JAVELIN BLADDER 100
A PHASE 3, MULTICENTER, MULTINATIONAL, RANDOMIZED, OPEN-LABEL, PARALLEL-ARM STUDY OF AVELUMAB (MSB0010718C) PLUS BEST SUPPORTIVE CARE VERSUS BEST SUPPORTIVE CARE ALONE AS A MAINTENANCE TREATMENT IN PATIENTS WITH LOCALLY ADVANCED OR METASTATIC UROTHELIAL CANCER WHOSE DISEASE DID NOT PROGRESS AFTER COMPLETION OF FIRST-LINE PLATINUM-CONTAINING CHEMOTHERAPY

STATISTICAL ANALYSIS PLAN – B9991001

Compounds:	MSB0010718C
Compound Name:	Avelumab
Version:	2.0
Date:	02-Apr-2019

[redacted]

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDICES

1. VERSION HISTORY

This Statistical Analysis Plan (SAP) for study B9991001 is based on the protocol amendment 4 dated 28-Mar-2019.

Table 1. Summary of Major Changes in SAP Amendments

Version	Version Date	Summary of Changes
1	23-Dec-2015	Not applicable (N/A)
2	02-Apr-2019	Section 2.1 "Study Objectives", Section 3.3 "Exploratory Endpoints", Section 6.3 "Other Endpoints" updated to remove irRECIST objectives and endpoints as per protocol amendment 4. Section 2.2 "Study Design" was updated to reflect changes made in protocol amendment 3. Section 3.2.2 "Efficacy endpoints" – corrected the definition of disease control to include best overall response (BOR) of non-CR/non-PD to take into account patients without measurable disease at baseline; clarified the criterion for SD must have been met at least 6 weeks after the date of randomization as per protocol. Section 3.4.1 "Study drug, study treatment and baseline definitions" – added definition of end date of study treatment; added definition of baseline for immunogenicity analyses; removed definition of baseline for ECG assessments as no summaries will be provided (ECGs are only collected post-treatment if clinically indicated). Section 3.4.2 "Baseline characteristics" – updated the stratification factor from "Best response to first-line chemotherapy per BICR assessment" to "Best response to first-line chemotherapy" to reflect changes made in protocol amendment 3. Section 3.5.1 "Adverse events" – updated the definition of treatment-emergent adverse events; clarified that start of new anti-cancer drug therapy is used in the definition of on-treatment period. Section 4.2 "Safety Analysis Set" – clarified that patients randomized to Arm B that incorrectly received avelumab will be classified in Arm A for the purpose of safety analyses. Section 4.3.1 "Per-protocol analysis set" – updated the definition of PP analysis set to align with the key inclusion/exclusion criteria. Section 5.1.2 "Decision rules" – added description of primary analysis of OS; updated Table 6 to reflect correct simulated operational characteristics of the study design and added a footnote to clarify that the actual interim and final analyses will occur at the same calendar time for both patient populations. Section 5.2.5 "Definition of start of new anti-cancer drug therapy", Section 5.2.6 "Definition of start of new anti-cancer therapy", Section 5.2.10 "Adequate baseline tumor assessment" and Section 5.2.11 "Adequate post-baseline tumor assessment" were added to provide the details associated with derivations used to define the on-treatment period and to implement censoring in efficacy analyses. Section 5.2.7 "Definition of on-treatment period" – clarified that start of new anti-cancer drug therapy is used in the definition of on-treatment period. Section 5.3.2.4 "Exposure" – clarified that in imputation derivations the cutoff date is used as last dosing date in the analyses of patients that have study drug ongoing at the time of the cutoff date. Section 5.3.3.1 "Date of last contact" – added withdrawal of consent date in the

		derivation of date of last contact. Section 5.3.3.2 "Death date" imputations – deleted the last step in the imputation "if the day is missing from the date of last contact it will be imputed to 1st day of the month and year of last contact only if derived from the 'Survival Follow-up' eCRF page" since not needed. Section 5.3.3.4 "Date of start of new anti-cancer therapy" was added to provide details of imputation rules for incomplete dates for start date of new anti-cancer therapy (drug therapy, radiation, surgery). Section 5.3.3.5 "PRO data" was added to provide details of imputation rules for missing PRO data. Section 6.1.1 "Overall Survival" – added that RCIs will be calculated for the analyses of OS; provided more details and hierarchy associated with the derivation of reasons for censoring; provided more details on the analysis of time of follow-up for OS. Section 6.2.2 "Efficacy Endpoints" – clarified that the efficacy endpoints will be summarized for both co-primary populations; provided more details on analysis of PFS and Time of Follow-Up for PFS; provided more details and hierarchy associated with the derivation of reasons for censoring for PFS; added details on the assessment of concordance for PFS between investigator and BICR assessments. Section 6.2.2.2 "Sensitivity analyses for overall survival" – deleted the sensitivity analysis based on CRF-derived randomization stratification factors since the data collected in the eCRF does not enable unequivocal derivation of the stratification factors; provided additional details for checking the proportional hazard assumption and added further details regarding implementation of restricted mean survival (RMST) methodology. Section 6.2.2.3 "Objective response and tumor shrinkage" – added definition of BOR of non-CR/non-PD for patients without measurable disease at baseline; added analysis to assess association of study treatment and objective response; added summary for BOR discrepancy between BICR and Investigator; provided further granularity for reasons for not evaluable BOR. Section 6.2.2.4 "Disease Control" – added non-CR/non-PD in the definition for disease control to take into consideration the patients without measurable disease at baseline. Removed the original planned analysis for DCR at 24 weeks since more adequately summarized based on Kaplan-Meier rates for PFS. Section 6.2.2.5 "Duration of response" – added details regarding censoring for duration of response since "no adequate baseline assessment" which is used in censoring for PFS analyses is not applicable to analyses of duration of response for patients with objective response. Section 6.2.3 "Pharmacokinetic endpoints" – provided additional details for the analysis of Ctrough, Cmax and plasma concentrations. Section 6.2.5 "Biomarker endpoints" – exploratory biomarker endpoints were updated to be consistent with protocol amendment 3. Section 6.2.6 "Endpoints for Immunogenicity Data of Avelumab" – added details for the analysis of ADA and nAb. Section 6.2.7 "PRO endpoints" and subsections – added details for the analysis of PRO data. Section 6.4 "Subset Analyses" – Added clarification that the subset analyses will be performed for both the co-primary populations; removed the subset analysis based on "response to first-line chemotherapy as assessed by BICR" to reflect the updates in protocol amendment 3; added subset analysis based on baseline liver lesions and baseline lung lesions; updated the p-value for the interaction test to

		be based on Wald's test; removed forest plots for median DR by subgroup. Section 6.5.1 "Baseline summaries" – Added clarification that the baseline summaries will be tabulated for both co-primary populations; updated geographic region classifications. Section 6.5.1.3 "Disease characteristics" – removed some analysis to align with data collection as per updates made to eCRF. Section 6.5.1.4 "Prior anti-cancer therapies" – clarified that the prior anti-cancer therapy information will be included in the listing of anti-cancer therapies with a flag to identify prior therapies. Section 6.5.1.5 "Exposure to first-line anti-cancer chemotherapy regimen" – removed summary for number of cycles, dose intensity and relative dose intensity since prior to randomization. Section 6.5.2.1 "Patient disposition" – removed summaries based on CRF-derived randomization stratification factors since the data collected in the eCRF does not enable unequivocal derivation of the stratification factors; removed the summary of reason for randomization not equal to treatment assignment since data collected in eCRF does not enable such analysis. Section 6.5.3 "Study Treatment Compliance and Exposure" – added additional details for summary of exposure to avelumab; removed by-cycle summaries; added details for derivation and summary of dose delays. Section 6.5.3.5 "Infusion interruptions" was added. Section 6.5.5 "Subsequent anti-cancer therapies" – added summary of follow-up anti-cancer drug therapies for patients with PD-L1-positive tumors. Section 6.6.1 "Adverse events" and subsections – updated definition of treatment-emergent adverse event. Given the nature of immune-related AEs (irAEs) and infusion related reactions (IRRs), removed the summaries of treatment-related irAEs and treatment-related IRRs; removed the requirement that all summaries by SOC and PT will be replicated by PT only. Added summaries of AEs leading to dose reduction, AEs leading to interruption of study treatment and AEs leading to both dose reduction and interruption of study treatment. Section 6.6.1.2 "Adverse events leading to dose reduction" and Section 6.6.1.3 "Adverse events leading to interruption of study treatment" were added. Section 6.6.2 "Death" – deleted references to primary reason for death since not collected in the eCRF. Section 6.6.4 "Other significant adverse events" – given the nature of the study and BSC as comparator arm, nominal p-values for risk difference for Tier-1 events will not be calculated. Section 6.6.5 "Laboratory data" and subsections – Changed the formula for corrected calcium to be expressed in SI units; clarified that the denominator to calculate percentages for each laboratory parameter is the number of patients evaluable for the particular summary; added summary for patients with newly occurring or worsening laboratory abnormities; updated lab parameters that will be summarized in shift tables. Section 6.6.7 "Electrocardiogram" – removed the analyses of ECGs that are not applicable to this study as ECGs are only collected when clinically indicated. Section for "Physical Examination" was removed as data not collected in the eCRF. Section 7 "INTERIM ANLAYSES" – added details for the calculation of the boundary for final efficacy analysis. Section 8 "REFERENCES" – references updated. Section 9 "APPENDICES" –added appendices with the irAE and IRR derivation

		rules; deleted the tabulation of cycle derivations since included with more details in the exposure section. Minor editorial and consistency changes throughout the document.

