-C30 functional scale Physical Functioning.

   o The QLQ-C30 functional scale Emotional Functioning.

4. The mean score changes from baseline to the primary analysis time point for other items/scales of QLQ-C30 and QLQ-BR23 as appropriate.

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 PRO analysis will be conducted in Part 2 subjects only. The analysis will be conducted in all subjects in subjects with CPS≥1 and CPS≥10.

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 [12], 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 [11] 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 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 [13]. 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 both treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t, X_i is the stratification stratum value (Protocol Section 5.4) for subject i, and β is the corresponding coefficient for stratum.

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 (e.g. mean and standard error) of change from baseline with no imputation for missing data of the following score/scales will also be plotted: QLQ-C30 global health status/QoL, QLQ-C30 physical functioning, QLQ-C30 emotional functioning, and EQ-5D VAS.

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, QLQ-C30 functional scales physical functioning and emotional functioning.

Of note, for all above ePRO analysis (Section 3.6.2.3.2 to Section 3.6.2.3.3) in subjects with PD-L1 positive tumors, only the stratification factors of chemotherapy on study (taxane vs gemcitabine/carboplatin) and prior treatment with same class of chemotherapy in the (neo)adjuvant setting (yes vs no) will be considered in the analysis models, i.e., not considering the stratification factor of PD-L1 tumor status.

3.6.2.3.4 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.

Part 1

Descriptive summary statistics (e.g., counts, percentage, mean, standard deviation) will be provided for safety endpoints by treatment for Part 1 as appropriate.

Part 2

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) will be classified as belonging to "Tier 2" or "Tier 3", based on the number or percent of subjects with events observed. Specific AEs occurring in ≥ 5% of subjects in one or more treatment groups will be considered Tier 2 endpoints. Specific Serious and Grade 3-5 AEs occurring in at least 8 subjects in the pembrolizumab + chemotherapy group, or at least 2 subjects in the placebo + chemotherapy group will also be considered Tier 2 endpoints. All other adverse experiences and predefined limits of change will belong to Tier 3.

The threshold of at least 5% of subjects with events in one or more treatment groups was chosen for specific AEs as Tier 2 endpoints because this incidence rate would allow meaningful statistical assessments for AEs in general. The threshold of at least 8 subjects in the pembrolizumab + chemotherapy group, or at least 2 subjects in the placebo + chemotherapy group was chosen for specific Serious and Grade 3-5 AEs because the 95% CI for the between-group difference in percent incidence will always include zero with 2 to 1 randomization ratio if there are less than 8 subjects with events in the treatment group and less than 2 subjects with events in the control group, and thus would add little to the interpretation of potentially meaningful differences. Serious and Grade 3-5 AEs are expected to occur less frequently but important for the overall safety assessment, as such the threshold to classify these AEs as Tier 2 endpoints are lower than that for general specific AEs. 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 [14].

Table 5 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	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, SOCs (incidence ≥ 5% of subjects in one or more treatment groups)	X	X
	Specific Serious AEs, SOCs (incidence ≥ 8 subjects in the pembrolizumab + chemotherapy group, or ≥ 2 subjects in the placebo + chemotherapy group)	X	X
	Specific Grade 3-5 AEs, SOCs (incidence ≥ 8 subjects in the pembrolizumab + chemotherapy group, or ≥ 2 subjects in the placebo + chemotherapy group)	X	X
Tier 3	Specific AEs, SOCs (incidence < 5% of subjects in both treatment groups) or PDLCs		X
	Specific Serious AEs, SOCs (incidence < 8 subjects in the pembrolizumab + chemotherapy group and < 2 subjects in the placebo + chemotherapy group)		X
	Specific Grade 3-5 AEs, SOCs (incidence <8 subjects in the pembrolizumab + chemotherapy group and < 2 subjects in the placebo + chemotherapy group)		X
	Change from Baseline Results (Labs, Vital Signs)		X
Note: SOC=System Organ Class; PDLC=Pre-Defined Limit of Change; 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 + chemotherapy 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

Part 1 and Part 2

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

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 PFS2 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

The study has one planned safety interim analysis for Part 1 and 3 planned efficacy interim analyses for Part 2. Results will be reviewed by the external DMC.

3.7.1 Part 1: Safety Interim Analysis

A safety interim analysis will be performed after all Part 1 subjects have completed the first 21 or 28 days (depending on chemotherapy treatment) of study treatment (unless early discontinued), i.e., 21 days after the first study treatment administration if the subject is in the pembrolizumab + gemcitabine/carboplatin arm or 28 days after the first study treatment administration if the subject is in either of the pembrolizumab + taxane arms. Interim safety data will be reviewed by the DMC. It is estimated that the safety interim analysis will occur approximately 3 months after the first subject is randomized (depending on enrollment rate).

In addition, continuous safety monitoring will be performed for Part 1 prior to the safety interim analysis by the study team. If a potential safety issue signal is observed before 10 subjects are enrolled in a treatment arm, the DMC will be notified to review data prior to the pre-specified safety interim analysis.

