Official Title of Study:

A Randomized Phase 3 Study of Nivolumab plus Ipilimumab or Nivolumab Combined with Fluorouracil plus Cisplatin versus Fluorouracil plus Cisplatin in Subjects with Unresectable Advanced, Recurrent or Metastatic Previously Untreated Esophageal Squamous Cell Carcinoma

NCT Number: NCT03143153

Document Date (Date in which document was last revised): February 23, 2021

STATISTICAL ANALYSIS PLAN
FOR CLINICAL STUDY REPORT

A RANDOMIZED PHASE 3 STUDY OF NIVOLUMAB PLUS IPILIMUMAB OR NIVOLUMAB COMBINED WITH FLUOROURACIL PLUS CISPLATIN VERSUS FLUOROURACIL PLUS CISPLATIN IN SUBJECTS WITH UNRESECTABLE ADVANCED, RECURRENT OR METASTATIC PREVIOUSLY UNTREATED ESOPHAGEAL SQUAMOUS CELL CARCINOMA

PROTOCOL CA209648

VERSION # 4.0

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 BACKGROUND AND RATIONALE

CA209648 is a randomized, global Phase 3 study of nivolumab plus ipilimumab or nivolumab in combination with fluorouracil plus cisplatin versus fluorouracil and cisplatin chemotherapy as first line-therapy in inoperable advanced, recurrent or metastatic esophageal squamous cell carcinoma (ESCC). This study will determine if nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin improve overall survival (OS) and/or progression free survival (PFS) over fluorouracil and cisplatin standard of care chemotherapy (further: chemotherapy) in subjects with ESCC whose tumors express PD-L1. Additional objectives include further characterization of the efficacy, adverse event profile, pharmacokinetics (PK), patient reported outcomes, and potential predictive biomarkers of nivolumab in combination with ipilimumab or nivolumab in combination with fluorouracil and cisplatin in subjects with ESCC.

Research Hypothesis:

- The administration of nivolumab plus ipilimumab will improve OS compared with fluorouracil and cisplatin combination in subjects with unresectable advanced, recurrent or metastatic ESCC with PD-L1 expression ≥ 1%.
- The administration of nivolumab combined with fluorouracil plus cisplatin will improve OS compared with fluorouracil and cisplatin combination alone in subjects with unresectable advanced, recurrent or metastatic ESCC with PD-L1 expression ≥ 1%.
- The administration of nivolumab plus ipilimumab will improve PFS as assessed by a blinded independent central review committee (BICR) compared with fluorouracil and cisplatin combination in subjects with unresectable advanced, recurrent or metastatic ESCC with PD-L1 expression ≥ 1%.
- The administration of nivolumab combined with fluorouracil plus cisplatin will improve PFS as assessed by a BICR compared with fluorouracil and cisplatin combination alone in subjects with unresectable advanced, recurrent or metastatic ESCC with PD-L1 expression ≥ 1%.

Schedule of Analyses:

In this study, two formal efficacy analyses are planned:

- Final PFS analysis and Interim OS analysis: For superiority of PFS in subjects who were randomized to nivolumab plus ipilimumab vs. subjects who were randomized to chemotherapy and in subjects who were randomized to nivolumab in combination with chemotherapy vs. subjects who were randomized to chemotherapy only. Analysis will be performed when 136 PFS events per BICR are observed or when a 12-month minimum follow-up (defined as the time from the date of the last patient was randomized to the clinical cutoff date) is reached if the planned number of events is unlikely to be reached among the PD-L1 expressing subjects in the chemotherapy arm. By design, 136 PFS events among the PD-L1 expressing subjects in the chemotherapy arm are expected to be reached approximately 33 months after first patient randomized. At the same time, formal interim analysis of OS will be also conducted (in case of significant results, study may be stopped). All the events observed in the locked database will be used in the final analysis of PFS, following intention-to-treat principle (Section 5).

- Final OS analysis: For superiority of OS in subjects who were randomized to nivolumab plus ipilimumab vs. subjects who were randomized to chemotherapy and in subjects who were randomized to nivolumab in combination with chemotherapy vs. subjects who were randomized to chemotherapy only. Analysis will be performed when 140 OS events are observed among the PD-L1 expressing subjects in the chemotherapy arm. This is expected to be reached approximately 49 months after first patient randomized. In the case that the observed number of OS events exceed the pre-planned number, all the events observed in the locked database will be used in the final analysis of OS, following intention-to-treat principle (Section 5).

Since each of the formal efficacy analyses will be triggered by the pre-planned number of events among the PD-L1 expressing subjects in the chemotherapy arm, the external independent statistical group supporting DMC review (Axio Research) will be utilized to conduct unblinded event tracking (the tracking plan is documented separately).

2 STUDY DESCRIPTION

2.1 Study Design

This is a randomized Phase 3 study of nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin compared with fluorouracil and cisplatin combination in adult (≥ 18 years) male and female subjects with unresectable advanced, recurrent or metastatic ESCC.

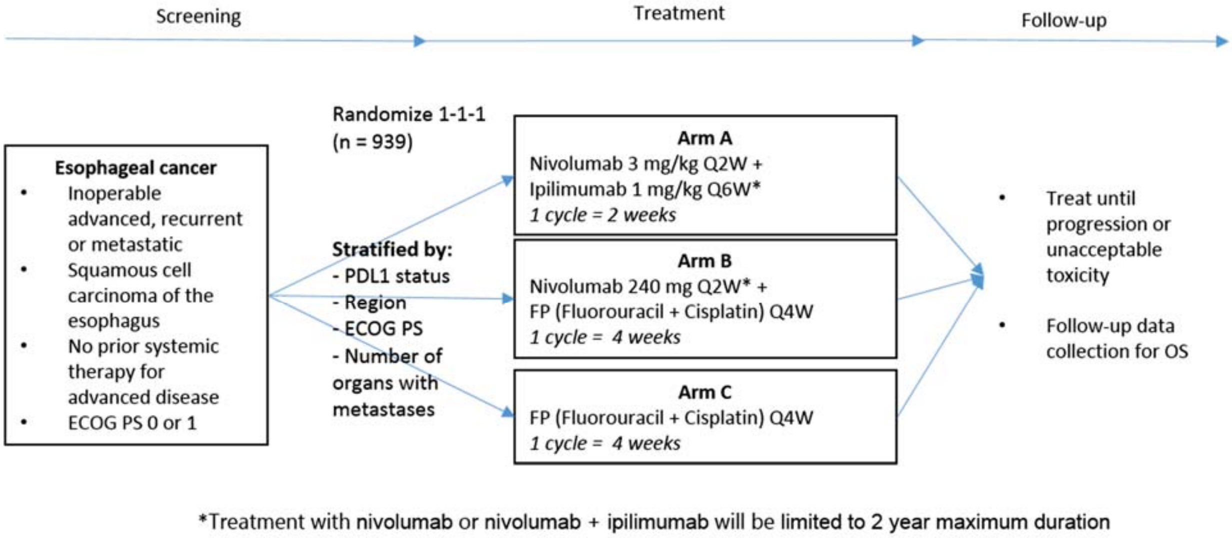
After signing the informed consent form, and upon confirmation of the subject's eligibility, subjects with unresectable advanced, recurrent or metastatic ESCC will be randomized in a 1:1:1 ratio to one of the following open-label treatments:

- Nivolumab-plus-ipilimumab (N+I) arm: Subjects will receive treatment with nivolumab 3mg/kg as a 30-minute infusion every 2 weeks and ipilimumab as a 30-minute infusion 1mg/kg every 6 weeks.
- Nivolumab-plus-chemotherapy (N+C) arm: Subjects will receive treatment with nivolumab 240mg as a 30-minute infusion on Day 1 and Day 15, fluorouracil 800mg/m^2/day as an IV continuous infusion on Day 1 through Day 5 (for 5 days), and cisplatin 80mg/m^2 as a 30 to 120-minute infusion on Day 1 of 4-week cycle.
- Chemotherapy (CT) arm: Subjects will receive treatment with fluorouracil 800mg/m^2/day as an IV continuous infusion from Day 1 through Day 5 (for 5 days), and cisplatin 80mg/m^2 as a 30- to 120-minute infusion on Day 1 of 4-week cycle.

Treatment with nivolumab or nivolumab with ipilimumab will be given for up to 24 months in the absence of disease progression (PD) or unacceptable toxicity. Chemotherapy will be given as per the study dosing schedule until PD or unacceptable toxicity.

Baseline and all subsequent scans will be submitted to a BICR for analysis and archiving, once the subject is randomized and throughout the study period.

Figure 2.1-1: Study Scheme


Screening
Treatment
Follow-up
Esophageal cancer
• Inoperable advanced, recurrent or metastatic
• Squamous cell carcinoma of the esophagus
• No prior systemic therapy for advanced disease
• ECOG PS 0 or 1
Randomize 1-1-1
(n = 939)
Stratified by:
- PDL1 status
- Region
- ECOG PS
- Number of organs with metastases
Arm A
Nivolumab 3 mg/kg Q2W +
Ipilimumab 1 mg/kg Q6W*
1 cycle = 2 weeks
Arm B
Nivolumab 240 mg Q2W* +
FP (Fluorouracil + Cisplatin) Q4W
1 cycle = 4 weeks
Arm C
FP (Fluorouracil + Cisplatin) Q4W
1 cycle = 4 weeks
• Treat until progression or unacceptable toxicity
• Follow-up data collection for OS
*Treatment with nivolumab or nivolumab + ipilimumab will be limited to 2 year maximum duration

This study will consist of 3 phases: screening, treatment, and follow-up:

Screening Phase

- Begins by establishing the subject's initial eligibility and signing of the informed consent.
- Subject is enrolled using the Interactive Response Technology (IRT) system.
- Subjects must have PD-L1 IHC testing, with evaluable results, performed by the central lab during the Screening period.
- Either a formalin-fixed, paraffin-embedded (FFPE) tissue block or unstained tumor tissue sections, with an associated pathology report if available, must be submitted for biomarker evaluation prior to randomization. The tumor tissue sample may be fresh or archival if obtained within 6 months prior to randomization, and there can have been no systemic therapy (eg, adjuvant) given after the sample was obtained.
- All subjects with PD-L1 result will be randomized, but only subjects with PD-L1 expression ≥ 1% will be assessed as part of the primary objective.
- Subject is assessed for study eligibility according to the inclusion and exclusion criteria. All screening assessments and procedures must be performed within 28 days prior to randomization unless otherwise specified.

Treatment Phase

- The treatment phase begins when the treatment is assigned in the IRT. Study treatment must begin within 3 calendar days of randomization.

- Treated subjects will be evaluated for tumor assessments every 6 weeks (± 7 days) starting on Week 7 up to and including Week 48, then every 12 weeks (± 7 days) thereafter.
- Treatment with nivolumab with ipilimumab or nivolumab (Arms A and B) will be given for up to 24 months in the absence of PD or unacceptable toxicity. Chemotherapy (Arms B and C) will be given as per the study dosing schedule until PD or unacceptable toxicity. Then, the subject will enter the Follow-up Phase.

Dosing in the N+I arm:

- Nivolumab 3mg/kg will be administered IV on Day 1 and then every 2 weeks.
- Ipilimumab 1mg/kg will be administered IV on Day 1 and then every 6 weeks following the administration of nivolumab.
- When both nivolumab and ipilimumab are to be administered on the same day, nivolumab is to be administered first. The second infusion will always be ipilimumab, and will start at least 30 minutes after completion of the nivolumab infusion.
- Treatment beyond initial, investigator-assessed, RECIST 1.1-defined progression is permitted if the subject has investigator-assessed clinical benefit and is tolerating treatment.

Dosing in the N+C arm:

- Nivolumab combined with fluorouracil and cisplatin will be administered IV in a 4-week cycle. Nivolumab 240mg will be administered IV every 2 weeks on Day 1 and Day 15 of each cycle. Fluorouracil 800mg/m²/day will be administered as an IV continuous infusion on Day 1 through Day 5 (for 5 days) and cisplatin 80mg/m² will be administered IV on Day 1 of 4-week cycle.
- When both nivolumab and fluorouracil and cisplatin combination are to be administered on the same day, nivolumab is to be administered first. Infusion of fluorouracil and cisplatin will start at least 30 minutes after completion of the nivolumab infusion.
- Treatment with nivolumab beyond initial, investigator-assessed, RECIST 1.1-defined PD is permitted if the subject has investigator-assessed clinical benefit and is tolerating treatment.

Dosing in the CT arm:

- Fluorouracil 800mg/m²/day will be administered as an IV continuous infusion on Day 1 through Day 5 (for 5 days) and cisplatin 80mg/m² will be administered IV on Day 1 of 4-week cycle.

Follow-up Phase

- Begins when the decision to discontinue a subject from study therapy is made (no further treatment with study therapy).
    - FU1 = 30 days from last dose (± 7 days) or coinciding with the date of discontinuation (± 7 days) if date of discontinuation is greater than 35 days after last dose

  - FU2 = 84 days from FU1 (± 7 days).

- Subjects who discontinue treatment for reasons other than tumor progression will continue to have tumor assessments as per the schedule for the assigned treatment arm until confirmed PD by BICR, withdrawal of consent or lost to follow-up. (For details on discontinuation of tumor assessments, see Section 2.5.)
- Subjects will be followed for all AEs until these resolve, return to baseline or are deemed irreversible. All AEs will be documented for a minimum of 100 days after the last dose of study treatment.
- After completion of the first two follow-up visits, subjects will be followed every 3 months (± 14 days) for survival via in-person visit or phone. Ad hoc survival data requests may be made during the study as well, particularly during database locks. Additional subsequent cancer therapy details such as regimen, setting of the regimen, line of therapy, start date and end date of each regimen, best response to the regimen and date of progression after second line therapy will be collected.
- The PRO questionnaires collection and biomarker sampling will continue as per time and event schedule.

2.2 Treatment Assignment

After the subject's initial eligibility is established and informed consent has been obtained, the subject must be enrolled into the study by entering information into the IRT system to obtain the subject number. Every subject that signs the informed consent form must be assigned a subject number in the IRT system.

Once enrolled in the IRT, subjects that have met all eligibility criteria (including the required tumor tissue received and result obtained by the central laboratory and the pathology report approved by the investigator) will be ready to be randomized through the IRT.

Subjects meeting all eligibility criteria will be randomized in a 1:1:1 ratio to either the N+I arm, or the N+C arm, or the CT arm. Randomization will be stratified by the following factors:

1) PD-L1 status (≥ 1% vs. < 1% [including indeterminate])*
2) Region (East Asia (Japan/Korea/Taiwan [J/K/T]) vs Rest of Asia vs rest of world [RoW])
3) ECOG performance status (0 vs. 1).
4) Number of organs with metastases (≤ 1 vs. ≥ 2)

* During enrolment, the proportion of subjects with or without PD-L1 tumor expression will be monitored, and may be re-assessed in case it does not reflect study assumptions (i.e., subjects with PD-L1 tumor expression ≥ 1% is approximately 50% of all comers).

2.3 Blinding and Unblinding

Not applicable.

2.4 Protocol Amendments

Global amendments incorporated in the protocol with relevant changes are described in Table 2.4-1.

Table 2.4-1: Relevant Protocol Amendments

Amendments	Date of Issue	Summary of Major Changes
Revised Protocol 05	29-Oct-2020	This revised protocol allows for the final PFS analysis to be triggered when 136 PFS events per BICR are observed among the PD-L1 expressing subjects in the chemotherapy arm or when at least 12 months minimum follow up is reached, if the target number of PFS events is unlikely to be reached.
Revised Protocol 04	12-Sep-2018	This revised protocol restricts study entry to participants of previous nivolumab clinical studies where overall survival was listed as a primary or co-primary endpoint. Live /attenuated vaccines were prohibited and the inclusion criterion related to the assessment of renal function was expanded to allow the consideration of measured creatinine clearance. Cisplatin infusion times longer than 120 minutes were allowed if deemed necessary by investigator per local standard of care/local label. PFS2/TSST was added as an exploratory endpoint. The section on biomarker assessments was revised. Program updates were added and internal inconsistencies were corrected.
Revised Protocol 03	02-Feb-2018	This revised protocol removed the procedures for the reinitiation of nivolumab ± ipilimumab treatment after disease progression for up to 1 additional year. In addition, it added clarification to the treatment beyond progression procedures to limit treatment to a maximum duration of 24 months.
Revised Protocol 02	25-Oct-2017	To align the protocol with the latest SmPC, simplify procedures, and provide clarifications in the protocol.
Revised Protocol 01	21-Dec-2016	Incorporates Amendment(s) 02
Amendment 02	21-Dec-2016	Expansion of the esophageal cohort into a 3-arm randomized Phase 3 study in first line squamous esophageal cancer. The study now includes a nivolumab plus chemotherapy arm (fluorouracil and cisplatin) and a chemotherapy alone arm in addition to the existing nivolumab and ipilimumab arm. The gastric cohort was removed.

2.5 Independent Review of Progression

At the time of investigator-assessed initial radiographic progression per RECIST 1.1 in any given subject, the site must request the blinded independent central review (BICR) of progression from the third party radiology vendor.

Tumor assessments for each subject should be submitted to the radiology vendor as they are performed on an ongoing basis. The blinded, independent radiologists will review all available tumor assessments for that given subject and determine if RECIST 1.1 criteria for progression have been met. The independent assessment of whether or not the given subject met RECIST 1.1 criteria for progression will be provided to the site. Subjects whose disease progression is not confirmed centrally will be required to continue tumor assessments (if clinically feasible) according to the protocol-specified schedule. **Subsequent tumor assessments** must be submitted to the third party radiology vendor for subsequent review and **may be discontinued when the investigator and independent radiologists both assess the subject to have met RECIST 1.1 criteria for progression**. In addition, subjects receiving treatment beyond progression must continue tumor assessments until such treatment has been discontinued.