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study B9991001. 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.

A separate SAP will cover the interim analyses for periodic safety review by the External Data Monitoring Committee (E-DMC).

Statistical analyses will be performed using cleaned eCRF data as well as non-CRF data (ie, pharmacokinetic [PK] concentration, anti-drug antibodies (ADA; neutralizing antibody [nAb]), supplemental biomarker data, and tumor assessment results by the Blinded Independent Central Review (BICR)).

The primary analysis will include all data up to a cut-off date which is determined by the number of events (deaths) required for the analysis of Overall Survival (OS) and a minimum follow-up of 12 months after the last patient is randomized. The cut-off date is determined once a data extract (before database lock) is available which indicates that the required number of events for OS in both co-primary populations and the minimum follow-up of 12 months is expected to occur by the cut-off date.

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

2.1. Study Objectives

Primary Objectives

- To demonstrate the benefit of maintenance treatment with avelumab plus best supportive care (BSC) vs. BSC alone in prolonging OS in patients with unresectable locally advanced or metastatic Urothelial Cancer (UC) whose disease did not progress on or following completion of first-line platinum-containing chemotherapy in each co-primary UC patient population: 1) patients determined to have PD-L1-positive tumors (including infiltrating immune cells) by a verified Good Manufacturing Practice (GMP) PD-L1 immunohistochemistry (IHC) test, and 2) all randomized patients.

Secondary Objectives

- To compare the PFS of avelumab plus BSC vs. BSC alone in patients determined to have PD-L1-positive tumors (including infiltrating immune cells) by a verified GMP PD-L1 IHC test, and in all randomized patients.

- To evaluate the anti-tumor activity of avelumab plus BSC and BSC alone according to Response Evaluation Criteria in Solid Tumors (RECIST) v1.1 in patients determined to

have PD-L1-positive tumors (including infiltrating immune cells) by a verified GMP PD-L1 immunohistochemistry (IHC) test, and in all randomized patients.

- To evaluate the overall safety profile of avelumab plus BSC and BSC alone.
- To evaluate the PK of avelumab in each of the co-primary UC patient populations treated with avelumab.
- To assess the immunogenicity of avelumab in each of the co-primary UC patient populations treated with avelumab.
- To evaluate candidate predictive biomarkers of sensitivity or resistance to avelumab in pre-treatment tumor tissue in each of the co-primary UC patient populations treated with avelumab.
- To evaluate the effect of avelumab plus BSC and BSC alone on patient-reported outcomes (PROs) in each of the co-primary UC patient populations.

Exploratory Objectives

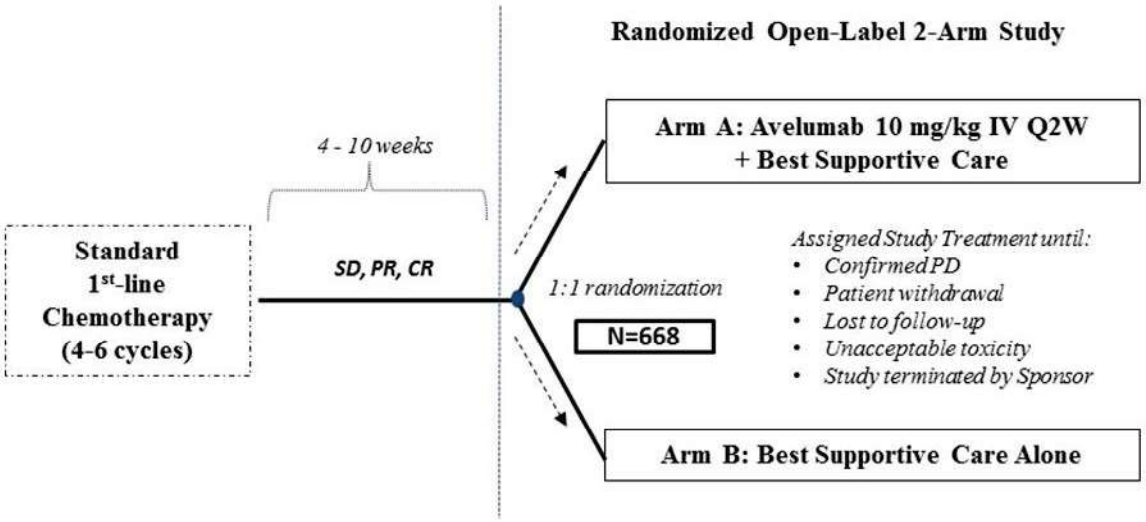
- To explore the predictive and/or pharmacodynamic (PD) characteristics of peripheral blood and additional tumor tissue biomarkers relevant to the mechanism of action of or resistance to avelumab.

2.2. Study Design

This is a Phase 3, multicenter, multinational, randomized, open-label, parallel-arm efficacy and safety study of avelumab plus BSC compared to BSC alone as a maintenance treatment after completion of first-line platinum-based chemotherapy (gemcitabine + cisplatin or gemcitabine + carboplatin) without evidence of disease progression in adult patients with unresectable locally advanced or metastatic UC.

The study design is illustrated in Figure 1.

Figure 1. Study Design Schema



a. Allowed first-line chemotherapy regimens are gemcitabine + cisplatin or gemcitabine + carboplatin.
b. Randomization must occur at least 4 and not more than 10 weeks after the last dose of first-line chemotherapy and will be stratified by: best response on 1st-line therapy (CR or PR vs. SD) and metastatic disease site (visceral vs. non-visceral).

CR = complete response; IV = intravenous; PD = progressive disease; PR = partial response; Q2W = every 2 weeks; SD = stable disease

- A total of approximately 668 patients without progressive disease (PD) as per RECIST v1.1 guidelines (ie, with ongoing CR, PR, or SD) after 1st line chemotherapy will be randomized in this study. It is estimated that at least 334 patients with confirmed PD-L1-positive tumors (including infiltrating immune cells) will be randomized in this study.

- Patients will be randomized in a 1:1 ratio into two study arms: avelumab + BSC (Arm A) or BSC alone (Arm B).

- Randomization will be stratified by (i) best response to first-line chemotherapy (CR/PR vs. SD), and (ii) metastatic disease site (visceral vs. non-visceral) at the time of initiating first-line chemotherapy.

- This study is designed with two co-primary populations: 1) patients determined to have PD-L1-positive tumors (including infiltrating immune cells) by a verified GMP PD-L1 IHC test, and 2) all randomized patients.

- Patients must have received at least 4 cycles, but not more than 6 cycles, of a first-line chemotherapy regimen consisting of either gemcitabine + cisplatin or gemcitabine + carboplatin before randomization into this study. No other chemotherapy regimen is allowed as the first line chemotherapy for inclusion in this clinical trial (please see Section 4.1 of the protocol, Inclusion Criteria 2).

- Randomization must occur at least 4 but not more than 10 weeks after the date of administration of the last dose of chemotherapy. Patients will initiate study treatment (Cycle 1 Day 1) within 3 days after randomization.

- Only patients without progressive disease as per RECIST v1.1 guidelines (ie, with ongoing CR, PR, or SD) after 4-6 cycles of chemotherapy will be allowed to be randomized in this study.

  - Post-chemotherapy confirmatory scan(s) (CT/MRI) for eligibility must be performed within 28 days prior to the date of randomization to assess response status following first-line chemotherapy.

  - Based on the post-chemotherapy confirmatory scan(s), the investigator should assess patient eligibility (ie, CR, PR, or SD) before randomization.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoint

- Overall Survival (OS).

  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) and laboratory abnormalities as graded by National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) v4.03; vital signs (blood pressure, pulse rate).

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

3.2.2. Efficacy endpoints

- Progression-free survival (PFS) based on BICR assessment per RECIST v1.1.

