RENAL CELL CARCINOMA - JAVELIN RENAL 101

A PHASE 3, MULTINATIONAL, RANDOMIZED, OPEN-LABEL, PARALLEL-ARM STUDY OF AVELUMAB (MSB0010718C) IN COMBINATION WITH AXITINIB (INLYTA®) VERSUS SUNITINIB (SUTENT®) MONOTHERAPY IN THE FIRST-LINE TREATMENT OF PATIENTS WITH ADVANCED RENAL CELL CARCINOMA

STATISTICAL ANALYSIS PLAN – B9991003

Compounds:	MSB0010718C, AG-013736, SU011248
Compound Name:	Avelumab, Axitinib, Sunitinib
Version:	5.0
Date:	16-Jul-2018

[redacted]

[redacted]

TABLE OF CONTENTS

[redacted]

LIST OF TABLES

LIST OF FIGURES

APPENDICES

CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

CCI

[redacted]

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study B9991003. 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 (i.e., tumor assessment results by the Blinded Independent Central Review (BICR), PK concentration, Anti-Drug Antibody (ADA), Neutralizing Antibody (nAb), biomarker data).

The primary analysis for progression-free survival (PFS) will include all data up to a cut-off date which is determined by the number of events required for the primary analysis of PFS in PD-L1 positive patients and a minimum follow-up of 12 months after the last patient is randomized.

The primary analysis for overall survival (OS) will include all data up to a cut-off date which is determined by the number of events required for the primary analysis of OS in PD-L1 positive patients.

The cut-off dates are determined once a data extract (before database lock) is available which indicates that the required number of events for PFS by Blinded Independent Central Review (BICR) assessment in PD-L1 positive patients and the minimum follow-up of 12 months, or the required number of events for OS in PD-L1 positive patients 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 Objective

- To demonstrate that avelumab in combination with axitinib is superior to sunitinib monotherapy in prolonging PFS or OS in the first-line treatment of PD-L1 positive patients with advanced renal cell carcinoma (aRCC).

Secondary Objectives

- To demonstrate that avelumab in combination with axitinib is superior to sunitinib monotherapy in prolonging PFS in the first-line treatment of patients with aRCC unselected for PD-L1 expression;
- To demonstrate that avelumab in combination with axitinib is superior to sunitinib monotherapy in prolonging OS in the first-line treatment of patients with aRCC unselected for PD-L1 expression;
- To evaluate other measures of efficacy of avelumab in combination with axitinib and sunitinib monotherapy in the first-line treatment of aRCC patients;
- To evaluate progression-free survival on next-line of therapy (PFS2);
- To evaluate the overall safety profile of avelumab in combination with axitinib and sunitinib monotherapy in the first-line treatment of aRCC patients;

[redacted]

- To evaluate the population pharmacokinetics of avelumab and axitinib when administered in combination;
- To evaluate the time to treatment discontinuation/failure due to toxicity;
- To evaluate the proportion of patients who discontinued treatment due to toxicity;
- To evaluate candidate predictive biomarkers in pre-treatment tumor tissue that may aid in the identification of a patient subpopulation most likely to benefit from treatment with avelumab in combination with axitinib and sunitinib monotherapy;
- To assess the immunogenicity of avelumab when combined with axitinib;
- To evaluate the effects of avelumab in combination with axitinib and sunitinib monotherapy on patient-reported outcomes.

CCI

[redacted]

2.2. Study Design

This is a Phase 3, multinational, multicenter, randomized (1:1), open-label, parallel 2-arm study in which approximately 830 patients, including a minimum of 580 PD-L1 positive patients, are planned to be randomized to receive either avelumab in combination with axitinib or sunitinib monotherapy.

- Arm A: avelumab plus axitinib;
- Arm B: sunitinib.

Patients will be stratified according to ECOG PS (0 vs 1) and region (United States vs Canada/Western Europe vs the rest of the world).

Crossover between treatment arms will not be permitted.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoint

- PFS based on BICR assessment per RECIST v1.1 for PD-L1 positive patients.

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.

- OS for PD-L1 positive patients.

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

3.2. Secondary Endpoints

3.2.1. Safety endpoints

- Adverse Events (AEs) as graded by National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) v4.03;

  AEs will be graded by the investigator according to the CTCAE v4.03 and coded using the Medical Dictionary for Regulatory Activities (MedDRA);

- Laboratory abnormalities as graded by NCI CTCAE v4.03;

- Vital signs (blood pressure, pulse rate);

- Time to treatment discontinuation/failure due to toxicity;

  Time to treatment discontinuation/failure due to toxicity is defined as the time from first dose of study treatment to discontinuation of study treatment (discontinuation of either avelumab or axitinib for arm A) due to an adverse event or death due to study treatment toxicity.

- Treatment discontinuation due to toxicity.

  Treatment discontinuation due to toxicity is defined as the discontinuation of study treatment (discontinuation of either avelumab or axitinib for arm A) due to an adverse event or death due to study treatment toxicity.

3.2.2. Efficacy endpoints

- PFS by BICR assessment per RECIST v1.1 for patients unselected for PD-L1 expression.

- OS for patients unselected for PD-L1 expression

- PFS based on Investigator assessment per RECIST v1.1.

- Objective Response (OR), Disease Control (DC), Time to Tumor Response (TTR) and Duration of Response (DR) based on BICR assessment and based on Investigator assessment, per RECIST v1.1.

  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.

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.

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

DR is defined, for patients with an 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.

- Progression-Free Survival on next-line therapy (PFS2).

  PFS2 is defined as the time from randomization to discontinuation of next-line treatment, second objective disease progression, or death from any cause, whichever occurs first.

- Measures of clinical outcome (PFS, OS, OR, DC, TTR, and DR) in biomarker-positive and biomarker-negative subgroups.

3.2.3. Patient reported outcomes

- FACT-Kidney Symptom Index (FKSI-19)

  The FKSI-19 was developed to be part of the Functional Assessment of Chronic Illness Therapy (FACIT) system. It is an instrument specifically designed to be a stand-alone instrument to measure symptoms and quality of life in patients with advanced kidney cancer. It contains 19 questions, most of which overlap with the FACT-G questions (Rao et al. 2009). The FKSI-19 was developed in order to represent the symptoms and issues most important to people receiving treatment for aRCC.

  A 9-item subscale of the FKSI known as the FKSI-Disease Related Symptoms subscale (FKSI-DRS) measuring advanced kidney cancer disease related symptoms was created using inputs from oncologists and patients (Cella 2005). This subscale includes the following 9 items: lack of energy, pain, losing weight, bone pain, fatigue, shortness of breath, coughing, bothered by fevers, and hematuria.

- EuroQoL 5-Dimension (EQ-5D)

  The EuroQol EQ-5D is a 6-item patient-completed questionnaire designed to assess health status in terms of a single index value or utility score (The EuroQoL Group, 1990) There are two components to the EQ-5D; a Health State Profile which has patients rate their level of problems (none, slight, moderate, severe, extreme/unable) in 5 areas (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), and a Visual Analogue Scale (VAS) in which patients rate their overall health status from 0 (worst imaginable) to 100 (best imaginable). Published weights are available that allow for the creation of a single summary score (Rabin and de Charro, 2001). Overall scores range from 0 to 1, with low scores representing a higher level of dysfunction.

3.2.4. Pharmacokinetic endpoints

- PK parameters including trough concentrations (C_{trough}) of avelumab and trough concentrations (C_{trough}) and maximum concentrations (C_{max}) of axitinib

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

Parameter	Definition	Method of Determination
C_{max} [a]	Maximum observed plasma concentration	Observed directly from data
C_{trough} [b]	Predose concentration during multiple dosing	Observed directly from data
$C_{max}(dn)$ [a]	Dose normalized C_{max}	C_{max} / Dose

[a] C_{max} will be calculated only for axitinib
[b] C_{trough} will be calculated for avelumab and axitinib

3.2.5. Immunogenicity endpoints

- Anti-drug antibodies (ADAs; nAb) of avelumab when in combination with axitinib.

3.2.6. Biomarker endpoints

- Tumor tissue biomarker status (i.e., positive or negative; based on, for example, PD-L1 expression and/or quantitation of tumor infiltrating CD8+ T lymphocytes as assessed by immunohistochemistry [IHC]; or based on expression of a gene/protein or group of genes/protein relating to, or indicative of an immune response within the periphery or tumor microenvironment).