If clinically acceptable, subsequent therapy should begin only after RECIST 1.1 progression has been assessed by BICR. For subjects who start palliative local therapy or subsequent therapy without prior assessment of RECIST 1.1 progression by central review, the Investigator must continue tumor assessments (if clinically feasible) according to the protocol-specified schedule and submit them to the third-party radiology vendor. When RECIST 1.1 progression is assessed by the investigator (whether assessed before or after the start of palliative local therapy or subsequent therapy), the BICR assessment must be requested. Tumor assessments may be discontinued when the independent radiologist assesses the subject to have met RECIST 1.1 criteria for progression.

2.6 Data Monitoring Committee

An independent Data Monitoring Committee (DMC) will be utilized. A DMC will be established to provide oversight of safety and efficacy considerations in CA209648. Additionally, the DMC will provide advice to the sponsor regarding actions the committee deems necessary for the continuing protection of subjects enrolled in the study. The DMC will be charged with assessing such actions in light of an acceptable benefit/risk profile for nivolumab plus ipilimumab and nivolumab plus fluorouracil and cisplatin. The DMC will act in an advisory capacity to BMS and will monitor subject safety and evaluate the available efficacy data for the study. The oncology therapeutic area of BMS has primary responsibility for design and conduct of the study.

DMC will review the safety data from the study when approximately (dependent on the accrual speed) 45 subjects (about 5% of subjects) have been treated and followed for at least 4 weeks.

The DMC will be reviewing safety data every approximately six months and will be also responsible for reviewing and making recommendations at Final PFS analysis and Interim OS analysis. (Section 1).

3 OBJECTIVES

3.1 Primary

- To compare the OS of nivolumab plus ipilimumab to fluorouracil and cisplatin combination in subjects with PD-L1 expression ≥ 1%.
- To compare the OS of nivolumab combined with fluorouracil plus cisplatin to fluorouracil and cisplatin combination in subjects with PD-L1 expression ≥ 1%.
- To compare the PFS of nivolumab plus ipilimumab to fluorouracil and cisplatin combination as assessed by a BICR in subjects with PD-L1 expression ≥ 1%.
- To compare the PFS of nivolumab combined with fluorouracil plus cisplatin to fluorouracil and cisplatin combination as assessed by a BICR in subjects with PD-L1 expression ≥ 1%.

3.2 Secondary

- To compare the OS of nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin to fluorouracil and cisplatin combination in all randomized subjects.
- To compare the PFS of nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin to fluorouracil and cisplatin combination as assessed by a BICR in all randomized subjects.
- To compare the objective response rate (ORR) of nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin to fluorouracil and cisplatin combination as assessed by a BICR in subjects with PD-L1 expression ≥ 1%.
- To compare the ORR of nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin to fluorouracil and cisplatin combination as assessed by a BICR in all randomized subjects.

3.3 Exploratory

- To assess PFS of nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin vs fluorouracil and cisplatin combination as assessed by investigators in subjects with PD-L1 expression ≥ 1% and in all randomized subjects.
- To assess ORR of nivolumab plus ipilimumab and nivolumab combined with fluorouracil plus cisplatin vs fluorouracil and cisplatin combination as assessed by investigators in subjects with PD-L1 expression ≥ 1% and in all randomized subjects.
- To assess Duration of Response (DOR) of nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin vs fluorouracil and cisplatin combination as assessed by BICR and by investigators in subjects with PD-L1 expression ≥ 1% and in all randomized subjects.
- To assess time from randomization to the date of investigator-defined documented second objective disease progression or start of second subsequent therapy or death due to any cause, whichever comes first (PFS2/TSST) of nivolumab plus ipilimumab (Arm A) or nivolumab combined with fluorouracil plus cisplatin (Arm B) vs fluorouracil and cisplatin combination (Arm C) as assessed by investigators in subjects with PD-L1 expression ≥ 1% and in all randomized subjects.
- To assess the overall safety and tolerability of treatment with nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin vs. fluorouracil and cisplatin combination.
- To characterize the PK of nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin.
- To characterize the immunogenicity of nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin.
- To characterize immune correlates of nivolumab plus ipilimumab, nivolumab combined with fluorouracil plus cisplatin, and fluorouracil and cisplatin combination.
- To evaluate the pharmacodynamic activity of nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin in the peripheral blood.

- To explore potential biomarkers associated with clinical efficacy (OS, PFS, and ORR) and/or with incidence of adverse events of nivolumab plus ipilimumab or nivolumab combined with fluorouracil plus cisplatin by analyzing biomarker measures within the tumor microenvironment and periphery (e.g., blood, serum, plasma) in comparison to clinical outcomes.
- To assess the subject's overall health status using the EQ-5D (EQ-5D-3L) index and visual analog scale.
- To assess the subject's cancer-related quality of life using the Functional Assessment of Cancer Therapy-Esophageal (FACT-E) questionnaire and selected components, including the Esophageal Cancer Subscale (ECS) and 7-item version of the FACT-General (FACT-G7).

4 ENDPOINTS

A summary of the efficacy endpoints is presented in Table 4-1.

Table 4-1: Overview of Major Efficacy Endpoints

	N+C vs CT		**N+I vs CT**	
	All PDL1-Expressing Subjects	**All Randomized Subjects**	**All PDL1-Expressing Subjects**	**All Randomized Subjects**
Primary Endpoints	OS, PFS by BICR		OS, PFS by BICR	
Secondary Endpoints	ORR by BICR	OS PFS by BICR ORR by BICR	ORR by BICR	OS PFS by BICR ORR by BICR
Exploratory Endpoints	PFS by investigator ORR by investigator DOR* PFS2/TSST by investigator	PFS by investigator ORR by investigator DOR* PFS2/TSST by investigator	PFS by investigator ORR by investigator DOR* PFS2/TSST by investigator	PFS by investigator ORR by investigator DOR* PFS2/TSST by investigator

* By BICR and by investigator

For the definition of All PD-L1 Expressing Subjects and All Randomized Subjects, see Section 6.3.

4.1 Primary Endpoints

Primary endpoints are OS and PFS in All PD-L1 expressing subjects.

4.1.1 Overall Survival in All PD-L1 Expressing Subjects

OS is defined as the time between the date of randomization and the date of death. For subjects without documentation of death, OS will be censored on the last date the subject was known to be alive.

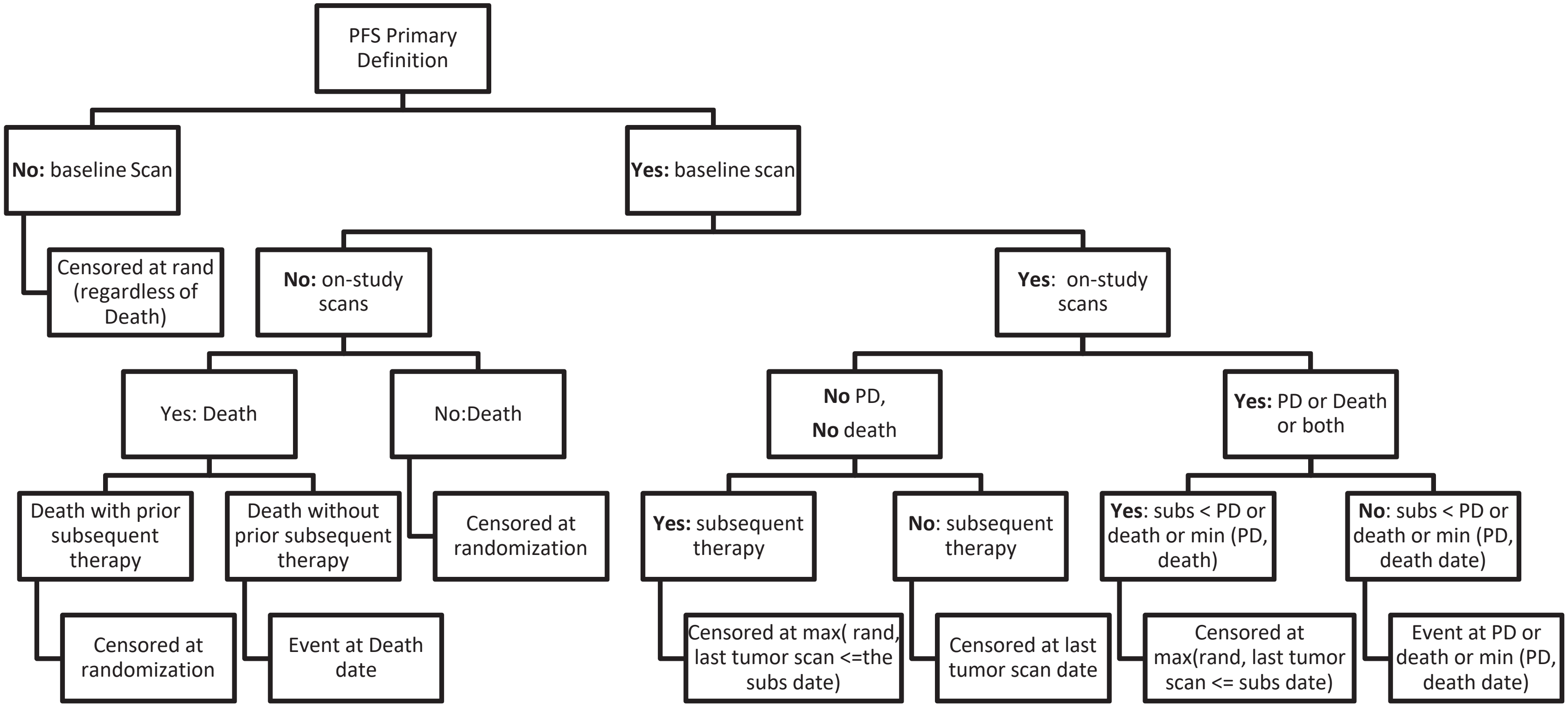
4.1.2 Progression-Free Survival per BICR in All PD-L1 Expressing Subjects

For the purposes of the primary endpoints, PFS is as assessed by BICR in subjects with PD-L1 expressing tumors.

PFS is defined as the time from randomization to the date of the first documented PD per BICR or death due to any cause. Subjects without any baseline scan will be censored at the randomization date (regardless of death). Subjects who did not have any on-study tumor assessments and did not die (or died after initiation of the subsequent anti-cancer therapy) will be censored at the randomization date. Subjects who die without a reported prior PD per BICR (and die without start of subsequent therapy) will be considered to have progressed on the date of death. Subjects who did not have documented PD per BICR per RECIST1.1 criteria and who did not die, will be censored at the date of the last evaluable tumor assessment on or prior to initiation of the subsequent anti-cancer therapy. Subjects who started any subsequent anti-cancer therapy without a prior reported PD per BICR will be censored at the last tumor assessment on or prior to initiation of the subsequent anti-cancer therapy.

See Figure 4.1.2-1 for the graphical representation of this definition.

Figure 4.1.2-1: Graphical Representation of the Primary PFS Definition


PFS Primary Definition
No: baseline Scan
Yes: baseline scan
Censored at rand (regardless of Death)
No: on-study scans
Yes: on-study scans
Yes: Death
No:Death
No PD, No death
Yes: PD or Death or both
Death with prior subsequent therapy
Death without prior subsequent therapy
Censored at randomization
Yes: subsequent therapy
No: subsequent therapy
Yes: subs < PD or death or min (PD, death)
No: subs < PD or death or min (PD, death date)
Censored at randomization
Event at Death date
Censored at max(rand, last tumor scan <=the subs date)
Censored at last tumor scan date
Censored at max(rand, last tumor scan <= subs date)
Event at PD or death or min (PD, death date)

rand = randomization, subs = subsequent therapy

4.2 Secondary Endpoints

4.2.1 Overall Survival in All Randomized Subjects

Overall survival in All Randomized subjects is defined the same way as for the primary endpoint (for subjects with PD-L1 expressing tumors).

4.2.2 Progression-Free Survival per BICR in All Randomized Subjects

Progression-free survival in All Randomized subjects is defined the same way as for the primary endpoint (for subjects with PD-L1 expressing tumors).

4.2.3 Objective Response Rate per BICR in All Randomized Subjects

Objective response rate as assessed by BICR in subjects with PD-L1 expressing tumors and in All Randomized subjects are secondary endpoints for this study.

It is defined as the number of subjects with a best overall response (BOR) of CR or PR divided by the number of randomized subjects in the population for each treatment group. BOR is defined as the best response designation as determined by BICR per RECIST 1.1, recorded between the date of randomization and the date of objectively documented progression or the date of subsequent anti-cancer therapy (including tumor-directed radiotherapy and tumor-directed surgery), whichever occurs first. For subjects without documented progression or subsequent anti-cancer therapy, all available response designations will contribute to the BOR determination. For subjects who continue treatment beyond progression, the BOR will be determined based on response designations recorded up to the time of the initial RECIST 1.1-defined progression.

4.3 Exploratory Endpoints

4.3.1 Efficacy Exploratory Endpoints

<u>Progression-Free Survival per Investigator</u>

Progression-free Survival as assessed by investigator in subjects with PD-L1 expressing tumors and in All Randomized subjects is defined similarly as the corresponding (primary) endpoints as assessed by BICR - except that only tumor assessments by investigator will be taken into account.

<u>Objective Response Rate per Investigator</u>

Objective response rate as assessed by investigator in subjects with PD-L1 expressing tumors and in All Randomized subjects is defined similarly as the corresponding (primary) endpoints as assessed by BICR - except that only tumor assessments by investigator will be taken into account.

<u>Duration of Response</u>

Duration of Response (as assessed by BICR and as assessed by investigator) is defined as the time between the date of first documented response (CR or PR) to the date of the first disease progression, per RECIST 1.1 or death due to any cause, whichever occurs first. For subjects who neither progress nor die, the duration of objective response will be censored at the same time they were censored for the primary definition of PFS. DOR will be evaluated only for subjects whose BOR is CR or PR.

Survival rate, PFS rate and duration of Stable Disease (SD) are not positioned as objectives per protocol, but will be evaluated.

Survival Rate

Overall survival rate at different timepoints, e.g. at 18, 24, and 36 months is defined as the probability that a subject is alive at 18, 24, and 36 months, respectively, following randomization. Overall survival rates at other timepoints are defined similarly.

PFS Rate

PFS rate at different timepoints, e.g. at 6, 12, 18 and 24 months is defined as the probability that a subject has not progressed and is alive at 6, 12, 18 and 24 months, respectively, following randomization. PFS rates at other timepoints are defined similarly.

Duration of Stable Disease

Duration of SD is defined as the time between the randomization date and the date of the first documented tumor progression (per RECIST 1.1) or death due to any cause, whichever occurs first. Censoring rules will be the same as for DOR analysis. It will be evaluated only for subjects whose BOR is SD.

PFS2/TSST

PFS on next-line therapy (PFS2) / Time To Second Subsequent Therapy (TSST) is defined as the time from randomization to the documented progression (radiographic or clinical progression) per investigator assessment after the next line of systemic therapy or start of the second next line systemic therapy or death from any cause, whichever occurs first. Subjects who were alive and without the second progression and without the second next line systemic therapy will be censored at last known alive date.

The following censoring rules will be applied for PFS2/TSST:

- Subjects who did not receive any subsequent anti-cancer systemic therapy:
  - Subjects who died, the death date is the event date;
  - Else the subject’s PFS2/TSST is censored at the last known alive date.
- Subjects who received at least one subsequent anti-cancer systemic therapy:
  - Subjects who had a disease progression after the start of the first subsequent anti-cancer systemic therapy, this disease progression date is the event date;
  - Else if a subject died or start of the second next line systemic therapy, the date of min (death, start date of the second next line systemic therapy) is the event date;
  - Else the subject’s PFS2/TSST is censored at the last known alive date.

4.3.2 Safety Exploratory Endpoints

The assessment of safety will be based on the incidence of adverse events (AEs), serious adverse events (SAEs), adverse events leading to discontinuation, adverse events leading to dose modification, select adverse events (select AEs) for EU/ROW Submissions, immune-mediated

AEs (IMAEs) for US Submission, other events of special interest (OEOSI), and deaths. The use of immune modulating concomitant medication will be also summarized. In addition clinical laboratory tests, and immunogenicity (i.e. development of anti-drug antibody) will be analyzed.

4.3.3 Outcomes Research Exploratory Endpoints

EQ-5D-3L

Subjects' reports of general health status will be assessed using the EuroQoL Group's EQ-5D-3L. EQ-5D-3L essentially has 2 components: the descriptive system and the visual analogue scale (VAS).

The instrument's descriptive system consists of 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 3 levels, reflecting "no health problems," "moderate health problems," and "extreme health problems." A dimension for which there are no problems is said to be at level 1, while a dimension for which there are extreme problems is said to be at level 3. Thus, the vectors 11111 and 33333 represent the best health state and the worst health state, respectively, described by the EQ-5D-3L. Altogether, the instrument describes 3^5 = 243 health states. Empirically derived weights can be applied to an individual's responses to the EQ-5D-3L descriptive system to generate an index measuring the value to society of his or her current health. Such preference-weighting systems have been developed for Japan, UK, US, Spain, Germany, and numerous other populations. For this study, EQ-5D-3L utility index values will be computed using a scoring algorithm based on the United Kingdom Time-Trade-Off (UK TTO) value set[1]

In addition, the EQ-5D-3L includes a VAS, which allows respondents to rate their own current health on a 101-point scale ranging from 0=”worst imaginable” health to 100="best imaginable" health state [2].