- Investigator-assessed PFS.

  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.

- Objective Response (OR), Time to Tumor Response (TTR), Duration of Response (DR), and Disease Control (DC), as assessed per RECIST v1.1 by BICR and investigator.

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

DR is defined, for patients with 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.

DC is defined as CR, PR, non-CR/non-PD or stable disease (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.

3.2.3. Patient Reported Outcomes endpoints

- Patient-reported bladder cancer symptom, functioning, global quality of life (QOL), and Time to Deterioration (TTD) using the NCCN-FACT FBlSI-18; and health status using the EQ-5D-5L.

The NCCN FACT FBlSI-18 (a later version of the FACT-Bl) was developed to be part of the Functional Assessment of Chronic Illness Therapy (FACIT) system and was specifically created with input from the Food and Drug Administration (FDA) and validated in bladder cancer patients. The NCCN FACT FBlSI-18 is designed to be a stand-alone instrument to measure symptoms and QoL in patients with UC and was created using inputs from patients and oncologists. The 'Disease Related Symptoms-Physical' subscale of the NCCN-FACT FBlSI-18 (FBlSI-DRS-P) uses a subset of physical symptoms which are considered to be specific to UC.

The EuroQol EQ-5D-5L is a 6-item patient-completed questionnaire designed to assess health status in terms of a single index value or utility score. There are 2 components to the EuroQol EQ-5D-5L, a Health State Profile which has individuals 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. Overall scores range from 0 to 1, with low scores representing a higher level of dysfunction.

3.2.4. Pharmacokinetic endpoints

- Maximum concentrations (C_{max}) and trough concentrations (C_{trough}) for avelumab.

C_{max} and C_{trough} are calculated after single and multiple dose and at steady-state.

Table 2. PK Parameters to be Determined for Avelumab

Parameter	Definition	Method of Determination
C_{max}	Maximum observed plasma concentration	Observed directly from data
C_{trough}	Pre-dose concentration during multiple dosing	Observed directly from data

3.2.5. Immunogenicity endpoints

- Incidence of anti-drug antibodies (ADA, nAb) against avelumab.

3.2.6. Biomarker endpoints

- Tumor tissue biomarkers including, but not limited to, PD-L1 expression and tumor-infiltrating CD8+ T lymphocytes.

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 inflammatory cells	Pathologist, assisted by image analysis
Tumor infiltrating CD8+ lymphocytes	The number of CD8+ cells per unit area and the percent of counted cells that are scored as CD8+	Pathologist, assisted by image analysis

3.3. Exploratory Endpoints

- *Biomarkers:* Peripheral blood and additional tumor tissue biomarkers consisting of the levels of cells, deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or proteins that may be related to anti-tumor immune response and/or response to or disease progression on avelumab, such as genes related to IFN-γ or transforming growth factor (TGF)-β.

3.4. Baseline Variables

3.4.1. Study drug, study treatment and baseline definitions

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

- Arm A = Avelumab + BSC;
- Arm B = BSC alone.

Start and end dates of study treatment:

- For Arm A:
  - The date/time of first dose of study treatment is the earliest date/time of non-zero dosing of the study drug.
  - The date/time of last dose of study treatment is the latest date/time of non-zero dosing of the study drug.

- For Arm B:
  - The date of start of study treatment is defined as the date of Cycle 1 Day 1 (C1D1); a patient is considered to be treated in Arm B if they complete the C1D1 visit.
  - 'End date of BSC' (last dose of study treatment) is the date of completion or discontinuation as collected in the 'Disposition - Treatment Phase' eCRF page.

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 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 (Arm A only)

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.

Definition of baseline for biomarker analyses

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

3.4.2. Baseline Characteristics

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

- Best response to first-line chemotherapy (CR or PR vs. SD)
- Metastatic disease site (visceral vs. non-visceral) at the time of initiating first-line chemotherapy. The 'non-visceral' stratum includes patients with locally-advanced disease as well as patients with only non-visceral disease.

The primary analyses of OS 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.

On-treatment period is defined as follows:

- Arm A: 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)
- Arm B: time from Cycle 1 Day 1 through minimum (30 days + end date of BSC, start day of new anti-cancer drug therapy – 1 day).

As defined in Section 3.4.1, 'End date of BSC' is the date of completion or discontinuation as collected in the 'Disposition - Treatment Phase' eCRF page.

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 Appendix 1 and Appendix 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	✓		
Efficacy: Exploratory	✓		
Safety			✓

4.1. Full Analysis Set

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

The FAS will be the primary analysis set for evaluating all efficacy endpoints, PRO endpoints, and patient characteristics within each of the co-primary populations:

1) all randomized patients, and

2) patients determined to have PD-L1-positive tumors (including infiltrating immune cells) by a verified GMP PD-L1 IHC test; these patients will be identified by a retrospective analysis of the mandatory recent FFPE tumor biospecimen (see Sections 4.1 and 6.1.1 of the study protocol) using an analytically-validated and GMP-verified PD-L1 IHC assay with a pre-specified scoring algorithm defining PD-L1-positive status. The PD-L1 IHC assay and scoring algorithm will be fully defined prior to initiation of sample analysis and will be documented separately.

4.2. Safety Analysis Set

The safety analysis set will include all patients who receive at least one dose of study drug on Arm A or completed C1D1 on Arm B. If a patient randomized to Arm B incorrectly receives avelumab, then the patient will be classified in Arm A for the purpose of safety analyses.

4.3. Other Analysis Set

4.3.1. Per-protocol analysis set

The 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.

- Patients randomized to Arm A who did not receive at least one dose of avelumab
- Patients randomized to Arm B who were treated with avelumab

- Baseline ECOG status ≥ 2
- Patient does not meet inclusion criteria 1, 2, or 3 and patient meets exclusion criterion 1

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.

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

4.3.3. Biomarker analysis set

The biomarker analysis set is a subset of the safety analysis set and will include patients who have at least one baseline biomarker assessment performed.

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.

5. GENERAL METHODOLOGY AND CONVENTIONS

5.1. Hypotheses and Decision Rules

5.1.1. Hypotheses and sample size determination

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

$$H_{0\,\text{All patients}}: \text{HR}_{\text{OS (Arm A/ArmB)}} \geq 1, \text{ versus } H_{a\,\text{All patients}}: \text{HR}_{\text{OS (Arm A/Arm B)}} < 1$$

$$H_{0\,\text{PD-L1+}}: \text{HR}_{\text{OS (Arm A/Arm B)}} \geq 1, \text{ versus } H_{a\,\text{PD-L1+}}: \text{HR}_{\text{OS (Arm A/Arm B)}} < 1$$

where HR_{OS} represents the hazard ratio for OS in the experimental arm vs the control arm.

Approximately 668 patients will be randomized to the treatment arms using a 1:1 randomization, stratified by best response to chemotherapy (CR or PR vs. SD) and site of metastasis (visceral vs. non-visceral) at time of initiating first-line chemotherapy. The type 1 error rate will be split between the co-primary populations to preserve the overall type 1 error rate for the study at 0.025 one-sided. It is estimated that at least 50% of the randomized patients will have adequate tissue for analysis and be determined to have PD-L1-positive tumors.

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

- For all patients and patients with PD-L1-positive tumors receiving BSC alone after first-line chemotherapy, the median OS is 12 months (Powles et al, 2015).
- For all patients receiving avelumab + BSC after first-line chemotherapy, the median OS is assumed to be 17.1 months.

- For patients with PD-L1-positive tumors receiving avelumab + BSC after first-line chemotherapy, the median OS is assumed to be 18.5 months.

This corresponds to a hazard ratio (HR) of 0.7 for all patients and 0.65 for patients with PD-L1-positive tumors under the exponential model assumption.

For all patients, if the true HR is 0.7 under the alternative hypothesis, a total of 425 OS events will be required to have 93% power to detect a HR of 0.7 using a one-sided log rank test at a significance level of 0.015 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 (-8) β-spending function to determine the non-binding futility boundary.

For patients with PD-L1-positive tumors, if the true HR is 0.65 under the alternative hypothesis, then a total of 219 OS events will be required to have 80% power to detect a HR of 0.65 using a one-sided log rank test at a significance level of 0.01 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 (-8) β-spending function to determine the non-binding futility boundary.

The sample size further assumes a 5% drop-out rate for OS on either treatment arm, a non-uniform patient accrual accomplished over a 28-month period. The data cut-off for the primary OS analysis will occur after the target number of events has been reached in both co-primary populations and the last patient randomized in the study has been followed for at least 12 months after randomization.