3.7.2 Part 2: Efficacy Interim Analyses

There are 3 planned efficacy interim analyses for Part 2 in this trial. Results will be reviewed by the DMC. Of note, the boundaries for the analyses in this section may be adjusted, as appropriate, using the graphical approach discussed in Section 3.8 – Multiplicity.

The boundaries provided in this section are calculated based on the estimated number of events at each analysis, and the actual interim boundaries will be determined using the actual observed and

the planned numbers of events at the time of interim analyses with the rule and spending functions specified in Section 3.8 – Multiplicity. The actual final boundaries will be adjusted accordingly. Any changes to the timing of the analyses, along with its rationale, will be documented in the sSAP or a memo to the study file before the database lock.

Of note, Protocol Amendment 05 occurred after the conduct of efficacy interim analysis 1 (IA1), and the following information regarding IA1 is based on the planned number of events and original multiplicity strategy specified in the protocol prior to Protocol Amendment 05.

Interim Analysis 1 (Final ORR, Interim PFS and Interim OS Analysis)

The primary purpose of efficacy IA1 is to perform the final ORR, interim PFS and interim OS analysis. The ORR analysis at IA1 is considered the final ORR analysis of the study. IA1 will be performed after: (1) enrollment is completed, and (2) ~ 9 months after the first 640 subjects randomized to Part 2.

At IA1, ORR analyses will be based on data from the first ~ 640 subjects randomized to Part 2 and be tested in 1) all subjects and, 2) subjects with CPS ≥1. All subjects randomized on or prior to the date the 640th Part 2 subject is randomized will be included in the ORR analysis.

The success boundary to demonstrate ORR superiority at IA1 approximately corresponds to an observed ORR difference of ~ 12.6 percentage points at $\alpha = 0.1\%$ (one-sided) for all subjects, and an observed ORR difference of ~ 14.1 percentage points at nominal $\alpha = \sim 0.145\%$ (one-sided) for subjects with CPS ≥1, if there are 640 subjects in all subjects and 480 subjects in subjects with CPS ≥1 available for analysis (assuming PD-L1 positivity prevalence CPS ≥1 of 75%), respectively.

The estimated boundaries for PFS and OS endpoints at IA1 are provided in Table 6, assuming the planned numbers of events are analyzed.

Interim Analysis 2 (Interim OS Analysis and Final PFS Analysis)

The primary purpose of efficacy interim analysis 2 (IA2) is to evaluate superiority of pembrolizumab + chemotherapy vs placebo + chemotherapy in OS, and to perform final PFS analysis. The analysis will be performed after ~ 185 OS events among subjects with CPS ≥10 have been observed. It is estimated that at IA2 ~ 523 OS events among all subjects and ~ 375 OS events among subjects with CPS ≥1 have been observed. The analysis may be delayed for up to 4 months if the planned number of OS events has not been reached in all subjects or in subjects with CPS ≥1. It is estimated that IA2 is expected to occur ~ 22 months after last subject randomized.

The estimated boundaries for PFS and OS endpoints at IA2 are provided in Table 6 and Table 7, assuming the planned numbers of events are analyzed.

Interim Analysis 3 (Interim OS Analysis)

The primary purpose of efficacy interim analysis 3 (IA3) is to evaluate superiority of pembrolizumab + chemotherapy vs placebo + chemotherapy in OS. The analysis will be performed after ~ 210 OS events among subjects with CPS ≥10 have been observed. It is estimated

at IA3 ~ 589 OS events among all subjects and ~ 424 OS events among subjects with CPS ≥1 have been observed. The analysis may be delayed for up to 4 months if the planned number of OS events has not been reached in all subjects or in subjects with CPS ≥1. It is estimated that IA3 is expected to occur ~ 30 months after last subject randomized.

The estimated boundaries for OS endpoints at IA3 are provided in Table 6 and Table 7, assuming the planned numbers of events are analyzed.

Final Analysis (Final OS Analysis)

The final analysis (FA) of the study is event driven and will be conducted after approximately ~ 664 OS events among all subjects, ~ 482 OS events among subjects with CPS ≥1, and ~ 240 OS events among subjects with CPS ≥10 have been observed. It is estimated that FA is expected to occur ~ 43 months after last subject randomized. If after 43 months after last subject randomized, the planned numbers of OS events still have not been observed, then the final OS analysis may be conducted at that time regardless. The success boundaries to demonstrate OS superiority at FA are presented in Table 6 and Table 7, if the planned numbers of OS events are analyzed.

Table 6 summarizes the timing, sample size and decision guidance of the 3 efficacy interim analyses and FA, assuming there is no alpha re-allocation among hypotheses. ORR boundaries are based on the assumptions of 640 randomized subjects and 75% PD-L1 CPS ≥1 prevalence, and may be updated at time of the analyses using the actual observed numbers. PFS and OS boundaries are based on planned number of events and may be updated at times of the analyses according to the actual observed number of events, spending functions, and the spending time approach as specified in Section 3.8 – Multiplicity.