A change from baseline of 0.08 for the EQ-5D-3L utility index score and of 7 for the EQ-5D-3L VAS are considered minimally important differences for the EQ-5D-3L[3].

FACT-E

The FACT-E questionnaire and selected components, including the FACT-G7 and ECS, will be used to assess the effects of underlying disease and its treatment on health-related quality of life (HRQL) for patients.

The FACT-E includes the 27-item FACT-General (FACT-G) generic cancer-related core measure, to assess symptoms and treatment-related effects impacting physical well-being (PWB; seven items), social/family well-being (SWB; seven items), emotional well-being (EWB; six items), and functional well-being (FWB; seven items). Seven of these items comprise the FACT-G7, an abbreviated version of the FACT-G that provides a rapid assessment of general HRQL in cancer patients.

In addition to the FACT-G, the FACT-E also includes a 17-item disease-specific Esophagus Cancer Subscale (ECS) that assesses concerns related to swallowing, vocalization, breathing, dry mouth, eating, disrupted sleep due to coughing, stomach pain, and weight loss. Each FACT-E item is rated on a five-point scale ranging from 0 (not at all) to 4 (very much). The FACT-E also includes the single-item GP5 item, which assesses the overall bother associated with the side effects of treatment.

Scores for the PWB, FWB, SWB, and EWB subscales can be combined to produce a FACT-G total score for each Cohort, which provides an overall indicant of generic HRQL. The FACT-G and ECS scores can be combined to produce a total score for the FACT-E, which provide a composite measure of general and targeted HRQL. The PWB, FWB and ECS scores can be combined to produce a Trial Outcome Index (TOI). Higher scores on all scales indicate reduced symptoms and better HRQL. The full FACT-E will be administered to subjects during the on-treatment phase and at follow up visits 1 and 2. However, to minimize subject response and administrative burden, only the FACT-G7 and ECS will be administered during the survival follow-up phase.

All questionnaires completed at baseline and on-study will be assigned to a time-point according to the windowing criteria in Table 4.3.3-1 and included in the analysis. In case a subject has two on-study assessments within the same window, the assessment closest to the time-point will be used. And, in the case of two assessments at a similar distance to the time-point, the latest one will be chosen. In the event where the subject has no assessment at all in a specific window, the observation will be treated as missing for that time-point.

Table 4.3.3-1: Time Windows for EQ-5D-3L and FACT-E Assessments

Nominal Time-Point	Time Window
all treatment groups	
Baseline[a]	On or prior to first dose on Day 1[b]
Week 3	Nominal Day 15 (Day 2 through Day 22, inclusive)
Week 5	Nominal Day 29 (+7 days/-6 days, inclusive)
Every Week 7	Nominal Day 43 (+21 days/-6 days, inclusive)
Every 6 weeks thereafter	Nominal Day (+21 days/-20 days, inclusive)
Follow-Up 1[a, c, d]	Post last dose and within 72 days after last dose Date (If date of discontinuation is within 35 days after last dose)
	Post last dose and within 42 days of the date of discontinuation (If date of discontinuation is after 35 days after last dose)
Follow-Up 2[a, c]	Post 72 days of last dose and within 159 days after last dose Date (If date of discontinuation is within 35 days after last dose)
	Post 42 days of the date of discontinuation and within 129 days of the date of discontinuation (If date of discontinuation is after 35 days after last dose)
Survival Visit i[e]	**Survival Follow-Up 1**

Table 4.3.3-1: Time Windows for EQ-5D-3L and FACT-E Assessments

Nominal Time-Point	Time Window
	Post 159 days of last dose and within 249 days after last dose Date (If date of discontinuation is within 35 days after last dose)
	Post 129 days of the date of discontinuation and within 219 days of the date of discontinuation (If date of discontinuation is after 35 days after last dose)
	Subsequent Survival Follow-Up visits i=2, 3....
	Post 249+(i-2)*90 days of last dose and within 249+(i-1)*90 days after last dose Date (If date of discontinuation is within 35 days after last dose)
	Post 219+(i-2)*90 days of the date of discontinuation and within 219+(i-1)*90 days of the date of discontinuation (If date of discontinuation is after 35 days after last dose)

a For subjects randomized and not treated: Baseline will be defined as "On or prior to randomization date". Any assessment after the randomization date will be considered as follow-up visit. "Last dose" date reference in Follow-Up 1 and Follow-Up 2 Time Window derivation will be replaced by randomization date.

b First dose date considers as Day 1 of Week 1. So Week 3 corresponds to 2 weeks of Treatment.

c These Time Window definitions are based on the definition of Follow-up visits definition in the study protocol (See Section 5, Table 5.1-5).

d Any assessment post last dose will be slotted to Follow-Up visit.

e Survival visits = every 3 months (i.e. 90 days assumed in calculation) from Follow-Up 2.

4.3.4 *Pharmacokinetics Exploratory Endpoints*

PK will be measured using serum concentration-time data.

4.3.5 *Immunogenicity Exploratory Endpoints*

Serum samples collected will be analyzed by a validated immunogenicity assay. Selected serum samples may be analyzed by an exploratory orthogonal method that measures Nivolumab and Ipilimumab.

In addition, ad hoc serum samples designated for pharmacokinetic or biomarker assessments may also be used for immunogenicity analysis if required (e.g., insufficient volume for complete immunogenicity assessment or to follow up on suspected immunogenicity related AE).

Further details on immunogenicity background and rationale, definitions, population for analyses and endpoints are described in APPENDIX 4.

4.3.6 *Biomarker Exploratory Endpoints*

The source of the biomarker data in this section will be from the lab.

Tumor Cell PD-L1 expression

For subjects with an available tumor biopsy specimen(s), the following will be considered:

- Tumor cell PD-L1 expression is defined as the percent of tumor cell membrane staining in a minimum of 100 evaluable tumor cells per validated Dako PD-L1 IHC assay unless otherwise specified. This is referred as quantifiable PD-L1 expression. If the PD-L1 staining could not be quantified, it is further classified as:
- Indeterminate: Tumor cell membrane staining hampered for reasons attributed to the biology of the tumor biopsy specimen and not because of improper sample preparation or handling
- Not evaluable: Tumor biopsy specimen was not optimally collected or prepared (e.g. PD-L1 expression is neither quantifiable nor indeterminate) (In this study, we do not expect to have such patients.)

PD-L1 status will consider three cutoffs 1%, 5% or 10% and PD-L1 status categories used for these analyses are:

- Each PD-L1 status subgroup
  - PD-L1 ≥ 1%: ≥ 1% PD-L1 expression
  - PD-L1 < 1%: < 1% PD-L1 expression
  - PD-L1 ≥ 5%: ≥ 5% PD-L1 expression
  - PD-L1 < 5%: < 5% PD-L1 expression
  - PD-L1 ≥ 10%: ≥ 10% PD-L1 expression
  - PD-L1 < 10%: < 10% PD-L1 expression
- PD-L1 status indeterminate, not evaluable or missing subgroup

Subjects without an available tumor biopsy specimen for PD-L1 evaluation will be considered as PD-L1 expression missing.

PD-L1 by Combined Positive Score (CPS)

PD-L1 by combined positive score (CPS) is defined as the number of PD-L1 staining cells (tumor cells, lymphocytes, macrophages) divided by the total number of viable tumor cells, multiplied by 100. Exploratory PD-L1 by CPS data is generated by rescoring PD-L1 slides previously stained at the time of randomization. The PD-L1 by CPS status subgroups will consider three cutoffs, 1, 5 and 10 and CPS categories used for these analyses are:

- CPS status subgroups
  - PD-L1 by CPS ≥ 1
  - PD-L1 by CPS < 1
  - PD-L1 by CPS ≥ 5
  - PD-L1 by CPS < 5
  - PD-L1 by CPS ≥ 10
  - PD-L1 by CPS < 10
- PD-L1 by CPS indeterminate, not evaluable or missing subgroup

Microsatellite Instability (MSI) Status

MSI refers to instability at the microsatellite tracts: gain or loss of nucleotides from DNA elements composed of short repeating motifs. MSI is determined by using whole exome sequencing. MSI status will be categorized to:

- MSI-H (High levels of Microsatellite Instability)
- MSS (Microsatellite Stable)
- unknown including not tested and invalid

5 SAMPLE SIZE AND POWER

It is by now established that time-to-events endpoints in I-O clinical trials tend to display specific characteristics. First, there may be long-term survival benefits in patients treated with immunotherapy - observed as a long lasting plateau towards the tail of the survival curve ("cure"). Second, some results also suggested a delayed effect - observed as late separation of survival curves between experimental and control arms. Both long-term survival and delayed onset of benefit may be linked to the mechanisms of action of I-O drugs. As these phenomena are expected to be observed in the current disease setting as well, a piecewise mixture cure rate model was used for the design setup.

Sample size calculations assume that the prevalence of subjects with PD-L1 tumor expression level ≥ 1% is approximately 50%. (During enrolment, the proportion of subjects with or without PD-L1 tumor expression will be monitored, and may be re-assessed in case it does not reflect study assumptions.)

The sample size is based on the primary objectives, i.e., on the comparisons of the PFS/OS distributions of subjects with PD-L1 expressing tumors between those who were randomized to receive nivolumab plus ipilimumab vs those randomized to receive chemotherapy, and between those who were randomized to receive nivolumab plus chemotherapy vs those randomized to receive chemotherapy.

For both experimental arms (N+I, N+C), the same OS distributions and the same PFS distributions are assumed (see Table 5-1). As a result, for each of the N+I vs CT and N+C vs CT comparisons:

- 250 PFS events in approximately 313 subjects with PD-L1 expressing tumors will provide approximately 90% power to detect an average Hazard Ratio (HR) of 0.62 with a Type I error of 1.5% (two-sided);
- 250 OS events in approximately 313 subjects with PD-L1 expressing tumors will provide approximately 90% power to detect an average HR of 0.6 with a Type I error of 1% (two-sided).

To have approximately 313 randomized subjects with PD-L1 expressing tumors for each comparison, approximately 470 subjects with PD-L1 expressing tumors need to be randomized in a 1:1:1 ratio in the 3 arms. Which therefore translates to a total of approximately 939 subjects (with any PD-L1 result) to be randomized in a 1:1:1 ratio to the N+I or N+C or CT arms. Assuming a piecewise constant accrual rate, it is estimated that these 939 subjects will be accrued within 29 months.

Although for the comparison of both experimental arms with the control arm, the same treatment effect is assumed, observed treatment effects may vary. Therefore, the events of interest (OS, PFS) observed in the CT arm *only* will be used for determining the timing of the interim and final efficacy analyses. (Note that the number of events in the CT arm will be monitored by an independent statistical group.) Final PFS analysis is planned when 136 events are observed among the PD-L1 expressing subjects in the CT arm. This is expected to be reached after approximately 33 months. If the planned number of PFS events per BICR is unlikely to be reached for any unforeseen reasons, the final PFS analysis may occur when at least 12 months minimum follow-up is reached. Should this scenario occur, corresponding to PFS events of 110 and 121, the power would be 80% and 85% respectively (details of this scenario are discussed in APPENDIX 7.

Final OS analysis is planned when 140 events are observed among the PD-L1 expressing subjects in the CT arm. This is expected to be reached after approximately 49 months (at the time of the final PFS analysis, a formal interim analysis for OS will be conducted).

Details of the sample size calculations for the subjects whose tumors express PD-L1 are provided in Table 5-1.

Table 5-1: Subjects whose Tumors Express PD-L1 - Summary of Sample Size Parameters and Schedule of Analyses (N+I vs CT and N+C vs CT)

	OS	PFS
# of randomized subjects	≈313 (total in the 3 arms: ≈470)	
Hypothesized delayed effect	3 months	1 month
Hypothesized cure rate in experimental arms (after delay)	15%	0
Hypothesized HR after delayed effect, in non-cure subjects	0.65	0.55
Hypothesized median in control arm	9 months[a]	4 months
Significance level	0.01	0.015
FINAL ANALYSIS b		
Criteria for time of LPLV	140 events are observed among the PD-L1 expressing subjects in the CT arm	136 events are observed among the PD-L1 expressing subjects in the CT arm
Projected time of LPLV (from first patient randomized)	49 months	33 months
Projected # of events[c]	250	250
Significance level	[0.009][d]	0.015

Table 5-1: Subjects whose Tumors Express PD-L1 - Summary of Sample Size Parameters and Schedule of Analyses (N+I vs CT and N+C vs CT)

	OS	**PFS**
Power	90%	90%
Overall HR	0.6	0.62
Median in experimental arm	14.4 months	6.4 months
Critical HR[e] / Minimal difference in median[f]	0.72 / 3.5 months	0.73 / 1.4 months
INTERIM ANALYSIS for OS b		
Criteria for time of LPLV	At the time of PFS final analysis	N/A
Projected # of events[c]	≈175 (70% of all events)	N/A
Significance level	[0.002][d]	N/A
Probability of crossing boundary	29%	N/A
Hypothesized overall HR	0.67	N/A
Critical HR[e] / Minimal difference in median[f]	0.62 / 5.6 months	N/A

[a] Asia: 10 months, RoW: 6 months

[b] Results based on simulations (except: Criteria for time of LPLV)

[c] Projected # of events for one comparison (i.e. for N+I vs CT; and for N+C vs CT)

[d] For OS, significance levels will be recalculated based on the actual number of deaths at interim analysis

[e] Largest observed HR at which a statistically significant difference would be observed

[f] Difference in median, corresponding to a minimal clinically significant effect size

Analyses on OS and PFS in all randomized subjects will be carried out at the time of the primary analysis (in subjects whose tumors express PD-L1); OS and PFS in all randomized subjects are expected to be mature by that time. They will be tested only if significance level is passed on them. Details of the related power calculations are provided in Table 5-2.

Table 5-2: All Randomized Subjects - Summary of Power Calculation Parameters and Schedule of Analyses (N+I vs CT and N+C vs CT)

	OS	**PFS**
# of randomized subjects	626 (total in the 3 arms: 939)	
Hypothesized delayed effect	3 months in PD-L1+, 4 months in PD-L1-	1 month in PD-L1+, 2 months in PD-L1-
Hypothesized cure rate in experimental arms (after delay)	15% in PD-L1+, 10% in PD-L1-	0

Table 5-2: All Randomized Subjects - Summary of Power Calculation Parameters and Schedule of Analyses (N+I vs CT and N+C vs CT)

	OS	PFS
Hypothesized HR after delayed effect, in non-cure subjects	0.65 in PD-L1+, 0.85 in PD-L1-	0.55 in PD-L1+, 0.75 in PD-L1-
Hypothesized median in control arm	9 months[a]	4 months
Significance level[b]	0.01	0.015
FINAL ANALYSIS [c]		
Criteria for time of LPLV	At the time of the PD-L1+ LPLV	At the time of the PD-L1+ LPLV
Projected # of events[d]	≈514	≈512
Significance level[b]	[0.009][e]	0.015
Power[f]	94%	90%
Hypothesized overall HR	0.68	0.72
Median in experimental arm	12.3 months	5.5 months
Critical HR[g] / Minimal difference in median[h]	0.80 / 2.3 months	0.81 / 0.9 months
INTERIM ANALYSIS for OS [b]		
Criteria for time of LPLV	At the time of PFS final analysis	N/A
Projected # of events[d]	≈362 (70.4% of all events)	N/A
Significance level[b]	[0.002] [e]	N/A
Probability of crossing boundary[f]	33%	N/A
Hypothesized overall HR	0.75	N/A
Critical HR [g] / Minimal difference in median[h]	0.74 / 3.1 months	N/A
Accrual rate per month	Gradually increasing accrual rates (max = 45 subjects per months)	
Accrual Duration	29 months	

[a] Asia: 10 months, RoW: 6 months

[b] In case the significance level from the corresponding primary endpoint is passed. (Note that endpoint-specific [i.e. initially allocated] significance level is 0.)

[c] Results based on simulations (except: Criteria for time of LPLV)

[d] Projected # of events for one comparison (i.e. for N+I vs CT; and for N+C vs CT)

[e] For OS, significance levels will be recalculated based on the actual number of deaths at interim analysis

[f] Not accounting for hierarchy

[g] Largest observed HR at which a statistically significant difference would be observed

[h] Difference in median, corresponding to a minimal clinically significant effect size

Simulations were conducted in R-v3.1.3.

6 STUDY PERIODS, TREATMENT REGIMENS AND POPULATIONS FOR ANALYSES

6.1 Study Periods

- Baseline period:
  - Baseline evaluations or pre-treatment events will be defined as evaluations or events that occur before the date and time of the first dose of study treatment. Evaluations (laboratory tests, vital signs, and biomarkers (excluding PD-L1)) on the same date and time of the first dose of study treatment will be considered as baseline evaluations. For the stratification factors and PD-L1 from lab, baseline evaluations will be defined as evaluations that occur before the date of the randomization. Events (AEs) on the same date and time of the first dose of study treatment will not be considered as pre-treatment events.
  - In cases where the time (onset time of event or evaluation time and dosing time) is missing or not collected, the following definitions will apply:
    - ♦ Pre-treatment AEs will be defined as AEs with an onset date prior to but not including the day of the first dose of study treatment;
    - ♦ Baseline evaluations (laboratory tests, vital signs and biomarkers) will be defined as evaluations with a date (and time if collected) on or prior to the date of first dose of study treatment.
  - If there are multiple valid observations in the baseline period, then the latest non-missing observation will be used as the baseline in the analyses. If multiple observations exist on the latest collection date (and time if collected), the record with the latest data entry date and time will be used. If multiple observations exist on the latest collection date (and time if collected) and data entry date and time, then the first observation is used as baseline, unless otherwise specified.
    - ♦ For PD-L1, non-missing is identified as those with quantifiable test result. After applying the rule above, if there are no records with a quantifiable test result, then select those with indeterminate result (“INDETERMINATE”). If there are no records with indeterminate test result, then select those with unavailable result (“NOT EVALUABLE”). If there are no records with unavailable test result, then select those which are not reported or not available result (all other records).
    - ♦ For Anti-Drug Antibody (ADA), the baseline record of Nivolumab and Ipilimumab immunoglobulin (IMG) evaluation must be less than the date and time of the first Nivolumab and Ipilimumab dose date and time.
- Post baseline period:
  - On-treatment AEs will be defined as AEs with an onset date and time on or after the date and time of the first dose of study treatment (or with an onset date on or after the day of first dose of study treatment if time is not collected or is missing). For subjects who are off

study treatment, AEs will be included if event occurred within a safety window of 30 days (or 100 days depending on the analysis) after the last dose of study treatment. No "subtracting rule" will be applied when an AE occurs both pre-treatment and post-treatment with the same preferred term and grade.