5.1.2. Decision rules

To protect the integrity of the study and to preserve the type I error rate, a fraction of alpha for efficacy will be spent at the interim analysis and accounted for in the overall type I error rate (if the interim analysis is performed exactly at the planned number of OS events, alpha spent at interim is 0.005 for all patients and 0.002 for patients with PD-L1-positive tumors). The significance levels for the interim and final efficacy analyses of OS will be determined by using the Lan-DeMets procedure with an O'Brien-Fleming stopping boundary. The overall significance level for the efficacy analysis of OS will be preserved at (one-sided) 0.015 for all patients and 0.01 for patients with PD-L1-positive tumors.

Two analyses will be performed for OS:

The interim analysis of OS will be performed after at least 315 of all randomized patients have died (74% of the target number of OS events for the 'all patients' population), and all of the following conditions have been met:

1) all initially planned patients have been randomized, and

2) at least 146 patients with PD-L1-positive tumors have died (approximately 66.7% of the total OS events expected in the 'patients with PD-L1-positive tumors' population).

The primary analysis of OS will be performed after at least 425 of all randomized patients and at least 219 patients with PD-L1-positive tumors have died, and the last patient randomized in the study has been followed for at least 12 months after randomization.

The study will have met its primary objective if the experimental arm is statistically significantly superior to the control arm for either one of the co-primary populations, either at the time of the interim analysis or at the time of the final analysis.

The efficacy and futility boundaries for the co-primary populations with the planned number of events at interim look are listed below in Table 5. The boundary values will be updated using the pre-specified α- and β-spending functions with the observed number of events at the time of the interim analysis. If the value of the test statistic for OS exceeds the efficacy boundary, then the experimental arm may be declared statistically significantly superior to the control arm (in the 'all patients' or 'patients with PD-L1-positive tumors' populations). If the value of the test statistic for OS exceeds the futility boundary, the study may be stopped for futility. Alternatively, as appropriate, the sample size of the study may be adjusted using the method outlined by Cui et al (1999). If the results of the interim analysis indicate serious safety concerns, the sponsor, in conjunction with the External Data Monitoring Committee (E-DMC), will communicate with the Health Authorities regarding stopping the clinical trial.

Table 5. Stopping Boundaries for Overall Survival

Population	Number of Events	Efficacy		Futility	
		Z scale	p-value (one-sided)	Z scale	p-value (one-sided)
All patients	315	$z < -2.595$	$p < 0.005$	$z > -0.789$	$p > 0.215$
Patients with PD-L1-positive tumors	146	$z < -2.947$	$p < 0.002$	$z > -0.397$	$p > 0.346$

Since the observed number of events at the interim analysis for each comparison may not be exactly equal to the planned number of OS events for that comparison, the efficacy and futility boundaries will be updated based on the actual number of observed events using the pre-specified α-and β-spending functions. Therefore, the observed Z-test statistic at the interim analysis will be compared with the updated efficacy and futility boundaries. If the study continues to final analysis, the p-value that will be used to declare statistical significance at the final analysis will be based on the actual number of OS events documented at the cut-off date for the final analysis and the α already spent at the interim analysis. Therefore, if the interim analysis occurs exactly after 315 events for all patients and 146 events for patients with PD-L1-positive tumors, and the study continues until the final analysis, the observed p-value for the comparison will have to be < 0.014 for all patients and <0.009 for patients with PD-L1-positive tumors to declare statistical significance for that population. If the number of events in the final analysis deviates from the expected number of events, the final analysis criterion will be determined so that the overall significance level across all analyses is maintained at one-sided 0.025.

Based on the stopping boundaries, the timing of interim analysis and information fractions defined above the design has the following operating characteristics.

Table 6. Simulated cumulative probabilities to stop for futility at the interim or final OS analysis

Scenario	Look	Number of OS events	Calendar Time (months)	P(Reject H_{0i})	P(Reject H_{ai})
$H_{0\ \text{All patients}}$ is true (HR=1)	Interim	315	29	0.0051	0.7817
	Final	425	36	0.0163	0.9837
$H_{a\ \text{All patients}}$ is true (HR=0.7)	Interim	315	31	0.7039	0.0098
	Final	425	39	0.9257	0.0743
Simulations performed in EAST 6 with number of simulations = 10,000 and seed (H_0, H_a) = 3637647, 15346244					
$H_{0\ \text{PD-L1+}}$ is true (HR=1)	Interim	146	28	0.0008	0.6576
	Final	219	37	0.0100	0.9900
$H_{a\ \text{PD-L1+}}$ is true (HR=0.65)	Interim	146	30	0.3670	0.0136
	Final	219	41	0.8066	0.1934
Simulations performed in EAST 6 with number of simulations = 10,000 and seed (H_0, H_a) = 15459942, 15509960.					
The actual Interim Analysis and Final Analysis will occur at the same calendar time for both patient populations corresponding to the latest of achieving the target number of events in each population.					

5.2. General Methods

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

- Arm A = Avelumab + BSC;
- Arm B = BSC alone.

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

5.2.1. Data handling after the cut-off date

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

5.2.2. Pooling of centers

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

5.2.3. Presentation of continuous and qualitative variables

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

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

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

5.2.4. Definition of study day

Start day of study treatment is the day of the first dose of study treatment in Arm A, and the day of Cycle 1 Day 1 in Arm B.

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

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

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

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

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

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

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

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

5.2.6. Definition of start of new anti-cancer therapy

Start date of new anti-cancer therapy (drug, radiation, surgery) is used for censoring in efficacy analyses (see Section 6.1.1 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 follows:

- Arm A: 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);
- Arm B: time from Cycle 1 Day 1 through minimum (30 days + end date of BSC, start day of new anti-cancer drug therapy – 1 day).

Here, 'end date of BSC' is the date of completion or discontinuation as collected in the 'Disposition - Treatment Phase' eCRF page.

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

5.2.9. Unscheduled visits

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

5.2.10. Adequate baseline tumor assessment

Adequate baseline is defined using the following criteria:

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

5.2.11. Adequate post-baseline tumor assessment

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

5.3. Methods to Manage Missing Data

5.3.1. Missing data

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

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

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

5.3.1.1. Pharmacokinetic concentrations

Concentrations Below the Limit of Quantification

For all calculations, figures and estimation of individual pharmacokinetic parameters, all concentrations assayed as below the level of quantification (BLQ) will be set to zero. The BLQ values will be excluded from calculations of geometric means and their confidence intervals (CIs). A statement similar to 'All values reported as BLQ have been replaced with zero' should be included as a footnote to the appropriate tables and figures.

Deviations, Missing Concentrations and Anomalous Values

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

1. A concentration has been reported as ND (ie, not done) or NS (ie, no sample);
2. A deviation in sampling time is of sufficient concern or a concentration has been flagged as anomalous by the clinical pharmacologist.

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

5.3.1.2. Pharmacokinetic parameters

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

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

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

5.3.2. Handling of incomplete dates

5.3.2.1. Disease history

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

- If the day is missing, it will be imputed to the 15th day of the month.
- If both day and month are missing and the year is prior to the year of the first study treatment, the month and day will be imputed as July 1st.
- If both day and month are missing and the year is same as the year of the first study treatment, the month and day will be imputed as January 1st.
- If the date is completely missing, no imputation will be performed.

5.3.2.2. Adverse events

Incomplete AE-related dates will be imputed as follows:

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

5.3.2.3. Prior and concomitant medications

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

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

treatment start date is 19/NOV/2014, then the imputed medication start date will be 19/NOV/2014.

- In all other cases the missing medication day or missing medication month will be replaced by 1.
- Incomplete stop date will be replaced by the last day of the month (if day is missing only), if not resulting in a date later than the date of patient's death. In the latter case the date of death will be used to impute the incomplete stop date.
- In all other cases the incomplete medication stop date will not be imputed.

5.3.2.4. Exposure

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

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

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

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

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

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

5.3.3.1. Date of last contact

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

- All patient assessment dates (blood draws (laboratory, PK), vital signs, performance status, ECG, tumor assessments)
- Start and end dates of anti-cancer therapies administered after study treatment discontinuation
- AE start and end dates
- Last date of contact collected on the 'Survival Follow-up' eCRF (do not use date of survival follow-up assessment unless status is 'alive')
- Study drug start and end dates
- Randomization date

- Withdrawal of consent date
- Date of discontinuation on disposition eCRF pages (do not use if reason for discontinuation is lost to follow-up).

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

5.3.3.2. Death date

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

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

5.3.3.3. Tumor assessments

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

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

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

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

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

5.3.3.4. Date of start of new anti-cancer therapy

Incomplete dates for start date of new anti-cancer therapy (drug therapy, radiation, surgery) will be imputed as follows and will be used for determining censoring dates for efficacy

analyses and in the derivation of the end of on-treatment period. PD date below refers to PD date by investigator assessment.