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

CCI [redacted]

CCI

[redacted]

3.4. Baseline Variables

3.4.1. Study drug, study treatment and baseline definitions

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

- Arm A = avelumab plus axitinib;
- Arm B = sunitinib.

Determination of PD-L1 status

Determination of PD-L1 status will be performed retrospectively in a central laboratory.

PD-L1 status will be assessed using a validated PD-L1 IHC test kit manufactured under GMP specifications, in conjunction with a pre-specified scoring algorithm and cutoff defining PD-L1 positive versus negative status that will be established prior to initiation of scoring of PD-L1.

Start and end dates of study treatment

For Arm A

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

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

For Arm B

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.

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.

3.4.2. Baseline characteristics

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

- ECOG PS (0 vs 1);
- Region (United States vs Canada/Western Europe vs the rest of the world).

The primary analyses of PFS and OS for PD-L1 positive patients, as well as the secondary analyses of PFS and OS for patients unselected for PD-L1 expression 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	✓	✓ (key secondary only*)	
CCI			
Safety			✓

* key secondary endpoints are OS and PFS for patients unselected for PD-L1 expression

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 set

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 key objectives of the study. Patients who meet any of the following criteria will be excluded from the PP analysis set.

- Patient did not receive at least one dose of the randomized study treatment;
- *Applicable only for PFS:* Patient without a tumor assessment ≥ 6 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);
- *Applicable only for PFS:* Patient without measurable disease at baseline by BICR;
- ECOG status > 1;
- Any prior therapies as defined by the exclusion criterion 1 of the study protocol:
  - Prior systemic therapy directed at advanced or metastatic RCC.
  - Prior adjuvant or neoadjuvant therapy for RCC if disease progression or relapse has occurred during or within 12 months after the last dose of treatment.

- Prior immunotherapy with IL-2, IFN-α, or anti-PD-1, anti-PD-L1, anti-PD-L2, anti-CD137, or anti-cytotoxic T-lymphocyte-associated antigen-4 (CTLA-4) antibody (including ipilimumab), or any other antibody or drug specifically targeting T-cell co-stimulation or immune checkpoint pathways.
- Prior therapy with axitinib and/or sunitinib as well as any prior therapies with other VEGF pathway inhibitors.

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

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 axitinib.

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 axitinib.

4.3.3. Biomarker analysis set

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 Arm A.

4.3.5. PRO analysis set

The PRO analysis set will include all randomized patients (FAS).

5. GENERAL METHODOLOGY AND CONVENTIONS

5.1. Hypotheses and Decision Rules

All references to PFS in Section 5.1.1 and Section 5.1.2 refer to PFS by BICR assessment even when not specifically stated.

5.1.1. Hypotheses and sample size determination

The following statistical hypotheses will be tested to address the primary objectives:

H_{01}: $HR_{PFS+} \geq 1$ vs H_{11}: $HR_{PFS+} < 1$

H_{02}: $HR_{OS+} \geq 1$ vs H_{12}: $HR_{OS+} < 1$

where HR_{PFS+} and HR_{OS+} are the hazard ratios (Arm A vs. Arm B) of PFS and OS, respectively, in PD-L1 positive patients.

In addition the following statistical hypotheses will be tested to address the secondary objectives:

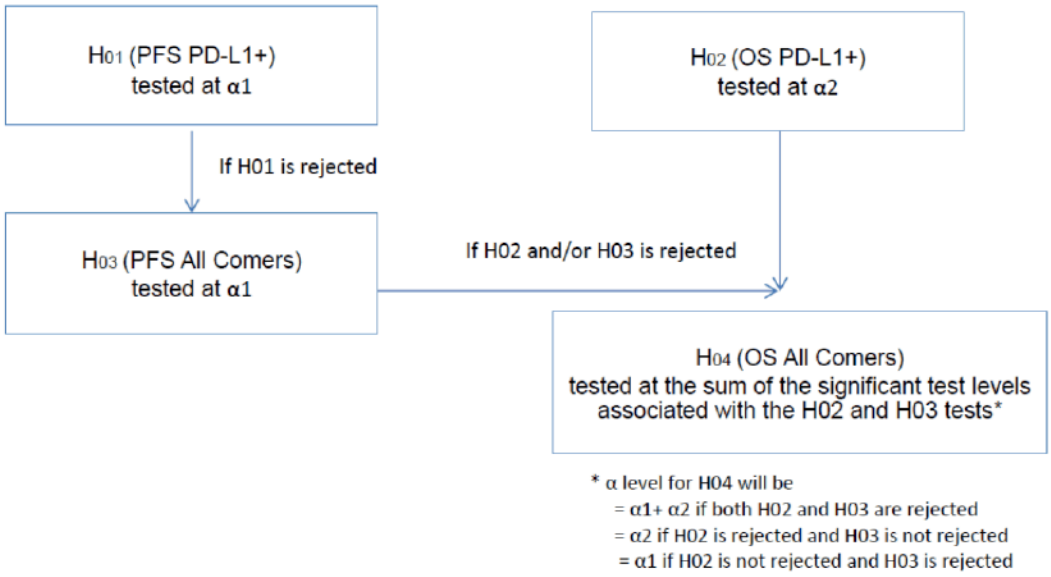
H_{03}: $HR_{PFS} \geq 1$ vs H_{13}: $HR_{PFS} < 1$

H_{04}: $HR_{OS} \geq 1$ vs H_{14}: $HR_{OS} < 1$

where HR_{PFS} and HR_{OS} are the hazard ratios (Arm A vs. Arm B) of PFS and OS, respectively, for patients unselected for PD-L1 expression.

Overall type I-error will be maintained at or below 1-sided 0.025 by allocating $\alpha=0.004$ (α_1) to the PFS comparison in PD-L1+ patients and by allocating $\alpha=0.021$ (α_2) to the OS comparison in PD-L1 positive patients. A gatekeeping procedure will be used to allow further testing of PFS and OS for patients unselected for PD-L1 expression 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



Approximately 830 patients, including a minimum of 580 PD-L1 positive patients, will be randomized to the treatment arms using a 1:1 randomization, stratified by ECOG PS (0 vs 1) and region (United States vs Canada/Western Europe vs the rest of the world).

For the primary PFS comparison, 336 PFS events by BICR assessment for PD-L1 positive patients will be required to have at least 90% power to detect a hazard ratio of 0.65 using a 1-sided log-rank test at a significance level of 0.004, 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 (-15) β-spending function to determine the non-binding futility boundary.

For the primary OS comparison, 368 OS events for PD-L1 positive patients will be required to have at least 90% power to detect a hazard ratio of 0.70 using a 1-sided log-rank test at a significance level of 0.021, and a 4-look group-sequential design with Lan-DeMets (O'Brien-Fleming) α- spending function to determine the efficacy boundary and a Gamma Family (-3) β-spending function to determine the non-binding futility boundary.

The sample size for this study is determined based on the following:

1. the median PFS for patients receiving sunitinib is 11 months (Sutent Summary of Product Characteristics) and the median PFS for patients receiving avelumab in combination with axitinib is 16.9 months for PD-L1 positive patients and 15.7 months for patients unselected for PD-L1 expression; this, corresponds to a hazard ratio (HR) of 0.65 and 0.7, respectively under the exponential model assumption;

2. the median OS for patients receiving sunitinib is 26.4 months (Sutent Summary of Product Characteristics) and the median OS for patients receiving avelumab in combination with axitinib is 37.7 months for PDL1 positive patients and 35.2 months for patients unselected for PD-L1 expression; this corresponds to a HR of 0.7 and 0.75, respectively under the exponential model assumption;

3. PFS drop-out rate of approximately 15% and OS drop-out rate of approximately 5%;
4. 70% of the randomized patients are PD-L1 positive;
5. non-uniform patient accrual accomplished over a 21 month period.

The sample size of approximately 830 patients will also allow an assessment of PFS and OS in patients unselected for PD-L1 expression.

If H_{01} is rejected then PFS in patients unselected for PD-L1 expression can be tested and 490 PFS events by BICR assessment will be required to have at least 90% power to detect a HR of 0.70 using a 1-sided log rank test at a significance level of 0.004, and a 2-look group sequential design with Lan-DeMets (O'Brien-Fleming) α-spending function to determine the efficacy boundary.