- On-treatment evaluations (laboratory tests, and vital signs) will be defined as evaluations taken after the day (and time, if collected and not missing) of first dose of study treatment. For subjects who are off study treatment, evaluations should be within a safety window of 30 days (or 100 days depending on the analysis) after the last dose of study treatment.

6.2 Treatment Regimens

The treatment group "**as randomized**" will be retrieved from the IRT system

- Arm A in the IRT system: Experimental arm N+I
- Arm B in the IRT system: Experimental arm N+C
- Arm C in the IRT system: Control arm CT

The treatment group "**as treated**" will be, in general, the same as the arm randomized by IRT. However, if a subject received the incorrect drug for **the entire period** of treatment, the subject's treatment group will be defined as the incorrect drug the subject actually received.

6.3 Populations for Analyses

The following definitions of populations will be applicable for subjects whose tumors express PD-L1 and also for subjects regardless of PD-L1 expression, except otherwise specified.

- All Enrolled Subjects: All subjects who signed an informed consent form and were registered into the IRT
- All Randomized Subjects: All enrolled subjects who were randomized to any treatment arm in the study
- All PD-L1 Expressing Subjects: All randomized subjects with tumors expressing PD-L1 (≥ 1%)
- All PD-L1 Negative Subjects: All randomized subjects with tumors PD-L1 expression < 1%
- All Treated Subjects: All randomized subjects who received at least one dose of study drug during the study
- All Treated PD-L1 Expressing Subjects: All treated subjects with tumors expressing PD-L1 (≥ 1%)
- PK Subjects: All randomized subjects with available serum time-concentration data.
- Outcome Research subjects: All randomized subjects who have an assessment at screening/baseline and at least 1 follow-up assessment while on treatment.
- Immunogenicity subjects: All randomized subjects who have an assessment at screening/baseline and at least 1 follow-up assessment

- Biomarker subjects: All randomized subjects with available biomarker data. (PD-L1 expression status and other assays).
  - All PD-L1 tested subjects: All enrolled subjects who had a tumor biopsy specimen available for assessment of PD-L1 expression.
  - All randomized PD-L1 subjects: All tested PD-L1 subjects who are randomized to the study and with baseline PD-L1 expression (includes non-evaluable [should be none per protocol] and indeterminate)
  - All evaluable PD-L1 subjects: All PD-L1 tested subjects who are randomized to the study and with quantifiable baseline PD-L1 expression

Unless otherwise specified, the safety analyses will include all treated subjects.

Unless otherwise specified, the efficacy analyses will include all randomized subjects.

7 STATISTICAL ANALYSES

7.1 General Methods

Unless otherwise noted, discrete variables will be tabulated by the frequency and proportion of subjects falling into each category, grouped by treatment (with total, as needed). Percentages given in these tables will be rounded and, therefore, may not always sum to 100%. Continuous variables will be summarized by treatment group (with total, as needed) using the mean, standard deviation, median, minimum and maximum values, unless specified otherwise.

Unless mentioned otherwise, **outputs will be provided separately for the N+I vs CT analyses and for the N+C vs CT analyses.**

All listings will include PD-L1 status as well.

This study has 4 stratification factors. Each factor has 2 to 3 levels, with 24 strata in total.

- region (J/K/T vs. rest of Asia vs. RoW)
- ECOG performance status (0 vs. 1)
- number of organs with metastases (< 1 vs. ≥ 2)
- PD-L1 expression level (≥ 1% vs. < 1% or indeterminate)

Anticipating the study population may not be equally distributed across levels in a given stratification factor, it is likely that the number of subjects in some strata may be small. With three treatment arms in this study, the stratum size with less than 20 subjects would likely cause unreliable estimates in the stratified analyses[4, 5, 6, 7].

Therefore the following rule will be used when applicable.

If the number of subjects in at least one stratum is less than 20 across all randomized subjects, the planned stratified analyses will be modified per the following steps.

- The marginal distribution of each level of a given stratification factor will be examined.
- The stratification factor with the lowest prevalence in a level will be excluded from the planned stratified analyses.

- After the above step, if there is still at least one stratum with less than 20 subjects, the above steps will be repeated for the rest of stratification factors until all the stratum size is at least 20.

7.1.1 General Methods: Efficacy

Protection of Type I Error

Family-wise Type I error will be protected in the strong sense across all primary and secondary endpoints. The p-values from sensitivity analyses for efficacy endpoints are for descriptive purpose only and not adjusted for multiplicity.

Stratification Factors

For all stratified analyses, stratification factors will be based on the randomization factors (Section 2.2). Specifically, stratification factors used for analyses are as follows:

- for Subjects Expressing PD-L1:
  - region (J/K/T vs rest of Asia vs. RoW)
  - ECOG performance status (0 vs. 1)
  - number of organs with metastases (≤ 1 vs. ≥ 2)
- for All Randomized Subjects
  - region (J/K/T vs. rest of Asia vs. RoW)
  - ECOG performance status (0 vs. 1)
  - number of organs with metastases (≤ 1 vs. ≥ 2)
  - PD-L1 expression level (≥ 1% vs. < 1% or indeterminate)

Unless otherwise specified, stratification factors are used as entered in the IRT.

In the situation that the stratification factors are adjusted as specified per Section 7.1 General Methods, the adjustment will apply to all the stratified analyses in this document.

General Methods for Time-to-event Variables

Time to event distribution (e.g. OS, PFS, DOR) will be estimated using Kaplan-Meier (KM) techniques.

Median survival time along with 95% or adjusted confidence interval (CI) will be constructed based on a log-log transformed CI for the survivor function S(t)[8],[9]. Rates at fixed timepoints (e.g. OS at 6 months) will be derived from the KM estimate and corresponding CI will be derived based on Greenwood formula[10] for variance derivation and on log-log transformation applied on the survivor function S(t)[11].

Unless otherwise specified, the stratified log-rank test will be performed to test the comparison between time to event distributions. Unless otherwise specified, the stratified HR between

2 treatment groups along with CI will be obtained by fitting a stratified Cox model with the treatment group variable as unique covariate.

General Methods for Bivariate Variables

The difference in rates between the two treatment arms along with their two-sided 95% CI will be estimated using the following Cochran-Mantel-Haenszel (CMH) method of weighting[12], adjusting for the stratification factors:

$$\hat{\theta} = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i} \sim N\left[\theta, \frac{\sum_i w_i^2 \left[\frac{p_{ix}(1-p_{ix})}{n_{ix}-1} + \frac{p_{iy}(1-p_{iy})}{n_{iy}-1}\right]}{\left(\sum_i w_i\right)^2}\right]$$

where $\hat{\theta} = p_{ix} - p_{iy}$ is the rate difference of the ith stratum, $w_i = \frac{n_{ix} n_{iy}}{n_{ix} + n_{iy}}$, and n_{ix} and n_{iy} are the number of subjects randomized to treatments x and y, respectively, in the ith stratum. Associated odds-ratio will be derived.

7.2 Study Conduct

7.2.1 Accrual

The accrual pattern will be summarized per region, country, investigational site and per month for all enrolled subjects. No separate outputs are provided for N+I vs. CT and for N+C vs. CT. Randomization date, first dosing date, country, investigational site will be presented in a by subject listing of accrual.

7.2.2 Relevant Deviations

The relevant Protocol Deviations will be summarized for all randomized subjects and for all PD-L1 expressing subjects, by treatment group and overall. The following programmable deviations from inclusion and exclusion criteria will be considered as relevant protocol deviations. Non-programmable relevant eligibility and on-treatment protocol deviations, as well as important (both programmable and nonprogrammable) eligibility and on-treatment protocol deviations will be reported via the Clinical Monitoring system (i.e. clinSIGHT).

At Entrance

- Subjects without squamous cell carcinoma or adenosquamous cell carcinoma of esophagus
- Subjects with no unresectable advanced, recurrent or metastatic ESCC
- Subjects who have received prior systemic therapy for advanced or metastatic disease (adjuvant, neoadjuvant, and definitive is allowed)
- Subject with baseline ECOG performance status > 1.
- Subjects without any measurable disease at baseline.

- Subjects without any tumor cell PD-L1 result. (“Indeterminate” is not a deviation.)

<u>On-treatment</u>

- Subjects receiving concurrent anti-cancer therapy (chemotherapy, hormonal therapy, immunotherapy, curative surgery, non-palliative radiation therapy, standard or investigational agents for treatment of gastric cancer).
- Subjects treated differently as randomized (subjects who received the wrong treatment, excluding the never treated).

A subject listing will also be produced.

7.3 Study Population

7.3.1 Subject Disposition

The total number of subjects enrolled (randomized or not randomized) will be presented along with the reason for not being randomized.

Number of subjects randomized but not treated along with the reason will be tabulated by treatment group as randomized. This analysis will be performed on all randomized subjects and on all PD-L1 expressing subjects.

Number of subjects who discontinued study treatment along with corresponding reason will be tabulated by treatment group as treated. This analysis will be performed on all treated subjects and on treated PD-L1 expressing subjects.

A subject listing for all randomized subjects will be provided showing the subject’s randomization date, first and last dosing date, off study date and reason for going off-study. A subject listing for subjects not randomized will also be provided, showing the subject’s race, gender, age, consent date and reason for not being randomized.

7.3.2 Demographics and Other Baseline Characteristics

The following baseline characteristics will be summarized by treatment arm as randomized. All baseline presentations identify subjects with missing measurements. This analysis will be performed on all randomized subjects and on PD-L1 expressing subjects. Listings will also be provided.

- Age (descriptive statistics)
- Age category (< 65, ≥ 65 and < 75, ≥ 75 and < 85, ≥ 85; ≥ 75, ≥ 65)
- Gender (male, female)
- Race (White, Black or African American, Asian Indian, Chinese, Japanese, Asian Other, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Other)
- Region (J/K/T, Rest of Asia, RoW) (*stratification factor*)
- Region (Asia, Non-Asia)
- ECOG PS (*stratification factor*)

- Weight (descriptive statistics)
- Disease stage at initial diagnosis (Stage IA, Stage IB, Stage IIA, Stage IIB, Stage IIIA, Stage IIIB, Stage IIIC, Stage IV)
- Histologic grade at initial diagnosis (Gx, G1, G2, G3, G4, unknown)
- Histological classification at initial diagnosis (squamous cell carcinoma, adenosquamous cell carcinoma, other)
- Location at initial diagnosis (upper thoracic, middle thoracic, lower thoracic, gastroesophageal junction)
- TNM classification at initial diagnosis:
  - Tumor (Tx, T0, Tis, T1, T2, T3, T4, unknown)
  - Nodes (Nx, N0, N1, N2, N3, unknown)
  - Metastasis (Mx, M0, M1, unknown)
- Disease status at current diagnosis (recurrent - loco-regional, recurrent - distant, de-novo metastatic**, unresectable advanced)
- Smoking status (current/former, never smoker, unknown)
- Alcohol use (current/former, never, unknown)
- All lesions at Baseline (Investigator and BICR): sites of disease, number of disease sites per subject, number of subjects with no measurable lesion, number of organs with metastases
- Target Lesions (Investigator and BICR): Presence of target lesions, site of target lesion, sum of diameters of target lesions
- Time from Initial Disease Diagnosis to Randomization (< 6months, 6months - < 1 year, 1 - < 2 year, 2 - < 3 year, 3 - < 4 year, 4 - < 5 year, ≥ 5 year)
- Time from randomization to first dose date (≤ 3days, 4 - 5days, 6 - 7days, 8 - 14days, 15 - 21days, >21days, Not Reported or Not Treated)

For the purposes of baseline characteristic summary, stratification factors will be retrieved from the CRF.

** The ‘de-novo metastatic’ implies cancer which is already metastatic (Stage IV) at first presentation.

7.3.3 Medical History

General medical history will be listed by subject and pretreatment events will be tabulated. This analysis will be performed on all randomized subjects and on PD-L1 expressing subjects.

7.3.4 Prior Therapy Agent

Prior therapy will be summarized for all randomized subjects and on all PD-L1 expressing subjects.

Prior anti-cancer therapy:

- Setting of prior systemic therapy regimen received (adjuvant, neo-adjuvant, definitive CRT).

- Time from completion of prior adjuvant/neo-adjuvant/definitive therapy to treatment (for subjects who received prior adjuvant/neo-adjuvant therapy), (< 6 months, 6 - < 12month, ≥ 12 months)
- Prior surgery related to current cancer (yes or no).
  - type of surgery
- Prior radiotherapy (yes or no).
- Prior systemic therapy classified by therapeutic class and generic name.

Other Prior therapy:

- Prior/current non-study medication classified by anatomic and therapeutic classes.

Agents and medication will be reported using the generic name. A listing by subject will also be provided.

7.3.5 Baseline Examinations

Subjects with abnormal baseline physical examination will be tabulated for all randomized subjects and on all PD-L1 expressing subjects, by examination criteria and by treatment group.

7.3.6 Discrepancies between IRT Stratification Factors and Other Datasets

Summary tables (cross-tabulations) by treatment group for stratification factor (except for region and Number of organs with metastases) will be provided to show any discrepancies between what was reported through IRT vs. other data sources at baseline:

- PD-L1 expression level (≥ 1% vs. < 1% or indeterminate) (IRT vs. clinical database)
- ECOG performance status (0 vs. 1) (IRT vs. CRF)

This analysis will be performed on all randomized subjects and on all PD-L1 expressing subjects.

7.4 Extent of Exposure

Analyses in this section will be performed in all treated subjects and in treated PD-L1 expressing subjects, by treatment group as treated.

7.4.1 Administration of Study Therapy

For details on the dosing schedule per protocol, see Section 2.1.

The following parameters will be summarized (descriptive statistics) by treatment group:

- Relative dose intensity (%) using the following categories: < 50%; 50 - < 70%; 70 - < 90%; 90 - < 110%; ≥ 110%.
- Number of doses/cycles received (summary statistics).
- Cumulative dose
- Duration of treatment

- using a KM curve whereby the last dose date will be the event date for those subjects who are off study therapy. Subjects who are still on study therapy will be censored on their last dose date. Median duration of treatment and associated 95% CI will be provided.

A by-subject listing of dosing of study medication (record of study medication, infusion details, dose change) and a listing of batch number will be also provided.

Key parameters used to characterize dosing data for the N+I arm are defined in Table 7.4.1-1; for the N+C and CT arms, in Table 7.4.1-2.

Table 7.4.1-1: Administration of Study Therapy in the N+I Arm - Definition of Parameters

	Nivolumab	**Ipilimumab**
Dosing schedule per protocol	3mg/kg Q2WK, IV	1mg/kg Q6WK, IV
Dose*	Total dose administered (mg)/most recent weight (kg)	Total dose administered (mg)/most recent weight (kg)
Cumulative Dose	The sum of all doses (mg/kg) administered to a subject during the treatment period	The sum of all doses (mg/kg) administered to a subject during the treatment period
Relative Dose Intensity (%)	Cum dose (mg/kg) / [(Last nivolumab dose date – nivolumab Start dose date + 14) x 3 / 14 x 100]	Cum dose (mg/kg) / [(Last ipilimumab dose date – ipilimumab Start dose date + 42) x 1 / 42 x 100]
Duration of Treatment	Last dose date (of the last administered study therapy) - Start dose date (of the first administered study therapy) + 1	

* Dose administered in mg at each dosing date and weight are collected on the CRF

Table 7.4.1-2: Administration of Study Therapy in the N+C and CT Arms

	Nivolumab	**Fluorouracil**	**Cisplatin**
Dosing schedule per protocol	240mg Q2WK, IV	800mg/m2/day, Q4WK continuous infusion on Day1-5 of the cycle	80mg/m2 Q4WK, IV
Dose*	Ratio of Total Volume Infused with Total Volume Prepared x 240 in mg	Total dose administered (mg) / most recent BSA	Total dose administered (mg) / most recent BSA
Cumulative Dose	The sum of all doses (mg) administered to a subject during the treatment period	The sum of all doses (mg/m2) administered to a subject during the treatment period	The sum of all doses (mg/m2) administered to a subject during the treatment period
Relative Dose Intensity (%)	Cumulative dose (mg) / [Last nivolumab dose date – nivolumab Start dose date + 14) x 240 / 14]	Cumulative dose (mg/m2) / [First dose of fluorouracil in the last cycle – fluorouracil Start dose date + 28) x 800 x 5 / 28]**	Cumulative dose (mg/m2) / [First dose of cisplatin in the last cycle – cisplatin Start dose date + 28) x 80 / 28]
Duration of Treatment	Last dose date (of the last administered study therapy) - Start dose date (of the first administered study therapy) + 1		

* Dose administered in mg at each dosing date and BSA (computed using recent weight and baseline height) are collected on the CRF.