Table 6 Summary of Timing, Sample Size and Decision Guidance of Efficacy Interim Analyses and Final Analysis (Part 2, at Initial Alpha)

Analysis	Criteria for Conduct of Analysis	Endpoint and Testing Population	Parameter	Efficacy Bar[a, b]
Interim Analysis 1: Final ORR, Interim PFS and Interim OS Analysis	IA1 occurred ~ 4 months after last subject randomized. IA1 was to be conducted when: (1) enrollment is completed, and (2) ~ 9 months after first 640 subjects are randomized in Part 2 It was estimated that at IA1: ~ 500 PFS events among all subjects, ~ 360 PFS events among subjects with CPS ≥1, ~ 260 OS events among all subjects, and ~ 185 OS events among subjects with CPS ≥1 have been observed.	ORR in all subjects	p-value (1-sided) at boundary ~ ORR difference at boundary	0.001 ~ 12.6 percentage points
		ORR in subjects with CPS ≥1	p-value (1-sided) at boundary ~ ORR difference at boundary	~ 0.00145[c] ~ 14.1 percentage points
		PFS in all subjects	p-value (1-sided) at boundary ~ HR at boundary	0.0005 ~ 0.73
		PFS in subjects with CPS ≥1	p-value (1-sided) at boundary ~ HR at boundary	0.0005 ~ 0.69

Analysis	Criteria for Conduct of Analysis	Endpoint and Testing Population	Parameter	Efficacy Bar[a, b]
		OS in all subjects	p-value (1-sided) at boundary ~ HR at boundary	0.0004 ~ 0.64
		OS in subjects with CPS ≥1	p-value (1-sided) at boundary ~ HR at boundary	0.0004 ~ 0.59
Interim Analysis 2: Interim OS Analysis/Final PFS Analysis	IA2 will be conducted after ~ 185 OS events among subjects with CPS ≥10 have been observed [d]. It is estimated that at IA2: ~ 523 OS events among all subjects, ~ 375 OS events among subjects with CPS ≥1, ~ 634 PFS events among all subjects, ~ 463 PFS events among subjects with CPS ≥1, and ~ 235 PFS events among subjects with CPS ≥10 have been observed.	PFS in subjects with CPS ≥10	p-value (1-sided) at boundary ~ HR at boundary	0.00411 ~ 0.69
		OS in subjects with CPS ≥1	p-value (1-sided) at boundary ~ HR at boundary	0.0022 ~ 0.73
		OS in subjects with CPS ≥10	p-value (1-sided) at boundary ~ HR at boundary	0.0034 ~ 0.66
Interim Analysis 3: Interim OS Analysis	IA3 will be conducted after ~ 210 OS events among subjects with CPS ≥10 have been observed [d]. It is estimated that at IA3: ~ 589 OS events among all subjects, ~ 424 OS events among subjects with CPS ≥1 have been observed.	OS in subjects with CPS ≥1	p-value (1-sided) at boundary ~ HR at boundary	0.0036 ~ 0.76
		OS in subjects with CPS ≥10	p-value (1-sided) at boundary ~ HR at boundary	0.0050 ~ 0.69
Final Analysis (FA): Final OS Analysis	FA will be conducted after [e]: ~ 664 OS events among all subjects, ~ 482 OS events among subjects with CPS ≥1, and ~ 240 OS events among subjects with CPS ≥10 have been observed.	OS in subjects with CPS ≥1	p-value (1-sided) at boundary ~ HR at boundary	0.0060 ~ 0.78
		OS in subjects with CPS ≥10	p-value (1-sided) at boundary ~ HR at boundary	0.0082 ~ 0.72

[a.] Efficacy bar represents boundary at which statistical significance supporting pembrolizumab + chemotherapy is superior to placebo + chemotherapy can be claimed.
[b.] Efficacy bars at IA1 are based on planned number of events and original multiplicity strategy specified in the protocol prior to Protocol Amendment 05. Efficacy bars at IA2/IA3/FA for OS in subjects with CPS ≥1 are based on the actual number of events at IA1 and planned numbers of events at IA2/IA3/FA.
[c.] Approximate nominal alpha based on the Spiessens and Debois method accounting for correlation between ORR in all subjects and ORR in subjects with CPS ≥1. The actual nominal alpha will be calculated based on the actual correlation between these two populations for testing.
[d.] IA2 and IA3 may be delayed for up to 4 months if the planned number of OS events in all subjects or in subjects with CPS ≥1 has not yet been reached.
[e.] FA may be conducted after 43 months post last patient randomized even if the planned numbers of events are not reached at that time.