If either H_{02} or H_{03} are rejected, then OS in patients unselected for PD-L1 expression can be tested at the sum of the significance levels associated with the significant H_{02} and H_{03} tests (see Figure 1). With 534 OS events, the power is 91% (if both H_{02} and H_{03} are rejected), 90% (if H_{02} is rejected and H_{03} is not rejected) or 74% (if H_{02} is not rejected and H_{03} is rejected) to detect a HR of 0.75 using a 1-sided log rank test at a significance level of 0.025, 0.021 or 0.004, respectively, and a 4-look group sequential design with Lan-DeMets (O'Brien-Fleming) α-spending function to determine the efficacy boundary.

The data cutoff for the primary PFS analysis will occur after the target number of PFS events by BICR assessment in PD-L1 positive patients has been reached and the last patient randomized in the study has been followed for at least 12 months after randomization.

The data cutoff for the primary OS analysis will occur after the target number of deaths in PD-L1 positive patients has been reached.

The study will be considered positive if the stratified log-rank test for either PFS or OS in PD-L1 positive patients is significant at the respective α levels.

5.1.2. Decision rules

The interim and the primary analyses for each endpoint will be performed based on the Full Analysis Set after all patients have been randomized in the study and the target number of events has occurred as described below. A maximum of 4 distinct analyses cutoffs are planned in the study:

- at the time when approximately 235 PFS events (70% of the expected 336 events) by BICR assessment have occurred in PD-L1 positive patients, and all patients have been randomized in the study (IA for PFS and IA1 for OS);
- at the time when 336 PFS events by BICR assessment have occurred in PD-L1 positive patients, and the last patient randomized in the study has been followed for at least 12 months after randomization (primary analysis for PFS and IA2 for OS);
- 15 months after the primary analysis for PFS (IA3 for OS);

- at the time when 368 deaths have occurred in PD-L1 positive patients (primary analysis for OS).

A maximum of two analyses will be performed for PFS (in PD-L1 positive patients and in patients unselected for PD-L1 expression):

1. an interim analysis after all patients have been randomized in the study and at least 235 (70%) of the 336 PFS events in PD-L1 positive patients have been documented by BICR, and
2. the final analysis for PFS after all patients randomized in the study have been followed for 12 months and at least 336 PFS events in PD-L1 positive patients have been documented by BICR.

A maximum of four analyses will be performed for OS (in PD-L1 positive patients and in patients unselected for PD-L1 expression):

1. an interim analysis at the time of the interim analysis for PFS;
2. an interim analysis at the projected time of the final analysis for PFS;
3. an interim analysis 15 months after the final analysis for PFS;
4. a final analysis for OS when 368 deaths in PD-L1 positive patients are observed.

To protect the integrity of the study and to preserve the type I error rate, a gatekeeping testing strategy is implemented and the significance levels for the interim and final efficacy analyses for each endpoint will be determined by using the Lan-DeMets procedure with an O'Brien-Fleming stopping boundary as described in Section 5.1.1.

Table 5 displays the maximum number of analyses expected for each of the primary endpoints and the associated efficacy and futility boundaries if the analyses are performed at the planned number of events as shown in the table. The futility boundaries are non-binding but the study may be stopped for futility if at the time of the first interim analysis, both PFS and OS in PD-L1 positive patients cross the corresponding futility boundaries. If the efficacy boundary is crossed, for either of the primary endpoints, at the time of an interim analysis or at the time of the primary analysis then the primary objective of the study will have been demonstrated.

Table 5. PFS and OS in PD-L1 positive patients – Efficacy and Futility Boundaries

Endpoint	**PFS in PD-L1positive patients**		**OS in PD-L1 positive patients**			
Analysis	**IA1**	**PA**	**IA1**	**IA2**	**IA3**	**PA**
Analysis cutoff trigger	*235 PFS events in PD-L1+*	*336 PFS events in PD-L1+*	*235 PFS events in PD-L1+*	*336 PFS events in PD-L1+*	*15 months after IA2*	*368 deaths in PD-L1+*
Number of events (Information fraction)	235 (70%)	336 (100%)	125 (34%)	195 (53%)	294 (80%)	368 (100%)
p-value (z-value) for efficacy	<0.0006 (<-3.2493)	<0.0038 (< -2.6682)	<0.00007 (< -3.7913)	<0.0015 (< -2.9686)	<0.0093 (<-2.3524)	<0.0180 (<-2.0975)
p-value (z-value) for futility [a]	>0.4059 (> -0.2380)	NA	>0.6411 (>0.3614)	>0.3672 (>-0.3394)	>0.0879 (>-1.3537)	NA

[a] Non-binding.

IA1 = interim analysis 1, IA2 = interim analysis 2, IA3 = interim analysis 3, PA = primary analysis

The observed number of events at the IAs may not match the planned number of 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.

Table 6 displays the analyses triggers for PFS and OS in patients unselected for PD-L1 expression, as well as the associated efficacy boundaries if the analyses are performed at the planned number of events as shown in the table. As described in Section 5.1.1, the significance level for the analyses of these endpoints is determined by the gatekeeping procedure.

Table 6. PFS and OS for patients unselected for PD-L1 expression – Efficacy Boundaries

Endpoint	**PFS in patients unselected for PD-L1 expression**		**OS in patients unselected for PD-L1 expression**			
Analysis	**IA1**	**PA**	**IA1**	**IA2**	**IA3**	**PA**
Analysis cutoff trigger	*235 PFS events in PD-L1+*	*336 PFS events in PD-L1+*	*235 PFS events in PD-L1+*	*336 PFS events in PD-L1+*	*15 months after IA2*	*368 deaths in PD-L1+*
Number of events [a] (Information	343 (70%)	490 (100%)	182 (34%)	283 (53%)	427 (80%)	534 (100%)

Endpoint	**PFS in patients unselected for PD-L1 expression**		**OS in patients unselected for PD-L1 expression**			
Analysis	**IA1**	**PA**	**IA1**	**IA2**	**IA3**	**PA**
Analysis cutoff trigger	*235 PFS events in PD-L1+*	*336 PFS events in PD-L1+*	*235 PFS events in PD-L1+*	*336 PFS events in PD-L1+*	*15 months after IA2*	*368 deaths in PD-L1+*
fraction)						
p-value (z-value) for efficacy [b]	<0.0006 (<-3.2478)	<0.0038 (< -2.6683)	<0.00012 (< -3.6656)	<0.0020 (< -2.8726)	<0.0115 (<-2.2727)	<0.0212 (<-2.0286)

[a] Number of events expected under H_{13} for PFS (assuming a HR of 0.7) and H_{14} for OS (assuming a HR of 0.75).

[b] The p-values and z-values noted for OS are those associated with the scenario when both H_{02} and H_{03} are rejected.

IA1 = interim analysis 1, IA2 = interim analysis 2, IA3 = interim analysis 3, PA = primary analysis

The observed number of events at the IAs may not match the planned number of events. The efficacy and futility boundaries will be updated based on the actual number of observed events using the pre-specified α-spending functions.

Since the observed number of events at the interim analysis may not be exactly equal to the planned number of 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, for the primary endpoints, futility boundaries. If the study continues to final analysis, the p-value that will be used to declare statistical significance at the final analysis for each endpoint will be based on the actual number of events documented at the cut-off date for the final analysis, the α already spent at the interim analysis and the gatekeeping testing strategy. Further details are provided in Section 7.

Based on the stopping boundaries defined above and the timing of the interim and final analyses as described above, the design has the following operating characteristics.

Table 7 Simulated cumulative probabilities to stop for efficacy or futility at the interim or final PFS analysis in PD-L1 positive patients

Scenario	**Look**	**Number of PFS events**	**Calendar Time (months)**	**P(Reject H_{01})**	**P(Reject H_{11})**
H_{01} is true (HR=1)	Interim	235	23	0.0008	0.5933
	Final	336	29	0.0049	0.9951
H_{11} is true (HR=0.65)	Interim	235	25	0.5088	0.0011
	Final	336	32	0.8918	0.1082

Simulations performed in EAST 6.3 with number of simulations = 10,000 and seed=1441014

Table 8. Simulated cumulative probabilities to stop for efficacy or futility at the interim or final OS analyses in PD-L1 positive patients

Scenario	Look	Number of OS events	Calendar Time (months)	P(Reject H_{02})	P(Reject H_{12})
H_{02} is true (HR=1)	Interim 1	125	23	0.0001	0.3668
	Interim 2	195	29	0.0013	0.6554
	Interim 3	294	41	0.0100	0.9156
	Final	368	53	0.0198	0.9802
H_{12} is true (HR=0.70)	Interim	125	25	0.0350	0.0098
	Interim 2	195	32	0.3095	0.0226
	Interim 3	294	46	0.7495	0.0538
	Final	368	61	0.8943	0.1057

Simulations performed in EAST 6.3 with number of simulations = 10,000 and seed=2564105

5.2. General Methods

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

- Arm A = avelumab plus axitinib;
- Arm B = sunitinib.