** 800 x 5 may be replaced by 1000 x 4 for sites that administer over 4 days.

7.4.2 Modification of Study Therapy

7.4.2.1 Dose Delays

Table 7.4.2.1-1: Dose Delays

	Ipilimumab	Nivolumab 3 mg/kg (a)	Nivolumab 240 mg (b)	Fluorouracil	Cisplatin
Specifications per Protocol					
Dose considered as actually delayed	If the delay is exceeding 5 days	If the delay is exceeding 3 days	If the delay is exceeding 3 days	If the delay is exceeding 3 days	If the delay is exceeding 3 days
Maximum delay allowed between doses	12 weeks	8 weeks	8 weeks	8 weeks	8 weeks
Definitions for the Analysis					
Dose Delay	duration of preceding cycle in days – 42	duration of preceding cycle in days – 14	duration of preceding cycle in days – 14	duration of preceding cycle in days – 28	duration of preceding cycle in days – 28
Categories of dose delays	on-time, 4 - 7 days, 8 - 14 days, 15 - 42 days, 42 - 56 days, > 56 days	on-time, 4 - 7 days, 8 - 14 days, 15 - 42 days, 42 - 56 days, > 56 days	on-time, 4 - 7 days, 8 - 14 days, 15 - 42 days, 42 - 56 days, > 56 days	on-time, 4 - 7 days, 8 - 14 days, 15 - 42 days, 42 - 56 days, > 56 days	on-time, 4 - 7 days, 8 - 14 days, 15 - 42 days, 42 - 56 days, > 56 days
(a) N+I arm (b) N+C arm					

The following parameters will be summarized by treatment arm:

- Number of dose delayed per subject, Length of Delay and Reason for Dose Delay.

Reason for dose delay will be retrieved from CRF dosing pages.

7.4.2.2 Dose Interruptions

Each study therapy may be interrupted. This information will be retrieved from CRF dosing pages. The following parameters will be summarized by treatment arm:

- Number of subject with at least one dose infusion interrupted along with reason for the interruptions and number of infusions interrupted per subject.

7.4.2.3 IV Rate Reductions

In this study, it is not allowed that IV rate is reduced.

7.4.2.4 Dose Reductions

Per protocol, there will be no dose escalations or reductions of nivolumab and ipilimumab. However, for chemotherapy, dose reduction is permitted (as specified in the protocol). This information will be retrieved from CRF dosing pages.

For subjects treated with chemotherapy, the following will be summarized by treatment group:

- Number of subjects with at least one dose reduction along with the reason of the dose reduction.

7.4.3 Concomitant Medications

Concomitant medications, defined as medications other than study medications which are taken at any time on-treatment (i.e. on or after the first day of study therapy and within 100 days following the last dose of study therapy), will be coded using the WHO Drug Dictionary.

The following summary tables will be provided:

- Concomitant medications (subjects with any concomitant medication, subjects by medication class and generic term).

A by-subject listing will accompany the table.

7.4.3.1 Subsequent Therapy

- Number and percentage of subjects receiving subsequent therapies will be summarized. Categories include:
- Immunotherapy including commercial Nivolumab (anti-PD1 agents, anti-PD-L1 agents, anti-CTLA-4 agents and others) by drug name
- Other anti-cancer agents excluding all immunotherapy (approved and investigational) by drug name
- Surgery for treatment of tumors
- Radiotherapy for treatment of tumors
- Any combination of the above

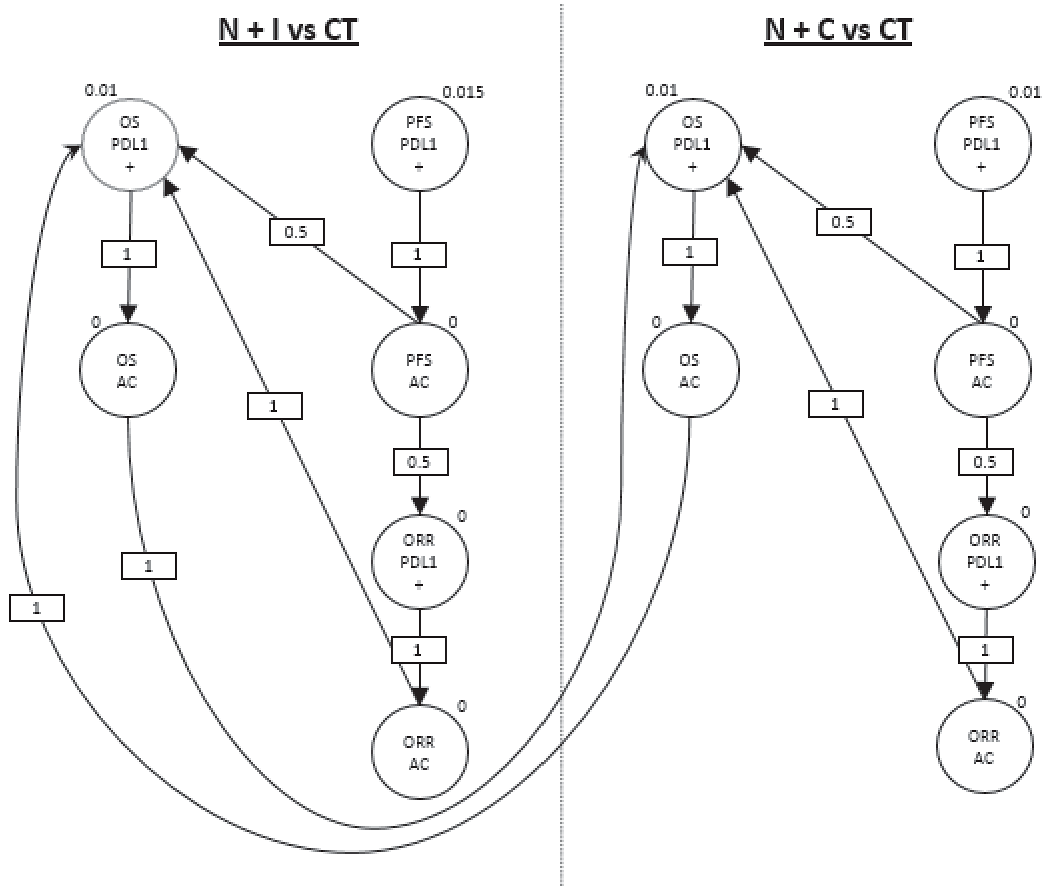
A subject listing of follow-up therapy will also be produced for subjects who had any subsequent therapy.

7.5 Efficacy

7.5.1 Protection of Type I Error Across Primary and Secondary Endpoints

The primary and secondary endpoints are tested using the Bonferroni-based graphical approach by Maurer and Bretz (2013)[13]: see Figure 7.5.1-1 for the graphical display of the multiple testing procedure. In this graph defining the test procedure, hypotheses together with their local significance levels are represented by weighted vertices and the propagation rule is represented by weighted directed edges. Algorithm 1 in Maurer and Bretz (2013) provides the rules for updating the local significance levels and the transition weights after rejecting an individual hypothesis. The resulting sequentially rejective testing procedure controls family-wise Type I error rate of 5% in the strong sense, and is uniquely determined by the graph in Figure 7.5.1-1.

Figure 7.5.1-1: Graphical Representation of the Testing Strategy for the Primary and Secondary Endpoints



OS = Overall survival, PFS = Progression-free survival, ORR = Objective response rate, PDL1+ = PDL1-expressing subjects, AC = all randomized subjects

In the graph in Figure 7.5.1-1, each vertex (circle) corresponds to a hypothesis to be tested. The vertices on the left hand side represent the hypotheses comparing N+I vs. CT; the vertices on the right hand side represent the hypotheses comparing N+C vs. CT. The vertices in the upper row correspond to the primary endpoints, i.e. OS and PFS in PD-L1 expressing subjects. The vertices in the rest of the rows correspond to the secondary endpoints, i.e. OS and PFS in all randomized subjects, and ORR in PD-L1 expressing and all randomized subjects. The numbers next to the vertices are the initially allocated (endpoint-specific) alpha levels. (Note that it is 0 for all secondary endpoints - meaning that they cannot be tested until alpha is passed to them from the primary endpoints.) The weight (in rectangular frame) associated with a directed edge (line) between any two vertices indicates the fraction of the (local) significance level at the initial vertex that is added to the significance level at the terminal vertex, if the hypothesis at the tail is rejected[14].

The test procedure is identical for the N+I vs. CT and for the N+C vs. CT comparisons. It is as follows.

At **the time of the PFS final analysis**, all 4 primary endpoints will be tested, with the following initially allocated (endpoint-specific) alpha levels:

- PFS in PD-L1 expressing subjects: 0.015
- OS in PD-L1 expressing subjects: the overall initially allocated (endpoint-specific) alpha of 0.01 will be distributed over the interim analysis (IA) and final analysis (FA) based on the actual number of deaths for each comparison at OS IA, using Lan-DeMets alpha spending function with O'Brien-Fleming boundaries. (If the observed number of OS events at IA for a comparison with a given experimental arm is exactly 175 [70% of the target number of OS events], the significance levels are 0.002 and 0.009 for OS IA and OS FA, respectively.)

If none of the primary endpoints can be rejected then no secondary endpoints are tested, and the study continues to proceed to the final OS analysis. If, however, any of these endpoints can be rejected, graph will be updated as per Algorithm 1 of Maurer and Bretz (2013), and those secondary endpoints to which alpha is passed can be tested. More specifically, the algorithm ensures that in case a primary endpoint (in PDL1-expressing subjects) is rejected, its entire alpha level is passed to the corresponding (OS or PFS) endpoint in all randomized subjects. Further, if for a given experimental arm (N+I or N+C), both OS endpoints (in PD-L1 expressing subjects and in all randomized subjects) are rejected, their significance level is passed to the primary OS endpoint of the other experimental arm. For a given experimental arm, if PFS in all randomized subjects is rejected, its significance level is split into two: half of it will be passed on to the primary OS endpoint, and half of it to ORR in PDL1-expressing subjects. (Note that ORR endpoints [in PDL1-expressing and in all randomized subjects] can be tested only in case PFS in all randomized subjects is rejected.) Further, there is no direct edge between any two endpoints within all randomized subjects: this reflects the intention for any given endpoint to first give chance to the primary population to be rejected, and only thereafter for all randomized subjects.

Each time a hypothesis is rejected, and the graph is updated, the alpha levels for IA and FA will be recalculated for each remaining OS endpoint. (O'Brien-Fleming alpha spending function will be used for the primary OS endpoints [in PDL1-expressing subjects], and Pocock for the secondary OS endpoints [in all randomized subjects].)

At the time of the final OS analysis, none of the PFS or ORR endpoints will be tested. (I.e. all remaining vertices and the related edges will be deleted from the graph). For the OS analyses, all events in the database at the time of the lock will be used.

To illustrate how the sequentially rejective procedure works, a hypothetical example is included in the appendix.

The final OS analysis will be triggered when 140 events are observed among the PD-L1 expressing subjects in the chemotherapy arm. At the time of interim analysis, the significance level will be calculated according to the actual pooled events number between the treatment arm and the control arm at the interim and the estimated pooled events number between the treatment arm and the control arms at the final. At the final analysis, the significance level will be calculated using the number of events in the database at time of database lock and will consider the alpha level already spent at the interim analysis.

7.5.2 Primary Efficacy Endpoints

Analyses in this section will be performed by treatment group as randomized.

7.5.2.1 Primary Analyses

Overall survival and PFS as assessed by BICR in All PD-L1 expressing subjects will be compared between N+I vs CT and between N+C vs CT using a two-sided log rank test, stratified by the stratification factors as specified in Section 7.1.1, i.e. by

- region (J/K/T vs rest of Asia vs RoW),
- ECOG performance status (0 vs. 1),
- number of organs with metastases (≤ 1 vs. ≥ 2).

For each comparison, the HR with its associated two-sided 100(1-α)% CIs will be estimated via a stratified Cox model with treatment arm as the only covariate in the model.

Overall survival and PFS for each treatment arm will be estimated and plotted using the KM product-limit method. Median survival time along with 95% CI will be constructed based on a log-log transformed CI for the survival function.[8,9]

Status of Censored Subjects

The status of subjects who are censored in the OS KM analysis will be tabulated for each randomized treatment group using following categories:

- On-study (on-treatment in follow-up);
- Off-study (lost to follow-up, withdraw consent, never treated etc.).

The status of subjects who are censored in the PFS KM analysis will be tabulated for each randomized treatment group using the following categories:

1) Censored on randomization date
   a) no baseline scan
   b) no on-study scan and no death (or death with prior subsequent therapy)
2) Censored on date of last tumor scan on-study
   a) Received subsequent anti-cancer therapy
   b) Still on treatment
   c) In follow-up

d) Off study
  i) lost to follow-up
  ii) subject withdrew consent
  iii) other

The source of PFS event (death vs. PD) will be summarized by treatment group.

7.5.2.2 Sensitivity Analyses

- **Overall Survival and Progression-Free Survival**
- *Stratified analysis using stratification factors as obtained from the baseline CRF pages* (instead of IRT) except for the number of organs with metastases at baseline which will be retrieved from IRT. This analysis will be performed only if at least one stratification factor at randomization (as per IRT) and baseline are not concordant for at least 10% of All PD-L1 expressing subjects.
  - region (J/K/T vs rest of Asia vs. RoW) - from CRF
  - ECOG performance status (0 vs. 1) - from CRF
  - number of organs with metastases (≤ 1 vs. ≥ 2) - from IRT
- *Analysis using a 2-sided, un-stratified log-rank tes*t, an un-stratified Cox proportional hazards model with treatment as the single covariate.
- *Analysis for subjects with no relevant deviation*. This analysis will be conducted only if there are more than 10% subjects with relevant protocol deviations.

Each of the above analysis will use the same significance level as the corresponding primary analysis. Estimate of the HR, its two sided CI (adjusted as for the corresponding primary analysis) and p-value will be presented.

- A multivariate adjusted, stratified (by factors as specified in Section 7.1.1) Cox model will be fitted to assess the treatment effect when adjusted for potential prognostic factors. The following potential prognostic factors will be included in the model. The list of covariates considered in the multivariate cox model may be modified if there are issues with multicollinearity among the covariates or other important prognostic factors are identified.
  - Age categorization (< 65 vs. ≥ 65)
  - Gender (Male vs. Female)
  - Race (Asian vs. non-Asian)
  - Weight (< 60kg vs. ≥ 60kg)
  - Disease status at current diagnosis (recurrent vs. de novo metastatic and unresectable advanced vs. de novo metastatic)
  - Smoking status (current/former vs. never/unknown)
  - Alcohol use (current/former vs. never/unknown)

For the multivariate analysis, HR and 95% CI will be provided for treatment variable and all covariates. Descriptive p-values will be provided.

- *Analysis for investigating non-proportionality in Kaplan-Meier curves.* If the Kaplan-Meier curves indicate the HR is not constant over time such as a clear delayed separation, the power of the standard logrank test is reduced. In this case, as suggested by the non-proportional hazards working group, the Max-Combo test is a promising approach with the flexibility of allowing different types of non-proportionality and robustness under model mis-specification. The R package nphsim can facilitate such testing procedures. We will apply the Max-Combo test to the PFS/OS data with the following parameter setting (rho, gamma) of Fleming-Harrington class of weights: (rho = 0, gamma = 0) attributes equal weights to all; (rho = 1, gamma = 0) emphasizes early difference in survival curves; (rho = 0, gamma = 1) weight more for late difference; (rho = 1, gamma = 1) emphasizes middle difference. In the current implementation of the R package nphsim, stratification is not available. Therefore, the unstratified Max-Combo test in the package will be applied and the resulting p-value will be reported, along with estimated HR and adjusted 95% CI. Since the stratification is not involved, the results of this sensitivity analysis should only be compared with the original results of the unstratified analysis.

<u>Sensitivity analyses for PFS to investigate alternative censoring schemes</u>

- *Analysis accounting for assessment on/after subsequent therapy:* PFS will be defined similarly to the primary definition except that events (progression or death) and disease assessments that occurred on or after subsequent anti-cancer therapy will be considered (no time point truncation). This definition is considered as primary definition by EMA. The PFS accounting for subsequent therapy (primary definition, Section 4.1.2) is the primary endpoint of the study and will be the basis of the primary test to determine statistical significance.
- *PFS accounting for two or more consecutively missing disease assessments prior to PFS event*: This analysis will be performed only if at least 10% of PFS events have missing prior disease assessments. In case a subject has two or more consecutively missing disease assessments, the subject will be censored at the last disease assessment date prior the PFS event.
- *Analysis in which progression-free subjects who are lost to follow-up for any cause* will be considered as having an event at the time of the last tumor assessment date prior to loss to follow-up.

Each of the above analysis will use the same significance level as the corresponding primary analysis. Estimate of the HR, its two sided 100(1-α)% CI and p-value will be presented.

7.5.2.3 Consistency of Treatment Effect in Subsets

To assess consistency of treatment effects in different subsets, a “forest” plot of the OS and PFS unstratified HR (and 95% CI) will be produced for the following subgroups.