- The end date of new anti-cancer therapy will be included in the imputations for start date of new anti-cancer therapy. If the end date of new anti-cancer therapy is
  - completely missing then it will be ignored in the imputations below
  - partially missing with only year (YYYY) available then the imputations below will consider 31DECYYYY as the end date of the new anti-cancer therapy
  - partially missing with only month and year available then the imputations below will consider the last day of the month for MMMYYYY as the end date of the new anti-cancer therapy
- For patients who have not discontinued study treatment at the analysis cut-off date, last dose of study treatment is set to the analysis cut-off date in the imputations below.
- If the start date of new anti-cancer therapy is completely or partially missing, then the imputed start date of new anti-cancer therapy is derived as follows:
  - Start date of new anti-cancer therapy is completely missing

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

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

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

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

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

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

    IF

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

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

    THEN

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

    ELSE IF

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

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

THEN

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

ELSE IF

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

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

THEN

imputed start date = 01 MMM YYYY;

ELSE IF

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

THEN

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

ELSE IF

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

THEN

imputed start date = 01 MMM YYYY.

5.3.3.5. PRO data

If there are missing items on the FBISI-18, subscale scores can be prorated. This is done by multiplying the sum of the subscale by the number of items in the subscale, then dividing by the number of items actually answered, as follows:

Prorated subscale score = [Sum of item scores] × [Number of items in subscale] ÷ [Number of items answered]

When there are missing data, prorating by subscale in this way is acceptable if > 50% of the items were answered. The total score is then calculated as the sum of the un-weighted subscale scores. In addition, a total score should only be calculated if all of the component subscales have valid scores.

For the EQ-5D-5L, questions not answered will be considered missing items and will not be

utilized. For EQ-5D-5L, the entire utility score for that cycle is deemed missing if the answer to any one of the 5 dimensions is missing.

6. ANALYSES AND SUMMARIES

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

6.1. Primary Endpoint

6.1.1. Overall Survival

6.1.1.1. Primary Analysis

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

Overall survival (OS) is defined as the time from 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 within each of the co-primary populations, and will be performed using a one-sided stratified log-rank test as described in Section 5.1.

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

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

In order to account for the group sequential design in this study, the repeated CI (RCI) method (Jennison and Turnbull, 2000), will be used to construct the 2-sided RCIs for the hazard ratio at the interim and the final analyses of 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 two-sided 95% CIs. In particular, the OS rates at 6, 12, 18, 24 and 30 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice (2002) (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 7 following the hierarchy shown.

Table 7. 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) = Patient 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 rates 6, 12, 18, 24 and 30 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 based on BICR assessment and based on Investigator assessment.

Efficacy endpoints defined below will be summarized separately for each of the co-primary populations.

6.2.2.1. Progression Free Survival

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 (ie ≤ 16 weeks after the date of randomization) in which case the death will be considered an event.

In this study antitumor activity will be assessed through radiological tumor assessments conducted at screening, at 8 weeks after randomization, then every 8 (±3 days) weeks for up to 1 year from randomization, and every 12 weeks (±3 days) thereafter until PD as assessed by BICR regardless of initiation of subsequent anti-cancer therapy.

The censoring and event date options to be considered for the PFS and DR analysis are presented in Table 8.

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

Table 8. Outcome and event dates for PFS analyses

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

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

[b] If there are no adequate post-baseline assessments prior to PD or death, then the time without adequate

Scenario	Date of event/censoring	Outcome
assessment should be measured from the date of randomization; if the criteria were met the censoring will be on the date of randomization.		

A stratified log-rank test (1-sided) will be used to compare the PFS time based on BICR assessment between the experimental arm and the control arm.

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

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

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 rates at 3, 6, 9, 12 and 15 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice (2002) (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 9 following the hierarchy shown.

Table 9. 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) = Patient 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 rates at 3, 6, 9, 12 and 15 months will be estimated with corresponding 2-sided 95% CIs.

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

Table 10. 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).

6.2.2.2. Sensitivity analyses for overall survival

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

- PP analysis set
- unstratified

Methods for evaluating the validity of model assumptions

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

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

If these show large departures from proportional hazards, then OS 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 (ie, the hazard ratio is not constant over time).

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

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

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

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

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.

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

See subgroups as defined in Section 6.4.

Multivariable Cox regression analysis will be carried out to assess and adjust the treatment effect for relevant baseline factors of potential prognostic impact. A stepwise selection procedure will serve to identify explanatory variables of potential prognostic values

additional to the randomization strata which will be included in all models during the selection procedure. The Cox's Proportional Hazard model is defined as:

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

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

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

6.2.2.3. Objective response and tumor shrinkage

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

BOR Based on Confirmed Responses:

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

- NE: all other cases.

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

- No baseline assessment
- No evidence of disease at baseline
- 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' (ie, 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 11 outlines the possible BOR outcomes by investigator and BICR.

Table 11. Possible BOR Outcomes for Investigator vs BICR

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

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

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

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

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

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

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

Tumor Shrinkage from Baseline:

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

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

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

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

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

6.2.2.4. Disease control

Disease Control (DC) is defined as BOR of CR, PR, non-CR/non-PD or SD. DC rate (DCR) is the proportion of patients with DC.

DCR will be summarized by frequency counts and percentages.

6.2.2.5. Duration of response

Duration of Response (DR) is defined, for patients with OR, as the time from the first documentation of objective response (CR or PR) to the date of first documentation of PD or death due to any cause. If a patient has not had an event (PD or death), DR is censored at the date of last adequate tumor assessment. The censoring rules for DR are described in Table 12.

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

Table 12. Outcome and Event Dates for DR Analyses

Scenario	Date of event/censoring	Outcome
PD or death - After at most one missing or inadequate post-baseline tumor assessment, OR - ≤ 16 weeks after the date of randomization	Date of PD or death	Event
PD or death - After 2 or more missing or inadequate post-baseline tumor assessments	Date of last adequate tumor assessment [a] 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 [a] 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 [a] documenting no PD before new anti-cancer therapy is given or missed tumor assessments	Censored

[a] 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.

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 rates at 3, 6 and 12 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice (2002) (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.

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 13 following the hierarchy shown.

Table 13. DR Censoring Reasons and Hierarchy

Hierarchy	Condition	Censoring Reason
1	Start of new anti-cancer therapy	Start of new anti-cancer therapy
2	Event after 2 or more missing or inadequate post-baseline tumor assessments/date of randomization	Event after 2 or more missing assessments [a]
3	No event and [withdrawal of consent date ≥ date of randomization OR End of study (EOS) = Patient refused further follow-up]	Withdrawal of consent
4	No event and lost to follow-up in any disposition page	Lost to follow-up
5	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
6	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.

6.2.2.6. Time to response

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

$$\text{TTR (in months)} = [\text{first date of OR} - \text{date of randomization} + 1]/30.4375$$

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

6.2.3. Pharmacokinetic endpoints

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

C_{trough} and C_{max} for avelumab will be summarized descriptively (n, mean, SD, CV, median, minimum, maximum, geometric mean, its associated CV, and 95% CI) by co-primary population, cycle, and day. Pharmacokinetic parameters for avelumab will be taken from observed plasma concentration-time data as described in Section 3.2.4.

C_{trough} will be summarized for all pre-dose samples collected prior to avelumab dosing on the same visit day. C_{max} will be summarized for all post-dose samples collected +/- 10% or +/- 30 minutes from the end of infusion, whichever is further from the end of infusion time.

Presentation of pharmacokinetic data will include:

- Descriptive statistics (n, mean, SD, %CV, median, minimum, maximum) of plasma concentrations will be presented in tabular, cycle, day and nominal time.

- Linear-linear plots of mean and median plasma concentrations by nominal time for avelumab will be presented for PK sampling days by cycle, and study day.

- Pharmacokinetic parameters for avelumab 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). Box plots for C_{max} and C_{trough} for avelumab will be generated. Individual data points, the geometric mean and the median of the parameter will be overlaid on the box plots. If Arm A has limited evaluable PK data (n<4), matchstick plots showing changes C_{max} and C_{trough} in individual patients will then be generated. The geometric mean of the parameter will be overlaid in the plots.

6.2.4. Population pharmacokinetic endpoints

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

6.2.5. Biomarker endpoints

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

Exploratory endpoints in the study include peripheral blood and additional tumor tissue biomarkers consisting the levels of cells, deoxyribonucleic acid (DNA), ribonucleic acid (RNA), or proteins that may be related to anti-tumor immune response and/or response to or disease progression on avelumab, such as genes related to IFN-γ or transforming growth factor (TGF)-β.

