A PHASE 3, MULTICENTER, RANDOMIZED, OPEN-LABEL STUDY OF AVELUMAB (MSB0010718C) ALONE OR IN COMBINATION WITH PEGYLATED LIPOSOMAL DOXORUBICIN VERSUS PEGYLATED LIPOSOMAL DOXORUBICIN ALONE IN PATIENTS WITH PLATINUM-RESISTANT/REFRACTORY OVARIAN CANCER

JAVELIN OVARIAN 200

STATISTICAL ANALYSIS PLAN

Compounds:	MSB0010718C
Compound Name:	Avelumab
Version:	3.0
Date:	13-Dec-2017

[redacted]

[redacted]

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

1. VERSION HISTORY

This Statistical Analysis Plan (SAP) for study B9991009 is based on the protocol amendment 01 dated 15-Dec-2016.

CCI

		CCI [redacted]
[redacted]	[redacted]	[redacted]

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study B9991009. This document may modify the plans outlined in the protocol; however, any major modifications of the primary endpoint definition or its analysis will also be reflected in a protocol amendment.

CCI [redacted]

Statistical analyses will be performed using cleaned eCRF data as well as non-CRF data (ie, PK concentration, ADA, nAb, CCI [redacted], and tumor assessment results by the Blinded Independent Central Review (BICR)).

The primary analysis will include all data up to a cut-off date which is determined by the number of events for Progression-Free Survival (PFS) by Blinded Independent Central Review (BICR) and for Overall Survival (OS) and minimum follow-up of 12 months after the last patient is randomized. The cut-off date is determined once a data extract (before database lock) is available which indicates that the required number of events for PFS by BICR and for OS and minimum follow-up of 12 months have occurred.

Due to data cleaning activities, the final number of events might deviate from the planned number. The data cut-off date will not be adjusted retrospectively in this case.

2.1. Study Objectives

Primary Objectives

- To demonstrate that avelumab given alone or in combination with Pegylated liposomal doxorubicin (PLD) is superior to PLD alone in prolonging OS in patients with platinum - resistant/platinum-refractory ovarian cancer.
- To demonstrate that avelumab given alone or in combination with PLD is superior to PLD alone in prolonging PFS in patients with platinum-resistant/platinum-refractory ovarian cancer.

Secondary Objectives

- To evaluate anti-tumor activity of avelumab given alone or in combination with PLD versus PLD alone in ovarian cancer patients.

- To evaluate the overall safety profile of avelumab alone or in combination with PLD versus PLD alone in ovarian cancer patients.
- To characterize the Pharmacokinetics (PK) of doxorubicin (PLD samples) and avelumab when administered in combination, and to assess the effect of avelumab on the PK of doxorubicin (PLD samples) and the effect of PLD on PK of avelumab.
- To assess the immunogenicity of avelumab.
- To evaluate candidate predictive biomarkers of sensitivity or resistance to avelumab alone or PLD in combination with avelumab in pre-treatment tumor tissue, that may aid in the identification of patient subpopulations most likely to benefit from treatment.
- To compare the effect of avelumab alone or in combination with PLD versus PLD alone on patient-reported outcomes (PRO) in patients with ovarian cancer.

Exploratory Objectives

- CCI [redacted]

2.2. Study Design

This is a Phase 3, multicenter, randomized, open-label, parallel 3-arm study in which approximately 550 patients will be randomized in a 1:1:1 ratio to receive avelumab alone, avelumab in combination with PLD, or PLD alone, as follows:

- Arm A: avelumab alone;
- Arm B: avelumab plus PLD;
- Arm C: PLD alone.

Patients will be stratified by platinum-refractory or platinum-resistant status, number of prior regimens (1 vs 2 or 3), and bulky disease (defined as presence of a tumor ≥5 cm) vs not. Cross-over will not be permitted.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoints

- Overall Survival (OS).

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

- Progression Free Survival as determined by BICR according to RECIST version 1.1.

  PFS is defined as the time from the date of randomization to the date of the first documentation of progressive disease (PD) or death due to any cause, whichever occurs first.

3.2. Secondary Endpoints

3.2.1. Safety endpoints

- Adverse Events (as graded by NCI CTCAE v.4.03);

  AEs will be graded by the investigator according to the Common Terminology Criteria for Adverse Events (CTCAE) version 4.03 and coded using the Medical Dictionary for Regulatory Activities (MedDRA)

- Laboratory abnormalities (as graded by NCI CTCAE v.4.03);

- Vital signs (blood pressure, pulse rate);

- Electrocardiograms (ECGs), ECHO or MUGA scans.

3.2.2. Efficacy endpoints

- PFS as determined by Investigator according to RECIST version 1.1.

- Objective Response (OR), Duration of Response (DR), and Disease Control (DC) as determined by Blinded Independent Central Review (BICR) and Investigator [as assessed by RECIST version 1.1].

  OR is defined as complete response (CR) or partial response (PR) according to RECIST v1.1 from the date of randomization until the date of the first documentation of PD. Both CR and PR must be confirmed by repeat assessments performed no less than 4 weeks after the criteria for response are first met.

  DR is defined, for patients with an objective response, as the time from first documentation of objective response (CR or PR) to the date of first documentation of PD or death due to any cause.

  DC is defined as CR, PR, non-CR/non-PD, or SD. Both CR and PR must be confirmed by repeat assessments performed no less than 4 weeks after the criteria for response are first met. Criteria for SD and non-CR/non-PD must have been met at least 6 weeks after the date of randomization.

3.2.3. Patient reported outcomes

- EORTC QLQ-C30, EORTC QLQ-OV28, and EQ-5D-5L.

  Patient-reported outcomes of HRQoL and ovarian cancer specific disease/treatment-related symptoms will be evaluated using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire – Core 30 (EORTC QLQ-C30) and its corresponding module for ovarian cancer, the Quality of Life Questionnaire – Ovarian Cancer 28 (QLQ-OV28). The EuroQoL 5 Dimensions 5 Levels (EQ-5D-5L) questionnaire will also be collected to assess general health status.

  The EORTC QLQ-C30 consists of 30 questions which assess five functional domains (physical, role, cognitive, emotional, and social), global health status/quality of life, disease/treatment related symptoms (fatigue, pain, nausea/vomiting dyspnea, appetite loss, sleep disturbance, constipation, and diarrhoea), and the perceived financial impact of disease (Aaronson et al. 1993, Aaronson et al. 1996, Osoba et al. 1997).

  The EORTC QLQ-OV28 is the ovarian cancer specific module of the EORTC quality of life questionnaire. The EORTC QLQ-OV28 is a 28 item instrument with seven (7) functional domain subscales. The 7 subscales include:

  (i) an abdominal/GI symptom subscale (7 items);

  (ii) a peripheral neuropathy subscale (3 items);

  (iii) a chemotherapy side effects subscale (7 items);

  (iv) a hormonal/menopausal symptoms subscale (2 items);

  (v) a body image subscale (2 items);

  (vi) an attitude to disease and treatment subscale (3 items); and

  (vii) a sexual function subscale (4 items).

  Higher scores are reflective of a greater presence of symptoms (Cull et al. 2001, Greimal et al. 2003).

  The EQ-5D-5L questionnaire consists of the EQ-5D-5L descriptive system and a visual analogue scale (the EQ VAS). The EQ-5D-5L descriptive system measures a patient's health state on 5 dimensions which include: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The respondent's self-rated health is assessed on a scale from 0 (worst imaginable health state) to 100 (best imaginable health state) by the EQ-VAS (The EuroQol Group, 2015).

3.2.4. Pharmacokinetic endpoints

- Pharmacokinetic parameters Ctrough for avelumab: Cmax, volume of distribution (Vd), clearance (CL), area under the concentration-time curve (AUC) for doxorubicin (PLD samples).

C_{max} and C_{trough} are calculated after single dose and at steady state. Other parameters will be calculated including, but not limited to T_{max}, AUC_{24}, AUC_{last}, T_{last}, $AUC_{sd,\tau}$, $AUC_{ss,\tau}$, $t_{1/2}$, CL, and Vd_z, $C_{ss,av}$ as data permit). Dose normalized parameters (eg, CDN-Cmax, DN-AUC) will be reported as appropriate.

Table 2. PK Parameters to be Determined for Avelumab and Doxorubicin

Parameter	Definition	Method of Determination
AUC_{24}	Area under the plasma concentration-time profile from time zero to 24 hours	Linear/Log trapezoidal method
AUC_{last}	Area under the plasma concentration-time profile from time zero to the time of the last quantifiable concentration (C_{last})	Linear/Log trapezoidal method
$AUC_{sd,\tau}$ $AUC_{ss,\tau}$	Area under the plasma concentration-time profile after single dose from time zero to the next dose (after single dose and at steady state)	Linear/Log trapezoidal method
C_{max}	Maximum observed plasma concentration	Observed directly from data
T_{max}	Time for C_{max}	Observed directly from data as time of first occurrence
$t_{½}$[a]	Terminal half-life	$Log_e(2)/k_{el}$, where k_{el} is the terminal phase rate constant calculated by a linear regression of the log-linear concentration-time curve. Only those data points judged to describe the terminal log-linear decline were used in the regression.
C_{trough}	Predose concentration during multiple dosing	Observed directly from data
CL	Clearance	Dose / AUCτ for steady state
Vd	Volume of distribution	Dose / (AUCτ ·kel) for steady state
AUC_{24} (dn)	Dose normalized AUC_{24}	AUC_{24} / Dose
AUC_{last}(dn)	Dose normalized AUC_{last}	AUC_{last} / Dose
AUC_{inf}(dn)[a]	Dose normalized AUC_{inf}	AUC_{inf} / Dose
C_{max}(dn)	Dose normalized C_{max}	C_{max} / Dose

[a] If data permit

3.2.5. Immunogenicity endpoints

- Anti-Drug antibodies and neutralizing antibodies against avelumab.

  Anti-Drug Antibody (ADA) / Neutralizing antibodies (nAb) titers for avelumab.

3.2.6. Biomarker endpoints

- Candidate predictive biomarkers in tumor tissue including, but not limited to, PD-L1 expression and tumor infiltrating CD8+ T lymphocytes as assessed by immunohistochemistry (IHC).

Table 3. Biomarker Definition and Determination

Parameter	Definition	Method of Determination
PD-L1 expression	The number of PD-L1 positive cells and/or qualitative assessment of PD-L1 staining on tumor and inflammatory cells in regions of interest that are defined by tumor cell morphology and the presence or absence of CD8+ cells	Pathologist, assisted by image analysis
Tumor infiltrating CD8+ lymphocytes	The number of CD8+ cells per unit area and the percent of counted cells that are scored as CD8+	Pathologist, assisted by image analysis

3.3. Exploratory Endpoints

CCI

- CA-125 levels.

3.4. Baseline Variables

3.4.1. Study drug, study treatment and baseline definitions

In this study, '**study drug**' refers to avelumab or PLD and '**study treatment**' (or '**treatment arm**') refers to one of the following:

- Arm A = avelumab alone;
- Arm B = avelumab plus PLD;
- Arm C = PLD alone.

Start and end dates of study treatment:

For Arm A and Arm C:

The date/time of first dose of study treatment is the earliest date/time of non-zero dosing of the study drug.

The date/time of last dose of study treatment is the latest date/time of non-zero dosing of the study drug.

For Arm B:

The date/time of first dose of study treatment in the combination arm is the earliest date/time of the first non-zero dose date/time for each of the study drugs.

The date/time of last dose of study treatment in the combination arm is the latest date/time of the last non-zero dose date/time for each of the study drugs.

Definition of baseline:

Definition of baseline for efficacy and PRO analyses

The last measurement prior to randomization will serve as the baseline measurement for efficacy and PRO analyses. If such a value is missing (since per protocol the first PRO assessment is planned to occur prior to dosing on Cycle 1 Day 1), the last measurement prior to the first dose of study treatment will be used as the baseline measurement except for analyses of tumor assessments data where the baseline assessment would be considered as missing.

Definition of baseline for immunogenicity analyses

The last available assessment prior to the start of treatment with avelumab is defined as 'baseline' result or 'baseline' assessment. If an assessment is planned to be performed prior to the first dose of avelumab in the protocol and the assessment is performed on the same day as the first dose of avelumab, it will be assumed that it was performed prior to avelumab administration, if assessment time point is not collected or is missing.

Definition of baseline for safety analyses

The last available assessment prior to the start of study treatment is defined as 'baseline' value or 'baseline' assessment for safety analyses. If an assessment is planned to be performed prior to the first dose of study treatment in the protocol and the assessment is performed on the same day as the first dose of study treatment, it will be assumed that it was performed prior to study treatment administration, if assessment time point is not collected or is missing. If assessment time points are collected, the observed time point will be used to determine pre-dose on study day 1 for baseline calculation. Unscheduled assessments will be used in the determination of baseline. However, if time is missing, an unscheduled assessment on study day 1 will be considered to have been obtained after study treatment administration.

Patients who start treatment and discontinue from the study on the same day may have two different sets of data collected on study day 1 (one during study and one in the End of Treatment (EOT) visit. Data reported at the EOT visit are not eligible for baseline selection.

If a scheduled pre-dose measurement actually occurred post-dose, then the corresponding measurement will be treated and analyzed similar to an unscheduled post-dose measurement.

Baseline for RR and QT/QTc interval assessments will be derived from the visit where both RR and QT are not missing. Triplicate ECGs are collected in the study and the baseline for each ECG measurement is the average of the pre-dose replicate measurements on the baseline day. Unscheduled assessments will not be included in the calculation of the average. QTcB and QTcF will be derived based on RR and QT. The average of the replicate measurements will be determined after the derivation of the individual parameter at each time point.

Definition of baseline for biomarker analyses

The last measurement prior to the start of study treatment will serve as the baseline measurement for biomarker analyses. For biomarkers that are planned to be measured on Cycle 1 Day 1 (eg, soluble proteins), it will be assumed that the measurement was performed prior to study treatment administration, if the assessment time point is not collected or is missing.

3.4.2. Baseline characteristics

Randomization is stratified by the following, as recorded in the Interactive Response Technology (IRT):

- Platinum-refractory or platinum-resistant status
- Number of prior regimens (1 vs 2 or 3)
- Bulky disease (defined as presence of a tumor ≥5 cm) vs not

The primary analyses of OS and PFS will be stratified by these randomization stratification factors.

Other baseline characteristics (including demographics, physical measurements, disease history and prior anti-cancer therapies) are described in Section 6.5.1. These baseline characteristics are not planned to be included as stratification variables or covariates in statistical models unless otherwise specified in Section 6.

3.5. Safety Endpoints

3.5.1. Adverse events

Treatment-Emergent Adverse Events

Treatment-emergent adverse events (TEAEs) are those events with onset dates occurring during the on-treatment period for the first time, or if the worsening of an event is during the on-treatment period.

On-treatment period is defined as the time from the first dose of study treatment through minimum (30 days + last dose of study treatment, start day of new anti-cancer drug therapy – 1 day). The start day of new anti-cancer drug therapy after the first dose of study treatment is derived as outlined in Section 5.2.5.

Adverse Events of Special Interest (AESIs)

AESIs are immune-related adverse events (irAE) and infusion-related reactions (IRRs). The criteria for classification of an AE as an irAE or IRR are described in Appendices 1 and 2, respectively.

3-Tier Adverse Events

It should be recognized that most studies are not designed to reliably demonstrate a causal relationship between the use of a pharmaceutical product and an adverse event or a group of adverse events. Except for select events in unique situations, studies do not employ formal adjudication procedures for the purpose of event classification. As such, safety analyses are generally considered as an exploratory analysis and its purpose is to generate hypotheses for further investigation. The 3-tier approach facilitates these exploratory analyses.

Adverse events and clusters of adverse events, of any causality and treatment-related, will also be summarized following a 3-tier approach. Under this approach, AEs are classified into 1 of 3 tiers.

Tier-1 events: These are pre-specified events or clusters of events of clinical importance and will be described in the Safety Review Plan.

Tier-2 events: These are events that are not Tier-1 but are "common". A MedDRA PT is defined as a Tier-2 event if it is reported by

a) at least 10% of patients with any grade in any treatment arm, or

b) at least 5% of patients with Grade 3, 4 or 5 in any treatment arm.

Tier-3 events: All other AEs that are classified neither as Tier-1 nor Tier-2.

4. ANALYSIS SETS

Data for all patients will be assessed to determine if patients meet the criteria for inclusion in each analysis population prior to releasing the database and classifications will be documented per Pfizer's standard operating procedures.

Only patients who signed informed consent will be included in the analysis sets below.

Table 4 summarizes the use of the analysis sets for efficacy, safety, baseline characteristics, and exposure.

Table 4. Statistical Analyses by Analysis Set

Endpoints	Full Analysis Set	Per Protocol Analysis Set	Safety Analysis Set
Baseline Characteristics	✓		✓
Prior and Concomitant Therapies	✓		✓
Exposure			✓
Efficacy: Primary	✓	✓	
Efficacy: Secondary	✓		
CCI [redacted]	[redacted]		
Safety			✓

4.1. Full Analysis Set

The full analysis set (FAS) will include all randomized patients. Patients will be classified according to the study treatment assigned at randomization.

4.2. Safety Analysis Set

The safety analysis set will include all patients who receive at least one dose of study drug. Patients will be classified according to the study treatment assigned at randomization unless the incorrect treatment(s) was/were received throughout the dosing period in which case patients will be classified according to the first study treatment received.

4.3. Other Analysis Sets

4.3.1. Per-protocol analysis sets

Per protocol (PP) analysis set is a subset of the FAS and will include patients who do not meet any of the following criteria that could impact the primary objectives of the study. Patients who meet any of the following criteria will be excluded from the PP analysis set for PFS by BICR.

- Patient did not receive at least one dose of the randomized study treatment
- Patient without a tumor assessment >7 weeks after randomization (unless PD by BICR or death is observed before that time in which case the patient will not be excluded from the PP analysis set)
- Patient without measurable disease at baseline by BICR
- ECOG status > 1
- Patient does not meet Inclusion Criteria 1, 2 or 3

    1. Histologically confirmed epithelial ovarian, fallopian tube, or peritoneal cancer, including malignant mixed Müllerian tumors with high grade serous component.
    2. Platinum-resistant/refractory disease, defined as disease progression within 180 days following the last administered dose of platinum therapy (resistant), or lack of response or disease progression while receiving the most recent platinum-based therapy (refractory), respectively.
    3. Received up to 3 lines of systemic anticancer therapy for platinum-sensitive disease, most recently platinum-containing, and no prior systemic therapy for platinum-resistant disease.

Note that the protocol only requires that patients have measurable disease at baseline as per investigator assessment. However, since PFS by BICR is the primary endpoint, "patient without measurable disease at baseline by BICR" will be used as a criterion to exclude patients from the PP analysis set for this sensitivity analysis of the PFS by BICR primary endpoint.

Patients who meet any of the following criteria will be excluded from the PP analysis set for OS.

- Patient did not receive at least one dose of the randomized study treatment
- ECOG status > 1
- Patient does not meet Inclusion Criteria 1, 2 or 3

4.3.2. PK analysis sets

The PK concentration analysis set is a subset of the safety analysis set and will include patients who have at least one post-dose concentration measurement above the lower limit of quantitation (LLQ) for avelumab or doxorubicin.

The PK parameter analysis set is a subset of the safety analysis set and will include patients who have at least one of the PK parameters of interest for avelumab or doxorubicin**.**

4.3.3. Biomarker analysis sets

The biomarker analysis set for biomarkers that are measured only at screening is a subset of the safety analysis set and will include patients who have at least one screening biomarker assessment.

The biomarker analysis set for biomarkers that are measured sequentially is a subset of the safety analysis set and will include patients who have at least one baseline and one on-treatment biomarker assessment for the same marker.

Analysis sets will be defined separately for blood-based and tumor tissue-based biomarkers.