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 hypotheses of PFS in all subjects, PFS in subjects with CPS ≥1, and OS in all subjects have initial alpha of 0% at IA2 (PFS) or IA2/IA3/FA (OS). As such, after IA1 they can only be tested after alpha re-allocation if relevant hypothesis(es) is supported. The efficacy decision guidance for these endpoints with respect to the re-allocated alpha from the support of other hypothesis(es) is summarized in Table 7 below (selected scenarios), assuming the planned numbers of events specified in Table 6 are available for analyses at each time point.

If an efficacy boundary is crossed at IA1 or IA2 for PFS, or at an interim analysis or the FA for OS, in either all subjects or subjects with CPS ≥1 or CPS ≥10, 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 to obtain mature OS data.

Of note, an assumption of 75% prevalence of PD-L1 CPS ≥1, and 38% prevalence of PD-L1 CPS ≥10 in mTNBC subjects were made in above calculations. The above timing, estimated event count, and criteria for interim and final analyses are subject to modification in the sSAP as needed based on emerging data on PD-L1 prevalence in mTNBC.

Table 7 Summary of Efficacy Decision Guidance after Alpha Re-Allocation (Part 2, Selected Scenarios)

Endpoint	Scenario	Total Alpha Allocated	Analysis	Efficacy Boundary (After Alpha Re-Allocation)	
				p-value (1-sided) at Boundary	Approx. HR at Boundary
H1: PFS in all subjects (IA2)	H2 and H3 supported	0.00111	IA2	0.00111	0.77
H2: PFS in subjects with CPS ≥1 (IA2)	H3 supported	0.00111	IA2	0.00111	0.74
H4: OS in all subjects (IA2/IA3/FA)	H5 supported	0.0075	IA2	0.0026	0.77
			IA3	0.0038	0.79
			FA	0.0060	0.81

3.8 Multiplicity

Part 1

Multiplicity adjustment is not applicable.

Part 2

The multiplicity strategy specified in this section will be applied to the 6 primary hypotheses and the 2 secondary hypotheses of Part 2: primary hypotheses of superiority of pembrolizumab + chemotherapy compared to placebo + chemotherapy in PFS and OS in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1 and CPS ≥10), and secondary hypotheses of superiority of

pembrolizumab + chemotherapy compared to placebo + chemotherapy in ORR in all subjects and in subjects with PD-L1 positive tumors (CPS ≥1).

Based on emerging biomarker data external to this study, the initial alpha allocation among the 6 primary hypotheses and 2 secondary hypotheses is revised in Protocol Amendment 05. The revision of the alpha allocation occurs after the conduct of efficacy IA1. The family-wise Type-I error rate for this study is strongly controlled at 2.5% (one-sided) across all 6 primary hypotheses on PFS and OS as well as 2 secondary hypotheses on ORR. Figure 1 displays the revised multiplicity strategy diagram of the study. The initial one-sided alpha allocation for each hypothesis is shown in the rectangle representing the hypothesis. The weights for re-allocation from each hypothesis to the others are represented in the numbers along the lines connecting hypotheses. Overall, a total of 0.5% alpha is allocated to PFS endpoints, a total of 1.8% alpha is allocated to OS endpoints, and a total of 0.2% alpha is allocated to ORR endpoints.

Table 8 also summarizes the revised initial alpha allocation before any alpha re-allocation.

Table 8 Initial Alpha Allocation

Hypothesis	Initial Alpha Allocation
H1: PFS in all subjects	• 0.043% allocated at IA1 (already spent at IA1). • 0% allocated at IA2.
H2: PFS in subjects with CPS ≥1	• 0.046% allocated at IA1 (already spent at IA1). • 0% allocated at IA2.
H3: PFS in subjects with CPS ≥10	• 0.411% allocated at IA2 only
H4: OS in all subjects	• 0.039% allocated at IA1 (already spent at IA1). • 0% allocated at IA2/IA3/FA (group sequential).
H5: OS in subjects with CPS ≥1	• 0.75% allocated to IA1/IA2/IA3/FA (group sequential), which includes 0.036% spent at IA1.
H6: OS in subjects with CPS ≥10	• 1.011% allocated to IA2/IA3/FA (group sequential).
H7: ORR in all subjects	• 0.1% allocated at IA1 only (already spent at IA1).
H8: ORR in subjects with CPS ≥1	• 0.1% allocated at IA1 only (already spent at IA1).