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 i.e., number of non-missing values and number of missing values [i.e., 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 (e.g., 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 (e.g., 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 6.1.1 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 'Concomitant Surgery', and 'Follow-up Surgery' eCRF pages when 'Surgery Outcome' = 'Resected' or 'Partially Resected'.

When 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', 'Concomitant Surgery', 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

For reporting conventions, mean and median should generally be displayed to 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. E.g., 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 (i.e. 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 6.2.2.5). 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, e.g. 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 (i.e., not done) or NS (i.e., 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 (i.e., 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 (i.e., 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.

5.3.2. Handling of incomplete dates

5.3.2.1. Disease history

Incomplete dates for disease history (e.g., 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 (e.g., X-ray, CT scan) must be completed with day, month and year.

If there are multiple scan dates associated with an evaluation, i.e., 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 (e.g., 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.

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. Progression-free survival as assessed by BICR per RECIST v1.1 in PD-L1 positive patients

6.1.1.1. Primary analysis

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

Progression-Free Survival (PFS) is defined as the time from the 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 (i.e., ≤ 12 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 6 weeks (+7 days window allowed) until PD regardless of initiation of subsequent anti-cancer therapy. After 18 months from the date of randomization, tumor assessments will be conducted less frequently, i.e., at 12-week intervals (+7 days window allowed).

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	Censored [a]
PD or death - After at most one missing or inadequate post-baseline tumor assessment, OR - ≤ 12 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 ≤12 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 the experimental arm 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), i.e. 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 6, 12, 18, 24, 30, 36 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley 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 (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 6, 12, 18, 24, 30, 36 months will be estimated with corresponding 2-sided 95% CIs.

6.1.2. Overall survival in PD-L1 positive patients

6.1.2.1. 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 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.

$$\text{OS (months)} = [\text{date of death or censoring} - \text{date of randomization} + 1]/30.4375$$

The primary analysis of OS will compare the OS time between the experimental arm 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), i.e. 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, 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, 30, 36, 42, 48, 60 months will be estimated with corresponding two-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley 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 (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 11 following the hierarchy shown.

Table 11. 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 > 16 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, 30, 36, 42, 48, 60 months will be estimated with corresponding 2-sided 95% CIs.

6.2. Secondary Endpoint(s)

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 separately based on BICR assessments 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 6.1.1.1 now referring to PD by investigator assessment instead of PD by BICR assessment.

Endpoints defined in Sections 6.2.2.5, 6.2.2.6, 6.2.2.7, 6.2.2.8 and 6.2.2.9 will be summarized separately for PD-L1 positive patients and for patients unselected for PD-L1 expression.

6.2.2.1. Overall survival for patients unselected for PD-L1 expression

The methodology described in Section 6.1.2 for OS in PD-L1 positive patients will be followed for the assessment of OS for patients unselected for PD-L1 expression. A stratified log-rank test(1-sided) stratified by randomization stratification factors will be used at the interim and/or final analyses at the significance level associated with the testing strategy outlined in Section 5.1.1 for the testing of H_{04}.

6.2.2.2. Progression-free survival as assessed by BICR per RECIST v1.1 for patients unselected for PD-L1 expression

The methodology described in Section 6.1.1 for PFS by BICR assessment in PD-L1 positive patients will be followed for PFS by BICR assessment for patients unselected for PD-L1 expression. A stratified log-rank test (1-sided) stratified by randomization stratification factors will be used at the interim and/or final analyses at the significance level associated with the testing strategy outlined in Section 5.1.1 for the testing of H_{03}.

6.2.2.3. Sensitivity analyses for progression-free survival

All sensitivity analyses will be performed for PD-L1 positive patients and for patients unselected for PD-L1 expression.

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 6.1.1.1 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
- 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 proportional hazards assumption will be checked visually by plotting log(-log(PFS)) 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 PFS by BICR assessment 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 (i.e., 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 (e.g., 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.

In this section, 'survival' is meant to denote PFS.

The treatment effect between the experimental arm and the control 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 the experimental arm and the control arm will be plotted against time τ.

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

Table 12. 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

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 PFS 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.

6.2.2.4. Sensitivity analyses for overall survival

All sensitivity analyses will be performed for PD-L1 positive patients and for patients unselected for PD-L1 expression.

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 6.1.2.1 with the modifications below:

- PP analysis set;

- unstratified;
- using strata derived according to eCRF data instead of that entered in IRT.

Methods for evaluating the validity of model assumptions

The same methodology described in Section 6.2.2.3 for PFS will be used for OS.

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

The same methodology described in Section 6.2.2.3 for PFS will be used for OS.

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

- 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) $\geq$ 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) $\geq$ 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 $\leq$ 12 weeks after the date of randomization (and not qualifying for CR, PR, SD or non-CR/non-PD).
- NE: all other cases.

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 BOR of complete response (CR) or partial response (PR) according to RECIST v1.1.

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 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 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 too early' (i.e., 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 13 outlines the possible BOR outcomes by investigator and BICR.

Table 13. 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.

6.2.2.6. 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.

6.2.2.7. 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, 18, 24, 30 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley 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 (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.

6.2.2.8. 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).

6.2.2.9. Progression-free survival on next-line therapy (PFS2)

PFS2 is defined as the time from the date of randomization to discontinuation of next-line treatment after first objective disease progression by investigator assessment, second objective disease progression by investigator assessment after initiation of next-line treatment, or death from any cause, whichever occurs first.

PFS2 (months) = [date of event or censoring – start date +1]/30.4375

A patient will be considered to have an event if

1) the patient had objective PD on or prior to start of next-line anti-cancer treatment, AND started next-line anti-cancer treatment, AND (had objective PD after start of next-line anti-cancer treatment OR discontinued next-line anti-cancer treatment)

   OR

2) the patient died.

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median PFS2 time with 2-sided 95% CIs. In particular, the PFS2 rate at 6, 12, 18, 24, 30, 36, 48, 60 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley 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 (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.

The censoring and event date options to be considered for PFS2, each corresponding censoring reason and its hierarchy are presented in Table 14. Frequency (number and percentage) of patients with an event (PD after next line treatment, discontinuation of next-line treatment or death) and censoring reasons will be presented by treatment arm.

Table 14. Outcome, Event Dates, Censoring Reasons and Hierarchy for PFS2 Analyses

Scenario	Date of event/ censoring	Event/ Censoring reason/ Censoring hierarchy
(No PD[a]) and (no death)	Date of last adequate tumor assessment[b] documenting no PD	Censored/ No PD/ 1
(No PD[a]) and death	Date of death	Event (Death)
(PD[a] date > NTX[c] start date) and (no death)	Start date of NTX[c]	Censored/ Start of new anti-cancer treatment before PD/ 2
(PD[a] date > NTX[c] start date) and death	Date of death	Event (Death)
(PD[a] date ≤ NTX[c] start date) and (PD2[d] date > NTX[c] start date)		
• If PD2 date ≤ NTX[c] end date	Date of PD2	Event (PD after start of next line anti-cancer treatment)
• If PD2 date > NTX[c] end date	End date of NTX[c]	Event (Discontinuation of next line anti-cancer treatment)
(PD[a] date ≤ NTX[c] start date) and (no PD2[d]) and (no death)		
• If NTX[c] end date is non-missing	End date of NTX[c]	Event (Discontinuation of next line anti-cancer treatment)
• Else if [withdrawal of consent date ≥ date of randomization OR End of study (EOS) = Subject refused further follow-up]	(Withdrawal of consent date) or (EOS visit date where subject refusal of further follow-up is recorded)	Censored/ Withdrawal of consent/ 3
• Else if [lost to follow-up in any disposition page]	Last contact date	Censored/ Lost to follow-up/ 4
• Else if no prior conditions are met	Last contact date	Censored/ Ongoing without PFS2 event/ 5
(PD[a] and no NTX[c]) and (no death)	Last contact date	Censored/ Ongoing without PFS2 event/ 6
(PD[a] date ≤ NTX[c] start date) and (no PD2[d]) and death		
• If NTX[c] end date is non-missing	End date of NTX[c]	Event (Discontinuation of next line anti-cancer treatment)
• Else if the prior condition is not met	Date of death	Event (Death)

Scenario	Date of event/ censoring	Event/ Censoring reason/ Censoring hierarchy
(PD[a] and no NTX[c]) and death	Date of death	Event (Death)

[a] PD is the first PD by investigator assessment per RECIST v1.1, without considering any censoring rules

[b] If there are no adequate post-baseline assessments, then the censoring date is the date of randomization. If patient has initiated next-line anti-cancer treatment, the last adequate post-baseline assessment on or prior to start date of next-line anti-cancer treatment will be considered.