4.3.4. Immunogenicity analysis set

The immunogenicity analysis set is a subset of the safety analysis set and will include patients who have at least one ADA/nAb sample collected for avelumab in the avelumab containing arms.

5. GENERAL METHODOLOGY AND CONVENTIONS

5.1. Hypotheses and Decision Rules

5.1.1. Hypotheses and sample size determination

The study is designed to test, in parallel, the following hypotheses:

H_{01}: $HR_{OS(avelumab/PLD)} \geq 1$, versus H_{a1}: $HR_{OS(avelumab/PLD)} < 1$

H_{02}: $HR_{OS(avelumab+PLD/PLD)} \geq 1$, versus H_{a2}: $HR_{OS(avelumab+PLD/PLD)} < 1$

H_{03}: $HR_{PFS(avelumab/PLD)} \geq 1$, versus H_{a3}: $HR_{PFS(avelumab/PLD)} < 1$

H_{04}: $HR_{PFS(avelumab+PLD/PLD)} \geq 1$, versus H_{a4}: $HR_{PFS(avelumab+PLD/PLD)} < 1$

where HR_{OS} represents the hazard ratio for OS in each of the experimental arms vs the control arm, and HR_{PFS} represents the hazard ratio for PFS by BICR assessment in each of the experimental arms vs the control arm.

The treatment effects on the primary PRO endpoint, time to deterioration (TTD), will also be tested.

H_{05}: $HR_{TTD(avelumab/PLD)} \geq 1$, versus H_{a5}: $HR_{TTD(avelumab/PLD)} < 1$

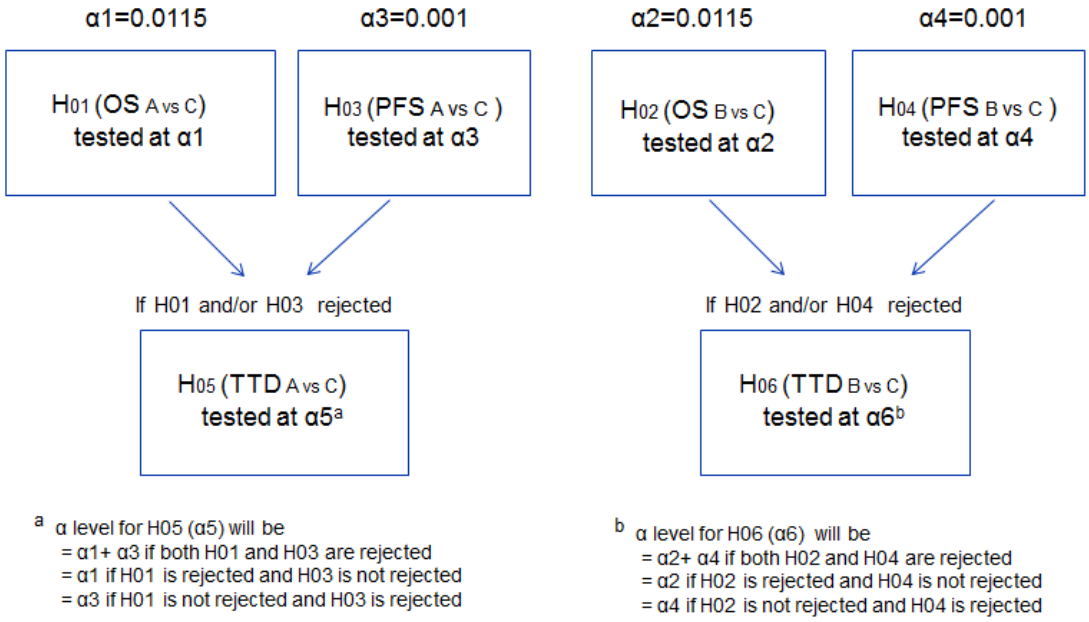
H_{06}: $HR_{TTD(avelumab+PLD/PLD)} \geq 1$, versus H_{a6}: $HR_{TTD(avelumab+PLD/PLD)} < 1$

where HR_{TTD} represents the hazard ratio for TTD in each of the experimental arms vs the control arm.

Approximately 550 patients will be randomized to the treatment arms using a 1:1:1 randomization, stratified by platinum-refractory or platinum-resistant status, number of prior regimens (1 vs 2 or 3), and bulky disease (defined as presence of a tumor ≥5 cm) vs not.

Overall type I-error will be maintained at or below 1-sided 0.025 by allocating 0.0115 alpha to each of the two OS comparisons and 0.001 to each of the two PFS comparisons. A gatekeeping procedure will be used to allow further testing of TTD for patients as described in Figure 1. The significance levels for each test will also take into account the group sequential nature of the design.

Figure 1 Testing Strategy



For each of the two OS comparisons, 196 OS events will be required to have at least 90% power to detect a hazard ratio of 0.6 using a 1-sided log-rank test at a significance level of 0.0115, and a 2-look group-sequential design with Lan-DeMets (O'Brien-Fleming) α-spending function to determine the efficacy boundary and a Gamma Family (-10) β-spending function to determine the non-binding futility boundary.

For each of the two PFS comparisons, 325 PFS events by BICR assessment will provide 93% power to detect a HR of 0.6 using a 1-sided log rank test at a significance level of 0.001, and a 2-look group sequential design with Lan-DeMets (O'Brien-Fleming) α-spending function to determine the efficacy boundary and a Gamma Family (-10) β-spending function to determine the non-binding futility boundary.

The analysis cut-off date for the test of the 4 primary hypotheses for OS and PFS will be the same for the IA and, if applicable, for the final analysis. The hypotheses of TTD will be tested at the time of the primary OS and PFS analyses.

The sample size for this study is determined based on the assumptions that the median OS and PFS in control arm is 12 months and 3.5 months, respectively (Gordon et al, 2001; Colombo et al, 2012), and that treatment with avelumab alone or avelumab in combination with PLD is expected to increase the median OS and median PFS to ≥ 20 months and ≥ 5.8 months, respectively, corresponding to a hazard ratio (HR) of ≤ 0.6 under the exponential model assumption. The sample size further assumes a 5% drop-out rate for OS and 15% drop-out rate for PFS within each treatment arm and a non-uniform patient accrual over a 19-month period. The data cut-off for the primary OS and PFS analyses will occur after the target number of events for both endpoints has been reached within each comparison and the last patient randomized in the study has been followed for at least 12 months after randomization.

5.1.2. Decision rules

OS

To protect the integrity of the study and to preserve the type I error rate, a fraction of alpha (0.002) for OS efficacy will be spent at the interim analysis and accounted for in the overall type I error rate if the interim analysis is performed at the planned number of OS events. The nominal significance levels for the interim and final efficacy analyses of OS will be determined by using the Lan-DeMets procedure with an O'Brien-Fleming stopping boundary. The overall significance level for the efficacy analysis of OS, within each comparison, will be preserved at 0.0115 (1-sided).

Two analyses will be performed for OS:

1. an interim analysis will be conducted after all patients have been randomized in the study and at least 131 (67%) of the 196 OS events for each of the OS comparisons have been documented and
2. the final analysis for OS will be conducted after all patients randomized in the study have been followed for 12 months and at least 196 OS events for each of the OS comparisons have been documented.

PFS based on BICR assessment

All references to PFS in this section pertain to PFS based on BICR assessment even when not specifically stated.

A fraction of alpha (0.0003) for PFS efficacy will be spent at the interim analysis and accounted for in the overall type I error rate if the interim analysis is performed at the planned number of PFS events. The nominal significance levels for the interim and final efficacy analyses of PFS will be determined by using the Lan-DeMets procedure with an O'Brien-Fleming stopping boundary. The overall significance level for the efficacy analysis of PFS, within each comparison, will be preserved at 0.001 (1-sided).

Two analyses will be performed for PFS:

1. an interim analysis will be conducted after all patients have been randomized in the study and at least 267 (82%) of the 325 PFS events for each of the PFS comparisons have been documented based on BICR assessment, and
2. the final analysis for PFS after all patients randomized in the study have been followed for 12 months and at least 325 PFS events for each of the PFS comparisons have been documented based on BICR assessment.

The interim analyses for OS and PFS will occur at the same time for both treatment comparisons, ie the analysis cut-off date for the test of the 4 hypotheses, H_{01}, H_{02}, H_{03} and H_{04}, will be the same and will be set after all patients have been randomized in the study and the target number of events across all endpoints and treatment comparisons have been reached. If the value of the test statistic for either of the primary endpoints (either OS or PFS) exceeds the associated efficacy boundary for a comparison, then the experimental treatment associated with that comparison (or treatments if this is met for both comparisons) may be declared statistically significantly superior to the control treatment for that endpoint. If the value of the test statistics for both OS and PFS exceed the associated futility boundaries, then the experimental treatment associated with that comparison (or the study if this is met for both comparisons) may be stopped for futility.

Table 5. Interim Analysis - Efficacy and Futility Boundaries

	Efficacy boundary		Futility boundary	
	z value	p -value	z value	p -value
OS Assuming 131 OS events for each comparison at the time of the IA	-2.8791	0.002	-0.2339	0.4075
PFS Assuming 267 PFS events for each comparison at the time of the IA	-3.4474	0.0003	-1.8938	0.0291

Since the observed number of events at the interim analysis for each comparison may not be exactly equal to the planned 131 OS events and 267 PFS events, the efficacy and futility boundaries will be updated based on the actual number of observed events using the pre-specified α-and β-spending functions. Therefore, the observed Z-test statistic at the interim analysis will be compared with the updated efficacy and futility boundaries.

If the study continues to final analysis, the p-value that will be used to declare statistical significance at the final analysis will be based on the actual number of OS and PFS events documented at the cut-off date for the final analysis and the α already spent at the interim analysis. Therefore, if the interim analysis occurs after exactly 131 OS events for each comparison, and the study continues until the final analysis, the observed p-value for the comparison will have to be < 0.011 to declare statistical significance; if the interim analysis

occurs after exactly 267 PFS events for each comparison, and the study continues until the final analysis, the observed p-value for the comparison will have to be < 0.001 to declare statistical significance. If the number of events in the final analysis deviates from the expected number of events, the final analysis criterion will be determined so that the overall significance level across all analyses and comparisons is maintained at 1-sided 0.025.

Based on the stopping boundaries defined above and the timing of interim analysis at 67% information fraction for OS and 82% information fraction for PFS the design has the following operating characteristics.

Table 6 presents the simulated cumulative probabilities to cross the futility or efficacy boundaries within each comparison for OS at the time of the interim and final analyses and Table 7 presents the simulated cumulative probabilities to cross the futility or efficacy thresholds within each comparison for PFS at the time of the interim and final analyses. Within each comparison, crossing of only one boundary (for either OS or PFS) at interim will not trigger early termination for futility but may trigger early declaration of efficacy.

Table 6. Simulated cumulative probabilities to cross the futility or efficacy boundaries at the interim or final OS analysis

Scenario	Look	Number of OS events [a]	Calendar Time (months)	P(Reject H_{0i}) for OS	P(Reject H_{ai}) for OS
H_{0i} is true (HR=1)	Interim	131	21.5	0.0025	0.5959
	Final	196	27	0.0113	0.9887
H_{ai} is true (HR=0.6)	Interim	131	23.5	0.5040	0.0040
	Final	196	31	0.9008	0.0992

Simulations performed in EAST with number of simulations = 10,000 and seed = 1234.
[a] For each comparison; i=1, 2.

Table 7. Simulated cumulative probabilities to cross the futility or efficacy boundaries at the interim or final PFS analysis

Scenario	Look	Number of PFS events [a]	Calendar Time (months)	P(Reject H_{0i}) for PFS	P(Reject H_{ai}) for PFS
H_{0i} is true (HR=1)	Interim	267 (PFS)	21.5	0.0001	0.9693
	Final	325 (PFS)	27	0.0011	0.9989
H_{ai} is true (HR=0.6)	Interim	267 (PFS)	23.5	0.7609	0.0117
	Final	325 (PFS)	31	0.9302	0.0698

Simulations performed in EAST with number of simulations = 10,000 and seed = 1234.
[a] For each comparison; i=3, 4.

5.2. General Methods

As described in Section 3.4, in this study '**treatment arm**' refers to one of the following:

- Arm A = avelumab alone;
- Arm B = avelumab plus PLD;
- Arm C = PLD alone.

Endpoints will be summarized based on the analysis sets described in Table 4 by treatment arm, unless otherwise specified.

5.2.1. Data handling after the cut-off date

Data after the cut-off date may not undergo the cleaning process and will not be displayed in any listings or used for summary statistics, statistical analyses or imputations.

5.2.2. Pooling of centers

In order to provide overall estimates of treatment effects, data will be pooled across centers. The 'center' factor will not be considered in statistical models or for subgroup analyses due to the high number of participating centers in contrast to the anticipated small number of patients randomized at each center.

5.2.3. Presentation of continuous and qualitative variables

Continuous variables will be summarized using descriptive statistics ie, number of non-missing values and number of missing values [ie, n (missing)], mean, median, standard deviation (SD), minimum, maximum and first and third quartile (Q1 and Q3).

Qualitative variables will be summarized by frequency counts and percentages. Unless otherwise specified, the calculation of proportions will include the missing category. Therefore counts of missing observations will be included in the denominator and presented as a separate category.

In case the analysis refers only to certain visits, percentages will be based on the number of patients still present in the study at that visit, unless otherwise specified.

5.2.4. Definition of study day

Start day of study treatment is the day of the first dose of study treatment.

The study day for assessments occurring on or after the start of study treatment (eg, adverse event onset, tumor measurement) will be calculated as:

Study day = Date of the assessment/event - start of study treatment + 1.

The study day for assessments occurring prior to the first dose of study treatment (eg, baseline characteristics, medical history) will be negative and calculated as:

Study day = Date of the assessment/event - start of study treatment.

The study day will be displayed in all relevant data listings.

5.2.5. Definition of start of new anti-cancer drug therapy

Start date of new anti-cancer drug therapy is used to determine the end of the on-treatment period (see Section 5.2.7).

The start date of new anti-cancer drug therapy is the earliest start date of anti-cancer drug therapy recorded in the 'Follow-up Cancer Therapy' eCRF pages that is after the first dose of study treatment. When the start date of anti-cancer drug therapy is missing or partially missing, the imputation rules described in Section 5.3.3.4 should be applied using only data from the 'Follow-up Cancer Therapy' eCRF pages.

5.2.6. Definition of start of new anti-cancer therapy

Start date of new anti-cancer therapy (drug, radiation, surgery) is used for censoring in efficacy analyses (see Section 6.1.2 and Section 6.2.2).

The start date of new anti-cancer therapy is the earliest date after randomization amongst the following:

- Start date of anti-cancer drug therapy recorded in the 'Follow-up Cancer Therapy' eCRF pages
- Start date of radiation therapy recorded in 'Concomitant Radiation Therapy', and 'Follow-up Radiation Therapy' eCRF pages with 'Treatment Intent' = 'Curative in intent'
- Surgery date recorded in 'Follow-up Surgery' eCRF pages when 'Surgery Outcome' = 'Resected' or 'Partially Resected'.

When the start date of anti-cancer therapy is missing or partially missing, the imputation rules described in Section 5.3.3.4 should be applied using 'Follow-up Cancer Therapy', 'Concomitant Radiation Therapy', 'Follow-up Radiation Therapy', and 'Follow-up Surgery' eCRF pages.

5.2.7. Definition of on-treatment period

Safety endpoints will be summarized based on the on-treatment period unless otherwise specified.

On-treatment period is defined as the time from the first dose of study treatment through minimum (30 days + last dose of study treatment, start day of new anti-cancer drug therapy – 1 day).

Safety data collected outside the on-treatment period as described above will be listed and flagged in listings but will not be summarized.

5.2.8. Standard derivations and reporting conventions

The following conversion factors will be used to convert days into weeks, months or years: 1 week = 7 days, 1 month = 30.4375 days, 1 year = 365.25 days.

Demographics and physical measurements:

- Age [years]:
  - (date of given informed consent - date of birth + 1) / 365.25
  - In case of missing day, day only: Age [years]: (year/month of given informed consent – year/month of birth)
  - In case only year of birth is given: Age [years]: (year of given informed consent - year of birth)

The integer part of the calculated age will be used for reporting purposes.

- BMI (kg/m^2) = weight (kg)/[height (m)]^2
- BSA (m^2) = ([height (cm) × weight (kg)] / 3600)$^{0.5}$

For reporting conventions, mean and median should generally be displayed one more decimal place than the raw data and standard deviation should be displayed to two more decimal places than the raw data. Percentages will be reported to one decimal place. The rounding will be performed to closest integer / first decimal using the common mid-point between the two consecutive values. Eg, 5.1 to 5.4 will be rounded to an integer of 5, and 5.5 to 5.9 will be rounded to an integer of 6.

5.2.9. Unscheduled visits

Generally, data collected at unscheduled visits will be included and analyzed for both safety and efficacy analyses in the same fashion as the data collected at scheduled visits except where otherwise noted in the sections that follow. Descriptive statistics (mean, SD, median, minimum, maximum, quartiles) by nominal visit or time point for safety endpoints such as laboratory measurements, ECGs and vital signs will include only data from scheduled visits.

5.2.10. Adequate baseline tumor assessment

Adequate baseline is defined using the following criteria:

- All baseline assessments must be within 28 days prior to and including the date of randomization.
- All documented lesions must have non-missing assessments (ie, non-missing measurements for target lesions and non-missing lesions assessment status at baseline for non-target lesions).

5.2.11. Adequate post-baseline tumor assessment

An adequate post-baseline assessment is defined as an assessment where a response of CR, PR, SD, non-CR/non-PD, or PD can be determined (see Section 6.2.2.3). Time points where the response is not evaluable (NE) or no assessment was performed will not be used for determining the censoring date.

5.3. Methods to Manage Missing Data

5.3.1. Missing data

Unless otherwise specified, all data will be evaluated as observed, and no imputation method for missing values will be used.

In all patient data listings imputed values will be presented. In all listings imputed information will be flagged.

Missing statistics, eg when they cannot be calculated, should be presented as 'ND' or 'NA'. For example, if N=1, the measure of variability (SD) cannot be computed and should be presented as 'ND' or 'NA'.

5.3.1.1. Pharmacokinetic concentrations

Concentrations Below the Limit of Quantification

For all calculations, figures and estimation of individual pharmacokinetic parameters, all concentrations assayed as below the level of quantification (BLQ) will be set to zero. In log-linear plots these values will not be represented. The BLQ values will be excluded from calculations of geometric means and their CIs. A statement similar to 'All values reported as BLQ have been replaced with zero' should be included as a footnote to the appropriate tables and figures.

Deviations, Missing Concentrations and Anomalous Values

In summary tables and plots of median profiles, concentrations will be set to missing if one of the following cases is true:

1. A concentration has been reported as ND (ie, not done) or NS (ie, no sample);

2. A deviation in sampling time is of sufficient concern or a concentration has been flagged as anomalous by the clinical pharmacologist.

Summary statistics will not be presented at a particular time point if more than 50% of the data are missing. For analysis of pharmacokinetic concentrations, no values will be imputed for missing data.

5.3.1.2. Pharmacokinetic parameters

Whether actual or nominal PK sampling time will be used for the derivation of PK parameters will be determined by the results of interim PK analyses. If a PK parameter cannot be derived from a patient's concentration data, the parameter will be coded as NC (ie, not calculated). NC values will not be generated beyond the day that a patient discontinues.

In summary tables, statistics will be calculated by setting NC values to missing. Statistics will not be presented for a particular treatment if more than 50% of the data are NC. For statistical analyses (ie, analysis of variance), PK parameters coded as NC will also be set to missing.

If an individual patient has a known biased estimate of a PK parameter (due for example to a deviation from the assigned dose level), this will be footnoted in summary tables and will not be included in the calculation of summary statistics or statistical analyses.