- Age category (< 65, ≥ 65 and < 75, ≥ 75, ≥ 65)
- Gender (male, female)
- Race (Asian, non-Asian)
- Region (J/K/T, Rest of Asia, RoW) (*stratification factor*)

- Region (Asia, Non-Asia)
- ECOG PS (*stratification factor*)
- Weight (< 60kg, ≥ 60kg)
- Disease stage at initial diagnosis (Stage I, Stage II, Stage III, Stage IV)
- Histologic grade at initial diagnosis (Gx, G1, G2, G3, G4, Not Otherwise specified)
- Histological classification at initial diagnosis (squamous cell carcinoma, adenosquamous cell carcinoma, other)
- Location at initial diagnosis (upper thoracic, middle thoracic, lower thoracic, gastroesophageal junction)
- Disease status at current diagnosis (recurrent - loco-regional, recurrent - distant, de-novo metastatic, unresectable advanced)
- Smoking status (current/former, never/unknown)
- Alcohol use (current/former, never/unknown)
- Number of organs with metastases at baseline (≤ 1 vs. ≥ 2) (*stratification factor*)
- Time from Initial Disease Diagnosis to Randomization (< 1 year, 1 - < 3 year, 3 -< 5 year, ≥ 5 year)
- Prior surgery (excluding biopsy) (Yes or No)
- Prior radiotherapy (Yes or No)

If subset category has less than 10 subjects per treatment group, HR will not be computed/displayed. Number of events and median OS/PFS along with 95% CI will be displayed for each treatment group.

For the purposes of the subset analyses, stratification factors will be retrieved from the CRF except for the number of organs with metastases at baseline which will be retrieved from IRT.

7.5.2.4 Subjects follow-up

The extent of follow-up is defined as the time between randomization date and last known date alive (for subjects who are alive) or death date (for subjects who died) will be summarized descriptively for all subjects randomized.

The currentness of follow-up for OS is defined as the time between last OS contact (i.e., last known date alive or death date) and data cut-off date. Subjects who died and subjects with last known alive date after data cut-off date will be considered as current for this analysis.

The currentness of follow-up for PFS is defined as the time between last tumor assessment date (regardless of initiation of subsequent therapy) and cut-off date. Subjects who have a PFS event (regardless of initiation of subsequent therapy) and subjects with last tumor assessment date on or after data cut-off will be considered as current for this analysis.

The currentness of follow-up for OS and PFS will be summarized by treatment group, and will be categorized in categories.

7.5.2.5 Survival Rates

Overall survival and PFS rates at 6, 12, 18, 24, 36, 48 months and at 5 year will also be estimated using KM estimates on the OS and PFS curves for each randomized arm. Minimum follow-up must be approximately longer than or equal to timepoint to generate the rate. Associated two-sided 95% CIs will be calculated using the Greenwood's formula.[8,9]

7.5.3 Secondary Efficacy Endpoints

Secondary endpoints are: OS and PFS by BICR in All Randomized Subjects, ORR by BICR in All PD-L1 expressing subjects and in All Randomized subjects.

Analyses for each of these endpoints will be performed by treatment group as randomized.

7.5.3.1 Overall Survival and PFS by BICR in All Randomized Subjects

As specified by Figure 7.5.1-1, if any of the primary endpoints is significantly superior, the corresponding secondary endpoint in all randomized subjects will be compared using a two-sided log rank test at the allocated significance level, stratified by the stratification factors as specified in Section 7.1.1. (I.e. by the same stratification factors as the primary endpoints plus by PD-L1 status [≥ 1% vs. < 1%] as recorded in IRT.)

For each comparison, the HR with its associated two-sided 95% CI (in case the given endpoint is formally tested, also with the 100(1-α)% CI) will be estimated via a stratified Cox model with treatment arm as the only covariate in the model. OS and PFS for each treatment arm will be estimated and plotted using the KM product-limit method. Median survival time along with 95% CI will be constructed based on a log-log transformed CI for the survival function.[8,9]

Additional Analyses for OS and PFS in All Randomized Subjects

The same additional analyses will be carried out for OS and PFS in All randomized Subjects as for OS and PFS in All PD-L1 expressing subjects, i.e.:

- Status of censored subjects (Section 7.5.2.1)
- All sensitivity analyses (Section 7.5.2.2)
- Analyses by subsets (Section 7.5.2.3)
- Extent of follow-up and currentness of follow-up (Section 7.5.2.4)
- Survival rates (Section 7.5.2.5)

7.5.3.2 Objective Response Rate by BICR

All analyses in this section will be carried out for All PD-L1 expressing subjects and for All Randomized subjects as well.

Best Overall Response will be summarized by response category for each treatment group.

ORR (as assessed by BICR) in subjects with PD-L1 expressing tumors and in all randomized subjects will be tested only if significance level is passed on them. For rules for passing on significance levels, see Section 7.5.1. Comparisons will use a two-sided CMH test (as detailed in Section 7.1.1).

ORR will be computed in each treatment group along with the exact 95% CI using Clopper-Pearson method[15]. An estimate of the difference in ORRs and corresponding 95% CI (in case the given endpoint is formally tested, also with the 100(1-α)% CI) will be calculated using CMH methodology and adjusted by the stratification factors as specified in Section 7.1.1. The stratified (source: IRT) odds ratios (Mantel-Haenszel estimator) between the treatments will be provided along with the 95% CI (in case the given endpoint is formally tested, also with the 100(1-α)% CI).

For additional analysis related to ORR by BICR, see Section 7.5.4.3.

7.5.4 Exploratory Efficacy Endpoints

Analyses in this section will be performed in all randomized subjects and in All PD-L1 expressing subjects, by treatment group as randomized.

7.5.4.1 Progression-Free Survival by Investigator

Hazard ratio (N+C vs CT and N+I vs CT) with its associated two-sided 95% CI will be estimated via a stratified Cox model with treatment arm as the only covariate in the model. PFS for each treatment arm will be estimated and plotted using the KM product-limit method. Median survival time along with 95% CI will be constructed based on a log-log transformed CI for the survival function.[8,9]

The PFS rates, the source of PFS event (death vs. PD), and the status of subjects who are censored in the PFS KM analysis will be evaluated the same way as for PFS by BICR.

7.5.4.2 PFS2/TSST

Hazard ratio (N+C vs CT; N+I vs CT) with its associated two-sided 95% CI will be estimated via a stratified Cox model with treatment arm as the only covariate in the model. PFS2/TSST for each treatment arm will be estimated and plotted using the KM product-limit method. Median PFS2/TSST time along with 95% CI will be constructed based on a log-log transformed CI for the survival function.[8,9]

The PFS2/TSST rates and the source of PFS2/TSST event (death vs. PD vs. second next line systemic therapy) will be evaluated.

7.5.4.3 Endpoints Related to Best Overall Response

Duration of Response and Duration of SD (as assessed by BICR and by investigator) in each treatment group will be estimated using KM product-limit method for subjects who achieve PR or CR. Median values along with two-sided 95% CI will be calculated.

For DOR, type of event (PD, death) and status of subjects censored (received subsequent therapy, ongoing follow-up [current, non-current], off study [withdrew consent] will be presented by treatment arm.

Time to objective response (TTR as assessed by BICR and by investigator): summary statistics will be provided for each treatment group for subjects who achieve PR or CR. Cumulative Response Rates will be tabulated for Week 6, Week 12, Month 4, 6, 8, and 12, and overall Response Rate will be provided for each treatment group.

7.6 Safety

Analyses in this section will be tabulated for all treated subjects by treatment group as treated, unless otherwise specified.

In addition, selected safety analyses (including any adverse events (AE), drug-related AE, serious AEs, drug-related SAE, AEs leading to discontinuation, drug-related AE leading to discontinuation, death summary) will be summarized by treatment group in all treated PD-L1 expressing subjects and all treated PD-L1 negative subjects. In all treated PD-L1 expressing subjects, the frequency of immune mediated adverse events and other events of special interest (within 100 days), select AEs and drug-related select AEs (within 30 days) will be summarized by treatment group.

7.6.1 Deaths

Deaths will be summarized by treatment group:

- All deaths, reasons for death.
- Deaths within 30 days of last dose received, reasons for death.
- Deaths within 100 days of last dose received, reasons for death.

A by-subject listing of deaths will be provided for the all enrolled subjects population.

7.6.2 Serious Adverse Events

Serious adverse events will be summarized by treatment group:

- Overall summary of SAEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.
- Overall summary of drug-related SAEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.

All analyses will be conducted using the 30-day safety window.

A by-subject SAE listing will be provided for the “enrolled subjects” population.

7.6.3 Adverse Events Leading to Discontinuation of Study Therapy

AEs leading to discontinuation will be summarized by treatment group:

- Overall summary of AEs leading to discontinuation by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.

- Overall summary of drug-related AEs leading to discontinuation by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.

The analyses will be conducted using the 30-day safety window.

A by-subject AEs leading to discontinuation listing will be provided.

7.6.4 Adverse Events Leading to Dose Modification

AEs leading to dose delay/reduction will be summarized by treatment group:

- Overall summary of AEs leading to dose delay/reduction by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.
- Overall summary of related AEs leading to dose delay/reduction by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.

The analysis will be conducted using the 30-day safety window.

A by-subject AEs leading to dose delay/reduction listing will be provided.

7.6.5 Adverse Events

Adverse events will be summarized by treatment group.

The following analyses will be conducted using the 30 days safety window only:

- Overall summary of any AEs by worst CTC grade (1, 2, 3, 4, 5, not reported) presented by SOC/PT.
- Overall summary of any AEs presented by worst CTC grade (any grade, grade 3-4, grade 5) by SOC/PT. This table will be restricted to events with an incidence greater or equal to 5% in any treatment group.
- Overall summary of any non-serious AEs presented by SOC/PT. This table will be restricted to events with an incidence greater or equal to 5% in any treatment group.
- Overall summary of any AEs that required immune modulating medication by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.
- Overall summary of drug-related AEs by worst CTC grade (1, 2, 3, 4, 5, not reported) presented by SOC/PT.

The following analyses will be conducted using the 30 days safety window and repeated using the 100 days safety window:

- Overall summary of drug-related AEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT.

A by-subject AE listing will be provided. A by-subject listing of any AE requiring immune modulating medications will also be provided.

7.6.6 Select Adverse Events

Unless otherwise specified, analyses will be performed by select AE category. Analyses will also be repeated by subcategory of endocrine events.

7.6.6.1 Incidence of Select AE

Select AEs will be summarized by treatment group for each category/subcategory.

The following analyses will be conducted using the 30-day safety window only:

- Overall summaries of any select AEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category or Subcategory/PT.
- Overall summaries of any drug-related select AEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category or Subcategory/PT.
- Overall summaries of any serious select AEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category or Subcategory /PT.
- Overall summaries of drug-related serious select AEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category or Subcategory /PT.
- Overall summaries of any select AEs leading to discontinuation by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category or Subcategory /PT.
- Overall summaries of drug-related select AEs leading to discontinuation by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category or Subcategory /PT.
- Summary of frequency of unique select AEs by Category.

A by-subject select AE listing will be provided.

7.6.6.2 Time-to Onset of Select AE

Time-to onset of drug-related select AEs (any grade, grade 3-5) will be summarized for each category/subcategory by treatment group.

Time-to onset analyses are restricted to treated subjects who experienced at least one drug-related select AE in the category/subcategory. The analyses will be conducted using the 30-day safety window.

Additional details regarding the time-to onset definition are described in time-to onset definition subsection of APPENDIX 3.

7.6.6.3 Time-to Resolution of Select AE

Time-to resolution of the following specific events will be summarized separately for each category/subcategory.

- Time-to resolution of drug-related select AE (any grade, grade 3-5) by treatment group
- Time-to resolution of drug-related select AE (any grade, grade 3-5) where immune modulating medication was initiated, by treatment group

Time-to resolution analyses are restricted to treated subjects who experienced the specific events. Time-to resolution where immune modulating medication was initiated analyses are restricted to treated subjects who experienced the specific events and who received immune modulating medication during the longest select AE.

The analyses will be conducted using the 30-day safety window.

The following summary statistics will be reported: percentage of subjects with resolution of the longest select AE, median time-to resolution along with 95% CI (derived from Kaplan-Meier estimation) and ranges.

See time-to resolution definition subsection of APPENDIX 3 for additional details.

7.6.7 Immune Modulating Medication

Immune modulating concomitant medications are medications entered on an immune modulating medication form or available from the most current pre-defined list of immune modulating medications. The list of anatomic class, therapeutic class and generic name used for the selection at the time of the database lock will be provided.

The percentage of subjects who received immune modulating concomitant medication for

- management of adverse event
- premedication
- other use
- any use
- management of drug-related select adverse event (any grade, grade 3-5) by select AE category/ subcategory (EU/ROW Submissions)
- management of IMAEs (any grade, grade 3-5) by IMAE category (US Submission) will be reported separately for each treatment group (percentages of treated subjects by medication class and generic term).

For each category/subcategory of drug-related select AEs (any grade, grade 3-5) and IMAEs (any grade, grade 3-5), the following will be reported for each treatment group:

- The total immune modulating medication treatment duration (excluding overlaps), duration of high dose of corticosteroid, initial dose of corticosteroid, and tapering duration (summary statistics)

Duration represents the total duration the subject received the concomitant medication of interest. If the subject took the medication periodically, then DURATION in the summation of all use. Initial dose represents the dose of the concomitant medication of interest received at the start of the event. In the case multiple medications started on the same date, the highest equivalent dose is chosen and converted to mg/kg by dividing by the subject’s recent weight.

These analyses, except the ones related to IMAEs will be conducted using the 30-day safety window. The analyses related to IMAEs will be conducted using the 100-day safety window.

7.6.8 Multiple Events

The following summary tables will be provided:

- A table showing the total number and rate (exposure adjusted) of occurrences for all AEs.
- A table showing the total number and rate (exposure adjusted) of occurrences for AEs occurring in at least 5% of subjects in any treatment group.

In addition, the rate (exposure adjusted) and its 95% CI evaluated for different time intervals will be displayed graphically for each treatment group. This analysis will be limited to the rate of all AEs and all drug-related AEs. The analyses will be conducted using the 30-day safety window.

A listing displaying the unique instances of all AEs, i.e., after duplicates have been eliminated and overlapping and contiguous occurrences of the same event (i.e. same PT) have been collapsed will be provided. No formal comparisons will be made between treatment groups.

7.6.9 Other Events of Special Interest

OEOSI will be summarized by treatment group for each category.

The following analyses will be conducted using the 100-day safety window:

- Overall summary of OEOSI by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category / PT
- Overall summary of drug-related OEOSI by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category / PT

A by-subject listing of OEOSI will be provided.

7.6.10 Immune Mediated Adverse Events

IMAEs will be summarized by treatment group for each immune-mediated category / PT using the 100-day safety window:

- Overall summary of non-endocrine IMAEs by worst CTC grade (any grade, grade 3-4, grade 5) where immune modulating medication was initiated presented by Category / PT.
- Overall summary of endocrine IMAEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category / PT.
- Overall summary of non-endocrine IMAEs leading to discontinuation by worst CTC grade (any grade, grade 3-4, grade 5) where immune modulating medication was initiated presented by Category / PT.
- Overall summary of endocrine IMAEs leading to discontinuation by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category / PT.

- Overall summary of non-endocrine IMAEs leading to dose delay or reduction by worst CTC grade (any grade, grade 3-4, grade 5) where immune modulating medication was initiated presented by Category / PT
- Overall summary of endocrine IMAEs leading to dose delay or reduction by worst CTC grade (any grade, grade 3-4, grade 5) presented by Category / PT.
- Summaries of time-to onset and time-to resolution of non-endocrine IMAEs where immune modulating medication was initiated presented by Category.
- Summaries of time-to onset and time-to resolution of endocrine IMAEs presented by Category.

A by-subject listing of IMAEs will be provided. By-subject listings of time-to resolution for longest IMAEs cluster (any grade and grade 3-5 in separate summaries) will also be provided. For new studies which collect investigator assessment of potential IMAE data, a by-subject listing of AEs considered as immune-mediated events per investigator but not qualified for IMAEs definition will also be provided.

In addition, for all Nivolumab and/or Ipilimumab treated subjects who experienced at least one IMAE, the following data presentation will be provided:

- Summary of subjects who were re-challenged with Nivolumab and/or Ipilimumab by IMAE category, with extended follow-up

For these, re-challenge is considered to have occurred when last Nivolumab and/or Ipilimumab infusion was administered after the onset of an IMAE.

7.6.11 Clinical Laboratory Evaluations

The analysis population for each laboratory test is restricted to treated subjects who underwent that laboratory test. Laboratory tests (in addition to the tests specified below) with CTC criteria collected in the specific studies may also be included in the summaries.

A by-subject listing of differences in categorization of SI and US laboratory test results will be provided.

7.6.11.1 Hematology

The following will be summarized by treatment group as worst CTC grade on-treatment per subject and as shift table of worst on-treatment CTC grade compared to baseline CTC grade per subject: hemoglobin (HB), platelets, white blood counts (WBC), absolute neutrophils count (ANC) and lymphocyte count (LYMPH).

The analyses will be conducted using the 30-day safety window.

A by-subject listing of these laboratory parameters will be provided.

7.6.11.2 Serum Chemistry

The following will be summarized by treatment group as worst CTC grade on-treatment per subject and as shift table of worst on-treatment CTC grade compared to baseline CTC grade per subject: ALT, AST, alkaline phosphatase (ALP), total bilirubin and creatinine.

The analyses will be conducted using the 30-day safety window.

A by-subject listing of these laboratory parameters will be provided.

7.6.11.3 Electrolytes

The following will be summarized by treatment group as worst CTC grade on-treatment per subject and as shift table of worst on-treatment CTC grade compared to baseline CTC grade per subject: sodium (high and low), potassium (high and low), calcium (high and low), and Glucose Serum.

The analyses will be conducted using the 30-day safety window.

A by-subject listing of these laboratory parameters will be provided.