An extension [1] of the graphical approach of Maurer and Bretz [2] will be used with the following figure to allocate and re-allocate alpha between hypotheses. Testing will first be performed in subjects with CPS ≥1 for a treatment effect on ORR (H8). If H8 is supported (ie, the null hypothesis is rejected), then the corresponding alpha can be added to that allocated for evaluating the treatment effect on ORR in all subjects (H7). If H7 is supported, then the corresponding alpha can be re-allocated to PFS in subjects with CPS ≥10 (H3). If H3 is supported, then the corresponding alpha can be re-allocated, 27% to PFS in subjects with CPS ≥1 (H2 at IA2 only), and 73% to OS in subjects with CPS ≥10 (H6). If H2 is supported, the alpha for that hypothesis

can be re-allocated to PFS in all subjects (H1 at IA2 only). If H1 is supported, the alpha for that hypothesis can be re-allocated to OS in subjects with CPS ≥10 (H6). If H6 is supported, the alpha for that hypothesis can then be re-allocated to OS in subjects with CPS ≥1 (H5). If H5 is supported, the alpha for that hypothesis can then be re-allocated to OS in all subjects (H4 at IA2/IA3/FA). If H4 is supported, at IA2/IA3/FA, the alpha for that hypothesis can be re-allocated back to H7.

In the above multiplicity strategy, actual alpha spent at IA1 was calculated based on the original pre-specified alpha allocation strategy prior to Protocol Amendment 05 by the pre-specified alpha spending functions, using the actual spending time calculated from the planned and observed information fractions at IA1. Under the revised alpha allocation, actual alpha spent at IA1 for PFS in all subjects (H1), PFS in subjects with CPS ≥1 (H2), and OS in all subjects (H4) will be kept intact by a Bonferroni approach to strongly control the family-wise Type-I error rate at one-sided 2.5%. After IA1, these alphas will no longer be re-allocated to other hypotheses under the graphical approach, nor can they be used to account for correlation among group sequential tests within each endpoint across different time points. As such, the revised total alpha after IA1 is now 2.5% - 0.043% (alpha spent at IA1 for H1) - 0.046% (alpha spent at IA1 for H2) - 0.039% (alpha spent at IA1 for H4) ~= 2.37% for all hypotheses (all numbers rounded, the actual alpha for testing will be calculated based on the actual alpha spent for H1, H2 and H4 at IA1 with high precision). The 0.036% alpha spent at IA1 for H5 will be part of the new initial alpha of 0.75% for H5. Of note, a 0% initial alpha is now assigned to PFS in all subjects at IA2, PFS in subjects with CPS ≥1 at IA2, and OS in all subjects at IA2/IA3/FA. These endpoints at these specific timepoints are now part of the graphical approach and can still be tested if a positive alpha can be re-allocated to them after the success of testing relevant hypothesis(es). For example, if PFS in CPS ≥10 is supported at IA2, a one-sided alpha of 0.111% (0.411% × 0.27) will be re-allocated to PFS in CPS ≥1 at IA2 for testing. In addition, if a positive alpha can be re-allocated to OS in all subjects (H4) at either IA2, IA3 or FA, then a Lan-DeMets O'Brien-Fleming alpha-spending function will be used to distribute the alpha among IA2, IA3, and FA for appropriate testing at each time point, respectively.

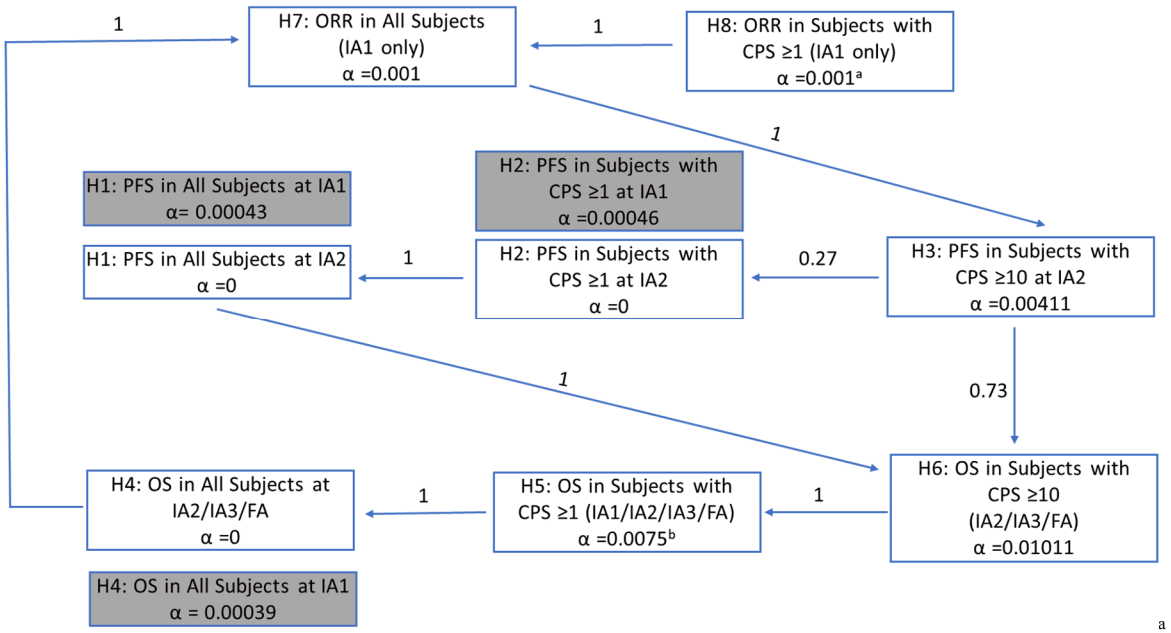
For OS endpoints, a Lan-DeMets O'Brien-Fleming approximation alpha-spending function is constructed to implement group sequential boundaries that control the Type-I error. Spending time will be plugged into the pre-specified spending function to calculate alpha spending. At the time of IA1(as applicable), IA2, and IA3 for OS, the spending time will be the minimum of the actual observed information fraction and the planned information fraction for each endpoint, respectively (with the exception of OS in subjects with CPS ≥1 at IA1, please see next paragraph). At the time of FA for OS endpoints, the spending time will be 1. Of note, while the spending time used for alpha-spending calculation will be the minimum of the actual observed information fraction and the planned information fraction, 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 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. [1]. Of note, prior to Protocol Amendment 05, a Hwang-Shih-DeCani alpha-spending function with gamma parameter (-4) was constructed to implement group sequential boundaries that control the Type-I error for PFS endpoints, and it is no longer applicable under Protocol Amendment 05.