5.3.1.3. Biomarker Parameters: Unintentional Duplicate Data

If an individual patient has unintentionally duplicated observations for a specific biomarker analyte(s) the following procedure will occur:

- For continuous data, the duplicate data points will be averaged, and the average added to the database;
- For non-continuous data (eg, listing of genes identified), the study team will discuss which, if any record, to select for analysis. A flag will be added to the database indicating which record is selected (including no record selected.)

For biomarker analytes for which it is intentional to have 2 samples (eg, denovo and archival tumor sample), these two samples will not be considered to be duplicate of each other, however, a variable in the database will identify the records.

5.3.2. Handling of incomplete dates

5.3.2.1. Disease history

Incomplete dates for disease history (eg, initial diagnosis date, date of documented, locally advanced, inoperable or metastatic disease diagnosis, date of response or progression in prior treatment) will be imputed as follows:

- If the day is missing, it will be imputed to the 15th day of the month.
- If both day and month are missing and the year is prior to the year of the first study treatment, the month and day will be imputed as July 1st.

- If both day and month are missing and the year is same as the year of the first study treatment, the month and day will be imputed as January 1st.
- If the date is completely missing, no imputation will be performed.

5.3.2.2. Adverse events

Incomplete AE-related dates will be imputed as follows:

- If the AE onset date is missing completely, then the onset date will be replaced by the start of study treatment.
- If only the day part of the AE onset date is missing, but the month and year are equal to the start of study treatment, then the AE onset date will be replaced by the start of study treatment. For example, if the AE onset date is --/JAN/2015, and study treatment start date is 15/JAN/2015, then the imputed AE onset date will be 15/JAN/2015.
- If both the day and month of the AE onset date are missing but the onset year is equal to the start of study treatment, then the onset date will be replaced by the start of study treatment. For example, if AE onset date is --/---/2014, and study treatment start date is 19/NOV/2014, then the imputed AE onset date will be 19/NOV/2014.
- In all other cases the missing onset day or missing onset month will be replaced by 1.
- Incomplete stop date will be replaced by the last day of the month (if day is missing only), if not resulting in a date later than the date of patient's death. In the latter case the date of death will be used to impute the incomplete stop date.
- In all other cases the incomplete stop date will not be imputed. If stop date of AE is after the date of cut-off, outcome of AE is ongoing at cut-off.

5.3.2.3. Prior and concomitant medications

Incomplete prior/concomitant medication dates will be imputed as follows:

- If the medication date is missing completely, then the medication date will be replaced by the start of study treatment.
- If the day of medication date is missing, but the month and year are equal to the start of study treatment, then the medication date will be replaced by the start of study treatment. For example, if the medication start date is --/JAN/2015, and study treatment start date is 15/JAN/2015, then the imputed medication start date will be 15/JAN/2015.
- If both the day and month of medication start date are missing but the start year is equal to the start of study treatment, then the medication date will be replaced by the start of study treatment. For example, if the medication start date is --/---/2014, and study treatment start date is 19/NOV/2014, then the imputed medication start date will be 19/NOV/2014.
- In all other cases the missing medication day or missing medication month will be replaced by 1.

- Incomplete stop date will be replaced by the last day of the month (if day is missing only), if not resulting in a date later than the date of patient's death. In the latter case the date of death will be used to impute the incomplete stop date.
- In all other cases the incomplete medication stop date will not be imputed.

5.3.2.4. Exposure

No imputation will be done for first dose date. Date of last dose of study drug, if unknown or partially unknown, will be imputed as follows:

- If the last date of study drug is completely missing and there is no End of Treatment eCRF page and no death date, the patient should be considered to be ongoing and use the cut-off date for the analysis as the last dosing date
- If the last date of study drug is completely or partially missing and there is EITHER an End of Treatment eCRF page OR a death date available (within the cut-off date), then imputed last dose date is:

  = 31DECYYYY, if only Year is available and Year < Year of min (EOT date, death date)

  = Last day of the month, if both Year and Month are available and Year = Year of min (EOT date, death date) and Month < the month of min (EOT date, death date)

  = min (EOT date, death date), for all other cases.

5.3.3. Imputation rules for date of last contact and efficacy assessments

5.3.3.1. Date of last contact

The date of last contact will be derived for patients not known to have died at the analysis cut-off using the latest complete date among the following:

- All patient assessment dates (blood draws (laboratory, PK), vital signs, performance status, ECG, tumor assessments)
- Start and end dates of anti-cancer therapies administered after study treatment discontinuation
- AE start and end dates
- Last date of contact collected on the ‘Survival Follow-up’ eCRF (do not use date of survival follow-up assessment unless status is ‘alive’)
- Study drug start and end dates
- Randomization date
- Withdrawal of consent date
- Date of discontinuation on disposition eCRF pages (do not use if reason for discontinuation is lost to follow-up).

Only dates associated with actual examinations of the patient will be used in the derivation. Dates associated with a technical operation unrelated to patient status such as the date a blood sample was processed will not be used. Assessment dates after the cut-off date will not be applied to derive the last contact date.

5.3.3.2. Death date

Missing or partial death dates will be imputed based on the last contact date:

- If the date is missing it will be imputed as the day after the date of last contact
- If the day or both day and month is missing, death will be imputed to the maximum of the full (non-imputed) day after the date of last contact and the following:
  - Missing day: 1st day of the month and year of death
  - Missing day and month: January 1st of the year of death

5.3.3.3. Tumor assessments

All investigation dates (eg, X-ray, CT scan) must be completed with day, month and year.

If there are multiple scan dates associated with an evaluation, ie, radiological assessments occur over a series of days rather than the same day, the choice of date of assessment could impact the date of progression and/or date of response. If there are multiple scan dates associated with an evaluation, the earliest of the scan dates associated with the evaluation will be used as the date of assessment.

If one or more investigation dates for an evaluation are incomplete but other investigation dates are available, the incomplete date(s) are not considered for calculation of the assessment date and assessment date is calculated as the earliest of all investigation dates (eg, X-ray, CT-scan).

If all measurement dates for an evaluation have no day recorded, the 1st of the month is used.

If the month is not completed, for any of the investigations for an evaluation, the respective assessment will be considered to be at the date which is exactly between the previous and the following assessment. If both a previous and following assessments are not available, this assessment will not be used for any calculations.

5.3.3.4. Date of start of new anti-cancer therapy

Incomplete dates for start date of new anti-cancer therapy (drug therapy, radiation, surgery) will be imputed as follows and will be used for determining censoring dates for efficacy analyses and in the derivation of the end of on-treatment period. PD date below refers to PD date by investigator assessment.

- The end date of new anti-cancer therapy will be included in the imputations for start date of new anti-cancer therapy. If the end date of new anti-cancer therapy is
  - completely missing then it will be ignored in the imputations below

- partially missing with only year (YYYY) available then the imputations below will consider 31DECYYYY as the end date of the new anti-cancer therapy
- partially missing with only month and year available then the imputations below will consider the last day of the month for MMMYYYY as the end date of the new anti-cancer therapy

- For patients who have not discontinued study treatment at the analysis cut-off date, last dose of study treatment is set to the analysis cut-off date in the imputations below.
- If the start date of new anti-cancer therapy is completely or partially missing then the imputed start date of new anti-cancer therapy is derived as follows:
  - Start date of new anti-cancer therapy is completely missing

    Imputed start date = min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy]
  - Only year (YYYY) for start of anti-cancer therapy is available

    IF YYYY < Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy] THEN imputed start date = 31DECYYYY;

    ELSE IF YYYY = Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy]

    THEN imputed start date = min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy]

    ELSE IF YYYY > Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy]

    THEN imputed start date = 01JANYYYY
  - Both Year (YYYY) and Month (MMM) for start of anti-cancer therapy are available

    IF

    YYYY = Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy], AND

    MMM < Month of min [max(PD date + 1 day, last dose of study treatment + 1 day), end date of new anti-cancer therapy]

    THEN

    imputed start date = DAY (Last day of MMM) MMM YYYY ;

    ELSE IF

    YYYY = Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy], AND

    MMM = Month of min [max(PD date + 1 day, last dose of study treatment + 1 day), end date of new anti-cancer therapy]

    THEN

imputed start date = min [max(PD date + 1 day, last dose of study treatment + 1 day), end date of new anti-cancer therapy]);

ELSE IF

YYYY = Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy], AND

MMM > Month of min [max(PD date + 1 day, last dose of study treatment + 1 day), end date of new anti-cancer therapy]

THEN

imputed start date = 01 MMM YYYY;

ELSE IF

YYYY < Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy]

THEN

imputed start date = DAY (Last day of MMM) MMM YYYY;

ELSE IF

YYYY > Year of min [max(PD date + 1, last dose of study treatment + 1), end date of new anti-cancer therapy]

THEN

imputed start date = 01 MMM YYYY.

5.3.3.5. PRO data

For the EORTC QLQ-C30 and QLQ-OV28, in cases where 2 answers are given to 1 item, the more severe answer (higher score for symptom domains, and lower score for the functioning/QoL domains of QLQ-C30; higher score for the QLQ- OV28) will be counted. If less than half of the constituent items on the QLQ-C30 and QLQ-OV28 have been answered for a multi-item subscale, that subscale will be considered missing. Single-item subscales will be considered missing if the constituent item is incomplete. Item 39 of the EORTC QLQ-Ov28 is a conditional question based on response to item 38. Item 39 should not be scored if the patient responded no hair loss (option 1) to item 38. For the EQ-5D-5L, in cases where 2 answers are given to 1 item, the item will be considered missing. For the EQ-5D-5L, questions not answered will be considered missing items and will not be utilized.

6. ANALYSES AND SUMMARIES

Refer to Section 4 for definitions of analysis sets and Section 5.2 for general methodology.

6.1. Primary Endpoints

6.1.1. Overall survival

6.1.1.1. Primary analysis

The following analyses will be based on the FAS using the strata assigned at randomization.

Overall survival (OS) is defined as the time from the date of randomization to the date of death due to any cause. Patients last known to be alive will be censored at date of last contact.

OS (months) = [date of death or censoring– date of randomization +1]/30.4375

The primary analysis of OS will compare the OS time between each of the experimental arms and the control arm, and will be performed using a 1-sided stratified log-rank test as described in Section 5.1.

The treatment effect, as measured by the hazard ratio, will be estimated using a Cox's Proportional Hazard model stratified by the randomization strata. Each stratum will define a separate baseline hazard function (using the 'STRATA' statement in SAS PROC PHREG), ie, for the i-th stratum the hazard function is expressed as: $h(i;t) = h(i,0;t) \exp(x\beta)$, where $h(i,0;t)$ defines the baseline hazard function for the i-th stratum and x defines the treatment arm (0=control arm, 1= experimental arm) and β is the unknown regression parameter.

Ties will be handled by replacing the proportional hazards model by the discrete logistic model (Ties=Discrete option in SAS PROC PHREG).

In order to account for the group sequential design in this study, the repeated CI (RCI) method (Jennison and Turnbull, 2000), will be used to construct the 2-sided RCIs for the hazard ratio at the interim and the final analyses of OS.

In addition, the unadjusted 95% CIs for the hazard ratio will also be reported at the interim and the final analyses for OS.

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median OS time with 2-sided 95% CIs. In particular, the OS rate at 6, 12, 18, 24 and 30 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice (2011) (conftype=loglog default option in SAS Proc LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

Frequency (number and percentage) of patients with an event (death) and censoring reasons will be presented by treatment arm. Reasons for censoring will be summarized according to the categories in Table 8 following the hierarchy shown.

Table 8. OS Censoring Reasons and Hierarchy

Hierarchy	Condition	Censoring Reason
1	No event and [withdrawal of consent date ≥ date of randomization OR End of study (EOS) = Subject refused further follow-up]	Withdrawal of consent
2	No event and [lost to follow-up in any disposition page OR data cut-off date – last contact date > 13 weeks]	Lost to follow-up
3	No event and none of the conditions in the prior hierarchy are met	Alive

The OS time or censoring time and the reasons for censoring will also be presented in a patient listing.

Time of Follow-Up for OS

A Kaplan-Meier plot for OS follow-up duration will also be generated to assess the follow-up time in the treatment arms reversing the OS censoring and event indicators. Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median time of follow-up for OS with 2-sided 95% CIs. In particular, the rate at 6, 12, 18, 24 and 30 months will be estimated with corresponding 2-sided 95% CIs.

6.1.2. Progression-free survival as assessed by BICR per RECIST v1.1

6.1.2.1. Primary analysis

The following analyses will be based on the FAS using the strata assigned at randomization. Progression of disease (PD) below refers to PD by BICR assessment.

Progression-Free Survival (PFS) is defined as the time from date of randomization to the date of the first documentation of PD or death due to any cause, whichever occurs first.

PFS data will be censored on the date of the last adequate tumor assessment for patients who do not have an event (PD or death), for patients who start a new anti-cancer therapy prior to an event (see Section 5.2.6) or for patients with an event after 2 or more missing tumor assessments. Patients who do not have an adequate baseline tumor assessment or who do not have an adequate post-baseline tumor assessment will be censored on the date of randomization unless death occurred on or before the time of the second planned tumor assessment (ie, ≤ 16 weeks after the date of randomization) in which case the death will be considered an event.

In this study antitumor activity will be assessed through radiological tumor assessments conducted at screening and every 8 weeks until disease progression regardless of initiation of subsequent anti-cancer therapy. The allowable time window for disease assessments is 5 days prior to dosing while on treatment and whenever disease progression is suspected (eg, symptomatic deterioration).

The censoring and event date options to be considered for the PFS and DR analysis are presented in Table 9.

PFS (months) = [date of event or censoring– date of randomization +1]/30.4375

Table 9. Outcome and event dates for PFS and DR analyses

Scenario	Date of event/censoring	Outcome
No adequate baseline assessment	Date of randomization [a]	Censored [a]
PD or death - After at most one missing or inadequate post-baseline tumor assessment, OR - ≤ 16 weeks after the date of randomization	Date of PD or death	Event
PD or death - after 2 or more missing or inadequate post-baseline tumor assessments	Date of last adequate tumor assessment [b] documenting no PD before new anti-cancer therapy is given or missed tumor assessments	Censored
No PD and no death	Date of last adequate tumor assessment [b] documenting no PD before new anti-cancer therapy is given or missed tumor assessments	Censored
Treatment discontinuation due to 'Disease progression' without documented progression	Not applicable	Information is ignored. Outcome is derived based on documented progression only.
New anti-cancer therapy given	Date of last adequate tumor assessment [b] documenting no PD before new anti-cancer therapy is given or missed tumor assessments	Censored

[a] However if the patient dies ≤16 weeks after the date of randomization the death is an event with date on death date

[b] If there are no adequate post-baseline assessments prior to PD or death, then the time without adequate assessment should be measured from the date of randomization; if the criteria were met the censoring will be on the date of randomization.

The primary efficacy analysis will compare the PFS time based on BICR assessment between each of the experimental arms and the control arm, and will be performed using a 1-sided stratified log-rank test as described in Section 5.1.

The treatment effect will be estimated using a Cox's Proportional Hazard model stratified by the randomization strata to calculate the hazard ratio. Each stratum will define a separate

baseline hazard function (using the 'STRATA' statement in SAS PROC PHREG), ie for the i-th stratum the hazard function is expressed as: $h(i;t) = h(i,0;t) \exp(x\beta)$, where $h(i,0;t)$ defines the baseline hazard function for the i-th stratum and x defines the treatment arm (0=control arm, 1= experimental arm) and β is the unknown regression parameter.

Ties will be handled by replacing the proportional hazards model by the discrete logistic model (Ties=Discrete option in SAS PROC PHREG).

In order to account for the group sequential design in this study, the repeated CI (RCI) method (Jennison and Turnbull, 2000), will be used to construct the 2-sided RCIs for the hazard ratio at the interim and the final analyses of PFS.

In addition, the unadjusted 95% CIs for the hazard ratio will also be reported at the interim and the final analyses for PFS.

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median PFS time with 2-sided 95% CIs. In particular, the PFS rate at 3, 6, 9, 12 and 15 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice (2011) (conftype=loglog default option in SAS Proc LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

Frequency (number and percentage) of patients with each event type (PD or death) and censoring reasons will be presented by treatment arm. Reasons for censoring will be summarized according to the categories in Table 10 following the hierarchy shown.

Table 10. PFS Censoring Reasons and Hierarchy

Hierarchy	Condition	Censoring Reason
1	No adequate baseline assessment	No adequate baseline assessment
2	Start of new anti-cancer therapy	Start of new anti-cancer therapy
3	Event after 2 or more missing or inadequate post-baseline tumor assessments/date of randomization	Event after 2 or more missing assessments [a]
4	No event and [withdrawal of consent date ≥ date of randomization OR End of study (EOS) = Subject refused further follow-up]	Withdrawal of consent
5	No event and lost to follow-up in any disposition page	Lost to follow-up
6	No event and [EOS present OR disposition page for any epoch after screening says patient will not continue into any subsequent phase of the study] and no adequate post-baseline tumor assessment	No adequate post-baseline tumor assessment
7	No event and none of the conditions in the prior hierarchy are met	Ongoing without an event

[a] 2 or more missing or inadequate post-baseline tumor assessments.

The PFS time or censoring time and the reasons for censoring will also be presented in a patient listing.

Time of Follow-Up for PFS

A plot will be generated to compare planned and actual relative day of tumor assessments by treatment arm. A Kaplan-Meier plot for PFS follow-up duration will also be generated to assess the follow-up time in the treatment arms reversing the PFS censoring and event indicators. Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median time of follow-up for PFS with 2-sided 95% CIs. In particular, the rate at 3, 6, 9, 12 and 15 months will be estimated with corresponding 2-sided 95% CIs.

6.2. Secondary Endpoints

6.2.1. Safety endpoints

Refer to Section 6.6.

6.2.2. Efficacy endpoints

The following analyses will be based on the FAS by treatment arm unless otherwise specified. Assessment of response will be made using RECIST v1.1.

Analyses for tumor-related endpoints will be performed based on BICR assessment and based on Investigator assessment. PFS by investigator assessment will be analyzed as a secondary endpoint using the same methodology that is described in Section 6.1.2.1 now referring to PD by investigator assessment instead of PD by BICR assessment. RCIs will not be constructed for the hazard ratio of PFS by investigator assessment.

6.2.2.1. Sensitivity analyses for overall survival

The following sensitivity analyses will be performed to explore the robustness of the primary analysis results for OS. These analyses are regarded as purely exploratory. The sensitivity analyses will repeat the primary analysis (p-value, HR and 95% CIs) described in Section 6.1.1.1 with the modifications below:

- PP analysis set for OS;
- unstratified;
- using strata derived according to eCRF data instead of those entered in IRT. For the stratification factor of bulky disease (presence of a tumor ≥5 cm vs not), if the investigator identifies in the eCRF a target lesion with diameter ≥5 cm, then this is considered bulky disease per eCRF data; since the diameter for non-target lesions is not collected in the eCRF, if there are no target lesions with diameter ≥5 cm and there is no protocol deviation for bulky disease status then the IRT stratification value for the bulky disease strata will be used for this assessment; if there are no target lesions with diameter

≥5 cm and there is a protocol deviation for bulky disease status then the opposite of the IRT stratification value for the bulky disease strata will be used for this assessment.

Methods for evaluating the validity of model assumptions

The proportional hazards assumption will be checked visually by plotting log(-log(OS)) versus log(time) within each randomization stratum.

Schoenfeld residuals for the stratified Cox proportional regression model will be plotted to investigate graphically violations from the proportional hazards (PH) assumption; a non-zero slope is evidence of departure from PH. The PH assumption will be formally tested using Schoenfeld's residual test (Schoenfeld, 1980; Therneau & Grambsch, 2000). Large departures from PH will be evidenced by a p-value <0.05.

If these show large departures from proportional hazards, then OS will also be analyzed based on restricted mean survival time (RMST) differences (Zhang, 2013).