7.6.11.4 Additional Analyses

In addition, further analyses on specific laboratory parameters will be performed by treatment group:

Abnormal Hepatic Function Test

The number of subjects with the following laboratory abnormalities from on-treatment evaluations will be summarized by treatment group:

- ALT or AST > 3 x ULN, > 5 x ULN, > 10 x ULN and > 20 x ULN
- Total bilirubin > 2 x ULN
- ALP > 1.5 x ULN
- Concurrent (within 1 day) ALT or AST > 3 x ULN and total bilirubin > 1.5 x ULN
- Concurrent (within 30 days) ALT or AST > 3 x ULN and total bilirubin > 1.5 x ULN
- Concurrent (within 1 day) ALT or AST > 3 x ULN and total bilirubin > 2 x ULN
- Concurrent (within 30 days) ALT or AST > 3 x ULN and total bilirubin > 2 x ULN

The analyses will be conducted using the 30-day safety window.

A by-subject listing of these specific abnormalities will be provided.

Abnormal Thyroid Function Test

The number of subjects with the following laboratory abnormalities from on-treatment evaluations will be summarized by treatment group:

- TSH value > ULN and
  - with baseline TSH value ≤ ULN

  - with at least one FT3/FT4 test value < LLN within 2-week window after the abnormal TSH test
  - with all FT3/FT4 test values ≥ LLN within 2-week window after the abnormal TSH test
  - with FT3/FT4 missing within 2-week window after the abnormal TSH test.

- TSH < LLN and
  - with baseline TSH value ≥ LLN
  - with at least one FT3/FT4 test value > ULN within 2-week window after the abnormal TSH test
  - with all FT3/FT4 test values ≤ ULN within 2-week window after the abnormal TSH test
  - with FT3/FT4 missing within 2-week window after the abnormal TSH test

The analyses will be conducted using the 30-day safety window.

A by-subject listing of these specific abnormalities will be provided.

7.6.12 Vital Signs

Vital signs collected on the CRF will be provided in separate listings.

7.6.13 Immunogenicity Analysis

Nivolumab-containing arms only. Further details on immunogenicity background and rationale, definitions, population for analyses and endpoints are described in APPENDIX 4.

7.6.14 Pregnancy

A by-subject listing of pregnancy tests results will be provided for randomized female subjects.

7.6.15 Adverse Events by Subgroup

Overall summary of any AEs and drug-related AEs by worst CTC grade (any grade, grade 3-4, grade 5) presented by SOC/PT and for each treatment group for the following subgroups:

- Sex (Male vs. Female)
- Race (Asian vs. non-Asian)
- Age (< 65 vs. 65 - < 75 vs. 75 - < 85 vs. ≥ 85 vs. ≥ 75 vs. ≥ 65)
- Region (J/K/T vs rest of Asia vs RoW)

These analyses will be conducted using the 30-day safety window only.

7.7 Pharmacokinetic Analysis

Pharmacokinetics analyses will be performed on the nivolumab-containing arms only.

The nivolumab and/or ipilimumab concentration vs time data obtained in this study may be combined with data from other studies in the clinical development program to develop population PK models. These models may be used to evaluate the effects of intrinsic and extrinsic covariates on the PK of nivolumab and/or ipilimumab and to determine measures of individual exposure (such as steady state peak, trough and time averaged concentration). Model determined exposures may be used for exposure response analyses of selected efficacy and safety endpoints. If the analyses are conducted, the results of population PK and exposure response analyses will be reported separately.

7.8 Biomarker Analysis

These analyses will be descriptive and not adjusted for multiplicity.

7.8.1 PD-L1 Expression

Analyses will be performed in all randomized subjects, by treatment group as randomized.

7.8.1.1 Distribution of PD-L1 Expression

Descriptive statistics of PD-L1 expression and PD-L1 status will be provided;

- Summary statistics of PD-L1 expression by treatment groups as randomized and overall
- Frequency of PD-L1 categorization: PD-L1 status (1%, 5% and 10%) by treatment group and overall, all randomized PD-L1 subjects

7.8.1.2 Association between PD-L1 Status and Efficacy

Analyses will be based on all PD-L1 randomized subjects if not otherwise specified. PD-L1 status categories (1%, 5% and 10%) will be used for these analyses.

For both OS and PFS per BICR, the following analyses will be provided:

- Within each category above, curves will be estimated using the KM product limit method for each treatment group. Two-sided, 95% CIs for median OS and PFS per BICR will be computed by Brookmeyer and Crowley method.
- Two-sided log-rank test comparing treatment arms
- HR (with corresponding two-sided 95% CI) will be estimated via a Cox model with treatment arm as the only covariate in the model (*note that for PDL1-expressing subjects, this is one of the sensitivity analyses, and won't be repeated; see Section 7.5.2.2*)
- Forest plot of HR with 95% CIs
- For ORR per BICR, the following analyses will be provided:
- Within each category above, ORR per BICR will be computed in each treatment group along with the exact 95% CI using Clopper-Pearson method[16].
- An estimate of the difference in ORRs per BICR and corresponding 95% CI will be calculated using CMH methodology.

7.8.1.3 Predictive Relationship between PD-L1 Status and Efficacy

Analyses will be based on all evaluable PD-L1 subjects.

For both OS and PFS per BICR, a Cox proportional hazards regression model will be fitted with treatment, PD-L1 status, and treatment by PD-L1 status interaction. Although the study is not designed to have appropriate power to formally test the interaction of the model, an interaction test at significance level of 0.2 will warrant further exploration and the following statistics will be reported:

- Interaction p-value
- HR of treatment vs. control and its associated 95% CI for each of the PD-L1 status subgroup
- HR PD-L1 ≥ 1% vs. < 1% and its associated 95% CI within each treatment group.

7.8.2 PD-L1 by Combined positive score (CPS) and Microsatellite Instability (MSI) Status

Upon data availability, analyses of exploratory biomarker PD-L1 by CPS and MSI status will be performed in all randomized subjects, by treatment group as randomized.

7.8.2.1 Descriptive Statistics of PD-L1 by CPS and MSI

- Summary of tumor specimen acquisition, characteristics by CPS and MSI respectively
- Summary of PD- L1 by CPS and MSI status by categories by CPS and MSI respectively

Below analyses are applicable for PD-L1 by CPS

- Summary statistics of PD-L1 by CPS expression.
- Waterfall plot of Individual PD-L1 by CPS expression.

7.8.2.2 Association between biomarker Status and Efficacy

OS and PFS per BICR will be analyzed for each biomarker category specified for PD-L1 by CPS and MSI:

- Curves will be estimated using the KM product limit method for each treatment group. Two-sided, 95% CIs for median PFS per BICR and OS will be computed.
- HR (with corresponding two-sided 95% CI) will be estimated via a Cox model with treatment arm as the only covariate in the model.
- Forest plot of HR with 95% CIs

7.9 Outcomes Research Analyses

The analysis of EQ-5D-3L and FACT-E (including FACT-G7 and ECS) data will be performed in all randomized subjects and all PD-L1 expressing subjects who have an assessment at baseline (assessment on or prior to first dose on Day 1) and at least 1 subsequent assessment while on treatment. The questionnaire completion rate, defined as the proportion of questionnaires actually received out of the expected number, will be calculated and summarized at each assessment point.

EQ-5D-3L data will be described by treatment group as randomized in the following ways:

- EQ-5D-3L index scores will be summarized at each assessment time point using descriptive statistics (i.e., N, mean with SD and 95% CI, median, first and third quartiles, minimum, maximum). The UK scoring algorithm will be applied as a reference case.
- EQ-VAS scores will be summarized at each assessment time point using descriptive statistics (i.e., N, mean with SD and 95% CI, median, first and third quartiles, minimum, maximum).
- The proportion (N) of subjects reporting no, moderate, or extreme problems will be presented for each of the 5 EQ-5D-3L dimensions at each assessment time point. Subjects with missing data will be excluded from the analysis.
- A line graph summarizing the mean changes from baseline for the EQ-5D-3L index and VAS scores will be produced.
- A by-subject listing of the level of problems in each dimension, corresponding EQ-5D-3L health state (i.e., 5-digit vector), EQ-5D-3L index score, and EQ-VAS score will be provided.

From the beginning of the on-treatment phase through follow-up Visit 2, data for the FACT-E will be described by treatment group as randomized in the following ways:

- FACT-G7 and FACT-E total and subscale (PWB, FWB, EWB, SWB, ECS, TOI) scores will be summarized at each assessment time point using descriptive statistics (i.e., N, mean with SD and 95% CI, median, first and third quartiles, minimum, maximum).
- Changes from baseline in FACT-G7, FACT-E total and subscale (PWB, FWB, EWB, SWB, ECS, TOI) scores will be summarized at each post-baseline assessment time point using descriptive statistics (i.e., N, mean with SD and 95% CI, median, first and third quartiles.
- The proportion (N) of subjects reporting each response category will be presented for the single-item GP5 item at each assessment time point. A stacked bar graph will be produced for GP5 which shows the proportion of subject with each response category at each timepoint. Subjects with missing data will be excluded from the analysis.
- A line graph summarizing the mean changes from baseline for the FACT-E total score, ECS and FACT-G7 scores will be produced.
- A by-subject listing of the responses to each item in the FACT-E will be provided (including the items also included in the ECS and FACT-G7 during the survival follow-up phase).

During the survival follow-up phase, data for the ECS and FACT-G7 will be described by treatment group as randomized in the following ways:

- ECS and FACT-G7 scores will be summarized at each assessment time point using descriptive statistics (i.e., N, mean with SD and 95% CI, median, first and third quartiles, minimum, maximum).
- Changes from baseline in ECS and FACT-G7 scores will be summarized at each post-baseline assessment time point using descriptive statistics (i.e., N, mean with SD and 95% CI, median, first and third quartiles, minimum, maximum).

7.10 COVID-19 Related Analyses

In this study, in order to evaluate the impact of COVID-19 pandemic, the following CRF pages were implemented.

- Disposition: The subjects who discontinue the study treatment or discontinue the study due to COVID-19
- Exposure: The subjects who have study therapy modification due to COVID-19

Listings will be provided based on the data collected on each CRF pages and additional analyses may be performed in order to evaluate the impact of COVID-19 on this study.

8 Analysis to Evaluate the Contribution of Ipilimumab Component in the N+I regimen in the Summary of Clinical Efficacy (SCE)

In order to evaluate the treatment effect between two experiment arms of N+I and N+C, similar analyses as described in Section 7.5 will be conducted for key efficacy endpoints. The below analyses will be conducted for all PD-L1 expressing subjects, all PD-L1 Negative Subjects, and all Randomized Subjects. These analyses will help to evaluate the contribution of ipilimumab component in the N+I regimen in the Summary of Clinical Efficacy (SCE) and Clinical Overview (CO) if applicable.

For Overall survival and PFS as assessed by BICR, the HR between N+I and N+C with its associated two-sided 95% CIs will be estimated via a stratified Cox model with treatment arm as the only covariate in the model. Overall survival and PFS as assessed by BICR for each of the three treatment arms will be estimated and overlayed using the KM product-limit method. Median survival time along with 95% CI will be constructed based on a log-log transformed CI for the survival function.[8,9]

ORR assessed by BICR will be computed in each treatment group along with the exact 95% CI using Clopper-Pearson method[15]. An estimate of the difference in ORRs and corresponding 95% CI between N+I and N+C will be calculated using CMH methodology and adjusted by the stratification factors as specified in Section 7.1.1. The stratified odds ratios (Mantel-Haenszel estimator) between N+I and N+C will be provided along with the 95% CI. The KM curves of the DOR in each of 3 treatment groups will be overlayed for subjects who achieve PR or CR. Median DOR along with two-sided 95% CI will be calculated.

Considering that the survival function of overall survival may potentially have different shapes in the N+I arm and N+C arm, the restricted mean survival time (RMST) will be evaluated for N+I and N+C arm. The difference of RMST between N+I and N+C with its associated two-sided 95% CIs will be estimated. The RMST will be estimated by the area under the KM curve up to a time point (1 year, 2 years, and 3 years, and maximum of the study follow-up).

9 CONTENT OF REPORTS

9.1 Within-Trial Analyses Performed To Date

Not applicable.

10 CONVENTIONS

The following conventions may be used for imputing partial dates for analyses requiring dates:

- For missing and partial adverse event onset dates, imputation will be performed using the Adverse Event Domain Requirements Specification[17].
- Missing and partial Non-Study Medication Domain dates will be imputed using the derivation algorithm described in 4.3.3 of BMS Non-Study Medication Domain Requirements Specification[18].
- Missing and partial radiotherapy and surgery dates will be imputed using algorithm described in APPENDIX 2.

For death dates, the following conventions will be used for imputing partial dates:

- If only the day of the month is missing, the 1st of the month will be used to replace the missing day. The imputed date will be compared to the last known date alive and the maximum will be considered as the death date.
- If the month or the year is missing, the death date will be imputed as the last known date alive.
- If the date is completely missing but the reason for death is present the death date will be imputed as the last known date alive

For date of progression, the following conventions will be used for imputing partial dates:

- If only the day of the month is missing, the 1st of the month will be used to replace the missing day.
- If the day and month are missing or a date is completely missing, it will be considered as missing.
- In case of the date of death is present and complete, the imputed progression date will be compared to the date of death. The minimum of the imputed progression date and date of death will be considered as the date of progression.

For other partial/missing dates, the following conventions may be used:

- If only the day of the month is missing, the 15th of the month will be used to replace the missing day.
- If both the day and the month are missing, “July 1” will be used to replace the missing information.
- If a date is completely missing, it will be considered as missing.

The following conversion factors will be used to convert days to months or years:

1 month = 30.4375 days and 1 year = 365.25 days.

Duration (e.g. DOR, etc) will be calculated as follows:

$$\text{Duration} = (\text{Last date - first date} + 1)$$

All statistical analyses will be carried out using SAS (Statistical Analysis System software, SAS Institute, North Carolina, USA) unless otherwise noted.

11 CONTENT OF REPORTS

All analyses describe in this SAP will be included in the Clinical Study Report except where otherwise noted. Refer to the Data Presentation Plan for mock-ups of all tables and listings.

12 DOCUMENT HISTORY

Document History

Version Number	Author(s)	Description
1.0	[redacted]	Initial version dated 18-Oct-2017
2.0	[redacted]	Version 2.0 dated 28-Aug-2020 • Updated the definition for the outcome research exploratory endpoint FACT-E and added the analysis window for outcome research assessments in Section 4.3.3. • Added Max-Combo method as a sensitivity analysis accounting for the potential non-proportionality observed from the OS/PFS KM curves in Section 7.5.2.2. • Added PFS2/TSST as an exploratory endpoint (Section 4.3.1) and the corresponding analysis (Section 7.5.4.2). • Removed the procedures for the reinitiation of nivolumab ± ipilimumab treatment after disease progression for up to 1 additional year. • Modified the Safety section (Section 7.6) to incorporate the contents from the safety section in the IO CORE SAP. • Added value "Gx" for histologic grade at initial diagnosis and added prior surgery and prior radiotherapy in Section 7.5.2.3. • Added 5% and 10% as additional subgroups for PD-L1 status in Section 7.8.1.2. • Added analysis to explore the association between ORR and PD-L1 in Section 7.8.1.2. • Added some additional analyses for EQ-5D-3L and FACT-E in Section 7.9. • Added APPENDIX 5 with analyses of data from China. • Added the imputation algorithm for missing and partial radiotherapy and surgery dates from IO CORE SAP (APPENDIX 2).
3.0	[redacted]	Version 3.0 dated 29-Oct-2020 • Added language and rationale to allow for the final PFS analysis to be triggered when 136 PFS events per BICR are observed among the PD-L1 expressing subjects in the chemotherapy arm or when at least 12 months minimum follow up is reached, if the target number of PFS events is unlikely to be reached in Section 1, Section 5 and APPENDIX 7.

Document History

Version Number	Author(s)	Description
4.0	[redacted]	Version 4.0 dated Feb 12, 2021 Main changes include • Added analysis to evaluate relevant treatment effect between N+I vs. N+C for key efficacy endpoints in Section 8. • Added PD-L1 by CPS and MSI analysis in Section 7.8.2. • Add language in section 7.1 to handle the small stratum in the stratified analyses. Other minor changes include • Aligned the PFS2/TSST definition with IO Core SAP in Section 4.3.1. • Modified the baseline definition for stratification factors as before the date of the randomization in Section 6.1. • Moved the definition of exploratory biomarker endpoints to Section 4.3.6. • Added definition for All PD-L1 Negative Subjects and All Treated PD-L1 Expressing Subjects in Section 6.3. • Added on-treatment curative Surgery as the relevant protocol deviation in Section 7.2.2. • Add in Section 7.5.2.2 'The list of covariates considered in the multivariate cox model may be modified if there are issues with multicollinearity among the covariates or other important prognostic factors are identified.' • Added limited safety analyses in PD-L1 expressing subjects and PD-L1 negative subjects in Section 7.6 to assess benefit-risk in PD-L1 subgroups. • Aligned the baseline definition in Section 7.9 with Table 4.3.3-1. • Added PRO analysis in All PD-L1 expressing subjects in addition to all treated subjects in Section 7.9. • Added a stacked bar graph for GP5 in Section 7.9. • Added COVID-19 Related Analysis in Section 7.10. • Added Section 9.1 there was no within-trial analyses performed to date. • Add definition of primary and secondary China subpopulation in APPENDIX 5. Includes 'Only unstratified analyses will be performed for efficacy endpoints and the sensitivity analysis of investigating non-proportionality in Kaplan-Meier curves will not be performed.' in APPENDIX 5. • Removed metastatic disease from prior anti-cancer therapy in Section 7.3.4. • Modified the subcategories for disease status at current diagnosis in the multivariate cox model in Section 7.5.2.2. • In the Section 7.5.2.3 consistency of treatment effect in subsets, added region (Asia, non-Asia) and excluded biopsy from prior surgery. • Removed safety summary in the 3 arms pooled (total) in Section 7.6 safety. • For the odds ratio, in addition to 95% CI, also added '(in case the given endpoint is formally tested, also with the 100(1-α)% CI)' in Section 7.5.3.2. • In Section 7.3.2 demographics and other baseline characteristics, added race (Asia vs. non-Asia)

Document History

Version Number	Author(s)	Description
		• Modified the relative dose intensity formula for fluorouracil in Table 7.4.1-2 • The time to response would analyzed descriptively only for subjects who achieved PR or CR. KM estimate of time to response was removed. • Removed alpha adjusted difference in ORRs in Section 7.8.1.2 Association between PD-L1 status and efficacy • Add in Section 7.3.2 For the purposes of baseline characteristic summary, stratification factors will be retrieved from the CRF. • Remove alpha adjusted CI for ORR in Section 7.8.1.2 association between PD-L1 status and efficacy.