[c] NTX is the next-line anti-cancer treatment.

[d] PD2 is the first PD after initiation of next-line anti-cancer treatment by investigator assessment per RECIST v1.1, without considering any censoring rules.

The PFS2 time or censoring time and the reasons for censoring will also be presented in a patient listing.

6.2.3. Pharmacokinetic endpoints

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

C_{trough} for avelumab and C_{trough} and C_{max} axitinib will be summarized descriptively (n, mean, SD, CV, median, minimum, maximum, geometric mean, its associated CV, and 95% confidence interval) by cycle and day.

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

Pharmacokinetic parameters for avelumab and axitinib 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 dose level, cycle, day and nominal time. Additionally similar descriptive statistics will also be generated for dose-normalized axitinib pharmacokinetic parameters, if applicable.

- Pharmacokinetic parameters for avelumab and axitinib will be listed and summarized by cycle and study day using descriptive statistics (n, mean, SD, %CV, median, minimum, maximum, geometric mean and its associated %CV, and 95% CI). PK parameters with zero values will be excluded from the calculation of geometric means and its associated %CV. If an intrapatient dose escalation or reduction occurs, dose-dependent PK parameters (C_{max}) 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 C_{max} parameters will be

summarized (as described above) using data in which different avelumab and axitinib doses were administered.

- Box plots for C_{max} and C_{trough} for axitinib 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 C_{max} and C_{trough} 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. In addition, box plots for dose-normalized axitinib C_{max} and C_{trough} parameters will be created using data from patients in Arm A.

- C_{trough} 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.

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 and axitinib exposure and biomarkers or significant safety/efficacy endpoints. The results of these analyses, if performed, may be reported separately.

6.2.5. Biomarker endpoints

Biomarkers will be summarized at all timepoints described below:

- Soluble proteins: screening, C1D1, C2D1, C3D1 and EOT.

- Tumor based biomarkers, archival FFPE Tumor Tissue Block: screening

- Tumor based biomarkers, de novo Tumor Biopsy: screening and EOT.

Summary statistics, using descriptive statistics (mean, SD, median, Q1, Q3, minimum, and maximum), at each timepoint and percent change from the value at screening if available (for continuous markers) will be provided separately for each treatment arm.

For each defined biomarker or group thereof (signature), patients may be classified as positive, negative, or unknown according to scoring algorithms and cutoffs established external to this study. Subset analyses for PFS by BICR assessment and OS will be performed for the subgroups defined by these cutoffs using the methodology described in Section 6.4. If there are fewer than 20 events in a subgroup for PFS by BICR assessment or OS, only descriptive summaries will be produced.

6.2.6. Endpoints for immunogenicity data of avelumab

All analyses described below are performed for Arm A only.

Blood samples for avelumab immunogenicity testing will be collected within 2 hours before the start of the avelumab infusion on Day 1, Day 15 and Day 29 of Cycle 1, on Day 1 and Day 29 of Cycles 2-4, on Day 1 of Cycle 6, and then every 2 cycles thereafter. 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 15.

Table 15 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 16. For nAb, treatment-boosted is not applicable since no titer result is available.

Table 16 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 ADA 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 endpoint pertinent to the immunogenicity analyses is C_{trough} .

Blood samples for avelumab PK will be collected pre-dose on Day 1, Day 15 and Day 29 of Cycle 1, on Day 1 and Day 29 of Cycles 2-4, Day 1 of Cycle 6, every 2 cycles after Cycle 6, and 1 month after end of avelumab treatment.

C_{trough} 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} 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} has 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 -42. Linear-linear and log-linear plots of mean and median for C_{trough} 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.

CCI

CCI

[redacted]

CCI

[redacted]

6.3.2. PRO endpoints

All PRO analyses will be performed for PD-L1 positive patients and for patients unselected for PD-L1 expression and will be based on the FAS.

6.3.2.1. PRO instruments

Patient-reported outcomes of HRQoL will be evaluated using FACT-Kidney Symptom Index (FKSI-19) and EQ-5D-5L. A 9-item subscale of the FKSI-19 known as the FKSI-Disease Related Symptoms subscale (FKSI-DRS) measures advanced kidney cancer disease related symptoms.

The primary PRO endpoint will be the time to symptom deterioration (TTD) which is defined as the time from date of randomization to the first time the patient's score shows a 3-point or greater decrease in FKSI-DRS.

6.3.2.2. Scoring procedure

The FKSI-19 and EQ-5D-5L will be scored according to their respective validation papers and user's guides. For the EQ-5D-5L, utility scores will be collated using the published weights (tariffs) for the United Kingdom (UK). Missing values in the instruments will also be handled following the guidance from their respective User Manuals.

6.3.2.3. Instrument completion rates

For each treatment arm and at each time point, the number and percentage of patients who complete the FKSI-19, FKSI-DRS 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.

6.3.2.4. Descriptive summaries over time

Separate tables will be used to summarize, for each treatment arm, the mean (and SD), and median (and range) of absolute scores and change from baseline of the FKSI-19, FKSI-DRS EQ-5D-5L, and EQ-5D-5LVAS 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.

6.3.2.5. Time-to-event endpoints

Time to deterioration (TTD) endpoint

A 3-point change from baseline in FKSI-DRS has been established as clinically meaningful.

Time to deterioration (TTD) is defined as the time from the date of randomization to the first time the patient's score shows a 3-point or greater decrease from baseline in FKSI-DRS.

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

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 for the experimental arm vs control arm.

Each stratum will define a separate baseline hazard function (using the 'STRATA' statement in SAS PROC PHREG), i.e. 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).

Time to Improvement (TTI) endpoint

A 3-point change from baseline in FKSI-DRS has been established as clinically meaningful.

Time to improvement (TTI) is defined as the time from the date of randomization to the first time the patient's score shows a 3-point or greater increase from baseline in FKSI-DRS.

TTI data will be censored at the last time when a patient completed the assessment for patients who have not improved.

TTI will be analyzed using the same methodology as TTD.

6.3.2.6. Binary endpoints

The proportion of patients who have improved from baseline at Cycle 5, defined as a 3-point or greater increase from baseline in FKSI-DRS, will be summarized using descriptive statistics as described in Section 5.2.3. Chi-square test will be used to compare the proportions between the experimental arm and control arm. Patients who did not complete the first 5 cycles of treatment and did not show improvement prior to EOT will be counted as not having improved.

6.3.2.7. Continuous endpoints

Random coefficient models will be carried out for the FKSI-19, FKSI-DRS EQ-5D-5L, and EQ-5D-5LVAS using PROC MIXED.

Outcomes are PRO post-baseline scores and the predictors are the corresponding baseline PRO score, treatment, time (treated as a continuous variable), and treatment-by-time interaction. Intercept and time are considered as random effects particular to each subject.

All available data for each patient will be used in analyses. 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).

6.3.2.8. Sensitivity Analyses

Sensitivity analyses based on a 2-point, 4-point and 5-point decrease from baseline in FKSI-DRS will be conducted for TTD; similarly, sensitivity analyses based on a 2-point, 4-point and 5-point increase from baseline in FKSI-DRS will be conducted for TTI. Cumulative distribution function plots of the change from baseline in the FKSI-DRS for Cycle 2, 4, 6, 8 and 10 will also be presented.

For TTD and TTI, the primary analyses will include data from baseline up to EOT assessment (not including EOT). Sensitivity analyses for these endpoints will include all assessments from baseline to the end of the safety follow-up period.

6.4. Subset analyses

Subset analyses will be performed for PD-L1 positive patients and for patients unselected for PD-L1 expression.