Restricted Mean Survival Time (RMST)

The hazard ratio estimate from the Cox proportional hazard model is routinely used to empirically quantify the between-arm difference under the assumption that the ratio of the two hazard functions is constant over time. When this assumption is plausible, such a ratio estimate may capture the relative difference between two survival curves. However, the clinical meaning of such a ratio estimate is difficult, if not impossible, to interpret when the underlying PH assumption is violated (ie, the hazard ratio is not constant over time).

The RMST is a robust and clinically interpretable summary measure of the survival time distribution. Unlike median survival time, it is estimable even under heavy censoring. There is a considerable body of methodological research (eg, Royston and Parmar, 2011; Uno, Wei, et al., 2014; Zhang, 2013) about the use of RMST to estimate treatment effects as an alternative to the hazard ratio approach.

The RMST methodology is applicable independently of the PH assumption and can be used, at a minimum, as a sensitivity analysis to explore the robustness of the primary analysis results. However, when large departures from the PH assumption are observed, the log-rank test is underpowered to detect differences between the survival distributions for the treatment arms, and a test of the difference between the RMST for the experimental arm and the control arm may be more appropriate to determine superiority of the experimental arm compared to the control arm with respect to the time-to-event endpoint.

In particular, as it pertains to the **cut-off point (τ)** to evaluate the RMST, it is noted that the cut-off point should not exceed the minimum of the largest observed time for both treatment arms so that the RMST of all treatment arms being evaluated can be adequately estimated and comparison between treatments is feasible; τ should be clinically meaningful and closer to the end of the study follow-up so that the majority of survival outcomes will be covered by the time interval. The RMST up to time τ can then be interpreted as the expected survival time restricted to the common follow-up time τ among all patients. The selection of τ should ensure that the RMST evaluation will not go beyond the maximum time point where the

evaluation can be performed while also taking into account a large period of time that is expected to provide a meaningful assessment of treatment effect. To avoid arbitrary selection of the common cut-off τ for both treatment arms, three sets of analyses will be performed:

- τ_1 = minimum of (largest observed survival time for experimental arm, largest observed survival time for control arm).
- τ_2 = minimum of (largest survival event time for experimental arm, largest survival event time for control arm).
- τ_3 = midpoint between τ_1 and τ_2.

The treatment effect between each of the avelumab containing arms and PLD arm will be assessed based on the difference in RMST. The associated 95% CI for the difference in means and 1-sided p-value will be generated. RMST as a function of τ and the associated treatment effect between each of the avelumab containing arms and PLD arm will be plotted against time τ .

Exploratory analyses to investigate the impact of potential prognostic or effect modifying (predictive) factors

See subgroups as defined in Section 6.4.

Multivariable Cox regression analysis will be carried out to assess and adjust the treatment effect for relevant baseline factors of potential prognostic impact. A stepwise selection procedure will serve to identify explanatory variables of potential prognostic values additional to the randomization strata which will be included in all models during the selection procedure. The Cox's Proportional Hazard model is defined as:

$$h(t) = h(0;t)\, e^{Xb}$$

where h(0;t) defines the baseline hazard function and X defines the vector of explanatory variables and b the unknown vector of regression parameters.

In the stepwise selection procedure, variables are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination. The level of significance for an explanatory variable to enter the model is set to 0.15 (p-value of Score test) and the significance level for removing it is set to 0.40 (p-value of Wald test). This analysis will be performed using the stepwise selection method in SAS (Proc PHREG). Once this procedure stops, the factor 'treatment arm' will be added to the last selected model in order to evaluate the effect of treatment on OS time when adjusted for the selected explanatory variables. The hazard ratios of all selected explanatory variables and of treatment effects will be reported including 2-sided 95% CIs. No interactions will be considered. Post-baseline factors will not be considered for the model.

6.2.2.2. Sensitivity analyses for progression-free survival

The following sensitivity analyses will be performed to explore the robustness of the primary analysis results. These analyses are regarded as purely exploratory. The sensitivity analyses will repeat the primary analysis (p-value, HR and 95% CIs) described in Section 6.1.2.1 with the modifications below:

- PFS based on BICR assessment and counting all PD and deaths as events regardless of missing assessments or timing of the event
- PFS based on BICR assessment on the PP analysis set for PFS
- PFS based on BICR assessment using an unstratified analysis
- PFS based on BICR assessment using strata derived according to eCRF data instead of those entered in IRT. See Section 6.2.2.1 for the derivation associated with the stratification factor of bulky disease.
- PFS based on BICR assessment modifying the censoring rules in Table 9 to consider all deaths as events
- PFS based on BICR assessment modifying the censoring rules in Table 9 with initiation of subsequent anti-cancer therapies not used as a censoring reason

Methods for evaluating the validity of model assumptions

The same methodology described in Section 6.2.2.1 for OS will be used for PFS based on BICR assessment.

BICR vs Investigator assessment

A summary of the BICR assessment versus investigator assessment will be provided including numbers of concordant and discordant assessments as well as the number of cases where PFS event was assessed at different timepoints based on BICR and investigator assessments.

Table 11 outlines the possible outcomes by investigator and BICR (Amit et al. 2011).

Table 11. Possible Outcomes for Investigator vs BICR

		BICR	
		Event	**No Event**
Investigator	**Event**	a = a1 + a2 + a3	b
	No Event	c	d

a1: number of agreements on timing and occurrence of event;
a2: number of times agreement on event but INV declares event later than BICR;
a3: number of times agreement on event but INV declares event earlier than BICR;
N=a+b+c+d.
The timing agreement of event is defined as a window of ± 7 days.

The following measure of discordance will be calculated for each treatment arm:

- Total Event Discrepancy Rate: (b+c) / N
- Early Discrepancy Rate (EDR): (a3+b) / (a+b)
- Late Discrepancy Rate (LDR): (a2+c) / (a2+a3+b+c)
- Overall Discrepancy Rate: (a2+a3+b+c) / N

The EDR represents the positive predictive value of investigator assessment and quantifies the frequency with which the investigator declares PFS event earlier than BICR within each treatment arm as a proportion of the total number of investigator assessed events.

The LDR quantifies the frequency with which the investigator declares PFS event later than BICR as a proportion of the total number of discrepancies within the treatment arm.

Discordance metrics are calculated for each treatment arm and, for each metric, the difference in discordance between the experimental and control arms is used to evaluate potential bias. If the discordance is similar across the treatment arms then this suggests the absence of evaluation bias favoring a particular treatment arm. A negative differential discordance for EDR and/or a positive differential discordance for LDR may be indicative of investigator evaluation bias in favor of the experimental arm (Amit et al, 2011).

Exploratory analyses to investigate the impact of potential prognostic or effect modifying (predictive) factors

The same methodology described in Section 6.2.2.1 for OS will be used for PFS based on BICR assessment.

6.2.2.3. Objective response and tumor shrinkage

Best overall response (BOR) will be assessed based on reported overall lesion responses at different evaluation time points from the date of randomization until the first documentation of PD, according to the following rules.

BOR Based on Confirmed Responses:

- CR = at least two determinations of CR at least 4 weeks apart and before first documentation of PD
- PR = at least two determinations of PR or better (PR followed by PR or PR followed by CR) at least 4 weeks apart and before first documentation of PD (and not qualifying for a CR)
- SD (applicable only to patients with measurable disease at baseline) = at least one SD assessment (or better) ≥ 6 weeks after the date of randomization and before first documentation of PD (and not qualifying for CR or PR).
- Non-CR/non-PD (applicable only to patients with non-measurable disease at baseline) = at least one non-CR/non-PD assessment (or better) ≥ 6 weeks after the date of randomization and before first documentation of PD (and not qualifying for CR or PR).
- PD = first documentation of PD ≤ 12 weeks after the date of randomization (and not qualifying for CR, PR, SD or non-CR/non-PD).
- NE: all other cases.

Only tumor assessments performed on or before the start of any further anti-cancer therapies will be considered in the assessment of BOR. Clinical deterioration will not be considered as documented disease progression.

An objective status of PR or SD cannot follow one of CR. SD can follow PR only in the rare case that tumor increases by less than 20% from the nadir, but enough that a previously documented 30% decrease from baseline no longer holds. If this occurs, the sequence PR-SD-PR is considered a confirmed PR. A sequence of PR – SD – SD – PD would be a best response of SD if the window for SD definition has been met.

Objective Response (OR) is defined as a confirmed BOR of CR or PR according to RECIST v1.1. Both CR and PR must be confirmed by repeat assessments performed no less than 4 weeks after the criteria for response are first met.

Patients who do not have a post-baseline radiographic tumor assessment due to early progression, who receive anti-tumor therapies other than the study treatments prior to reaching a CR or PR, or who die, progress, or drop out for any reason prior to reaching a CR or PR will be counted as non-responders in the assessment of OR. Each patient will have an objective response status (0: no OR; 1: OR). OR rate (ORR) is the proportion of patients with OR in the analysis set.

ORR by treatment arm will be calculated along with the 2-sided 95% CI using the Clopper-Pearson method (exact CI for a binomial proportion as computed by default by the FREQ procedure using the EXACT option).

In addition, the frequency (number and percentage) of patients with a confirmed BOR of CR, PR, SD, non-CR/non-PD (applicable only to patients with non-measurable disease at baseline), PD and NE will be tabulated. Patients with confirmed BOR of NE will be summarized by reason for having NE status. The following reasons will be used:

- No baseline assessment
- No post-baseline assessments due to death
- No post-baseline assessments due to other reasons
- All post-baseline assessments have overall response NE
- New anti-cancer therapy started before first post-baseline assessment
- SD of insufficient duration (<6 weeks after the date of randomization without further evaluable tumor assessments)
- PD too late (>12 weeks after the date of randomization)

Special and rare cases where BOR is NE due to both SD of insufficient duration and late PD will be classified as 'SD of insufficient duration'.

The association of study treatment and OR will be tested by the General Association Statistic of the Cochran-Mantel-Haenszel test (CMH) with the randomization strata taken into account. The null hypothesis of no association in any of the randomization strata is tested against the alternative, which specifies that there is an association between study treatment and tumor response at least in one randomization stratum. The CMH test will be performed at 1-sided alpha level of 0.025.

The stratified odds ratio in terms of OR will also be estimated along with its 95% CI to compare study treatments. The odds ratio is defined as the odds of OR with experimental treatment divided by the odds of OR with control treatment. The Breslow-Day test will be used to check the homogeneity of the odds ratio across the randomization strata. It tests the null hypothesis that odds ratios in all strata are equal against the alternative hypothesis that at least in one stratum the odds ratio is different.

In case the null hypothesis of homogeneity of odds ratios across strata is not rejected at the 2-sided alpha level of 0.05, the common odds ratio will be determined using the Mantel-Haenszel estimate (by the FREQ procedure using CMH option in SAS); if the null hypothesis of homogeneity of odds ratio across all strata is rejected, the odds ratio per stratum will be calculated with the corresponding exact CI.

BICR vs Investigator Assessment:

Table 12 outlines the possible BOR outcomes by investigator and BICR.

Table 12. Possible BOR Outcomes for Investigator vs BICR

BOR		BICR Assessment					
		CR	**PR**	**SD**	**Non-CR/ non-PD**	**PD**	**NE**
Investigator Assessment	**CR**	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}
	PR	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}
	SD	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{36}
	Non-CR/ non-PD	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}
	PD	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	n_{56}
	NE	n_{61}	n_{62}	n_{63}	n_{64}	n_{65}	n_{66}

$\sum_{i=1}^{6}(n_{ii})$ is the number of agreements on BOR between BICR and Investigator

$\sum_{i,j=1}^{6}(n_{ij})$ for i ≠ j is the number of disagreements on BOR between BICR and Investigator

N= $\sum_{i,j=1}^{6}(n_{ij})$

The following measures of concordance will be calculated for each treatment arm:

- Concordance rate for BOR = $\sum_{i=1}^{6}(n_{ii})\ /N$
- Concordance rate for response = $[\sum_{i,j=1}^{2}(n_{ij}) + \sum_{i,j=3}^{6}(n_{ij})]/N$

Concordance rates are calculated for each treatment arm and, for each metric, the difference in concordance between the experimental and control arms are used to evaluate potential bias. If the concordance is similar across the treatment arms then this suggests the absence of evaluation bias favoring a particular treatment arm.

Tumor Shrinkage from baseline:

Tumor shrinkage will be summarized as the percent change from baseline in target lesions (sum of longest diameter for non-nodal lesion and short axis for nodal lesion) per time point. It will be derived as:

- ((Sum of target lesions at week XX – sum of target lesions at baseline)/sum of target lesions at baseline) × 100

The maximum reduction in target lesions from baseline will be derived across all the post-baseline assessments until documented disease progression, excluding assessments after start of subsequent anti-cancer therapy, as:

- Minimum of ((sum of target lesions at week XX – sum of target lesions at baseline)/sum of target lesions at baseline) × 100

A waterfall plot of maximum percent reduction in the sum of longest diameter for non-nodal lesions and short axis for nodal lesions from baseline will be created by treatment arm. These plots will display the best percentage change from baseline in the sum of the diameter of all target lesions for each patient with measurable disease at baseline and at least one post-baseline assessment.

6.2.2.4. Disease control

Disease Control (DC) is defined as BOR of CR, PR, non-CR/non-PD or SD. Both CR and PR must be confirmed by repeat assessments performed no less than 4 weeks after the criteria for response are first met. DC rate (DCR) is the proportion of patients with DC.

DCR will be summarized by frequency counts and percentages.

6.2.2.5. Duration of response

Duration of Response (DR) is defined, for patients with OR, as the time from the first documentation of objective response (CR or PR) to the date of first documentation of PD or death due to any cause. If a patient has not had an event (PD or death), DR is censored at the date of last adequate tumor assessment. The censoring rules for DR are as described for PFS in Table 9.

DR (months) = [date of event or censoring– first date of OR +1]/30.4375

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median DR time with 2-sided 95% CIs. In particular, the DR rate at 3, 6, 9, 12 and 15 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice (2011) (conftype=loglog default option in SAS Proc LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

DR will be displayed graphically and analyzed using Kaplan-Meier methodology. If the number of patients with OR is small, the Kaplan-Meier method may not provide reliable estimates. In this case, only descriptive statistics or listings will be provided.

6.2.2.6. Time to response

Time to response (TTR) is defined, for patients with OR, as the time from the date of randomization to the first documentation of objective response (CR or PR) which is subsequently confirmed.

TTR (in months) = [first date of OR – date of randomization +1]/30.4375

TTR will be summarized using simple descriptive statistics (mean, SD, median, min, max. Q1, Q3).

6.2.3. Pharmacokinetic endpoints

The following pharmacokinetic analyses will be based on the PK analyses set by treatment arm.

C_{trough} and C_{max} for avelumab and doxorubicin will be summarized descriptively (n, mean, SD, CV, median, minimum, maximum, geometric mean, its associated CV, and 95% confidence interval) by treatment arm, cycle, and day. Other standard parameters will be calculated including, but not limited to $AUC_{24,}$ T_{max}, AUC_{last}, T_{last}, $AUC_{sd,\tau}$, $AUC_{ss,\tau}$, $t_{1/2}$, CL, Vd_z, $C_{ss,av}$ as data permit.

Dose normalized parameters (eg, CDN- C_{max}, DN-AUC) will be reported as appropriate. The trough concentrations for avelumab and doxorubicin (for each dose as needed) will be plotted using a box whisker plot by cycle and day in order to assess the attainment of steady state.

Pharmacokinetic parameters for avelumab and doxorubicin will be taken from observed values or derived from plasma concentration-time data as described in Section 3.2.4.

Presentation of pharmacokinetic data will include:

- Descriptive statistics (n, mean, SD, %CV, median, minimum, maximum) of plasma concentrations will be presented in tabular form by treatment arm, dose level (for doxorubicin as needed), cycle, day and nominal time. Additionally similar descriptive statistics will also be generated for dose-normalized avelumab and doxorubicin pharmacokinetic parameters.
- Linear-linear and log-linear plots of mean and median plasma concentrations by nominal time for avelumab and doxorubicin will be presented for PK sampling days by treatment arm, cycle, and study day. Similar plots will be presented for each individual patient concentrations. Patients who have undergone intrapatient dose reduction or escalation will be excluded from the median plasma concentration-time plots.
- Pharmacokinetic parameters for avelumab and doxorubicin will be listed and summarized by treatment arm/dose level (for doxorubicin as needed), cycle and study day using descriptive statistics (n, mean, SD, %CV, median, minimum, maximum, geometric mean and its associated %CV, and 95% confidence interval). For Tmax, the range (min, max) will also be provided. PK parameters with zero values will be excluded from the calculation of geometric means and their associated %CV. If an intrapatient dose

escalation or reduction occurs, dose-dependent PK parameters (AUC and Cmax) for that patient may be dose-normalized when it is known that the drug exhibits linear PK within the dose range and other PK parameters will be reported as estimated; or may only be included in descriptive statistics and summary plots up to the time of the dose change. In addition, dose-normalized Cmax and AUC parameters will be summarized (as described above) using data pooled across treatment arms in which different avelumab and doxorubicin doses were administered.

- Box plots for AUC and Cmax for avelumab and doxorubicin will be generated. Individual data points, the geometric mean and the median of the parameter in each treatment will be overlaid on the box plots. If a treatment arm has limited evaluable PK data (n<4), matchstick plots showing changes in AUC and Cmax for each drug in individual patients will then be generated. The geometric mean of the parameter in each treatment will be overlaid in the plots.

- Ctrough and Cmax for avelumab will be plotted for each treatment arm using a box-whisker plot by cycle and day within cycle in order to assess the attainment of steady-state.

Assessment of drug-drug interaction

Effect of Avelumab on Doxorubicin (PLD Samples) Pharmacokinetics

The effect of avelumab dosing on doxorubicin PK will be evaluated based on overall assessment of the geometric mean ratios and associated 90% CI for C_{max}, AUC_{inf}, and AUC_{τ} of PLD on Day 1 of Cycle 2 in Arm B (avelumab plus PLD) compared to those on Day 1 of Cycle 2 of Arm C.

Effect of PLD on Avelumab Pharmacokinetics

The effect of PLD dosing on avelumab PK will be evaluated based on the overall assessment of the geometric mean ratios of C_{max} and Ctrough of Arm B on Day 1 of Cycle 2 compared to those on Day 1 of Cycle 2 of Arm A.

6.2.4. Population pharmacokinetic endpoints

Pharmacokinetic and pharmacodynamic data from this study may be analyzed using modeling approaches and may also be pooled with data from other studies to investigate any association between avelumab exposure and biomarkers or significant safety/efficacy endpoints. The results of these analyses, if performed, may be reported separately.

6.2.5. Biomarker endpoints

Secondary endpoints in the study are candidate predictive biomarkers in tumor tissue including, but not limited to, PD-L1 expression and tumor infiltrating CD8+ T lymphocytes as assessed by immunohistochemistry (IHC).

Biomarker data will be analyzed based on the biomarker analysis sets as defined in Section 4.3.3, by treatment arm.

For PD-L1 expression, patients will be classified as positive and negative according to scoring algorithms and cut-offs established from internal or external sources. Patients whose status cannot be determined are not considered to have screening biomarker assessment per the biomarker analysis set definition, and therefore will be excluded.

For CD8+, if at the time of reporting an internal or external standard becomes available, the above mentioned approach for PD-L1 will be used. Otherwise, data will be categorized using quartiles to define CD8+ subgroups.

The following analyses will be performed for each biomarker secondary endpoint.

Biomarker subgroups as defined above will be used to perform subgroup analyses for efficacy endpoints (BOR and PFS by BICR assessment, OS) using the methodology outlined in Section 6.4. In addition, for PFS by BICR assessment and OS, the hazard ratio for the biomarker subgroup level comparisons and the unadjusted 95% CIs for the hazard ratio will be reported for each treatment arm.