For OS in subjects with CPS ≥1 (H5), the revised initial alpha is 0.75% which includes 0.036% already spent at IA1. In order to account for the 0.036% alpha already spent at IA1 for H5 under initial alpha allocation, the spending time at IA1 will be fixed at 56.1% such that the corresponding

alpha distributed at IA1 using the Lan-DeMets O'Brien-Fleming spending function with a total initial alpha of 0.75% remains 0.036%.

The Spiessens and Debois method [3] will be used to calculate the nominal alpha of ORR in subjects with CPS ≥1 after accounting for the correlation between ORR endpoints in these two populations, while fixing the nominal alpha at 0.001 for ORR in all subjects. The actual observed correlation will be used for this adjustment in the analysis. Of note, if at IA2, IA3 or FA, additional alpha can be reallocated to H7 (e.g., H4 is not successful at IA1 but successful at IA2), the p-value of ORR in all subjects obtained from IA1 will be re-evaluated using the updated alpha threshold at that time. Similarly, if at IA3 or FA additional alpha can be re-allocated to PFS endpoint(s), the p-value of PFS endpoint(s) obtained from IA2 will be re-evaluated using the updated alpha threshold at that time.

Figure 1 Multiplicity Strategy



[a] Nominal alpha for testing will be calculated based on Spiessens and Debois method accounting for correlation between ORR in all subjects and ORR in subjects with CPS ≥1. Of note, while the nominal alpha will be calculated and used for testing, the allocated alpha (0.001) will be passed to H7 when applicable.
[b] H5 α=0.0075 which includes 0.00036 already spent at IA1.
Note: The shaded boxes in this figure represent alpha that has already been spent at IA1 and will be considered lost for future analyses. These alphas will no longer be re-allocated to other hypotheses under the graphical approach, nor can they be used to account for correlation among group sequential tests within each endpoint across different time points.

3.9 Sample Size and Power Calculations

Part 1

Part 1 of the study will enroll approximately 30 subjects. For each treatment with 10 subjects available for analysis, if the underlying incidence rate of a given type of AE is 5% or 10%, there is 40% or 65% chance of observing at least one AE among the 10 subjects, respectively. If no AE of a given type is observed among the 10 subjects, this study part will provide 80% confidence that the underlying percentage of subjects with the AE is <14.9% (90% confidence that the underlying rate of the AE is <20.6%) in the study population treated with the study treatment.

Part 2

Part 2 of the study will randomize approximately 828 subjects in a 2:1 ratio between the pembrolizumab + chemotherapy and the placebo + chemotherapy arms.

Randomization will be implemented centrally using IVRS and will be monitored on a regular basis. When IVRS alerts study is approaching the desired enrollment, screening should be stopped in time. However, subjects already in screening phase may be enrolled even after the maximum sample size has been reached.

Power considerations for each endpoint are described below.

PFS

The PFS power calculation is based on the following assumptions: 1) PFS follows an exponential distribution with a median of 5.5 months in the placebo + chemotherapy arm in all populations (all subjects, subjects with CPS ≥1, and subjects with CPS ≥10); 2) An enrollment period of 17 months for Part 2; 3) A yearly drop-out rate of 30%; 4) the true HR is 0.70, 0.62, and 0.60 for PFS in all subjects, subjects with CPS ≥1, and subjects with CPS ≥10, respectively. In addition, prior to Protocol Amendment 05, a Hwang-Shih-DeCani alpha-spending function with gamma parameter (-4) was constructed to implement group sequential boundaries that control the Type-I error for PFS endpoints.

Any change to the timing of the PFS analyses, along with its rationale, will be documented in the sSAP or a memo to the study file before the database lock.