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

The following subgroups will be defined and used for analyses:

- Randomization stratification factors
  - ECOG PS 0 (Reference) vs 1
  - Geographical Region (United States (Reference) vs Canada/Western Europe vs Rest of the World)
- Age
  - Age < 65 years (Reference)
  - Age ≥ 65 years
- Gender
  - Male (Reference)
  - Female
- 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)
- Nephrectomy at baseline
  - Yes (Reference)
  - No
- MSKCC prognostic criteria at baseline (detailed in Section 6.5.1.3)
  - Favorable (Reference)
  - Intermediate
  - Poor
- Heng prognostic criteria at baseline (detailed in Section 6.5.1.3)
  - Favorable (Reference)
  - Intermediate
  - Poor

Subset analyses for PFS, OS and DR will use the primary censoring rules described in Sections 6.1.1.1 and 6.1.2.1. All the subgroup analyses are exploratory. Treatment arms will be compared for PFS and OS 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 for the interaction model parameter.

The HR for PFS and OS 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 performed for PD-L1 positive patients and for patients unselected for PD-L1 expression and 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)

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), 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 diagnosis of local/regional/distant recurrence of disease (months), defined as (date of randomization – date of diagnosis of local/regional/distant recurrence of disease)/30.4375
- Type of disease recurrence
- Histopathological and Furhman grade
- TNM and AJCC TNM, both at initial diagnosis and current stage
- Prior nephrectomy
- Memorial Sloan-Kettering Cancer Center (MSKCC) prognostic criteria (Motzer et al., 2002)

  MSKCC prognostic criteria are based on the detection of five risk factors: ECOG PS ≥ 2, high LDH (> 1.5 times the ULN), serum hemoglobin < LLN, high corrected serum calcium (> 10 mg/dL), time from initial diagnosis to study treatment start less than one year. Patients will be assigned to one of the following risk groups:

  - Favorable (0 risk factors)
  - Intermediate (1 or 2 risk factors)
  - Poor (3 or more risk factors)

- Heng prognostic criteria (Heng et al., 2009)

  Heng criteria are based on the detection of the same risk factors evaluated by the MSKCC criteria, except for LDH, with the addition of: absolute neutrophils count > ULN, platelets count > ULN. Patients will be assigned to one of the following risk groups:

  - Favorable (0 risk factors)
  - Intermediate (1 or 2 risk factors)

- Poor (3 or more risk factors)

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.

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 A)
- 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 A)
- 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 (Arm A/Arm B/none) vs patients treated (Arm A/Arm B/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 axitinib in the combination arm 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. I.e., 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 CRF exposure page, if the patient received study treatment on that visit (i.e., 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 (i.e., 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 + 42 – 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 B) or at least one of the study drugs (Arm A) should be included.

Exposure will 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 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 6-week cycles.
- Axitinib administered BID PO at a dose of 5 mg.
- Sunitinib administered 50 mg PO QD on Schedule 4/2.

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

- Avelumab - total dose / weight
- Axitinib - total dose
- Sunitinib - total dose.

Exposure and compliance will be summarized by cycle and overall for each study drug in each treatment arm.

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 avelumab +1)/7,

where end date = start date of last cycle with non-zero dose of avelumab + 42 – 1

Duration of exposure to avelumab (weeks) = (last dose date of avelumab – first dose date of avelumab + 14)/7

Cumulative dose in a cycle or overall is the sum of the actual doses of avelumab received in a cycle or overall, respectively.

Actual Dose Intensity (DI)

- By cycle actual DI (mg/kg/6-week cycle) = [cumulative dose in the cycle (mg/kg)]/[cycle duration (weeks)/6]
- Overall actual DI (mg/kg/6-week cycle) = [overall cumulative dose (mg/kg)] / [intended duration of treatment with avelumab (weeks)/6].

Relative Dose Intensity (RDI)

- Intended DI (mg/kg/6-week cycle) = [intended cumulative dose per cycle] / [intended number of 6-weeks in a cycle] = [30 (mg/kg)] / [1(6-week cycle)] = 30 (mg/kg/6-week cycle)
- By cycle RDI (%) = 100 × [by cycle actual DI] / [intended DI]
  = 100 × [by cycle actual DI] / [30 (mg/kg/6-week cycle)]
- Overall RDI (%) = 100 × [overall actual DI] / [intended DI]
  = 100 × [overall actual DI] / [30 (mg/kg/6-week cycle)]

6.5.3.2. Exposure to axitinib

The dose level is calculated as actual dose administered (mg/day).

Intended duration of treatment with axitinib (weeks)=

(last dose date of axitinib – first dose date of axitinib + 1)/7

Duration of exposure to axitinib (weeks) =

(last dose date of axitinib – first dose date of axitinib + 1)/7

Cumulative dose is the sum of the actual doses of axitinib received in the study.

Actual Dose Intensity (DI)

- Overall actual DI (mg/week) = [overall cumulative dose (mg)] / [intended treatment duration (weeks)]

Relative Dose Intensity (RDI)

- RDI (%) = 100 × [overall cumulative dose] / [intended cumulative dose per week × number of weeks from first dose of study drug to last dose of axitinib]

  = 100 × [overall cumulative dose] / [7 ×2 × 5 × duration of exposure to axitinib]

6.5.3.3. Exposure to sunitinib

The dose level is calculated as actual dose administered (mg/day).

Intended duration of treatment with sunitinib (weeks) =

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

where end date = start date of last cycle with non-zero dose of sunitinib + 42 – 1

Duration of exposure to sunitinib (weeks) = (last dose date of sunitinib – first dose date of sunitinib + d)/7, where d = 1 if the patient discontinues sunitinib prior to the end of the last cycle or d = 14 otherwise (i.e. d=1 if date of last dose of sunitinib – date of first dose of sunitinib in the last cycle < 28).

Cumulative dose in a cycle or overall is the sum of the actual doses of sunitinib received in a cycle or overall, respectively.

Actual Dose Intensity (DI)

- By cycle actual DI (mg/6-week cycle) = [cumulative dose in the cycle (mg)]/[cycle duration (weeks)/6]
- Overall actual DI (mg/6-week cycle) = [overall cumulative dose (mg)] / [intended duration of treatment with sunitinib (weeks)/6].

Relative Dose Intensity (RDI)

- Intended DI (mg/6-week cycles) = [intended cumulative dose per cycle] / [intended number of 6-weeks in a cycle] = [7×50×4 (mg)] / [1 (6-week cycle)] = 1400 (mg/6-week cycle)
- By cycle RDI (%) = 100 × [by cycle actual DI] / [intended DI]
  = 100 × [by cycle actual DI] / [1400 (mg/6-week cycle)]
- Overall RDI (%) = 100 × [overall actual DI] / [intended DI]
  = 100 × [overall actual DI] / [1400 (mg/6-week cycle)]

6.5.3.4. Dose reductions

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

Applicable to axitinib and sunitinib. Dose reduction is defined as a change to a non-zero dose level lower than that planned in the protocol.

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

6.5.3.5. Dose escalations

Applicable to axitinib only. Dose escalation is defined as a change to a dose level higher than that planned in the protocol.

Number and percentage of patients with at least one dose escalation as well as a breakdown of dose escalations (1, 2, ≥3) will be summarized.

6.5.3.6. Dose interruptions

Applicable to axitinib and sunitinib.

An interruption is defined a 0 mg dose administered on one or more days for axitinib or within the dosing period of a cycle for sunitinib. What follows defines how dose interruptions will be counted in the case of multiple dose interruptions.

- If an interruption occurs consecutively for at least two days due to the same reason, then it will be counted only once (example: If the actual dose on days 1-3 is 10 mg and actual dose on days 4-5 is 0 mg and dose interruption on days 4-5 is due to AE, then the total number of dose interruptions is 1).
- If an interruption occurs consecutively for at least two days due to different reasons, then it will be counted for each reason (example: If the actual dose on days 1-3 is 10 mg and actual dose on days 4-5 is 0 mg and dose interruption on day 4 is due to AE and dose interruption on day 5 is due to dosing error, then the total number of dose interruptions is 2).
- If an interruption occurs for more than one day due to the same reason, but the days are not consecutive, i.e. there is at least one dosing day in between, then each dose interruption will be counted as a different occurrence (example: If the actual dose on days 1, 3 and 5, is 10 mg and actual dose on days 2 and 4 is 0 mg, and dose interruptions on day 2 and 4 are both due to dosing error, the total number of dose interruptions is 2).