CCI [redacted]

CA-125 levels. CCI [redacted]

6.2.6. Endpoints for immunogenicity data of avelumab

All analyses described below are performed by treatment arm containing avelumab (Arm A and Arm B), and for the avelumab-containing treatment arms combined unless otherwise specified.

Blood samples for avelumab immunogenicity testing will be collected within 2 hours before the start of the avelumab infusion on Day 1 of Cycles 1 through 4. Thereafter, a sample will be collected on Day 1 prior to avelumab infusion in Cycles 5, 6, 9, 12, 15, 18, 21 24 and at the End of Treatment. An additional sample must be collected at 30 days after the last dose of avelumab.

Samples positive for ADA will be analyzed for titer and may be analyzed for nAb. As of the finalization of this SAP, the nAb assay is not yet available, therefore the analyses of nAb data described in the following sections will only be conducted contingent upon assay and data availability at the time of reporting.

Patients will be characterized into different ADA categories based on the criteria defined in Table 13.

Table 13 Patients Characterized Based on Anti-Drug Antibody Results (ADA Status)

Category	Definition	Subjects at Risk (Denominator for Incidence)
ADA never-positive	No positive ADA results at any time point; ADA-negative patients (titer < cutpoint)	Number of patients with at least one valid ADA result at any time point
ADA ever-positive	At least one positive ADA result at any time point; ADA-positive patients (titer ≥ cutpoint)	Number of patients with at least one valid ADA result at any time point
Baseline ADA positive	A positive ADA result at baseline	Number of patients with valid baseline ADA result
Treatment-boosted ADA	A positive ADA result at baseline and the titer ≥ 8×baseline titer at least once after treatment with avelumab	Number of patients with valid baseline ADA results and at least one valid post-baseline ADA result
Treatment-induced ADA	Patient is ADA-negative at baseline and has at least one positive post-baseline ADA result; or if patient does not have a baseline sample, the patient has at least one positive past-baseline ADA result	Number of patients with at least one valid post-baseline ADA result and without positive baseline ADA result (including missing, NR)
Transient ADA response	If patients with treatment-induced ADA have (a single positive ADA result or duration between first and last positive result <16 weeks) and ADA result at the last assessment is not positive.	Number of patients with at least one valid post-baseline ADA result and without positive baseline ADA result (including missing, NR)
Persistent ADA response	If patients with treatment-induced ADA have duration between first and last positive ADA result ≥16 weeks or a positive ADA result at the last assessment	Number of patients with at least one valid post-baseline ADA result and without positive baseline ADA result (including missing, NR)

ADA: anti-drug antibody, NR = not reportable.

Patients will be characterized into different nAb categories based on the criteria in Table 14. For nAb, treatment-boosted is not applicable since no titer result is available.

Table 14 Patients Characterized Based on Neutralizing Antibody Results (nAb Status)

Category	Definition	Subjects at Risk (Denominator for Incidence)
nAb never-positive	No positive nAb results at any time point	Number of patients with at least one valid ADA result at any time point
nAb ever-positive	At least one positive nAb result at any time point	Number of patients with at least one valid ADA result at any time point
Baseline nAb positive	A positive nAb result at baseline	Number of patients with valid baseline ADA result
Treatment-induced nAb	Patient is not nAb positive at baseline and has at least one positive post-baseline nAb result; or if patient does not have a baseline sample, the patient has at least one positive past-baseline ADA result	Number of patients with at least one valid post-baseline ADA result and without positive baseline nAb result (including missing, NR)
Transient nAb response	If patients with treatment-induced nAb have (a single positive nAb result or duration between first and last positive result <16 weeks) and nAb result at the last assessment is not positive.	Number of patients with at least one ADA valid post-baseline result and without positive baseline nAb result (including missing, NR)
Persistent nAb response	If patients with treatment-induced nAb have duration between first and last positive nAb result ≥16 weeks or a positive nAb result at the last assessment	Number of patients with at least one valid post-baseline ADA result and without positive baseline nAb result (including missing, NR)

ADA = antidrug antibody, nAb = neutralizing antibody, NR = no result.

The number and percentage of patients in each ADA and nAb category will be summarized.

6.2.6.1. Time to and Duration of ADA and nAb response

The ADA and nAb analyses described below will include patients with treatment-induced ADA or nAb, respectively.

Time (weeks) to ADA response is defined as:

(Date of first positive ADA result – date of first dose of avelumab + 1)/7.

Time to ADA response will be summarized using simple descriptive statistics (mean, SD, median, min, max. Q1, Q3).

Duration (weeks) of ADA response is defined as:

(Date of last positive ADA result – date of first positive ADA result + 1)/7.

Duration of ADA response will be censored if:

- the last ADA assessment is positive AND patient is ongoing treatment with avelumab, or

- the last ADA assessment is positive AND patient discontinued treatment with avelumab AND the last planned ADA assessment (30 days after the last dose of avelumab) is after the cut-off date.

Time to nAb response and duration of nAb response are defined similarly based on first and last positive nAb result.

Kaplan-Meier estimates (product-limit estimates) will be presented together with a summary of associated statistics including the median ADA response time with 2-sided 95% CIs. ADA response rates at different timepoints will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates will be derived using the log-log transformation according to Kalbfleisch and Prentice (2011) (conftype=loglog default option in SAS Proc LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

Duration of ADA response will be displayed graphically and analyzed using Kaplan-Meier methodology. If the number of patients with ADA response is small, the Kaplan-Meier method may not provide reliable estimates. In this case, only descriptive statistics or listings will be provided

As data permit, the analyses described above will be repeated for patients with treatment-induced nAb.

6.2.6.2. ADA titer

For patients who are ADA ever positive, the maximum observed ADA titer for a patient will be summarized, overall and by ADA subcategories (baseline AEA positive, treatment-boosted ADA, treatment-induced ADA, transient ADA response, persistent ADA response) of patients having each discrete maximum titer value will be tabulated. The denominator to calculate the percentages will be the total number of patients in the associated ADA subcategory.

For patients with treatment-induced ADA, a cross tabulation of duration of ADA response and maximum ADA titer will be provided. The following categories for duration of ADA response will be used: ≤1, >1 to ≤3, >3 to ≤5, >5 to ≤7, >7 to ≤13, >13 to ≤16, >16 to ≤25, >25 weeks. In this categorization, the censoring in duration of ADA response is ignored.

6.2.6.3. Analysis of PK, safety and efficacy by immunogenicity status

The following ADA and nAb status will be used for the analyses described below.

ADA

- ADA ever-positive versus ADA never-positive
- ADA: treatment-induced ADA versus ADA never-positive or baseline ADA positive

nAb

- nAb ever-positive versus nAb never-positive
- nAb: treatment-induced nAb versus nAb never-positive or baseline nAb positive

Data listings will include immunogenicity data together with relevant PK, safety and efficacy data.

PK parameters and immunogenicity status

The following analyses will include patients in both the immunogenicity analysis set and in the PK parameter analysis set. The PK endpoints pertinent to the immunogenicity analyses are C_{trough} and C_{max}.

Blood samples for avelumab PK will be collected pre-dose and at the end of infusion (immediately before the end of avelumab infusion) on Days 1 and 15 of Cycles 1 to 4. Pre-dose samples should be collected on Days 1 of Cycles 5, 6, 9, 12, 15, 18, 21 24;, at the End of Treatment and at Day 30 Safety Follow-up visit.

C_{trough} and C_{max} will be summarized descriptively (n, mean, SD, CV, median, minimum, maximum, geometric mean, its associated CV, and 95% CI) by nominal time and ADA status. Linear-linear and log-linear plots of mean and median for C_{trough} and C_{max} over nominal time and by ADA status will be presented.

Among patients with treatment-induced ADA, analyses will be conducted to assess whether C_{trough} and C_{max} have any changes before and after the first positive ADA assessment. To be included in this analysis, patients must have the same PK parameter available both before and after the first positive ADA assessment. Relative PK day will be calculated as:

(PK assessment nominal day) – (first positive ADA assessment nominal day).

Nominal day is the protocol scheduled timing for an assessment. For example, if C_{trough} is collected on Day 1 of Cycle 2 and the first positive ADA result is observed on Day 1 of Cycle 3, then the relative PK day for this C_{trough} is -28. Linear-linear and log-linear plots of mean and median for C_{trough} and C_{max} over relative PK day will be presented.

As data permit, the analyses described above will be repeated for nAb.

Safety and immunogenicity status

The following analyses will include patients in the immunogenicity analysis set.

The frequency (number and percentage) of patients with each of the following will be presented by ADA status.

- TEAEs, by SOC and PT

- TEAEs leading to dose reduction of avelumab, by SOC and PT
- TEAEs leading to discontinuation of avelumab, by SOC and PT
- TEAEs leading to discontinuation of study treatment by SOC and PT
- Grade ≥ 3 TEAEs, by SOC and PT
- SAEs, by SOC and PT
- IRRs, by PT

For patients who had at least one IRR and have treatment-induced ADA, time related to first onset of an IRR (infusion 1, infusion 2, infusion 3, infusion 4 or later) will be summarized taking into account whether the IRR occurred on or after the first ADA positive assessment or whether the IRR occurred before the first ADA positive assessment.

As data permit, the analyses described above will be repeated for nAb.

Efficacy and immunogenicity status

For the ADA ever-positive patients, a listing will be prepared with patient ID, start and stop of avelumab treatment, date of first positive ADA result, time to ADA response, duration of ADA response, date of last ADA positive result, BOR, DR, PFS time or censoring time and reason for censoring, and OS time or censoring time and reason for censoring. If applicable, date of first positive nAb result, time to nAb response, duration of nAb response, date of last nAb positive result will also be presented. Tumor-related endpoints will be presented based on BICR assessment and based on Investigator assessment.

For the ADA ever-positive patients, the percent change from baseline in target lesions as well as the first occurrence of a new lesion and patient off avelumab treatment will be displayed against time point (weeks) in a line plot. Additional symbols will indicate the first and last ADA positive result and, if applicable, the first and last nAb positive result. Plot will be presented separately based on BICR assessment and based on Investigator assessment.

6.2.7. PRO endpoints

All PRO analysis will be based on FAS.

Patient-reported outcomes of HRQoL and ovarian cancer specific disease/treatment-related symptoms will be evaluated using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire – Core 30 (EORTC QLQ-C30) and its corresponding module for ovarian cancer, the Quality of Life Questionnaire – Ovarian Cancer 28 (QLQ-OV28). The EuroQoL (EQ-5D-5L) questionnaire will also be collected to assess general health status.

The primary PRO endpoint will be the time to deterioration analysis of the abdominal/GI domain of the EORTC QLQ-OV28 based on the ≥15 points threshold, since abdominal/GI symptoms are predominant and burdensome in this patient population (Cella, 2003).

6.2.7.1. PRO Scoring Procedure

The subscales of the EORTC QLQ-C30 and the QLQ-OV28 will be scored based on the EORTC scoring manual (Fayer et al. 2001). In summary, each scale of the EORTC QLQ-C30 and the QLQ-OV28 will be transformed so that scale scores will range from 0 to 100. The transformation will proceed in two steps. First, the average of the items contributing to a subscale will be calculated to compute the raw score of the scale. Next, a linear transformation will be applied to 'standardize' the raw score. After scores are transformed, higher scores on the EORTC QLQ-C30 or the QLQ-OV28 will represent higher ("better") levels of functioning and/or a higher ("worse") level of symptoms.

The EQ-5D-5L will be scored according to its scoring manual and UK weights will be applied. (EuroQol Group, 2015).

6.2.7.2. Instrument Completing Rate

For each treatment arm and at each time point, the number and percentage of patients who complete the QLQ-C30, QLQ-OV28, and EQ-5D-5L will be summarized, as will the reasons for non-completion of these measures. An instrument is considered complete if at least one item was answered by the patient.

For each treatment arm and at each time point, the number and percentage of patients who complete a specific subscale of the QLQ-C30 and QLQ-OV28 will also be summarized.

The summary will include a tabulation of the number and percentage of patients in each treatment arm who completed the summary scales at each time point.

For QLQ-OV28, item 39 is a conditional question based on response to item 38 regarding hair loss. Item 39 is only to be completed if patients had indicated hair loss for item 38. Therefore, the proportions of those who missed item 39 at each time point will also be reported by their response to item 38 (hair loss if response is 2-4, no hair loss if response is 1, or missing if no response provided). The proportion of those who completed item 39, but responded no hair loss (option 1) to item 38 will also be reported even though their response to item 39 will not be scored. Items 57 and 58 are also conditional questions based on response to item 56 regarding being sexually active. Items 57 and 58 are only to be completed if patients had indicated being sexually active for item 56. Therefore, the proportions of those who missed items 57 and 58 at each time point will also be reported by their response to item 56 (sexually active if response is 2-4, not sexually active if response is 1, or missing if no response provided). The proportion of those who completed items 57 and/or 58, but responded not sexually active (option 1) to item 56 will also be reported even though their response to items 57 and 58 will not be scored.

6.2.7.3. Descriptive summaries over time

The mean (and SD), and median (and range) of absolute scores and change from baseline of the QLQ-C30 and QLQ-OV28 multiple-item and single-item scale scores, and the EQ-VAS will be summarized for each treatment arm at each time point. A line chart depicting the means along with standard error (SE) error bars over time will be provided for each scale in each treatment arm. For the EQ-5D-5L health state profiles, the proportions of patients

having none, slight, some or moderate, severe, extreme/unable problems at each time point will be reported for each of the 5 dimensions.

6.2.7.4. Time-to-event endpoints

Time to deterioration (TTD) is defined as the time from randomization to the first time the patient's score shows a 15 point or higher increase in patient's EORTC QLQ-OV28 abdominal/GI symptom subscale.

TTD data will be censored at the last time when a patient completed a sub-scale assessment for patients who have not deteriorated.

Osoba et al, established that a 10-point or a greater clinical minimally important difference (MID) from baseline on the scales of the EORTC QLQ-C30 would correlate with significant (moderate) change in disease symptoms and functioning (Osoba, 1998)). However, Stockler et al, in an Ovarian-specific clinical trial, established a more conservative primary PRO hypothesis threshold of at least a 15% (≥15-point) absolute difference on the QLQ-OV28 abdominal/GI symptom subscale (Stockler et al, 2014). Hence, in the analysis of time to deterioration, a minimally important difference (MID) of 15 points or greater is proposed for the abdominal/ GI symptom subscale of the OV-28. Sensitivity TTD analysis of a ≥10 point increase will also be conducted to assess the robustness of the endpoints on the QLQ-OV28 subscale.

TTD of the abdominal/GI symptom sub-scales will be summarized as follows.

The treatment effect on TTD will be estimated using a Cox's Proportional Hazard model stratified by the randomization strata to calculate the hazard ratio and p-value for each of the experimental arms vs control arm. To explore the effect of the open-label nature of the study design on the PRO endpoints, an additional analysis will estimate the hazard ratio between the two experimental arms.

Each stratum will define a separate baseline hazard function (using the 'STRATA' statement in SAS PROC PHREG), ie for the i-th stratum the hazard function is expressed as: $h(i;t) = h(i,0;t) \exp(x\beta)$, where $h(i,0;t)$ defines the baseline hazard function for the i-th stratum and x defines the treatment arm (0=control arm, 1= experimental arm) and β is the unknown regression parameter.

Ties will be handled by replacing the proportional hazards model by the discrete logistic model (Ties=Discrete option in SAS PROC PHREG).

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median TTD time with 2-sided 95% CIs. The CIs for the median TTD will be calculated according to Brookmeyer and Crowley (1982).

6.2.7.5. Binary and categorical endpoints

The proportion of patients who improved from baseline, defined as a ≥10 point decrease on the abdominal/GI subscale (items 31-37) of the EORTC QLQ-OV28, will be summarized

and compared between each of the treatment arms at Cycle 3 (beginning of Week 9). Patients with missing questionnaires at Cycle 3 will be counted as not having improved.

The number and proportion of patients who improved, worsened, or remained stable for all of the symptom and functional domains, global QOL and single items of the EORTC QLQ C30, the QLQ OV28 and will be summarized and compared between each of the experimental arms and the control arm. The definition of improved, worsened, stable will use the 10-point MID for all domains and items and considered as an average per patient (as opposed to the 15-point used in the TTD analyses above).

6.2.7.6. Continuous endpoints

Longitudinal random intercept random slope mixed-effect model will be carried out for the total and subscales of the EORTC QLQ-C30, QLQ-OV28, EQ-5D-5L and EQ VAS and will compare each of the experimental arms with the control arm. Outcomes are PRO post-baseline scores and the predictors are the corresponding baseline PRO score, treatment, randomization stratification factor, time (treated as a continuous variable in days), and treatment-by-time interaction. Intercept and time are considered as random effects particular to each subject. All parameter estimates will be obtained using restricted maximum likelihood. The unstructured covariance structure will be used to define covariance between random effects (using option “Type=UN” as a part of the RANDOM statement in PROC MIXED). For the degrees-of-freedom calculations the Kenward and Roger algorithm will be used (using option “ddfm = kr” as a part of the MODEL statement in PROC MIXED).

6.2.7.7. Sensitivity Analysis

For the TTD and continuous endpoints, the main analysis will include data from baseline up to EOT assessment (not including EOT). A sensitivity analysis for these endpoints will be conducted including all assessments from baseline to end of the safety follow-up period.

6.3. Other Endpoints

CCI

[redacted]

CCI

[redacted]

CCI [redacted]

[redacted]

[redacted]

[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]

[redacted]

[redacted]s

CCI

[redacted]

6.3.2. CA-125

CA-125 values and changes from baseline will be summarized using descriptive statistics (mean, SD, median, Q1, Q3, minimum, and maximum).

6.4. Subset Analyses

Subset analyses will be performed for OS and for PFS, OR and DR per BICR assessment based on the FAS for the subgroups defined below.

The following subgroups will be defined and used for analyses:

- Randomization stratification factors
  - Platinum-refractory or platinum-resistant status,
  - Number of prior regimens (1 vs 2 or 3)
  - Bulky disease (defined as presence of a tumor ≥5 cm) vs not
- Age
  - Age < 65 years (Reference)
  - Age ≥ 65 years
- Race
  - Caucasian / White (Reference)
  - Asian
  - Black/African American
  - Other
- Ethnicity
  - Hispanic/Latino
  - Non-Hispanic/Latino (Reference)
- Pooled Geographical Region
  - North America

- Europe (Reference)
- Asia
- Rest of the World (Australasia, Latin America, Africa, and/or Middle East will be included as additional subgroups if including > 10% of the overall randomized population)

- BRCA
  - Positive (either BRCA1 or BRCA2 positive, ie, positive for known deleterious mutation(s) or genetic variant suspected to be deleterious)
  - Negative (Reference)
- Prior Bevacizumab
  - Yes
  - No (Reference)
- Bevacizumab-eligible as collected in the eCRF
  - Yes (Reference)
  - No
- CA-125 at baseline
  - ≤2xULN (Reference)
  - >2xULN
- PD-L1 status at baseline
  - Positive (Reference)
  - Negative

  Different cutoffs may be considered for defining PD-L1 status at baseline, including:
  - 1%, 5%, 50% and 75% on tumor cells;
  - 5%, 25%, 50% and 75% on immune cells;
  - 5%, 25%, 50% on tumor or immune cells

  Here, for each x=5, 25, 50, or 75, PD-L1 positive status is defined as tumor cell or immune cells, as appropriate, ≥x% and PD-L1 negative status is defined as tumor cell or immune cells, as appropriate, <x%.

Subset analyses for OS, PFS and DR will use the primary censoring rules described in Sections 6.1.1 and 6.1.2. All the subgroup analyses are exploratory. Treatment arms will be compared for OS and PFS using a 2-sided unstratified log-rank test for each subgroup level and the unstratified HR and its corresponding 95% CI will be computed per subgroup level.