PFS in all subjects

At IA1 it was expected that approximately 500 PFS events would have been accumulated among all subjects. An alpha of ~0.05% was to be allocated to PFS in all subjects at this analysis (subject to change according to the actual number of PFS events at IA1, based on the alpha-spending function with a total of 0.1% alpha originally allocated to PFS in all subjects). If the planned number of events of 500 was analyzed, this analysis had ~ 67% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy, if the underlying HR is 0.70.

At IA2 of the study, PFS in all subjects will only be tested if both hypotheses of PFS in subjects with CPS ≥1 and PFS in subjects with CPS ≥10 are supported. It is expected that approximately 634 PFS events will be observed among all subjects at IA2. A final PFS analysis will be performed at IA2. This analysis has ~ 89% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy for PFS in all subjects at a one-sided 0.111% alpha level, if the underlying HR is 0.70.

PFS in subjects with CPS ≥1

At IA1 it was expected that approximately 360 PFS events would have been observed among subjects with CPS ≥1. An alpha of ~0.05% was to be allocated to PFS in subjects with CPS ≥1 at this analysis (subject to change according to the actual number of PFS events at IA1, based on the alpha-spending function with a total of 0.1% alpha originally allocated to PFS in subjects with CPS ≥1). If the planned number of events of 360 was analyzed, this analysis had ~ 83% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy, if the underlying HR is 0.62.

At IA2 of the study, PFS in subjects with CPS ≥1 will only be tested if the hypothesis of PFS in subjects with CPS ≥10 is supported. It is expected that approximately 463 PFS events will be observed among subjects with CPS ≥1 at IA2. A final PFS analysis will be performed at IA2. This analysis has ~ 97% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy for PFS in subjects with CPS ≥1 at a one-sided 0.111% alpha level, if the underlying HR is 0.62.

PFS in subjects with CPS ≥10

At IA2 of the study, it is expected that approximately 235 PFS events would have been observed among subjects with CPS ≥10. The only PFS analysis in subjects with CPS ≥10 will be performed at IA2. The analysis has ~ 86% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy for PFS in subjects with CPS ≥10 at a one-sided 0.411% alpha level, if the underlying HR is 0.60.

OS

The sample size and OS power calculation is based on the following assumptions: 1) OS follows an exponential distribution with a median of 17.5 months in the placebo + chemotherapy arm in all populations (all subjects, subjects with CPS ≥1 and subjects with CPS ≥10); 2) An enrollment period of 17 months for Part 2 and a minimum of 43 months follow-up after enrollment completion; 3) A yearly dropout rate of 3%; 4) the true HR is 0.80, 0.71, and 0.65 for OS in all subjects, subjects with CPS ≥1, and subjects with CPS ≥10, respectively. In addition, a Lan-DeMets O'Brien-Fleming approximation alpha-spending function was constructed to implement group sequential boundaries that control the Type-I error for OS endpoints.

Any change to the timing of the OS analyses, along with its rationale, will be documented in the sSAP or a memo to the study file before the database lock.

OS in all subjects

After IA1, OS in all subjects can be tested if hypothesis of OS in subjects with CPS ≥1 is supported. With ~ 664 OS events among all subjects at the final OS analysis, the trial has ~ 60% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy at a one-sided 0.75% alpha level, if the underlying HR is 0.80. If the planned numbers of events are analyzed, success boundary for OS in all subjects at FA approximately corresponds to an observed HR of ~ 0.81 (~ 4.0 months improvement over a control median OS of 17.5 months).

OS in subjects with CPS ≥1

It is expected that ~ 482 OS events will have occurred among subjects with CPS ≥1 at the final OS analysis. For OS in subjects with CPS ≥1, the trial has ~ 87% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy at a one-sided 0.75% alpha-level, if the underlying HR is 0.71. If the planned numbers of events are analyzed at IA2/IA3/FA, success boundary for OS in subjects with CPS ≥1 at FA approximately corresponds to an observed HR of ~ 0.78 (~ 4.8 months improvement over a control median OS of 17.5 months).

OS in subjects with CPS ≥10

It is expected that ~ 240 OS events will have occurred among subjects with CPS ≥10 at the final OS analysis. For OS in subjects with CPS ≥10, the trial has ~ 79% power to demonstrate that pembrolizumab + chemotherapy is superior to placebo + chemotherapy at a one-sided 1.011% alpha-level, if the underlying HR is 0.65. If the planned numbers of events are analyzed, success boundary for OS in subjects with CPS ≥10 at FA approximately corresponds to an observed HR of ~ 0.72 (~ 6.8 months improvement over a control median OS of 17.5 months).

ORR

The ORR power calculation is based on the following assumptions: 1) under initial alpha allocated to ORR hypotheses; 2) the underlying ORR is 29% in the placebo + chemotherapy arm, and there is 15 or 18 percentage points increase in ORR in the pembrolizumab + chemotherapy arm (ORR of 44% or 47%), in both all subjects and in subjects with CPS ≥1, respectively. The power for ORR endpoints is summarized in Table 9.