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

Biomarkers will be summarized at all timepoints described below:

- Blood biospecimens for exploratory assessments: C1D1, C1D15 (Arm A only), C2D1, C3D1, C5D1 and EOT.

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

- Tumor based biomarkers, de novo Tumor Biopsy: screening and End of Treatment.

Descriptive summary statistics (mean, SD, median, Q1, Q3, coefficient of variation, minimum, and maximum), at each timepoint and ratio to baseline at each visit (if appropriate) will be provided for avelumab for continuous variables. Non-continuous variables will be summarized using n/N and percent at each timepoint and treatment arm.

For each biomarker or set of biomarkers, patients may be classified as positive, negative, or some other category according to scoring algorithms and cut-offs established from external

sources. If no external standards exist, biomarkers may be stratified using the median: ≤ median and > median. Percent of patients in each category will be tabulated and tested using methods for categorical variables.

Biomarkers may be included as a covariate in a survival analysis model; biomarkers will be stratified on the above-mentioned classifications. Using a Cox proportional hazard model, HR, 95% CI, and p-values will be calculated for OS (primary endpoint) and PFS by BICR assessment (secondary endpoint). Further analyses may include the biomarker status by treatment arm interaction. If there are fewer than 20 events in a subgroup, or the HR is not estimable, only descriptive summary statistics will be provided.

Biomarker value or classification will also be provided for percent responders. For analyses using the actual value, the value will be summarized by response category (yes or no to being a responder, defined as CR or PR from best overall response). The Wilcoxon Rank-Sum test will be used as discussed above. For analyses using the biomarker classification, a contingency table will be produced, and the treatment arms compared using a Fisher's exact test. If there are less than 5 observations in a cell, no p-value will be produced for either the Fisher's exact test.

For all biomarker analytes, it is not possible to pre-specify all possible analyses of interest. Thus, the biomarker results will follow the 2-Step Methodology approach. The analyses of initial interest will be performed as described above. A second set of analyses will be defined and performed based on the results from the first set of analyses. It is anticipated that if there are, for example, 100 analyses run in the first step, no more than 20 analyses will be run in the second step; estimates will get larger or smaller depending on the number of analyses run in the first step.

6.2.6. Endpoints for Immunogenicity Data of Avelumab

All analyses described below are performed for Arm A.

Blood samples (3.5 mL SST tube per timepoint) for evaluation of avelumab immunogenicity will be collected from Arm A patients within 2 hours before the start of the avelumab infusion on Day 1 and Day 15 of Cycles 1 through 3 and then on Day 1 of Cycle 5, 7, 9, 11 and 13. Immunogenicity blood samples will be assayed for anti avelumab antibodies using a validated analytical method. All of the samples that are positive for ADA may also undergo characterization for nAb.

Patients will be characterized into different ADA categories based on the criteria defined in Table 14.

Table 14. Patients Characterized Based on Anti-Drug Antibody Results (ADA Status)

Category	Definition	Patients 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 15. For nAb, treatment-boosted is not applicable since no titer result is available.

Table 15. Patients Characterized Based on Neutralizing Antibody Results (nAb Status)

Category	Definition	Patients 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 (2002) (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 and safety by immunogenicity status

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

ADA

- ADA ever-positive versus ADA never-positive

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

nAb

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

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

PK parameters and immunogenicity status

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

Blood samples for avelumab PK will be collected pre-dose and at the end of infusion (immediately before the end of avelumab infusion) on Day 1 and Day 15 of Cycles 1 – 3, and then pre-dose and at the end of avelumab infusion) on Day 1 of Cycle 5, 7, 9, 11 and 13.

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

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

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

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

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

Safety and immunogenicity status

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

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

- TEAEs, by SOC and PT
- TEAEs leading to discontinuation of avelumab, by SOC and PT

- TEAEs leading to dose reduction of avelumab, 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.

6.2.7. PRO endpoints

6.2.7.1. PRO Instruments

Patient-reported outcomes of HRQoL will be evaluated using NCCN-FACT Bladder Symptom Index (FBlSI-18) and EQ-5D-5L. A 9-item subscale of the FBlSI-18 known as the FBlSI-Disease Related Symptoms-Physical subscale (FBlSI-DRS-P) measuring advanced bladder cancer disease related symptoms will be used to define time to deterioration.

All PRO analyses will be based on FAS. Descriptive summaries will include all the assessments per schedule of assessment from baseline to the last PRO assessment. For time - to-event analysis and random coefficient models will include assessments from baseline up to EOT (not including EOT).

6.2.7.2. Scoring Procedure

The NCCN-FACT FBlSI-18 and EQ-5D-5L will be scored according to their respective validation papers and user's guides. Missing values in the instruments will also be handled following the guidance from their respective User Manuals. The NCCN-FACT FBlSI-18 questionnaire allows for the calculation of an overall score and scores for four subscales; Disease Related Symptoms – Physical (DRS-P), Disease Related Symptoms – Emotional (DRS-E), Treatment Side Effects (TSE), and Functional and Well-being (F/WB). For the NCCN-FACT FBlSI-18 multi-item scales, the score may be imputed as the mean of the non-missing questions if at least half the questions in that scale are answered. For EQ-5D-5L, the entire score for that cycle is deemed missing if the answer to any one of the 5 dimensions is missing. In addition to items for the 5 dimensions, the EQ-5D-5L contains a single visual analog scale (EQ-VAS) item that is used to assess the respondent's self-reported health status.

6.2.7.3. Instrument Compliance Rates

For each treatment arm and at each time point, the number and percentage of patients who complete the NCCN-FACT FBlSI-18, EQ-VAS 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.

6.2.7.4. Descriptive Summary

Separate tables will be used to summarize, for each treatment arm, the mean (and SD), and median (and range) of absolute scores, and the mean (and 95% CI) of change from baseline of the NCCN-FACT FBlSI-18, and all subscales (DRS-P, DRS-E, TSE, and F/WB) of the NCCN-FACT FBlSI-18, the EQ-5D-5L, and EQ-VAS at each time point. A line chart depicting the means along with 95% CIs over time will be provided for each scale in each treatment arm.

For the EQ-5D-5L, EQ-VAS, NCCN-FACT FBlSI-18, and subscales of the NCCN-FACT BlSI-18 (DRS-P, DRS-E, TSE, FW/B), for each post-baseline visit, a means score and mean change from baseline scores will be generated for each treatment arm, adjusting for baseline differences in the treatment arms. The adjusted means (ie LS-MEANS) will be obtained from ANCOVA model that included baseline score and treatment as independent variables. For each per-visit analysis, the n, LS-mean, standard error, and 95% CI will be reported.

A figure will be provided for the cumulative patient disposition by treatment arm per PRO assessment window (ie every 28-day cycle) for all scheduled assessments (Day 1 of each cycle) using the following categories:

- PRO assessment expected;
- PRO assessment not expected due to progression;
- PRO assessment not expected due to death;
- PRO assessment not expected due to other reasons.

6.2.7.5. Time-to-Deterioration Analysis

A 3-point change in FBlSI-DRS-P 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 FBlSI-DRS-P for 2 consecutive assessments.

Patients will be censored at the last time when they completed the assessment if they have not deteriorated.

TTD will be analyzed by a log-rank test stratified by randomization stratification factors. The time to deterioration in each treatment arm will be summarized using the Kaplan-Meier method and displayed graphically where appropriate. Confidence intervals (CIs) for the 25th, 50th and 75th percentiles will be reported. The Cox proportional hazards model will be fitted to compute the treatment hazard ratios and the corresponding 95% CI.

Sensitivity analyses based on a 2- and 4-point decrease from baseline in FBlSI-DRS-P will also be conducted to assess the robustness of the TTD analyses. Cumulative distribution plots of FBlSI-DRS-P will also be presented on Day 1 of Cycles 2, 4 and 6.

6.2.7.6. Random Coefficient Models

Random coefficient models will be carried out for the NCCN-FACT FBlSI-18 and all subscales (DRS-P, DRS-E, TSE, and F/WB), EQ-5D-5L, and EQ-VAS 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 patient. All available data for each patient should be used in the analyses. All parameter estimates should be obtained using restricted maximum likelihood. The unstructured covariance structure should 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 should be used (using option "ddfm = kr" as a part of the MODEL statement in PROC MIXED).

6.3. Other Endpoints

Not applicable.

6.4. Subset Analyses

Subset analyses will be performed for patients with PD-L1-positive tumors and for patients irrespective of PD-L1 expression.