A dose interruption is not considered a dose reduction or a dose delay.

The number and percentage of patients with dose interruptions and the corresponding reasons will be summarized.

6.5.3.7. Dose delays

Applicable to avelumab and sunitinib.

Avelumab

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.

In Cycle 1,

Dose Delay (days) = day of first dose of avelumab – 1.

After the first dose of Cycle 1

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

Sunitinib

Dose delay is the difference between the actual time between the last non-zero dose of a cycle and the actual first non-zero dose of the next cycle and the planned time between the

same two doses. Dose delay is only calculated for the first dose of sunitinib administered in each cycle.

In Cycle 1,

Dose Delay for Dose 1 (days) = day of first dose of sunitinib – 1.

In Cycle >1

Dose Delay for Dose x (days) = Date of first dose of sunitinib in Cycle x – Date of first dose of sunitinib in Cycle (x-1) – 42

Dose Delays will be grouped into the following categories

- No delay
- 1-3 days delay
- 4-6 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 1-3 days delay.

No delay and 1-3 days delay will also be summarized together.

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

6.5.3.8. Infusion rate reductions

Applicable to avelumab.

The number and percentage of patients with at least one infusion rate reduction of ≥50% 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.

6.5.3.9. Infusion interruptions

Applicable to avelumab.

An infusion interruption is defined as an infusion that is stopped and re-started on the same day (i.e., 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', 'Concomitant Surgery' 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 (i.e., no answer to the question 'Relationship with study treatment'). Related AEs are those related to at least one of the study drugs (in Arm A).
- **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 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 axitinib
  - TEAEs leading to dose reduction of sunitinib
  - TEAEs leading to interruption of avelumab
  - TEAEs leading to interruption of axitinib
  - TEAEs leading to interruption of sunitinib
  - TEAEs leading to discontinuation of avelumab
  - TEAEs leading to discontinuation of axitinib
  - TEAEs leading to discontinuation of sunitinib
  - 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 axitinib
  - Related TEAEs leading to discontinuation of sunitinib
  - 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 axitinib by SOC and PT
- TEAEs leading to dose reduction of sunitinib 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 6.5.3.7). 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 6.5.3.7).

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 axitinib by SOC and PT
- TEAEs leading to interruption of sunitinib 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 axitinib by SOC and PT
- TEAEs leading to both interruption and dose reduction of sunitinib by SOC and PT

This summary will take into account PTs with both actions as defined in Section 6.6.1, even though 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 axitinib by SOC and PT
- Related TEAEs leading to discontinuation of axitinib by SOC and PT
- TEAEs leading to discontinuation of sunitinib by SOC and PT
- Related TEAEs leading to discontinuation of sunitinib 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 axitinib, by Cluster and PT
- irAEs leading to discontinuation of sunitinib, 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
- Serious IRRs, by PT
- Time related to first onset of an IRR (infusion 1, infusion 2, infusion 3, infusion 4 or later)

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
  - 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
- 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. Time to treatment discontinuation or failure due to toxicity

Time to treatment discontinuation or failure due to toxicity (here onwards referred to as time to treatment failure (TTF), for simplicity) is defined as the time from first dose of study treatment to discontinuation of study treatment (discontinuation of either avelumab or axitinib for arm A) due to an adverse event or death due to study treatment toxicity.

TTF (in months) = [Date of study treatment discontinuation – start date of study treatment +1]/30.4375

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

6.6.6. Treatment discontinuation due to toxicity

Treatment discontinuation due to toxicity is defined as the discontinuation of study treatment (discontinuation of either avelumab or axitinib for arm A) due to an adverse event or death due to study treatment toxicity. The number and percentage of patients who discontinue treatment due to toxicity in each treatment arm will be estimated by dividing the number of patients who meet the above criteria by the number of patients in the safety set.

The corresponding exact 2-sided 95% CIs will be provided for each treatment arm, using the Clopper-Pearson method (exact CI for a binomial proportion as computed by default by the FREQ procedure using the EXACT option).

6.6.7. Laboratory data

6.6.7.1. 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 (e.g., 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 (i.e., 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 (e.g., hypokalemia) grades at baseline and post baseline will be set to 0 when the variables are derived for summarizing high direction toxicity (e.g., 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:

$$\text{Derived differential absolute count} = (\text{WBC count}) \times (\text{Differential \%value} / 100)$$

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, i.e., 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 (i.e. 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, i.e.:

- 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), B-type natriuretic peptide (BNP), troponin, and CK-MB

6.6.7.2. 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), prothrombin time (PT), international normalized ratio (INR), and PTT.
- Urinalysis: all urinalysis parameters
- HCV, HBV testing
- Hormone tests: TSH, free T4, ACTH
- ANCA, ANA, RF
- 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.

6.6.8. 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. 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.

6.6.9. 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 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% during the on-treatment period
- ≥15-point decrease from baseline in LVEF% to a post-baseline value < LLN during the on-treatment period.

Clinically significant findings will be listed.

6.6.10. 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 (e.g. >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.

6.6.11. Physical examination

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

6.6.12. 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 interim analyses will be performed by an independent statistician. The goals of the interim analyses are to allow early stopping of the study for futility or efficacy. The interim analysis of PFS and OS will be performed as described in Sections 5.1.1 and 5.1.2 using the methodology described in Section 6.1.1 for PFS and Section 6.1.2 for OS.

At the time of the interim analysis for PFS in PD-L1 positive patients, the first interim analysis for OS in PD-L1 positive patients will be performed.

If the interim analysis for PFS in PD-L1 positive patients crosses the pre-specified boundary for efficacy, the interim analysis for PFS for patients unselected for PD-L1 expression will be performed at the same time. If the interim analysis for PFS for patients unselected for PD-L1 expression or the first interim analysis for OS in PD-L1 positive patients cross the pre-specified boundaries for efficacy, then the first interim analysis for OS for patients unselected for PD-L1 expression will be performed at the same time.

Unblinded results from the interim analyses 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) PFS in PD-L1 positive patients analysis has crossed the pre-specified boundary for efficacy or (ii) OS in PD-L1 positive patients analysis has crossed the pre-specified boundary for efficacy or (iii) 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 final analysis for PFS in PD-L1 positive patients, the second interim analysis for OS in PD-L1 positive patients will be performed. If the final analysis for PFS in PD-L1 positive patients crosses the pre-specified boundary for efficacy, the final analysis for PFS in

patients unselected for PD-L1 expression will be performed at the same time. If the final analysis for PFS in patients unselected for PD-L1 expression or the second interim analysis for OS in PD-L1 positive patients cross the pre-specified boundaries for efficacy, then the second interim analysis for OS for patients unselected for PD-L1 expression will be performed at the same time.

The third interim analysis for OS in PD-L1 positive patients will be performed 15 months after the final analysis for PFS. If the third interim analysis for OS in PD-L1 positive patients crosses the pre-specified boundary for efficacy, the third interim analysis for OS for patients unselected for PD-L1 expression will be performed at the same time.

All patients will continue to be followed for OS until the final OS analysis (or earlier if OS in patients unselected for PD-L1 expression reaches statistical significance at an interim analysis).

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 PFS in PD-L1 positive patients

Throughout this section, references to PFS pertain to PFS by BICR assessment in PD-L1 positive patients.

Let $u(t_1)$ and $u(t_F)$ denote the upper critical boundaries based on the test statistics Z_1 and Z_F for efficacy at the interim and the final analysis, respectively, and let $l(t_1)$ and $l(t_F)$ denote the lower critical boundary for futility at the interim and final analysis, respectively. For the final analysis, $l(t_F)=u(t_F)$.

The critical values $u(t_1)$ and $l(t_1)$ for the interim analysis of PFS are determined such as

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1) \quad \text{and} \quad P_a(Z_1 \leq l(t_1)) = \beta(t_1) \ ,$$

where P_0 and P_a denote the probabilities under the null hypothesis and the alternative hypothesis, respectively, and where $\alpha(t_1)$ and $\beta(t_1)$ denote the α and β spent, respectively, based on the predefined spending functions at information fraction t_1 (t_1 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 will be calculated such that

$$\alpha(t_1) + P_0(Z_1 < u(t_1), Z_F \geq u_F) = 0.004$$

As described in Section 5.1.2, if the number of PFS events in the final analysis deviates from the target number of PFS events, the final analysis criteria will be determined as above taking into account the actual alpha spent at the interim analysis and the actual correlation between

the two test statistics Z_1 and Z_F, so that the overall 1-sided significance level across all analyses and comparisons is preserved at 0.004.