All the subgroup analyses will be exploratory; no adjustment for multiplicity will be performed. In the case of a low number of patients within a category (<5% of the randomized population), the categories will be pooled.

To assess the heterogeneity of treatment effects for PFS and OS across the subgroup levels, two Cox regression model will be fitted with PFS or OS, respectively, as the dependent variable and subgroup, treatment, and with and without the treatment-by-subgroup interaction as explanatory variables.

- Model 1: factors + treatment + subgroup
- Model 2: factors + treatment + subgroup + treatment×subgroup-variable

A p-value for the interaction test (Likelihood Ratio test) will be provided together with the HR and corresponding 95% CI of the interaction model parameter.

The HR for OS and PFS and corresponding 95% CIs for all subgroups will also be presented in a forest plot.

The ORR odds ratio for each subgroup and corresponding 95% CIs will also be presented in a forest plot for each treatment arm.

6.5. Baseline and Other Summaries and Analyses

6.5.1. Baseline summaries

The following analyses will be based on the FAS overall and separately by treatment arm.

6.5.1.1. Demographic characteristics

Demographic characteristics and physical measurements will be summarized by treatment arm using the following information from the 'Screening/Baseline Visit' eCRF pages.

- Demographic characteristics
  - Gender: Male, Female
  - Race: White, Black or African American, Asian, American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, Other, Unknown
  - Ethnic origin: Hispanic/Latino (Yes/No)
  - Age (years): summary statistics
  - Age categories :
    - < 65 years, ≥ 65 years
    - < 65, 65-<75, 75-<85, ≥ 85 years
  - Pooled Geographical Region (as applicable):
    - North America

  - Europe
  - Asia
  - Rest of the World (Australasia, Latin America, Africa, and/or Middle East will be included as additional pooled geographical regions if including > 10% of the overall randomized population)
- Geographic Region (as applicable):
  - North America
  - Latin America
  - Western Europe
  - Eastern Europe
  - Middle East
  - Australasia
  - Asia
  - Africa
- Eastern Cooperative Oncology Group (ECOG) Performance Status: 0, 1, 2, 3, and 4

- Physical measurements
  - Height (cm)
  - Weight (kg)
  - Body Mass Index (BMI) (kg/m^2)
  - Body Surface Area (BSA) (m^2)

Center codes will be used for the determination of the patient's geographic region.

The listing of demographics and baseline characteristics will include the following information: patient identifier, treatment arm, age, sex, race, ethnicity, height (cm), weight (kg), BMI (kg/m^2), BSA (m^2) and ECOG performance status.

6.5.1.2. Medical history

Medical history will be coded using the most current available version of Medical Dictionary for Regulatory Activities (MedDRA) and will be summarized from the 'Medical History' eCRF page. Medical history will be summarized as the numbers and percentages of patients by MedDRA preferred term (PT) as event category and MedDRA primary system organ class (SOC) as summary category. Each patient will be counted only once within each PT or SOC.

Medical history will be displayed in terms of frequency tables: ordered by primary SOC and PT in alphabetical order.

6.5.1.3. Disease characteristics

Information on disease characteristics collected on 'Primary Diagnosis', 'Substance Abuse' and RECIST eCRF pages will be summarized overall and by treatment arm. Summary statistics will be presented for the following.

From the 'Primary Diagnosis' eCRF page:

- Site of primary tumor
- Primary diagnosis (summarize all categories collected in the 'Primary Diagnosis' eCRF page)
- Time since initial diagnosis to date of randomization (months), defined as (date of randomization – date of initial diagnosis)/30.4375
- Time since histopathological diagnosis (months), defined as (date of randomization – date of histopathological diagnosis)/30.4375
- Platinum-free interval (defined as the interval from the date of the last platinum dose to the date of documented disease progression): 0-3 months, >3-6 months, > 6 months.

From the RECIST eCRF page (summary will be presented separately based on investigator assessment and BICR assessment):

- Measurable disease (lesions) at baseline (Yes, No)
- Involved tumor sites at baseline

From the 'Substance Use' eCRF page:

- Smoking history
  - Never smoker vs current vs former smoker
  - Smoking exposure (pack-years): 0, <20, 20-<40, ≥40 and summary statistics
  - Years since quitting: never smoker, current smoker, <5, 5-<10, ≥10 and summary statistics

Specifications for computation:

- Cigarette equivalents are calculated as follows: one cigar is regarded equivalent to 5 cigarettes and 1 pipe is regarded equivalent to 3 cigarettes
- Duration of nicotine consumption [years]:

  (end of nicotine consumption – start of nicotine consumption + 1) / 365.25

- Pack-years:
  - calculate cigarette equivalents per day using the conversion factors given above
  - convert to packs per day where 20 cigarettes are regarded as 1 pack
  - pack-years = packs per day × duration of nicotine consumption [years]

Listing of disease history will be provided with all relevant data (as collected on the 'Primary Diagnosis' and 'Substance Use' eCRF pages) and derived variables as above.

6.5.1.4. Prior anti-cancer therapies

The prior anti-cancer therapies are collected under the 'Prior Cancer Therapy', 'Prior Radiation Therapy' and 'Prior Surgery' eCRF pages.

The number and percentage of patients in each of the following anti-cancer therapy categories will be tabulated:

- Patients with at least one type of prior anti-cancer therapy
- Patients with at least one prior anti-cancer drug therapy
- Patients with at least one prior anti-cancer radiotherapy
- Patients with at least one prior anti-cancer surgery

Prior anti-cancer drug therapy will be summarized as follows based on the number and percentage of patients with the following:

- At least one prior anti-cancer drug therapy
- Number of prior anti-cancer drug therapy regimens: missing, 1, 2, 3, ≥4
- Prior bevacizumab (Yes/No)
- Intent of Drug Therapy: Neo-Adjuvant, Adjuvant, Advanced – Metastatic, Local regional Disease-Recurrence
- Best response: CR, PR, SD, PD, Unknown, Not applicable. Best response is derived from the last treatment regimen.

The prior anti-cancer drug therapies will also be summarized based on the number and percentage of patients by the drug class and preferred term. A patient will be counted only once within a given drug class and within a given drug name, even if he/she received the same medication at different times. The summary will be sorted on decreasing frequency of drug class and decreasing frequency of drug name in a given drug class. In case of equal frequency regarding drug class (respectively drug name), alphabetical order will be used.

Prior anti-cancer therapies will be included in the listings that follow with a flag to identify prior therapies. These will include the patient identification number, and all the relevant collected data-fields on the corresponding eCRF pages.

- Listing of anti-cancer drug therapies
- Listing of anti-cancer radiotherapy
- Listing of anti-cancer surgeries

6.5.2. Study conduct and patient disposition

The following analyses will be performed based on the FAS overall and separately by treatment arm.

6.5.2.1. Patient disposition

The percentages below will be calculated based on the number of patients in the FAS.

- Total number of patients screened overall
- Number of patients who discontinued from the study prior to randomization overall and by the main reason for discontinuation
- Number and percentage of randomized patients in each of the analysis sets defined in Section 4
- Number and percentage of randomized patients with study drug ongoing (separately for each study drug in Arm B)
- Number and percentage of randomized patients who discontinued study drug overall and by the main reason for discontinuation of study drug (separately for each study drug in Arm B)
- Number and percentage of patients who entered follow-up
- Number and percentage of patients who discontinued follow-up overall and by the main reason for discontinuation
- Number and percentage of patients who entered long-term follow-up
- Number and percentage of patients who discontinued long-term follow-up overall and by the main reason for discontinuation

The results of the randomization algorithm (according to IRT) will be summarized as follows:

- Number and percentage of randomized patients overall, by region (Europe, EEA (required by EudraCT), North America, Latin America, Middle East, Asia, Australasia, Africa), by country within region
- Number and percentage of randomized patients by center
- Number and percentage of randomized patients by randomization strata (IRT)
- Number and percentage of randomized patients by randomization strata (eCRF)

- Cross tabulation: stratum by IRT vs. stratum by eCRF
- Cross tabulation: patients randomized (avelumab/avelumab+PLD/PLD/none) vs. patients treated (avelumab/avelumab+PLD/PLD/none)

In addition, a cross tabulation of patients who have discontinued/are ongoing treatment with avelumab vs patients who have discontinued/are ongoing treatment with PLD in Arm B will also be provided.

6.5.2.2. Protocol deviations

All protocol violations that impact the safety of the patients and/or the conduct of the study and/or its evaluation will be reported. These include:

- Patients who are dosed on the study despite not satisfying the inclusion criteria;
- Patients who develop withdrawal criteria whilst on the study but are not withdrawn;
- Patients who receive the wrong treatment or an incorrect dose;
- Patients who receive an excluded concomitant medication;
- Deviations from GCP.

The identification of these and other CSR-reportable deviations will be based on the inclusion/exclusion criteria or other criteria presented in the protocol.

6.5.3. Study treatment compliance and exposure

The following analyses will be based on the safety analysis set by treatment arm.

Cycle definitions for study drugs that are administered in combination apply to all the study drugs in the combination. Ie, cycle is patient-dependent, rather than study-drug-dependent when study drugs are administered in combination.

For Cycle X, actual cycle start date for each patient is

- the earliest start date of dosing in the Cycle X day 1 visit eCRF exposure page, if the patient received study treatment on that visit (ie, any study drug with dose>0 at that visit)
- the first day of assessments in the Cycle X day 1 visit, if the patient did not receive study treatment on that visit (ie, all study drugs had dose=0 at that visit). Use start date in the exposure page if available; if start date is not available then use date of collection of vital signs on Cycle X day 1 visit.

Actual cycle end date for each patient is,

- for all cycles X except the last cycle, actual cycle end date = actual cycle (X+1) start date – 1 day;

- for the last cycle, actual cycle end date = actual cycle start date + 28 days – 1 day

Cycle duration (weeks) = (actual cycle end date – actual cycle start date + 1)/7

When summarizing exposure for each study drug, only cycles from first dose of study treatment until the last cycle with non-zero dose of study drug (Arm A and C) or at least one of the study drugs (Arm B) should be included.

Exposure may be summarized (per cycle and/or overall) as dose received (cumulative dose, actual dose intensity) and as dose received relative to intended dose (relative dose intensity [RDI]).

The information that will be summarized depends on how the study drug is dosed (eg, infusion cyclical, oral daily, oral cyclical).

The formulae below should be applied to each study drug separately even when study drugs are administered in combination.

The derivations below are provided for the following:

- Avelumab administered as a 1-hour IV infusion at a dose of 10 mg/kg once every 2 weeks in 4-week cycles.
- PLD administered as 1-hour IV infusion at a dose of 40 mg/m^2 once every 4 weeks in 4-week cycles.

Analysis of exposure will be based on the calculated actual dose levels

- Avelumab - total dose / weight
- PLD - total dose / m^2

6.5.3.1. Exposure to avelumab

The dose level for avelumab is calculated as actual dose administered/weight (mg/kg). The last available weight of the patient on or prior to the day of dosing will be used.

Intended duration of treatment with avelumab (weeks) =

(end date–date of first dose of study drug +1)/7,

where end date = start date of last cycle with non-zero dose of study drug + 28 – 1

Duration of exposure to avelumab (weeks) =

(last dose date of avelumab – first dose date of avelumab + 14)/7

The cumulative dose (mg/kg) of avelumab per patient in a cycle or overall is the sum of the actual doses of avelumab received in a cycle or overall, respectively

Each cycle for avelumab is defined by a 4-week period. The dose intensity (DI) and the relative dose intensity (RDI) will be calculated for each patient across all cycles. The actual dose intensity per cycle (mg/kg/4-week cycle) is defined as

- By cycle actual DI (mg/kg/4-week cycle) = Cumulative dose in the cycle (mg/kg)/[cycle duration (weeks)/4]
- Overall actual DI (mg/kg/4-week cycle) = [overall cumulative dose (mg/kg)] / [intended duration of treatment with avelumab (weeks)/4].

The relative dose intensity (RDI) is defined as the actual dose intensity divided by the intended dose intensity and expressed in %.

- Intended DI (mg/kg/4-week cycle) = [intended cumulative dose per cycle] / [intended number of 4-weeks in a cycle] = [20 (mg/kg)] / [1 (4-week cycle)] = 20 (mg/kg/4-week cycle)
- By cycle RDI (%) = 100 × [by cycle actual DI] / [intended DI]
  = 100 × [by cycle actual DI] / [20 (mg/kg/4-week cycle)]
- Overall RDI (%) = 100 × [overall actual DI] / [intended DI]
  = 100 × [overall actual DI] / [20 (mg/kg/4-week cycle)]

The summary of treatment exposure and compliance for avelumab will include the following information:

- Treatment duration (weeks)
- Total number of infusions received
- Cumulative dose (mg/kg)
- Dose intensity (mg/kg/cycle)
- Relative dose intensity (%).

6.5.3.2. Exposure to PLD

The dose level for PLD is calculated as actual dose administered/m^2 (mg/m^2). The last available weight and height of the patient on or prior to the day of dosing will be used to calculate the BSA. The duration of PLD treatment (in weeks) during the study for a patient is defined as:

Intended duration of treatment with PLD (weeks) =

(end date–date of first dose of PLD +1)/7,

where end date = start date of last cycle with non-zero dose of PLD + 28 – 1.

Duration of exposure to PLD (weeks) =

(last dose date of PLD – first dose date of PLD + 28)/7

The cumulative dose (mg/m^2) of PLD per patient in a cycle or overall is the sum of the actual doses of PLD received in a cycle or overall, respectively. Each cycle for PLD is defined by a 4-week period. The dose intensity (DI) and the relative dose intensity (RDI) of PLD will be calculated for each patient across all cycles. The actual dose intensity per cycle (mg/m^2/4-week cycle) is defined as

- By cycle actual DI (mg/m^2/4-week cycle) = Cumulative dose in the cycle (mg/m^2)/[cycle duration (weeks)/4]
- Overall actual DI (mg/m^2/4-week cycle) = overall cumulative dose (mg/m^2) / [intended duration of treatment with PLD (weeks)/4].

The relative dose intensity (RDI) is defined as the actual dose intensity divided by the intended dose intensity and expressed in %.

- Intended DI (mg/ m^2/4-week cycle) = [intended cumulative dose per cycle] / [intended number of 4-weeks in a cycle] = [40 (mg/m^2)] / [1 (4-week cycle)] = 40 (mg/m^2/4-week cycle)
- By cycle RDI (%) = 100 × [by cycle actual DI] / [intended DI]
  = 100 × [by cycle actual DI] / [40 (mg/m^2/4-week cycle)]
- Overall RDI (%) = 100 × [overall actual DI] / [intended DI]
  = 100 × [overall actual DI] / [40 (mg/m^2/4-week cycle)]

The summary of treatment exposure and compliance for PLD will include the following information:

- Treatment duration (weeks)
- Total number of infusions received
- Cumulative dose (mg/m^2)
- Dose intensity (mg/m^2/cycle)
- Relative dose intensity (%).

6.5.3.3. Dose reductions

Applicable to avelumab and PLD.

Dose reduction is defined as actual non-zero dose < 90% of the planned dose.

The number and percentage of patients with at least one dose reduction as well as a breakdown of the number of dose reductions (1, 2, 3, ≥4) will be summarized.

6.5.3.4. Dose delays

Applicable to avelumab and PLD.

Dose Delay is the difference between the actual time between two consecutive non-zero doses and the planned time between the same two consecutive non-zero doses.

Dose Delay for Dose x (days) = Date of Dose x – Date of Dose (x-1) – 14, for avelumab.

Dose Delay for Dose x (days) = Date of Dose x – Date of Dose (x-1) – 28, for PLD.

Delays will be grouped into the following categories based on the deviation of the actual to the planned treatment administration day (relative to the previous treatment administration date):

- No delay
- 1-2 days delay
- 3-6 days delay
- 7 or more days delay

For example, for avelumab, administered on a 2-week schedule, if one patient receives avelumab on Day 1, then the next avelumab administration date will be on Day 15; however, if the patient receives avelumab at Day 16 or 17, this is considered as 1-2 days delay.

No delay and 1-2 days delay will also be summarized together.

Number and percentage of patients with delayed study drug administration and maximum length of delay, ie, the worst case of delay if patients have multiple dose delays will be summarized.

6.5.3.5. Infusion rate reductions

Applicable to avelumab and PLD.

The number and percentage of patients with at least one infusion rate reduction of 50% or more compared to the first infusion rate reported in the eCRF as well as the frequency of patients with 1, 2, 3, or ≥4 infusion rate reductions of 50% or more will be summarized by treatment arm.

6.5.3.6. Infusion interruptions

An infusion interruption is defined as an infusion that is stopped and re-started on the same day (ie, for a visit more than one infusion start time and infusion end time are recorded).

The number and percentage of patients with at least one infusion interruption as well as the frequency of patients with 1, 2, 3, or ≥4 infusion interruptions will be summarized.

6.5.4. Concomitant medications and non-drug treatments

The following analyses will be based on the safety analysis set by treatment arm.

Concomitant medications are medications, other than study medications, which started prior to first dose date of study treatment and continued on on-treatment period as well as those started during the on-treatment period. **Prior medications** are medications, other than study medications and pre-medications for study drug, which are started before the first dose of study treatment.

Prior and concomitant medications will be summarized from the 'General Concomitant Medications' eCRF page. Pre-medications for study drug will also be summarized separately from the 'Pre-Medication Treatment' eCRF page.

Summary of prior medications, summary of concomitant medications and summary of pre-medications will include the number and percentage of patients by Anatomical Therapeutic Chemical (ATC) Classification level 2 and preferred term. A patient will be counted only once within a given drug class and within a given drug name, even if he/she received the same medication at different times. If any prior or concomitant medication is classified into multiple ATC classes, the medication will be summarized separately under each of these ATC classes. The summary tables will be sorted on decreasing frequency of drug class and decreasing frequency of drug name in a given drug class. In case of equal frequency regarding drug class (respectively drug name), alphabetical order will be used. In case any specific medication does not have ATC classification level 2 coded term, it will be summarized under 'Unavailable ATC classification' category.

A listing of prior medications and a listing of concomitant medications will be created with the relevant information collected on the 'General Concomitant Medications' eCRF page. A listing of pre-medications will be created with the relevant information collected on the 'Pre-Medication Treatment' eCRF page.

All concurrent procedures, which were undertaken any time during the on-treatment period, will be listed according to the eCRF page 'General Non-drug Treatments'.

A listing of concurrent procedures will be created with the relevant information collected on the 'General Non-drug Treatments' eCRF page.

6.5.5. Subsequent anti-cancer therapies

The following analyses will be based on the FAS by treatment arm.

Anti-cancer treatment will be provided in a data listing with data retrieved from 'Follow-up Cancer Therapy', 'Concomitant Radiation Therapy', 'Follow-up Radiation Therapy', and 'Follow-up Surgery' eCRF pages. Number and percentage of patients with any anti-cancer therapy after discontinuation will be tabulated overall and by type of therapy based on the data collected from the 'Follow-up Cancer Therapy', 'Follow-up Radiation Therapy' and 'Follow-up Surgery' eCRF pages.

6.6. Safety Summaries and Analyses

The Safety Analysis Set will be the primary population for safety evaluations. Summaries of AEs and other safety parameters will be based on the safety analysis set by treatment arm.

6.6.1. Adverse events

Treatment-emergent adverse events (TEAEs) are those events with onset dates occurring during the on-treatment period for the first time, or if the worsening of an event is during the on-treatment period as defined in Section 3.5.1.

All analyses described will be based on TEAEs (started during the on-treatment period) if not otherwise specified. The AE listings will include all AEs (whether treatment-emergent or not). AEs outside the on-treatment period will be flagged in the listings.

- **Related Adverse Events:** adverse events with relationship to study treatment (as recorded on the AE eCRF page, Relationship with study treatment = Related) reported by the investigator and those of unknown relationship (ie, no answer to the question 'Relationship with study treatment'). Related AEs are those related to any study drug (ie, at least one of the study drugs).