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

The following subgroups will be defined and used for analyses:

- Randomization stratification factors
  - Best response to first-line chemotherapy (CR or PR vs. SD)
  - Metastatic disease site (visceral vs. non-visceral) at the time of initiating first-line chemotherapy.
- Age
  - Age < 65 years (Reference)
  - Age ≥ 65 years
- Gender
  - Male (Reference)
  - Female
- Race
  - White (Reference)
  - Asian

- Black or African American
- Other

- Ethnicity
  - Hispanic or Latino
  - Not Hispanic or 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)
- PD-L1 status at baseline
  - Positive (Reference)
  - Negative
- First-line chemotherapy regimen
  - Gemcitabine+cisplatin (Reference)
  - Gemcitabine+carboplatin
- ECOG Performance Status
  - 0 (Reference)
  - ≥ 1
- Creatinine Clearance at baseline
  - ≥60 mL/min (Reference)
  - <60 mL/min
- Liver lesions at baseline
  - Yes
  - No
- Lung lesions at baseline
  - Yes

- No

Subset analyses for OS, PFS and DR will use the primary censoring rules described in Sections 6.1.1.1, 6.2.2.1 and 6.2.2.5. All the subgroup analyses are exploratory. Treatment arms will be compared for OS and PFS using a 2-sided unstratified log-rank test for each subgroup level and the unstratified HR and its corresponding 95% CI will be computed per subgroup level.

All the subgroup analyses will be exploratory; no adjustment for multiplicity will be performed. In the case of a low number of patients within a category (<5% of the randomized population), the categories will be pooled, or in cases where no meaningful pooling can be performed, the category may not be summarized.

To assess the heterogeneity of treatment effects for OS and PFS across the subgroup levels, two Cox regression models will be fitted with OS or PFS, 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 (Wald's test) will be provided together with the HR and corresponding 95% CI for the interaction model parameter.

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

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

6.5. Baseline and Other Summaries and Analyses

6.5.1. Baseline summaries

The following analyses will be performed for patients with PD-L1-positive tumors and for patients irrespective of PD-L1 expression, 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 or Latino
  - Not Hispanic or Latino
- 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

Separately from the RECIST eCRF page based on Investigator assessment and based on BICR assessment:

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

From the 'Substance Use' eCRF page:

- Smoking history (Never smoker, Current smoker, Former smoker)

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

6.5.1.4. Prior anti-cancer therapies

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

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

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

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

- At least one prior anti-cancer drug therapy
- Number of prior anti-cancer drug therapy regimens: missing, 1, 2, 3, ≥4
- Intent of Drug Therapy: Neo-Adjuvant, Adjuvant, Advanced-Metastatic, Locoregional Disease-Recurrence
- Best response: CR, PR, SD, PD, Unknown, Not applicable. Best response is derived from the last treatment regimen.

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

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

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

6.5.1.5. Exposure to first-line anti-cancer chemotherapy regimen

Exposure to first-line anti-cancer agents will be summarized separately for gemcitabine, cisplatin and carboplatin.

The duration of treatment (in weeks) for each agent is defined as:

> Treatment duration (weeks) = (last dose date of gemcitabine or cisplatin or carboplatin – first dose date of gemcitabine or cisplatin or carboplatin + 21)/7

The cumulative dose of each agent is the sum of the actual dose levels that the patient received.

The summary of treatment exposure for first-line chemotherapy regimen will include the following information:

- Total number of infusions received for each agent
- Duration of treatment
- Cumulative dose (mg/m^2 or AUC, as appropriate) for each agent

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;
- Number and percentage of randomized patients who discontinued study drug overall and by the main reason for discontinuation of study drug;
- 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);
- Cross tabulation: patients randomized (avelumab + BSC/BSC/none) vs. patients treated (avelumab + BSC/BSC/none).

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.

All dosing calculations and summaries will be based on 'Parenteral – Drug avelumab' eCRFs pages. A listing of study drug administration will be created with the information collected on the 'Parenteral – Drug avelumab' eCRF pages.

For Arm B, the duration of BSC treatment will be summarized based on the time from C1D1 through date of discontinuation of BSC as collected in the Disposition eCRF page.

The derivations below are provided for the following:

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

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

- Avelumab - total dose / weight

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.

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

- the start date of dosing in the Cycle X day 1 visit eCRF exposure page, if the patient received avelumab on that visit (ie, avelumab with dose>0 at that visit)

- the first day of assessments in the Cycle X day 1 visit, if the patient did not receive avelumab on that visit (ie, dose=0 for avelumab at that visit). Use start date in the exposure page if available; if start date is not available then use date of collection of vital signs on Cycle X day 1 visit.

Actual cycle end date for each patient is,

- for all cycles X except the last cycle, actual cycle end date = actual cycle (X+1) start date – 1 day;
- for the last cycle, actual cycle end date = actual cycle start date + 28 – 1 day

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

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 + 28 – 1

Duration of exposure to avelumab (weeks) =

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

Cumulative dose is the sum of the actual doses of avelumab received.

Actual Dose Intensity (DI)

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

Relative Dose Intensity (RDI)

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

6.5.3.2. Dose reductions

Applicable to avelumab.

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

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

6.5.3.3. Dose delays

Applicable to 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.

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, 17 or 18, this is considered as 1-3 days delay.

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

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

6.5.3.4. 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% will be summarized for Arm A.

6.5.3.5. Infusion interruptions

Applicable to avelumab.

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

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

6.5.3.6. Duration of BSC

The duration of BSC (in weeks) during the study for a patient is defined as:

Treatment duration (weeks) = (end date of BSC – start date of BSC + 1)/7

where end and start dates of BSC are defined in Section 3.4.1.

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 drugs, which started prior to first dose date of study treatment and continued during the on-treatment period as well as those started during the on-treatment period. **Prior medications** are medications, other than study drugs 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. The number and percentage of patients with anti-cancer drug therapies based on the data collected from the 'Follow-up Cancer Therapy' will also be summarized for patients with PD-L1-positive tumors, by treatment arm.

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 as defined in Section 3.5.1.

All analyses described below 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 with onset outside the on-treatment period will be flagged in the listings.

- **Related Adverse Events:** adverse events with relationship to study treatment (as recorded on the AE eCRF page, Relationship with study treatment = Related) reported by the investigator and those of unknown relationship (ie, no answer to the question 'Relationship with study treatment').
- **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 [SRP] 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 SRP 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 recorded on the 'Adverse Event' eCRF page. Action taken with BSC for patients in Arm A will not be summarized.

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 interruption of avelumab
- TEAEs leading to discontinuation of avelumab
- TEAEs leading to discontinuation of BSC
- Related TEAEs leading to discontinuation of avelumab
- Related TEAEs leading to discontinuation of BSC
- 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

Applicable to Arm A only.

The frequency (number and percentage) of patients with each of the following will be presented for TEAEs leading to dose reduction of avelumab for Arm A:

- TEAEs leading to dose reduction of avelumab 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

Applicable to Arm A only.

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 6.5.3.3). 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 6.5.3.3).

The frequency (number and percentage) of patients with each of the following will be presented for TEAEs leading to interruption of avelumab for Arm A:

- TEAEs leading to interruption of avelumab 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 avelumab for Arm A:

- TEAEs leading to both interruption and dose reduction of avelumab 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 Treatment Discontinuation

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 BSC by SOC and PT
- Related TEAEs leading to discontinuation of BSC 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 BSC, 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 irAEs (as defined in Appendix 1) and Tier-2 events separately. 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
- CIs reported are not adjusted for multiplicity and should be used for screening purposes only. The 95% CIs are provided to help gauge the precision of the estimates for the risk difference and should be used for estimation purposes only.

6.6.5. Laboratory data

6.6.5.1. Hematology and chemistry parameters

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

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

Abnormalities classified according to NCI-CTCAE toxicity grading version 4.03 will be described using the worst grade. For those parameters which are graded with two toxicities such as potassium (hypokalemia/hyperkalemia), the toxicities will be summarized separately. Low direction toxicity (eg, hypokalemia) grades at baseline and post baseline will be set to 0 when the variables are derived for summarizing high direction toxicity (eg, hyperkalemia), and vice versa.

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

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

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 $\geq$ 3×ULN, ALT $\geq$ 5×ULN, ALT $\geq$ 10×ULN, ALT $\geq$ 20×ULN
- AST $\geq$ 3×ULN, AST $\geq$ 5×ULN, AST $\geq$ 10×ULN, AST $\geq$ 20×ULN
- (ALT or AST) $\geq$ 3×ULN, (ALT or AST) $\geq$ 5×ULN, (ALT or AST) $\geq$ 10×ULN, (ALT or AST) $\geq$ 20×ULN
- TBILI $\geq$ 2×ULN
- Concurrent ALT $\geq$ 3×ULN and TBILI $\geq$ 2×ULN
- Concurrent AST $\geq$ 3×ULN and TBILI $\geq$ 2×ULN
- Concurrent (ALT or AST) $\geq$ 3×ULN and TBILI $\geq$ 2×ULN
- Concurrent (ALT or AST) $\geq$ 3×ULN and TBILI $\geq$ 2×ULN and ALP > 2×ULN
- Concurrent (ALT or AST) $\geq$ 3×ULN and TBILI $\geq$ 2×ULN and (ALP $\leq$ 2×ULN or missing)

Concurrent measurements are those occurring on the same date.