7.2.2. Interim analysis for PFS in patients unselected for PD-L1 expression

Throughout this section, references to PFS pertain to PFS by BICR assessment in patients unselected for PD-L1 expression.

The interim analysis for PFS for patients unselected for PD-L1 expression will be performed if H_{01} is rejected. No futility analysis will be performed for PFS in patients unselected for PD-L1 expression.

Let $u(t_1)$ and $u(t_F)$ denote the upper critical boundaries based on the test statistics Z_1 and Z_F for efficacy at the interim and the final analysis, respectively.

The critical values $u(t_1)$ for the interim analysis of PFS are determined such as

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1),$$

where P_0 denote the probabilities under the null hypothesis, and where $\alpha(t_1)$ denote the α spent, based on the predefined spending functions at information fraction t_1 (t_1 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 will be calculated such that

$$\alpha(t_1) + P_0(Z_1 < u(t_1), Z_F \geq u_F) = 0.004$$

As described in Section 5.1.2, if the number of PFS events in the final analysis deviates from the target number of PFS events, the final analysis criteria will be determined as above taking into account the actual alpha spent at the interim analysis and the actual correlation between the two test statistics Z_1 and Z_F, so that the overall 1-sided significance level across all analyses and comparisons is preserved at 0.004.

7.2.3. Interim analysis for OS in PD-L1 positive patients

Throughout this section, references to OS pertain to OS in PD-L1 positive patients.

Let $u(t_i)$ and $u(t_F)$ denote the upper critical boundaries based on the test statistics Z_i and Z_F for efficacy at the i^{th} interim and the final analysis, respectively, and let $l(t_i)$ and $l(t_F)$ denote the lower critical boundary for futility at the i^{th} interim and final analysis, respectively, where i=1, 2, 3. For the final analysis, $l(t_F)=u(t_F)$.

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

The critical values $u(t_1)$ and $l(t_1)$ for the 1st interim analysis of OS are determined such as

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1) \text{ and } P_a(Z_1 \leq l(t_1)) = \beta(t_1) ,$$

Critical boundaries for the additional interim analyses and the final analysis of OS are calculated recursively as follows

u(t2) is derived such that $\alpha(t_1) + P_0(Z_1 < u(t_1), Z_2 \geq u(t_2)) = \alpha(t_2)$,

l(t2) is derived such that $\beta(t_1) + P_a(Z_1 > l(t_1), Z_2 \leq l(t_2)) = \beta(t_2)$,

u(t3) is derived such that $\alpha(t_2) + P_0(Z_1 < u(t_1), Z_2 < u(t_2), Z_3 \geq u(t_3)) = \alpha(t_3)$,

l(t3) is derived such that $\beta(t_2) + P_a(Z_1 > l(t_1), Z_2 > l(t_2), Z_3 \leq l(t_3)) = \beta(t_3)$

The boundary for the final efficacy analysis is derived such that

$$\alpha(t_3) + P_0(Z_1 < u(t_1), Z_2 < u(t_2), Z_3 < u(t_3), Z_F \geq u_F) = \alpha'$$

where $\alpha' = 0.021$ for the analysis of OS in PD-L1 positive patients.

As described in Section 5.1.2, if the number of OS events in the final analysis deviates from the target number of OS events, the final analysis criteria will be determined as above taking into account the actual alpha spent at the interim analyses and the actual correlation between the three test statistics Z_i (i=1, 2, 3) and Z_F, so that the overall 1-sided significance level across all analyses and comparisons is preserved at 0.021.

7.2.4. Interim analysis for OS in patients unselected for PD-L1 expression

Throughout this section, references to OS pertain to OS in patients unselected for PD-L1 expression.

The interim analysis for OS for patients unselected for PD-L1 expression will be performed according to the testing strategy described in Section 7.1 and in Section 5.1.1. No futility analysis will be performed for OS for patients unselected for PD-L1 expression.

Let $u(t_i)$ and $u(t_F)$ denote the upper critical boundaries based on the test statistics Z_i and Z_F for efficacy at the i^{th} interim and the final analysis, respectively, where i=1, 2, 3.

In what follows P_0 denote the probability under the null hypothesis and $\alpha(t_i)$ denote the α spent based on the predefined spending functions at information fraction t_i (t_i is calculated as the ratio of the number of OS events observed at the time of the cut-off for the i^{th} interim analysis and the total number of OS events targeted for the final analysis).

The critical value $u(t_1)$ for the 1st interim analysis of OS are determined such as

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1) ,$$

[redacted]

Critical boundaries for the additional interim analyses and the final analysis of OS are calculated recursively as follows

u(t2) is derived such that $\alpha(t_1) + P_0(Z_1 < u(t_1), Z_2 \geq u(t_2)) = \alpha(t_2)$,

u(t3) is derived such that $\alpha(t_2) + P_0(Z_1 < u(t_1), Z_2 < u(t_2), Z_3 \geq u(t_3)) = \alpha(t_3)$

The boundary for the final efficacy analysis is derived such that

$$\alpha(t_3) + P_0(Z_1 < u(t_1), Z_2 < u(t_2), Z_3 < u(t_3), Z_F \geq u_F) = \alpha'$$

where α' = 0.025 or 0.021 or 0.004 for the analysis of OS in patients unselected for PD-L1 expression as determined according to the testing strategy described in Section 5.1.1.

8. REFERENCES

1. 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.

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

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

4. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy. 1990 Dec;16 (3): 199-208.

5. Heng DYC, Xie W, Regan MM, Warren MA, Golshayan AR, Sahi C, Eigl BJ, Ruether JD, Cheng T, North S, Venner P, Knox JJ, Chi KN, Kollmannsberger C, McDermott DF, Oh WK, Atkins MB, Bukowski RM, Rini BI, Choueiri TK. Prognostic Factors for Overall Survival in Patients With Metastatic Renal Cell Carcinoma Treated With Vascular Endothelial Growth Factor–Targeted Agents: Results From a Large, Multicenter Study. Journal of Clinical Oncology. 27: 5794-5799, 2009.

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

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

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

9. Motzer RJ, Bacik J, Murphy BA, Russo P, Mazumdar M. Interferon-Alfa as a comparative treatment for clinical interferon-alfa as a comparative treatment for clinical cell carcinoma. Journal of Clinical Oncology. 20: 289-296, 2002.

10. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med. 2001 Jul; 33 (5): 337-43. Review.

11. Rao D1, Butt Z, Rosenbloom S, Robinson D Jr, Von Roenn J, Kuzel TM, Cella D. A. Comparison of the Renal Cell Carcinoma-Symptom Index (RCC-SI) and the Functional Assessment of Cancer Therapy-Kidney Symptom Index (FKSI). J Pain Symptom Manage. 2009 Aug;38(2):291-8doi:10.1016/j.jpainsymman.2008.08.013. Epub 2009

12. 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. Statistics in medicine, 2011 30 (19),pp. 2409-2421

13. Santner TJ, Snell MK. (1980). 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

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

15. Sutent Summary of Product Characteristics, Pfizer Inc. http://www.ema.europa.eu/docs/en_GB/document_library/EPAR-Product_Information/human/000687/WC500057737.pdf.

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

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

18. 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 18. This case definition is hierarchical, i.e., each step is only checked for patients and events that have already met the prior step.

Table 18. 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) OR C) below are met then the AE is in for the next step A) conmed ATC code is in (H02A, H02B, D07, A01AC, S01BA, S01BB, L04AA, L04AB, L04AC, L04AD, L04AX, A07EA) 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" C) conmed ATC code is A10A and AE PT is in the irAE cluster associated with "Immune-mediated endocrinopathies: Type I Diabetes Mellitus"

5	A) No clear etiology (other than immune mediated etiology) B) Histopathology / biopsy consistent with immune-mediated event 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]	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) 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)

The data set associated with irAEs may be refined based on medical review. The final data set including any changes based on medical review (e.g., 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 avelumab 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 avelumab infusion (during or after the infusion) or the day after the avelumab 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 19 and will be identified for IV drugs only.

Table 19. Case Definition for IRRs – IV Study Drugs Administered Alone Or In Combination With Non-IV Study Drugs

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