Table 9 Power for ORR

Population	N	Nominal Alpha	~ ORR Difference at Success Boundary	True ORR Difference	Power
All Subjects	640	0.001	12.6 percentage points	15 percentage points	72%
				18 percentage points	91%
Subjects with CPS ≥1	480	0.00145[a]	14.1 percentage points	15 percentage points	58%
				18 percentage points	80%

[a] Approximate nominal alpha based on the Spiessens and Debois method accounting for correlation between ORR in all subjects and ORR in subjects with CPS ≥1. The actual nominal alpha will be calculated based on the actual correlation between these two populations.
Assume 29% ORR in the placebo + chemotherapy arm.
Assume 75% PD-L1 CPS ≥1 prevalence.

The assumptions for a median PFS of 5.5 months, median OS of 17.5 months, and an ORR of 29% in the placebo + chemotherapy arm are based on the estimates from Miles et al., 2013 [15], and the Phase III trial reported in O'Shaughnessy et al., 2014 [16].

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

Of note, the assumptions of 75% prevalence of PD-L1 CPS ≥1, 38% prevalence of PD-L1 CPS ≥10 in mTNBC subjects and low discrepancy rate (e.g., <10%) between central and local determinations of baseline measurable disease were made in above calculations. The above assumptions and sample size calculations are subject to modification as needed based on emerging data on PD-L1 prevalence in mTNBC as well as correlation between PD-L1 expression and treatment effect, and/or emerging data on the discrepancy rate between central and local determinations of baseline measurable disease in mTNBC subjects.

Although safety issues are not expected for any of the pembrolizumab and chemotherapy combinations in this study, if one or more of the chemotherapy options (i.e., nab-paclitaxel, paclitaxel, or gemcitabine/carboplatin) is stopped because of a safety issue, then the Part 2 primary analyses will be restricted to the remaining chemotherapy option(s) and the Part 2 sample size will be based on only those remaining option(s).

3.10 Subgroup Analyses and Effect of Baseline Factors

Part 1

There is no planned subgroup analysis for Part 1.

Part 2

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 endpoints will be estimated and plotted within each category of the following classification variables:

- Chemotherapy on study (nab-paclitaxel vs paclitaxel vs gemcitabine/carboplatin; taxane vs gemcitabine/carboplatin).
- Tumor PD-L1 status (CPS ≥1 vs CPS <1; CPS ≥5 vs CPS <5; CPS ≥10 vs CPS <10; CPS ≥15 vs CPS <15; CPS ≥20 vs CPS <20). Note: these subgroup analyses will only be conducted in the all subjects population.
- Prior treatment with same class of chemotherapy in the (neo)adjuvant setting (yes vs no).
- Prior (neo)adjuvant chemotherapy (yes vs no)
- Prior (neo)adjuvant taxane treatment (yes vs no)
- Prior (neo)adjuvant platinum treatment (yes vs no)
- Menopausal status (for females only; pre- vs post-menopausal)
- Age (<65 years vs ≥65 years)
- Geographic region (Europe/Israel/North America/Australia vs Asia vs Rest of World)
- Ethnic origin (Hispanic vs Non-Hispanic)
- ECOG status (0 vs 1)
- HER2 status (2+ by IHC vs 0-1+ by IHC)
- Disease-free interval (de novo metastasis vs <12 months vs ≥12 months)
- Number of metastatic sites (<3 vs ≥3)
- Visceral disease (yes vs no)
- LDH (≥2.0 x Upper Limit of Normal [ULN] vs <2.0 x ULN)

3.11 Compliance (Medication Adherence)

Part 1 and Part 2

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

Part 1 and Part 2

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

4. REFERENCES

[1] K. Anderson, L. Wang and C. Gause, "Application of graphical multiplicity methods in complex group sequential trials," Merck & Co., Inc. (US), Kenilworth (NJ), 2017.

[2] 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.

[3] B. Spiessens and M. Debois, "Adjusted significance levels for subgroup analyses in clinical trials," *Contemp Clin Trials,* vol. 31, no. 6, pp. 647-56, 2010.

[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] J. Robins 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. 2609-2631, 1991.

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

[10] 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.

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

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

[13] 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.

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

[15] D. Miles, V. Dieras, J. Cortes, A. Duenne, J. Yi and J. OShaughnessy, "First-line bevacizumab in combination with chemotherapy for HER2-negative metastatic breast cancer: pooled and subgroup analyses of data from 2447 patients," *Annals of Oncology,* vol. 24, no. 11, pp. 2773-2780, 2013.

[16] J. O'Shaughnessy, L. Schwartzberg, M. Danso, K. Miller, H. Rugo and M. Neubauer, "Phase III study of iniparib plus gemcitabine and carboplatin versus gemcitabine and carboplatin in patients with metastatic triple-negative breast cancer," *Journal of Clinical Oncology,* vol. 32, no. 34, pp. 3840-3847, 2014.