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

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

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

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

Parameters with NCI-CTC grades available:

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

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

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

- Hematology:

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

- Serum Chemistry:

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

Parameters with NCI-CTC grades not available:

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

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

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

In this study, these apply to the following parameters:

- Thyroid function and ACTH tests: Free T4, Thyroid Stimulating Hormone (TSH), ACTH

6.6.5.2. Other laboratory parameters

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

- Coagulation: activated partial thromboplastin time (aPTT), prothrombin time (PT) and International Normalized Ratio (INR).
- HCV, HBV, HIV serology
- Urinalysis: Protein, glucose and blood
- Other parameters: Absolute Monocytes, Absolute Eosinophils, Absolute Basophils, Chloride, Total Urea, Uric Acid, Total Protein, C-Reactive Protein, Lactate Dehydrogenase (LDH)
- Cardiac biomarkers: Troponin (cTnT or cTnI)
- Pregnancy test

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

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

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

6.6.6. Vital signs

Weight for the purposes of dose calculation will be recorded at screening and within 3 days pre-dose Day 1 and Day 15 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.

6.6.7. Electrocardiogram

ECGs in the B9991001 study are collected only when clinically indicated. Patients with qualitative ECG abnormalities will be listed.

6.6.8. ECOG Performance Status

The ECOG shift from baseline to highest score during the on-treatment period will be summarized by treatment arm.

7. INTERIM ANALYSES

7.1. Introduction

The goals of the interim analyses are to allow early stopping of treatment arm(s) for futility or efficacy and to potentially adjust the sample size. The interim analysis of OS will be performed as described in Sections 5.1.1 and 5.1.2 using the methodology described in Section 6.1.1.1 for OS.

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

7.2. Interim Analyses and Summaries

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

7.2.1. Interim analysis for OS

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 OS 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 $\alpha(t_1)$ 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 OS events observed at the time of the cut-off for the interim analysis and the total number of OS events targeted for the final analysis).

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

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

τ=0.015 for 'all patients' population,

τ=0.01 for 'patients with PD-L1-positive tumors' population

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 analysis and the actual correlation between the two test statistics Z_1 and Z_F, so that the overall one-sided significance level across all analyses and comparisons is preserved at 0.025.

7.2.2. Sample size re-estimation

As appropriate, the sample size of the study may be adjusted using the method outlined by Cui et al (1999). Specific details regarding the method are described in Appendix 3.

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. Cella DF, Tulsky DS. Quality of life in cancer: definition, purpose, and method of measurement. Cancer Investigation 1993;11(3):327-36.

4. Cella DF, Tulsky DS, Gray G, et al. The Functional Assessment of Cancer Therapy scale: development and validation of the general measure. J Clin Oncol 1993;11(3):570-9.

5. Cella DF. FACT Symptom Indexes for regulatory and other purposes. FACIT Organization Manual 2014.

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

7. Cui L, Hung HMJ, Wang S, et al. Modification of sample size in group-sequential clinical trials. Biometrics. 1999; 55: 853-7.

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

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

10. Jensen SE, Beaumont JL, Jacobsen PB, et al. Measuring priority symptoms in advanced bladder cancer: development and initial validation of a brief symptom index. The Journal of Supportive Oncology 2013; 11(92):86-93.

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

12. Kalbfleisch JD, Prentice, RL. The Statistical Analysis of Failure Time Data, 2nd Edition. Hoboken, Wiley Interscience, 2002.

13. Kim K. Study duration for group sequential clinical trials with censored survival data adjusting for stratification. Statistics in Medicine 1992: 11, 1477-1488

14. Powles T, Huddart RA, Elliott T, et al. A phase II/III, double-blind, randomized trial comparing maintenance lapatinib versus placebo after first line chemotherapy in HER1/2 positive metastatic bladder cancer patients. J Clin Oncol 2015;33 (suppl:Abstract 4505).

15. 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.

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

17. 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

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

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

20. Tsiatis A.A. The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. Biometrika 1981: 68, 1,311-315

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

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

9. APPENDICES

Appendix 1. Immune-Related Adverse Events

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

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

Table 16. Case Definition for irAEs

Step	Selection Criteria	Additional Notes
1	Event selected based on a list of pre-specified MedDRA PTs within clusters. These are included in the SRP as Tier1 events (Immune-mediated xxxx). If AE matches the list, then it is in for the next step	
2	AE onset during 1st study drug administration or anytime thereafter through 90 days after last dose of study treatment.	This is regardless of start of new anti-cancer drug therapy and regardless of TEAE classifications
3	Answer in the AE eCRF page to 'Was another treatment given because of the occurrence of the event' is 'YES'	
4	AE treated with corticosteroids or other immunosuppressant therapy. For endocrinopathies only: AE required hormone replacement	Look in the conmed pages for AE identifiers that match the AEs from Step 3. For each of such AEs if A) OR B) 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)	A) From the AE eCRF page. Is the AE clearly related to an etiology other than immune-mediated etiology? Yes / No If answer is Yes, check all that apply: • Underlying malignancy / progressive disease. • Other medical conditions. • Prior or concomitant medications / procedures. • Other. Specify.
	B) Histopathology / biopsy consistent with immune-mediated event	B) From the AE eCRF page. B1) Was there a pathology / histology evaluation performed to investigate the AE? Y/N B2) If answer to the above is Yes, does the pathology/histology evaluation confirms an immune mediated mechanism for the AE? Y/N B3) If pathology / histology evaluation performed to investigate the AE, provide summary of relevant findings of the pathology /histology report. (Free Text)
	Event is in if [Answer to 5B1 and 5B2 is YES (regardless of answer to 5A)] OR [Answer to 5B1 is YES AND answer to 5B2 is NO AND answer to 5A is NO] OR [Answer to 5B1 is NO AND answer to 5A is NO]	

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

Appendix 2. Infusion Related Reactions

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

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

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

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

Table 17. 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

Appendix 3. Technical Details For Sample Size Re-Estimation

In 1999, Cui et al. published the form of the test statistic at the final analysis if a sample size adjustment is performed in an adaptive design with normal endpoints. Using the proposed weighted test statistic ensures that the overall Type 1 error is maintained at the pre-specified level.

In 2007, Mehta presented an extension of Cui's method to time to event (TTE) endpoints reported using the two-sample log-rank test (LRT). This extension is based on results published in 1981 by Tsiatis (1981) who showed that the sequentially computed scores

statistics of the proportional hazards model converges asymptotically to a multivariate normal process with independent increments. Cui's method for normal endpoints is applied to the LRT, as a special case for the efficient scores test for the proportional hazards model.

The proposed weighted LRT statistic (T_2) calculated at the final analysis after a sample size increase preserves the Type 1 error.

$$T_2 = \sqrt{t_1}Z_1 + \sqrt{1-t_1}\left(\frac{\sqrt{t_2^*}Z_2^* - \sqrt{t_1}Z_1}{\sqrt{t_2^* - t_1}}\right) \quad (1)$$

where:
t_1 = Information fraction at the interim analysis
t_2* = Information fraction for LRT statistic at the final analysis after a sample size increase
Z_1 = LRT statistic at interim analysis
Z_2* = LRT statistic at final analysis after a sample size increase.

In 1992, Kim showed that the sequentially computed stratified log-rank test (SLRT) statistics has an asymptotic multivariate normal distribution with independent increments. Based on this result and using a similar approach as presented by Mehta, Cui's method can be further extended to TTE endpoints reported using the SLRT. Similarly, the proposed weighted SRLT statistic (U_2) calculated at the final analysis after a sample size increase preserves the Type 1 error.

$$U_2 = \sqrt{u_1}S_1 + \sqrt{1-u_1}\left(\frac{\sqrt{u_2^*}S_2^* - \sqrt{u_1}S_1}{\sqrt{u_2^* - u_1}}\right) \quad (2)$$

where:
$u_1 = D_1/D_{max}$ = Information fraction at the interim analysis; D_1 = total number of events at the time of interim analysis and D_{max} = 200
u_2* = Information fraction at the final analysis after a sample size increase
S_1 = LRT, at interim analysis, stratified for baseline stratification factors
S_2* = LRT, at final analysis after a sample size increase, stratified for baseline stratification factors