- **Serious Adverse Events (SAE):** serious adverse events (as recorded on the AE eCRF page, Serious Adverse Event = Yes).

- **Adverse Events Leading to Dose Reduction:** adverse events leading to dose reduction of study treatment (as recorded on the AE eCRF page, Action taken with study treatment = Dose reduced).

- **Adverse Events Leading to Interruption of Study Treatment**: adverse events leading to interruption of study treatment (as recorded on the AE eCRF page, Action taken with study treatment = Drug interrupted). The eCRF does not allow for a clear separation between interruption of an infusion and delays of administration for a parenteral drug as both are recorded using the same term on the eCRF ("Drug interrupted"). IRRs will be excluded in the analysis of AEs leading to Drug Interruption in case they only led to an interruption of the infusion.

- **Adverse Events Leading to Permanent Treatment Discontinuation:** adverse events leading to permanent discontinuation of study treatment (as recorded on the AE eCRF page, Action taken with study treatment = Drug withdrawn).

- **Adverse Events Leading to Death:** adverse event leading to death (as recorded on the AE eCRF page, Outcome = Fatal, as well as AEs of Grade 5).

- **Immune-related Adverse Events (irAE):** irAEs (as identified according to the methodology outlined in Appendix 1 for a pre-specified search list of MedDRA PTs, documented in the Safety Review Plan and finalized for analysis of the current studies data prior to DB lock)

- **Infusion-related Reactions (IRR)**: IRRs (as identified according to the methodology outlined in Appendix 2 for a pre-specified search list of MedDRA PTs documented in the

Safety Review Plan and finalized for analysis of the current studies data prior to DB lock).

Unless otherwise specified, AEs will be summarized by number and percentage of patients with the AE in the category of interest as described above, by treatment arm, primary SOC and PT in decreasing frequency based on the frequencies observed for the avelumab treatment arm (Arm A).

Each patient will be counted only once within each SOC or PT. If a patient experiences more than one AE within a SOC or PT for the same summary period, only the AE with the strongest relationship or the worst severity, as appropriate, will be included in the summaries of relationship and severity.

6.6.1.1. All adverse events

Adverse events will be summarized by worst severity (according to NCI-CTCAE version 4.03) per patient, using the latest version of MedDRA preferred term (PT) as event category and MedDRA primary system organ class (SOC) body term as Body System category.

In case a patient has events with missing and non-missing grades, the maximum of the non-missing grades will be displayed. No imputation of missing grades will be performed.

The following tables will be created:

- The overall summary of AEs table will include the frequency (number and percentage) of patients with each of the following by treatment arm:
  - TEAEs
  - TEAEs, Grade ≥ 3
  - Related TEAEs
  - Related TEAEs, Grade ≥ 3
  - TEAEs leading to dose reduction of avelumab
  - TEAEs leading to dose reduction of PLD
  - TEAEs leading to interruption of avelumab
  - TEAS leading to interruption of PLD
  - TEAEs leading to discontinuation of avelumab
  - TEAEs leading to discontinuation of PLD
  - TEAEs leading to discontinuation of any study drug
  - TEAEs leading to discontinuation of all study drugs
  - Related TEAEs leading to discontinuation of avelumab

- Related TEAEs leading to discontinuation of PLD
- Related TEAEs leading to discontinuation of any study drug
- Related TEAEs leading to discontinuation of all study drugs
- Serious TEAEs
- Related Serious TEAEs
- TEAEs leading to death
- Related TEAEs leading to death
- irAEs
- IRRs

- TEAEs by SOC and PT and worst grade
- Related TEAEs by SOC and PT and worst grade
- TEAEs leading to death by SOC and PT
- Related TEAEs leading to death by SOC and PT
- TEAEs Excluding SAEs, with frequency ≥ 5% in any treatment arm by SOC and PT

6.6.1.2. Adverse events leading to dose reduction

The frequency (number and percentage) of patients with each of the following will be presented for TEAEs leading to dose reduction of each study drug by treatment arm:

- TEAEs leading to dose reduction of avelumab by SOC and PT
- TEAEs leading to dose reduction of PLD by SOC and PT

The listing of all AEs leading to dose reduction will also be provided with the relevant information.

6.6.1.3. Adverse events leading to interruption of study treatment

The eCRF does not allow for a clear separation between interruption of an infusion and delays of administration for a parenteral drug as both are recorded using the same term on the eCRF ("Drug interrupted"). IRRs will be excluded in the analysis of AEs leading to Drug Interruption in case they only led to an interruption of the infusion (ie, did not lead to a dose reduction or a dose delay).

As such, AEs leading to interruption will be defined as AEs identified in the AE eCRF page with an action taken with study treatment of 'drug interrupted' excluding

- IRRs that occurred on the day of infusion with ≥90% of the planned dose given (ie IRRs that did not lead to a dose reduction) and subsequent administration of study drug had no

delay (as defined in Section 6.5.3.4). These IRRs will be considered as IRRs leading to interruption of infusion.

- IRRs occurring on the day after infusion and subsequent dose administration had no delay (as defined in Section 6.5.3.4).

The frequency (number and percentage) of patients with each of the following will be presented for TEAEs leading to interruption of each study drug by treatment arm:

- TEAEs leading to interruption of avelumab by SOC and PT
- TEAEs leading to interruption of PLD by SOC and PT

The listing of all AEs leading to interruption of study treatment will also be provided with the relevant information.

In addition, the frequency (number and percentage) of patients with each of the following will be presented for TEAEs leading to interruption of each study drug by treatment arm:

- TEAEs leading to both interruption and dose reduction of avelumab by SOC and PT
- TEAEs leading to both interruption and dose reduction of PLD by SOC and PT

This summary will take into account PTs with both actions as defined in Section 6.6.1, eventhough the actions may be captured for different PT records (ie, different onset for the PT with action “drug interrupted” and the PT with action “dose reduced”.

6.6.1.4. Adverse events leading to discontinuation of study treatment

The frequency (number and percentage) of patients with each of the following will be presented for TEAEs leading to permanent discontinuation of each study drug and study treatment, by treatment arm:

- TEAEs leading to discontinuation of avelumab by SOC and PT
- Related TEAEs leading to discontinuation of avelumab by SOC and PT
- TEAEs leading to discontinuation of PLD by SOC and PT
- Related TEAEs leading to discontinuation of PLD by SOC and PT
- TEAEs leading to discontinuation of any study drug by SOC and PT
- Related TEAEs leading to discontinuation of any study drug by SOC and PT
- TEAEs leading to discontinuation of all study drugs by SOC and PT
- Related TEAEs leading to discontinuation of all study drugs by SOC and PT

The listing of all AEs leading to treatment discontinuation will also be provided with the relevant information.

6.6.2. Deaths

The frequency (number and percentage) of patients in the safety analysis set who died and who died within 30 days after last dose of study treatment as well as the reason for death, will be tabulated based on information from the 'Notice of Death' and 'Survival Follow-Up' eCRFs, by treatment arm.

- All deaths
- Deaths within 30 days after last dose of study treatment
- Reason for Death
  - Disease progression
  - Study treatment toxicity
  - AE not related to study treatment
  - Unknown
  - Other.

In addition, date and cause of death will be provided in individual patient data listing together with selected dosing information (study treatment received, date of first / last administration, dose) and will include the following information:

- AEs with fatal outcome (list preferred terms of AEs with outcome=Fatal, as well as AEs of Grade 5),
- Flag for death within 30 days of last dose of study treatment.

6.6.3. Serious adverse events

The frequency (number and percentage) of patients with each of the following will be presented for treatment-emergent SAEs by treatment arm:

- SAEs by SOC and PT
- Related SAEs by SOC and PT.

The listings of all SAEs will also be provided with the relevant information with a flag for SAEs with onset outside of the on-treatment period.

6.6.4. Other significant adverse events

The frequency (number and percentage) of patients with each of the following will be presented for irAEs, by treatment arm:

- irAEs leading to death, by Cluster and PT
- irAEs, by Cluster and PT

- irAEs, Grade ≥ 3, by Cluster and PT
- irAEs leading to discontinuation of avelumab, by Cluster and PT
- irAEs leading to discontinuation of PLD, by Cluster and PT
- irAEs leading to discontinuation of any study drug, by Cluster and PT
- irAEs leading to discontinuation of all study drugs, by Cluster and PT
- Serious irAEs, by Cluster and PT

The listing of all irAEs will also be provided with the relevant information with a flag for irAEs with onset outside of the on-treatment period.

The frequency (number and percentage) of patients with each of the following will be presented for IRRs, by treatment arm:

- IRRs leading to death, by PT
- IRRs, by PT
- IRRs, Grade ≥ 3, by PT
- IRRs leading to discontinuation of avelumab, by PT
- IRRs leading to discontinuation of PLD, by PT
- IRRs leading to discontinuation of any study drug, by PT
- IRRs leading to discontinuation of all study drugs, by PT
- Serious IRRs, by PT
- Time related to first onset of an IRR (infusion 1, infusion 2, infusion 3, infusion 4 or later). For IV study drugs administered in combination the infusion numbers are those associated with the regimen, rather than the individual study drugs.

The listing of all IRRs will also be provided with the relevant information with a flag for IRRs with onset outside of the on-treatment period.

In addition, the following analyses will be presented for Tier-1 (irAEs and IRRs as defined in Appendices 1 and 2) and Tier-2 events separately. P-values and CIs for risk difference will be calculated based on the unconditional exact method by Santner and Snell (1980). No additional analyses will be presented for Tier-3 AEs.

- Frequency (number and percentage) of patients with each of the following by treatment arm and PT or Clustered Term:
  - Tier-2 AEs
  - Tier-2 AEs Grade ≥ 3

- Point estimate for risk difference and 95% CI for risk difference (experimental arm vs comparator arm) for each of the following by PT or Clustered Term:
  - irAEs
  - irAEs Grade ≥ 3
  - IRRs
  - IRRs Grade ≥ 3
  - Tier-2 AEs
  - Tier-2 AEs Grade ≥ 3
- 2-sided p-value associated with risk difference (experimental arm vs comparator arm) for each of the following by PT or Clustered Term:
  - irAEs
  - irAEs Grade ≥ 3
  - IRRs
  - IRRs Grade ≥ 3
- The p-values and CIs reported are not adjusted for multiplicity and should be used for screening purposes only. The 95% CIs are provided to help gauge the precision of the estimates for the risk difference and should be used for estimation purposes only.

6.6.5. Laboratory data

6.6.5.1. Hematology and chemistry parameters

Laboratory results will be classified according to the NCI-CTCAE criteria version 4.03. Non-numerical qualifiers (with the exception of fasting flags) will not be taken into consideration in the derivation of CTCAE criteria (eg, hypokalemia Grade 1 and Grade 2 are only distinguished by a non-numerical qualifier and therefore Grade 2 will not be derived). Additional laboratory results that are not part of NCI-CTCAE will be presented according to the categories: below normal limit, within normal limits and above normal limit (according to the laboratory normal ranges).

Quantitative data will be summarized using simple descriptive statistics (mean, SD, median, Q1, Q3, minimum, and maximum) of actual values and changes from baseline for each nominal visit over time (unscheduled measurements would therefore not be included in these summaries as described in Section 5.2.9). End of Treatment visit laboratory results will be summarized separately. The changes computed will be the differences from baseline. Qualitative data based on reference ranges will be described according to the categories (ie, Low, Normal, High).

Abnormalities classified according to NCI-CTCAE toxicity grading version 4.03 will be described using the worst grade. For those parameters which are graded with two toxicities

such as potassium (hypokalemia/hyperkalemia), the toxicities will be summarized separately. Low direction toxicity (eg, hypokalemia) grades at baseline and post baseline will be set to 0 when the variables are derived for summarizing high direction toxicity (eg, hyperkalemia), and vice versa.

For **WBC differential counts** (total neutrophil [including bands], lymphocyte, monocyte, eosinophil, and basophil counts), the absolute value will be used when reported. When only percentages are available (this is mainly important for neutrophils and lymphocytes, because the CTCAE grading is based on the absolute counts), the absolute value is derived as follows:

Derived differential absolute count = (WBC count) × (Differential %value / 100)

If the range for the differential absolute count is not available (only range for value in % is available) then Grade 1 will be attributed to as follows:

- Lymphocyte count decreased:
  - derived absolute count does not meet Grade 2-4 criteria, and
  - % value < % LLN value, and
  - derived absolute count $\geq 800/mm^3$
- Neutrophil count decreased
  - derived absolute count does not meet Grade 2-4 criteria, and
  - % value < % LLN value, and
  - derived absolute count $\geq 1500/mm^3$

For **calcium**, CTCAE grading is based on Corrected Calcium and Ionized Calcium (CALCIO). Corrected Calcium is calculated from Albumin and Calcium as follows

Corrected calcium (mmol/L) = measured total Calcium (mmol/L) + 0.02 (40 - serum albumin [g/L])

Liver function tests: Alanine aminotransferase (ALT), aspartate aminotransferase (AST), and total bilirubin (TBILI) are used to assess possible drug induced liver toxicity. The ratios of test result over upper limit of normal (ULN) will be calculated and classified for these three parameters during the on-treatment period.

Summary of liver function tests will include the following categories. The number and percentage of patients with each of the following during the on-treatment period will be summarized by treatment arm:

- ALT ≥ 3×ULN, ALT ≥ 5×ULN, ALT ≥ 10×ULN, ALT ≥ 20×ULN
- AST ≥ 3×ULN, AST ≥ 5×ULN, AST ≥ 10×ULN, AST ≥ 20×ULN

- (ALT or AST) ≥ 3×ULN, (ALT or AST) ≥ 5×ULN, (ALT or AST) ≥ 10×ULN, (ALT or AST) ≥ 20×ULN
- TBILI ≥ 2×ULN
- Concurrent ALT ≥ 3×ULN and TBILI ≥ 2×ULN
- Concurrent AST ≥ 3×ULN and TBILI ≥ 2×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 2×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 2×ULN and ALP > 2×ULN
- Concurrent (ALT or AST) ≥ 3×ULN and TBILI ≥ 2×ULN and (ALP ≤ 2×ULN or missing)

Concurrent measurements are those occurring on the same date.

Categories will be cumulative, ie, a patient with an elevation of AST ≥10×ULN will also appear in the categories ≥5×ULN and ≥3×ULN. Liver function elevation and possible Hy's Law cases will be summarized using frequency counts and percentages.

An evaluation of Drug-Induced Serious Hepatotoxicity (eDISH) plot will also be created, with different symbols for different treatment arms, by graphically displaying

- peak serum ALT(/ULN) vs peak total bilirubin (/ULN) including reference lines at ALT=3×ULN and total bilirubin=2×ULN.
- peak serum AST(/ULN) vs peak total bilirubin (/ULN) including reference lines at AST=3×ULN and total bilirubin=2×ULN.

In addition, a listing of all TBILI, ALT, AST and ALP values for patients with a post-baseline TBILI ≥ 2×ULN, ALT≥ 3×ULN or AST ≥ 3×ULN will be provided.

Parameters with NCI-CTC grades available:

The laboratory toxicities will be tabulated using descriptive statistics (number of patients and percentages) during the on-treatment period. The denominator to calculate percentages for each laboratory parameter is the number of patients evaluable for CTCAE grading (ie those patients for whom a Grade 0, 1, 2, 3 or 4 can be derived).

- The summary of laboratory parameters by CTCAE grade table will include number and percentage of patients with Grade 1, 2, 3, 4, Grade 3/4 and any grade (Grades 1-4), laboratory abnormalities during the on-treatment period.
- The shift table will summarize baseline CTCAE grade versus the worst on-treatment CTCAE grade. The highest CTCAE grade during the on-treatment period is considered as the worst grade for the summary.

The above analyses apply to hematology and chemistry evaluations which can be graded per CTCAE, ie:

- Hematology:

  Hemoglobin (HB), Leukocytes (white blood cell decreased), Lymphocytes (lymphocyte count increased/decreased), Neutrophils / Absolute Neutrophils Count (ANC) (neutrophil count decreased), Platelet Count (PLT) (platelet count decreased).

- Serum Chemistry:

  Albumin (hypoalbuminemia), Alkaline Phosphatase (alkaline phosphatase increased), Alanine Aminotransferase (ALT) (ALT increased), Amylase (serum amylase increased), Aspartate Aminotransferase (AST) (AST increased), Total Bilirubin (blood bilirubin increased, Cholesterol (cholesterol high), Creatinine (creatinine increased), Creatine Kinase (CPK increased), Potassium (hypokalemia/ hyperkalemia), Sodium (hyponatremia/ hypernatremia), Magnesium (hypomagnesemia/hypermagnesemia), Calcium (hypocalcemia/ hypercalcemia), Glucose (hypoglycemia/hyperglycemia), Gamma Glutamyl Transferase (GGT) (GGT increased), Lipase (lipase increased), Phosphates (hypophosphatemia), Triglycerides (hypertriglyceridemia).

Parameters with NCI-CTC grades not available:

Hematology and chemistry evaluations which cannot be graded per CTCAE criteria will be summarized as frequency (number and percentage) of patients with:

- shifts from baseline normal to at least one result above normal during on-treatment period
- shifts from baseline normal to at least one result below normal during on-treatment period

In this study, these apply to the following parameters:

- Hematology: Absolute Monocytes, Absolute Eosinophils, Absolute Basophils
- Serum Chemistry: Chloride, Total Urea, Uric Acid, Total Protein, C-Reactive Protein, Lactate Dehydrogenase (LDH)

6.6.5.2. Other laboratory parameters

All other parameters collected on the eCRF will be listed in dedicated listings presenting all corresponding collected information on the eCRF.

- Coagulation: activated partial thromboplastin time (aPTT) and prothrombin time (INR).
- Urinalysis: all urinalysis parameters
- Other parameters: hormone, and immunology parameters
- Pregnancy test

The listings of laboratory results will be provided for all laboratory parameters. The listings will be sorted by parameters and assessment dates or visits for each patient. Laboratory values that are outside the normal range will also be flagged in the data listings, along with corresponding normal ranges. A listing of CTCAE grading will also be generated for those laboratory tests.

In addition, listings of abnormal values will be provided for hematology, chemistry, urinalysis, coagulation parameters. If there is at least one abnormal assessment for any parameter, all the data for that laboratory parameter will be included into the listing.

For all tests not mentioned above but present in the clinical data, a listing of patients with at least one result for the relevant test will be provided.

6.6.6. Vital signs

Weight for the purposes of dose calculation will be recorded at screening and within 3 days pre-dose Day 1 of each cycle. Weight will not be collected at End of Treatment. Height will be measured at screening only.

Vital sign summaries will include all vital sign assessments from the on-treatment period. All vital sign assessments will be listed, and those collected outside the on-treatment period will be flagged in the listing.

All vital sign parameters will be summarized using descriptive statistics (mean, SD, median, Q1, Q3, minimum, and maximum) of actual values and changes from baseline for each visit over time. End of Treatment visit will be summarized separately. The changes computed will be the differences from baseline.

6.6.7. Electrocardiogram

ECG summaries will include all ECG assessments from the on-treatment period. All ECG assessments will be listed, and those collected outside the on-treatment period will be flagged in the listing. QTcB and QTcF will be derived based on RR and QT (see below). The average of the replicate measurements should be determined after the derivation of the individual parameter at each time point.

Selecting Primary QT Correction for Heart Rate

The analysis of QT data is complicated by the fact that the QT interval is highly correlated with heart rate. Because of this correlation, formulas are routinely used to obtain a corrected value, denoted QTc, which is independent of heart rate. This QTc interval is intended to represent the QT interval at a standardized heart rate. Several correction formulas have been proposed in the literature. For this analysis we will use some of those methods of correction, as described below. The QT interval corrected for heart rate by the Bazett's formula, QTcB, is defined as

$$QTcB = \frac{QT}{\sqrt{RR}},$$

the QT interval corrected for heart rate by the Fridericia's formula, QTcF, is defined as

$$QTcF = \frac{QT}{\sqrt[3]{RR}},$$

where RR represents the RR interval of the ECG, in seconds, and can be estimated as 60/Heart Rate.

Although Bazett's correction is the historical standard, it does not perform well when heart rate fluctuates. Fridericia's formula may perform better under these conditions. If QTcB and QTcF methods do not adequately correct for HR and there are a sufficient number of patients (eg >30) with baseline ECGs, an alternate correction to achieve the goal of getting uncorrelated QTc and RR is based on a linear regression methods which yields, theoretically, uncorrelated QTc and RR.

Linear regression method:

- Fit a model QT = a + b × RR to baseline data
- Use the estimated slope, $\hat{b}$, to correct QT
- Corrected QT for heart rate will be computed as follows:

$$QTcP = QT + \hat{b} \times (1 - RR)$$

Data will be summarized using QTcF and QTcB. However, if these are not appropriate for the data set due to an observed large correlation between corrected QT and HR using the baseline assessments, the results will also be summarized using QTcP.

ECG Summaries

The following analyses will be performed for each applicable ECG parameters (RR, PR, QRS, QT, ventricular rate -denoted as HR in what follows-, and QTc) by treatment arm, during the on-treatment period. The denominator to calculate percentages for each category is the number of patients evaluable for the category.

- Pearson correlation between QT and HR, QTc (QTcB, QTcF and, if applicable, QTcP) and HR using individual (non-averaged) baseline assessments
- For each of the ECG parameters (HR, and QT, QTc, QRS, PR intervals), descriptive statistics at baseline, at each post-baseline time point and changes from baseline at each post-baseline time point
- Frequency (number and percentage) of patients with notable ECG values according to the following categories:
  - QT/QTc increase from baseline >30 ms, >60 ms
  - QT/QTc > 450 ms, > 480 ms, > 500 ms

- HR ≤ 50 bpm and decrease from baseline ≥ 20 bpm
- HR ≥ 120 bpm and increase from baseline ≥ 20 bpm
- PR ≥ 220 ms and increase from baseline ≥ 20 ms
- QRS ≥ 120 ms

Patients with notable ECG interval values and qualitative ECG abnormalities will be listed for each patient and time point and the corresponding notable values and abnormality findings will be included in the listings.

Unscheduled ECG measurements will not be used in computing the descriptive statistics for change from baseline at each post-baseline time point. However, they will be used in the analysis of notable ECG changes and the shift table analysis of notable QT parameters.

6.6.8. MUGA/ECHO

LVEF% will be summarized using simple descriptive statistics (mean, SD, median, Q1, Q3, minimum, and maximum) of actual values and changes from baseline for each nominal visit over time. In addition, LVEF% will be summarized as frequency (number and percentage) of patients with:

- a shift from baseline normal to at least one result below the institutional lower limit of normal (LLN) during the on-treatment period
- ≥10-point decrease from baseline in LVEF% during the on-treatment period
- ≥10-point decrease from baseline in LVEF% to a post-baseline value < LLN during the on-treatment period
- ≥15-point decrease from baseline in LVEF%
- ≥15-point decrease from baseline in LVEF% to a post-baseline value < LLN during the on-treatment period.

Clinically significant findings will be listed.

6.6.9. Physical examination

Number and percentage of patients with abnormal findings in physical examination will be summarized by body system.

6.6.10. ECOG performance status

The ECOG shift from baseline to highest score during the on-treatment period will be summarized by treatment arm. ECOG performance status with shift from ECOG=0 or 1 to ECOG 2 or higher will also be presented in a data listing.

7. INTERIM ANALYSES

7.1. Introduction

The goals of the interim analyses are to allow early stopping of treatment arm(s) for futility or declare an experimental treatment (or treatments) to be statistically significantly superior to the control treatment arm. The interim analysis of OS and PFS will be performed as described in Sections 5.1.1 and 5.1.2 using the methodology described in Section 6.1.1 for OS and Section 6.1.2 for PFS.

The interim analyses for PFS and OS will occur at the same time and will be performed by an independent statistician.

Unblinded results from the interim analysis for OS and PFS will not be communicated to the Sponsor's clinical team or to any party involved in the study conduct (apart from the independent statistician and E-DMC members) until the E-DMC has determined that either (i) OS or PFS analysis has crossed the pre-specified boundary for efficacy or (ii) the study needs to be terminated due to any cause, including futility or safety reasons. Further details will be described in the E-DMC charter.

At the time of the final PFS and OS analyses, both PFS and OS analysis will be performed by the Sponsor's clinical team.

7.2. Interim Analyses and Summaries

At each analysis time point, the critical boundaries for the group sequential test will be derived from the predefined spending function(s) as described in Section 5.1. The calculations of boundaries will be performed using EAST.

7.2.1. Interim analysis for OS and PFS

Let $u_{OS}(t_1)$ and $u_{OS}(t_F)$ denote the upper critical boundaries based on the test statistics Z_{1_OS} and Z_{F_OS} for efficacy at the interim and the final analysis of OS, respectively, and let $l_{OS}(t_1)$ and $l_{OS}(t_F)$ denote the lower critical boundary for futility at the interim and final analysis of OS, respectively. For the final analysis of OS, $l_{OS}(t_F)=u_{OS}(t_F)$.

Let $u_{PFS}(t_1)$ and $u_{PFS}(t_F)$ denote the upper critical boundaries based on the test statistics Z_{1_PFS} and Z_{F_PFS} for efficacy at the interim and the final analysis of PFS, respectively, and let $l_{PFS}(t_1)$ and $l_{PFS}(t_F)$ denote the lower critical boundary for futility at the interim and final analysis of PFS, respectively. For the final analysis of PFS, $l_{PFS}(t_F)=u_{PFS}(t_F)$.

The critical values $u_{OS}(t_1)$ and $l_{OS}(t_1)$ for the interim analysis of OS are determined such as

$$P_0(Z_{1\_OS} \geq u_{OS}(t_1)) = \alpha(t_1^{OS}) \quad \text{and} \quad P_a(Z_{1\_OS} \leq l_{OS}(t_1)) = \beta(t_1^{OS}) \ ,$$

and critical values $u_{PFS}(t_1)$ and $l_{PFS}(t_1)$ for the interim analysis of PFS are determined such as

$$P_0(Z_{1\_PFS} \geq u_{PFS}(t_1)) = \alpha(t_1^{PFS}) \quad \text{and} \quad P_a(Z_{1\_PFS} \leq l_{PFS}(t_1)) = \beta(t_1^{PFS}) \ ,$$

where P_0 and P_a denote the probabilities under the null hypothesis and the alternative hypothesis, respectively, and $\alpha(t_1^{OS})$, $\beta(t_1^{OS})$ and $\alpha(t_1^{PFS})$, $\beta(t_1^{PFS})$ denote the α and β spent for interim analysis of OS and PFS, respectively, based on the predefined spending functions at information fraction t_1^{OS} and t_1^{PFS}, respectively. (t_1^{OS} is calculated as the ratio of the number of OS events observed at the time of the cut-off for the interim analysis and the total number of OS events targeted for the final analysis; t_1^{PFS} is calculated as the ratio of the number of PFS events observed at the time of the cut-off for the interim analysis and the total number of PFS events targeted for the final analysis).

The boundary for the final efficacy analysis of OS will be calculated such that

$$\alpha(t_1^{OS}) + P_0(Z_{1\_OS} < u_{OS}(t_1), Z_{F\_OS} \geq u_{OS}(t_F)) = 0.0115$$

The boundary for the final efficacy analysis of PFS will be calculated such that

$$\alpha(t_1^{PFS}) + P_0(Z_{1\_PFS} < u_{PFS}(t_1), Z_{F\_PFS} \geq u_{PFS}(t_F)) = 0.001$$

As described in Section 5.1.2, if the number of OS or PFS events in the final analysis deviates from the target number of OS or PFS events, the final analysis criteria will be determined as above taking into account the actual alpha spent at the interim analysis, the actual correlation between the two test statistics for OS, Z_{1_OS} and Z_{F_OS} and the actual correlation between the two test statistics for PFS, Z_{1_PFS} and Z_{F_PFS}, so that the overall 1-sided significance level across all analyses and comparisons is preserved at 0.025.

8. REFERENCES

1. Aaronson NK et al. The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. J Natl Cancer Inst 85:365-76, 1993.

2. Aaronson NK et al. The European Organization for Research and Treatment of Cancer (EORTC) modular approach to quality of life assessment in oncology: an update., in Quality of Life and Parmacoeconomics in Clinical Trials. 2nd Ed. S. B, Editor. Raven Press: New York. 179-89, 1996.

3. Amit O, et al. Blinded independent central review of progression in cancer clinical trials: Results from a meta-analysis and recommendation from a PhRMA working group. European Journal of Cancer 47:1772-1778, 2011.

4. Brookmeyer R, Crowley JJ. A confidence interval for the median survival time. Biometrics. 38: 29-41, 1982.

5. Cella D, Paul D, Yount S, et al. What are the most important symptom targets when treating advanced cancer? A survey of providers in the National Comprehensive Cancer Network (NCCN). Cancer Invest 21:526-535, 2003.

6. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika; 26, 404-413, 1934.

7. Colombo N, Kutarska E, Dimopoulos M, et al. Randomized, open-label, phase III study comparing patupilone (EPO906) with pegylated liposomal doxorubicin in platinum-refractory or -resistant patients with recurrent epithelial ovarian, primary fallopian tube, or primary peritoneal cancer. J Clin Oncol.Nov 1;30(31):3841-7, 2012.

8. Cull A, Howat S, Greimel E, Waldenstrom AC, Arraras J, Kudelka A, Chauvenet L, Gould A; EORTC Quality of Life Group [European Organization for Research and Treatment of Cancer]; Scottish Gynaecological Cancer Trials Group. Development of a European Organization for Research and Treatment of Cancer questionnaire module to assess the quality of life of ovarian cancer patients in clinical trials: a progress report. Eur J Cancer. 37(1):47-53, 2001.

9. The EuroQol Group. EQ-5D-5L User Guide: Basic Information on how to use the EQ-5D-5L: User Guide. V 2.1. The EuroQol Group, Rotterdam. 2015.

10. Fayers P et al. EORTC QLQ-C30 Scoring Manual. EORTC, Brussels. 2001.

11. Gordon A, Fleagle JT, Guthrie D, et al. Recurrent Epithelia Ovarian Carcinoma: A Randomized Phase III Study of Pegylated Liposomal Doxorubicin Versus Topotecan. J Clin Oncol,19: 3312-22, 2001.

12. Greimel E, Bottomley A, Cull A, Waldenstrom AC, Arraras J, Chauvenet L, Holzner B, Kuljanic K, Lebrec J, D'haese S; EORTC Quality of Life Group and the Quality of Life

Unit. An international field study of the reliability and validity of a disease-specific questionnaire module (the QLQ-OV28) in assessing the quality of life of patients with ovarian cancer. Eur J Cancer. 2003 Jul; 39(10):1402-8. Erratum in: Eur J Cancer. 39(17):2570, 2003.

13. Jennison C, Turnbull BW. In: Group Sequential Methods with Applications to Clinical Trials. Chapman & Hall/CRC, Boca Raton, 2000.

14. Kalbfleisch JD, Prentice, RL. Statistical Analysis of Failure Time Data, 2nd Edition. Hoboken, Wiley Interscience, 2011.

15. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 53: 457-81, 1958.

16. Osoba D, Aaronson N, Zee B, Sprangers M, te Velde A. Modification of the EORTC QLQ-C30 (version 2.0) based on content validity and reliability testing in large samples of patients with cancer. The Study Group on Quality of Life of the EORTC and the Symptom Control and Quality of Life Committees of the NCI of Canada Clinical Trials Group. Qual Life Res. 6(2):103-8, 1997.

17. Osoba D et al. Interpreting the significance of changes in health related quality of life scores. J. Cin Oncology. 16:139-144, 1998.

18. Royston P, Parmar MK: The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med 30:2409-2421, 2011.

19. Santner TJ, Snell MK. Small-sample confidence intervals for p1-p2 and p1/p2 in 2 x 2 contingency tables. Journal of the American Statistical Association, 75:386–394, 1980.

20. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 67:145-53, 1980.

21. Stockler MR et al. Patient-Reported Outcome Results From the Open-Label Phase III AURELIA Trial Evaluating Bevacizumab-Containing Therapy for Platinum-Resistant Ovarian Cancer. J Clin Oncol 32:1309-1316, 2014.

22. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health, 2000.

23. Uno H, Wei LJ, et al. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol 32, 2014.

24. Zhang X. Comparison of restricted mean survival times between treatments based on a stratified Cox model. DOI 10.1515/bams-2013-0101 Bio-Algorithms and Med-Systems 2013; 9(4): 183–9.

9. APPENDICES

Appendix 1. Immune-Related Adverse Events

The MedDRA PTs and clusters for irAEs are defined in the Safety Review Plan (SRP) for avelumab.

Immune-related AEs (irAEs) will be programmatically identified as outlined in Table 16. This case definition is hierarchical, ie, each step is only checked for patients and events that have already met the prior step.

Table 16. Case Definition for irAEs

Step	Selection Criteria	Additional Notes
1	Event selected based on a list of pre-specified MedDRA PTs within clusters. These are included in the SRP as Tier1 events (Immune-mediated xxxx). If AE matches the list then it is in for the next step	
2	AE onset during 1st study drug administration or anytime thereafter through 90 days after last dose of study treatment.	This is regardless of start of new anti-cancer drug therapy and regardless of TEAE classifications
3	Answer in the AE eCRF page to 'Was another treatment given because of the occurrence of the event' is 'YES'	
4	AE treated with corticosteroids or other immunosuppressant therapy. For endocrinopathies only: AE required hormone replacement	Look in the conmed pages for AE identifiers that match the AEs from Step 3. For each of such AEs if A) OR B) below are met then the AE is in for the next step A) conmed ATC code is in (H02A, H02B, H02C, D07, A01AC, S01BA, S01BB, L04AA, L04AB, L04AC, L04AD, L04AX) and AE PT is in any of the irAE clusters. B) conmed ATC code is in (H03A, H03B) and AE PT is in one of the irAE clusters associated with "Immune-mediated endocrinopathies"

5	A) No clear etiology (other than immune mediated etiology)	A) From the AE eCRF page Is the AE clearly related to an etiology other than immune-mediated etiology? Yes / No If answer is Yes, check all that apply: • Underlying malignancy / progressive disease. • Other medical conditions. • Prior or concomitant medications / procedures. • Other. Specify.
	B) Histopathology / biopsy consistent with immune-mediated event	B) From the AE eCRF page B1) Was there a pathology /histology evaluation performed to investigate the AE? Y/N B2) If answer to the above is Yes, does the pathology/histology evaluation confirms an immune mediated mechanism for the AE? Y/N B3) If pathology / histology evaluation performed to investigate the AE, provide summary of relevant findings of the pathology /histology report. (Free Text)
	Event is in if [Answer to 5B1 and 5B2 is YES (regardless of answer to 5A)] OR [Answer to 5B1 is YES AND answer to 5B2 is NO AND answer to 5A is NO] OR [Answer to 5B1 is NO AND answer to 5A is NO]	

The data set associated with irAEs may be refined based on medical review. The final data set including any changes based on medical review (eg, addition of cases that are not selected by the programmatic algorithm) will be the basis of the irAE analyses.

Appendix 2. Infusion Related Reactions

For defining an AE as IRR the onset of the event in relation to the infusion of study drug and time to resolution of the event will be considered.

- All AEs identified by the MedDRA PT query describing signs and symptoms will be considered potential IRRs when onset is on the day of study drug infusion (during or after infusion) and the event resolved with end date within 2 days after onset.
- All AEs identified by the MedDRA PTs of Infusion related reaction, Drug hypersensitivity, Anaphylactic reaction, Hypersensitivity, Type 1 hypersensitivity, will be considered potential IRRs when onset is on the day of study drug infusion (during or after the infusion) or the day after the study drug infusion (irrespective of resolution date).

The list of MedDRA PTs for 'IRRs SIGNS and SYMPTOMS' and PTs 'IRRs CORE' are defined in the SRP for avelumab.

Infusion-related reactions (IRRs) will be programmatically identified as outlined in Table 17 or Table 18 (identified for IV drugs only).

Table 17. Case Definition for IRRs – IV study drugs Administered Alone

Condition	Selection criterion
If AE meets [1 AND 2] OR [3 AND (4A OR 4B)] then AE is classified as an IRR	
1	PT is included in the 'IRRs SIGNS and SYMPTOMS' list
2	• AE onset date = date of infusion of study drug AND • AE timing related to study drug ('DURING', 'AFTER') AND • AE outcome in ('RECOVERED/RESOLVED', 'RECOVERED/RESOLVED WITH SEQUELAE', 'RECOVERING/RESOLVING') AND • AE end date – AE onset date <=2
3	PT is included in the 'IRRs CORE' list
4A	• AE onset date = date of infusion of study drug AND • AE timing related to study drug in ('DURING', 'AFTER')
4B	AE onset on the day after infusion

Table 18. Case Definition for IRRs – IV Study Drugs Administered in Combination (eg, Doublets or Triplets)

Condition	Selection criterion
IRR can be associated with the first IV drug and/or subsequent IV drugs that are administered in combination. Without loss of generality assume triplet IV with D_1 administered first then D_2 then D_3. The IV study drug or drugs associated with the IRR need to be identified in the analysis data set to enable subsequent analysis. The following are not sequential and an AE can be classified as an IRR associated with multiple D_J from one or more of I, II, III, IV, V below: I - If the AE meets [1 AND 2A1] for a D_J then the AE is classified as an IRR associated with the D_J that meets the 2A1 criterion II - If the AE meets [1 AND 2A2] for a D_J then the AE is classified as an IRR associated with the D_J and associated with D_{J+1} that meets the 2A2 criterion III - If the AE meets [3 AND 4B] for any D_J then the AE is classified as an IRR associated with all D_J that meet the 4B criterion. IV- If the AE meets [3 AND 4A1] for a D_J then the AE is classified as an IRR associated with the D_J that meets the 4A1 criterion V- If the AE meets [3 AND 4A2] for a D_J then the AE is classified as an IRR associated with the D_J and associated with D_{J+1} that meets the 4A2 criterion	
1	PT is included in the 'IRRs SIGNS and SYMPTOMS' list
2A1	• AE onset date = date of infusion of study drug D_J AND • AE timing related to study drug D_J ('DURING', 'AFTER') AND • [AE timing related to study drug D_{J+1} ('BEFORE') OR AE onset date < date of infusion of study drug D_{J+1}]AND • AE outcome in ('RECOVERED/RESOLVED', 'RECOVERED/RESOLVED WITH SEQUELAE', 'RECOVERING/RESOLVING') AND • AE end date – AE onset date <=2
2A2	• AE onset date = date of infusion of study drug D_J AND • AE timing related to study drug D_J ('DURING', 'AFTER') AND • AE timing related to study drug D_{J+1} ('DURING', 'AFTER') AND • AE outcome in ('RECOVERED/RESOLVED', 'RECOVERED/RESOLVED WITH SEQUELAE', 'RECOVERING/RESOLVING') AND • AE end date – AE onset date <=2
3	PT is included in the 'IRRs CORE' list
4A1	• AE onset date = date of infusion of study drug D_J AND • AE timing related to study drug D_J ('DURING', 'AFTER') AND • [AE timing related to study drug D_{J+1} ('BEFORE') OR AE onset date < date of infusion of study drug D_{J+1}]

4A2	• AE onset date = date of infusion of study drug D_J AND • AE timing related to study drug D_J ('DURING', 'AFTER') AND • AE timing related to study drug D_{J+1} ('DURING', 'AFTER')
4B	AE onset on the day after infusion of study drug D_J