APPENDIX 1 APPENDIX: EXAMPLE FOR SEQUENTIALLY REJECTIVE PROCEDURE

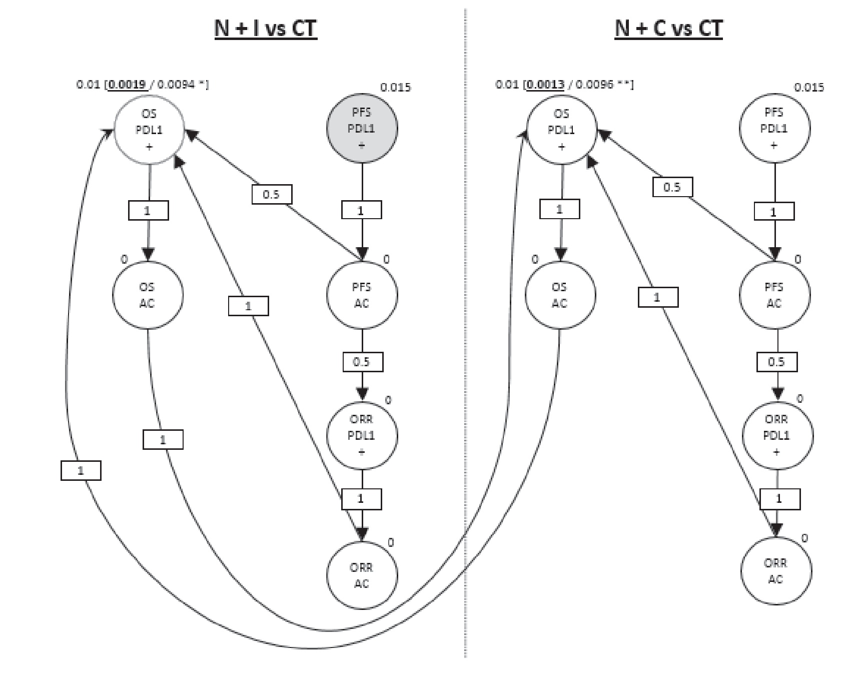
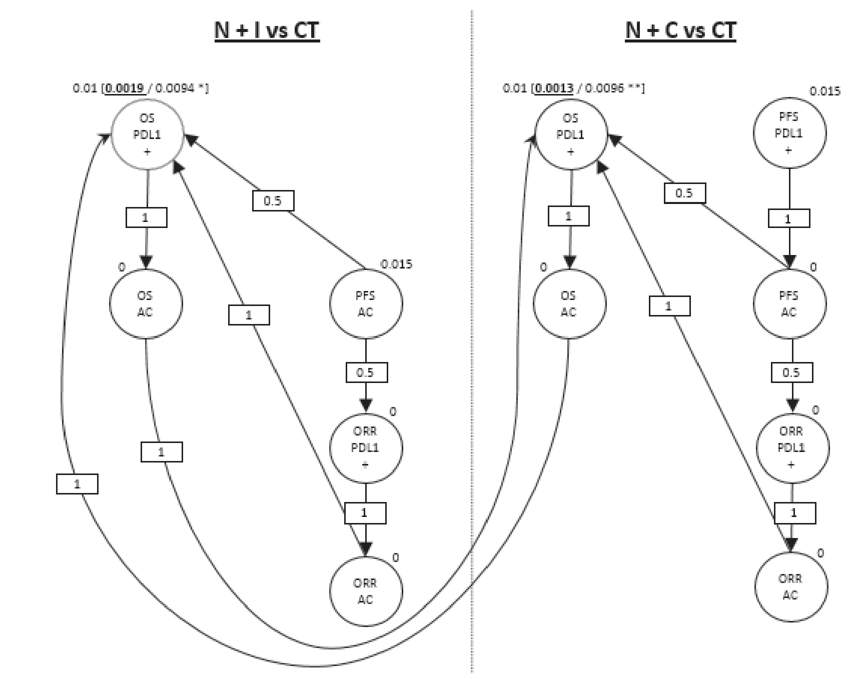
Section 7.5.1 of this SAP specifies the multiple testing procedure that controls family-wise Type I error rate of 5% in the strong sense across all primary and secondary endpoints. After rejecting an individual hypothesis, the local significance levels and the transition weights of the edges are updated: the rules for that update are determined by Algorithm 1 in Maurer and Bretz (2013). The resulting sequentially rejective testing procedure is uniquely determined by the graph in Figure 7.5.1-1. In order to illustrate the principles of this iterative procedure (summarized in Section 7.5.1), a simple hypothetical example is provided in this section.

This example considers a hypothetical outcome of the primary and secondary endpoints at the time of the PFS analysis (Time 1) and at the final OS analysis (Time 2). At Time 1, per protocol, all primary endpoints are formally tested (and thus all observed p-values are available): we assume that PFS in PDL1-expressing subjects is significant in the N+I vs. CT comparison at its initially allocated alpha. The 4 steps below demonstrate how the alpha will then be passed to the next testing based on the pre-specified testing strategy. At each step, a hypothesis is rejected (part A), its corresponding vertex and all edges will be removed; the remaining edges and nominal significance levels will be updated (part B). In the following, shaded vertices represent those hypotheses that are rejected at their nominal significance level at the given step. (For the sake of simplicity, examples of observed p-values are not provided.) (Notice the following: although the design assumes the same treatment effect and thus observed event numbers for the N+I vs. CT and N+C vs. CT comparisons, in the example, the OS information fractions observed at Time 1 are slightly different for the two comparisons. This reflects the fact that although the true treatment effect is assumed to be the same across comparisons, the observed number of events per comparison [N+I vs. CT and N+C vs. CT] is a random number.)

Step 1: At Time 1, reject primary endpoint PFS in PDL1-expressing subjects for N+I vs. CT (shaded vertex). As a result, the graph is updated as follows (see also Part B in table below):

- the rejected hypothesis and its outgoing edge is removed;
- the entire significance level of the rejected hypothesis is passed to the corresponding secondary endpoint in all randomized subjects (within the N+I vs. CT comparison).

Table 10-1: Step1: Time 1, Rejection of PFS PDL1+ for N+I vs. CT

Part A: Reject Hypothesis	Part B: Update Graph after Rejection
	

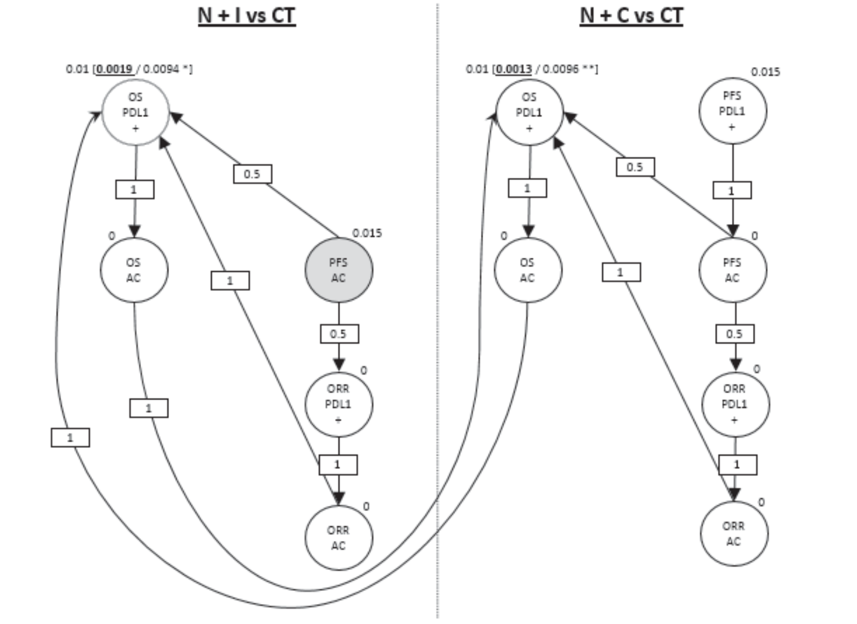
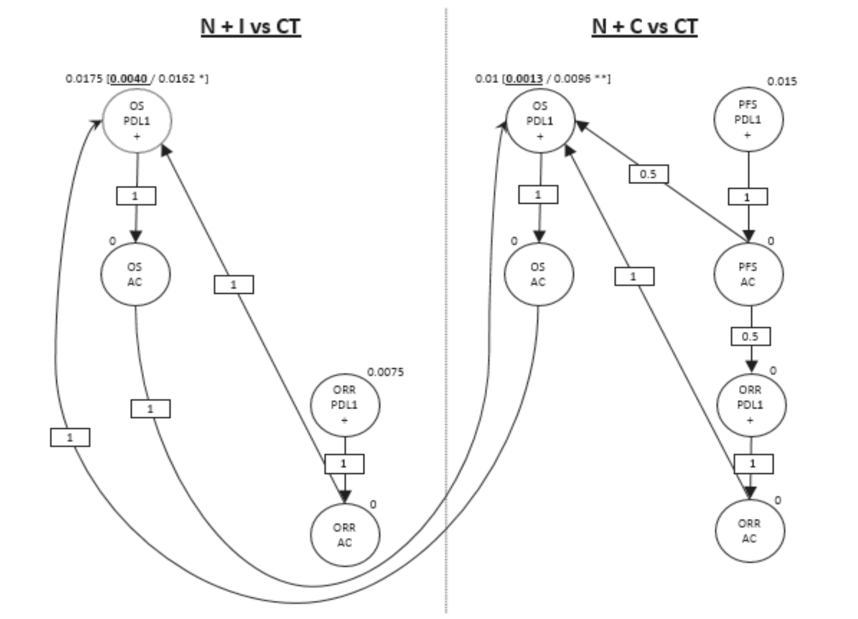
* Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 72% observed information fraction at Time 1

** Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 68% observed information fraction at Time 1

Step 2: Now assume that the secondary endpoint PFS in all randomized subjects for N+I vs. CT (shaded vertex) is significant at Time 1 at its nominal alpha. As a result, the graph is updated as follows (see also Part B in table below):

- the rejected hypothesis and its outgoing edges are removed;
- half of the rejected hypothesis' significance level is passed to ORR in PDL1-expressing subjects (secondary endpoint) - and thus it can be now tested at 0.0075;
- half of the rejected hypothesis' significance level is passed to another primary endpoint, OS in in PDL1-expressing subjects - and thus latter's local significance level is updated from 0.01 to 0.0175. The alpha spending function (O'Brien-Fleming with 72% information fraction at Time 1) is then re-applied to this updated local significance level.

Table 10-2: Step2: Time 1, Rejection of PFS AC for N+I vs. CT

Part A: Reject Hypothesis	Part B: Update Graph after Rejection
 N + I vs CT 0.01 [0.0019 / 0.0094 *] OS PDL1 + 1 0.5 0 OS AC 1 0.015 PFS AC 0.5 0 ORR PDL1 + 1 1 1 0 ORR AC N + C vs CT 0.01 [0.0013 / 0.0096 **] OS PDL1 + 0.015 PFS PDL1 + 1 0.5 1 0 OS AC 1 0 PFS AC 0.5 0 ORR PDL1 + 1 0 ORR AC	 N + I vs CT 0.0175 [0.0040 / 0.0162 *] OS PDL1 + 1 0 OS AC 1 0.0075 ORR PDL1 + 1 1 1 0 ORR AC N + C vs CT 0.01 [0.0013 / 0.0096 **] OS PDL1 + 0.015 PFS PDL1 + 1 0.5 1 0 OS AC 1 0 PFS AC 0.5 0 ORR PDL1 + 1 0 ORR AC

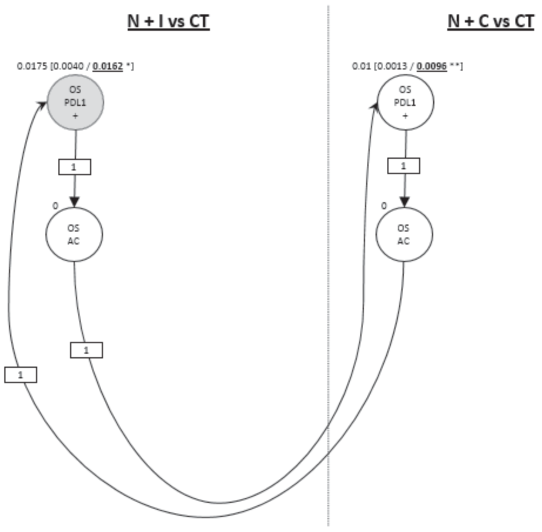
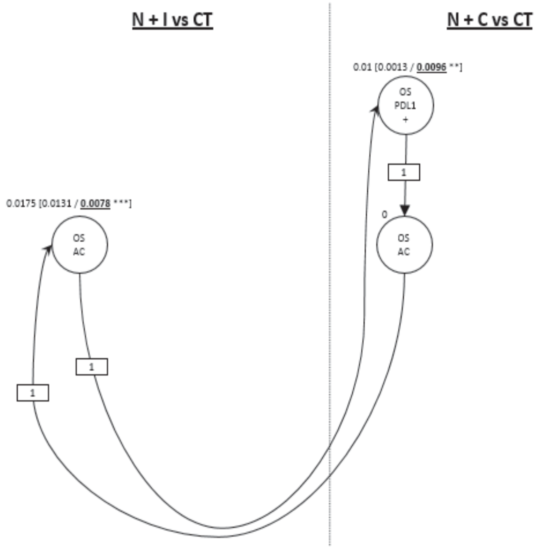
* Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 72% observed information fraction at Time 1

** Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 68% observed information fraction at Time 1

Step 3: For the sake of illustration, assume that no more endpoints can be rejected at Time 1, and thus the study continues to Time 2. As the PFS and ORR endpoints were not planned to be tested at Time 2, all related vertices and edges are removed from the graph. All OS endpoints will now use the nominal significance levels for Time 2, as determined by their alpha spending functions. Assume that we can reject primary endpoint OS in PDL1-expressing subjects for N+I vs. CT (shaded vertex) at its nominal significance level of 0.0162. As a result, the graph is updated as follows (see also Part B in the table below):

- the rejected hypothesis and its outgoing edge are removed;
- The entire local significance level of the rejected hypothesis is passed to its corresponding secondary endpoint in all randomized subjects in N+I vs. CT - and thus the secondary endpoint's local significance level (across timepoints) is updated to 0.0175. The alpha spending function (Pocock with 65% information fraction at Time 1) is then applied to this updated alpha level, and it can be tested at 0.0078.

Table 10-3: Step3: Time 2, Rejection of OS PDL1+ for N+I vs. CT

Part A: Reject Hypothesis	Part B: Update Graph after Rejection
 N + I vs CT N + C vs CT 0.0175 [0.0040 / 0.0162 *] OS PDL1 + 1 0 OS AC 1 1 0.01 [0.0013 / 0.0096 **] OS PDL1 + 1 0 OS AC	 N + I vs CT N + C vs CT 0.0175 [0.0131 / 0.0078 ***] OS AC 1 1 0.01 [0.0013 / 0.0096 **] OS PDL1 + 1 0 OS AC

* Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 72% observed information fraction at Time 1

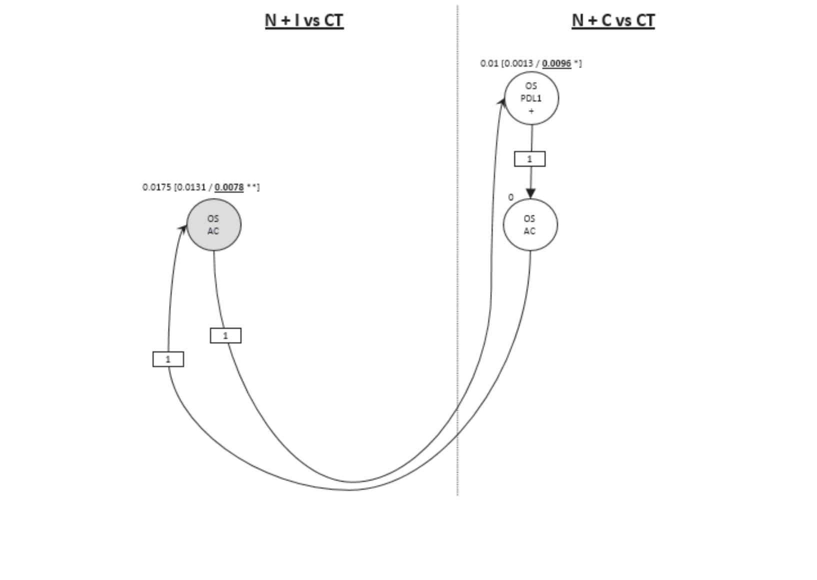
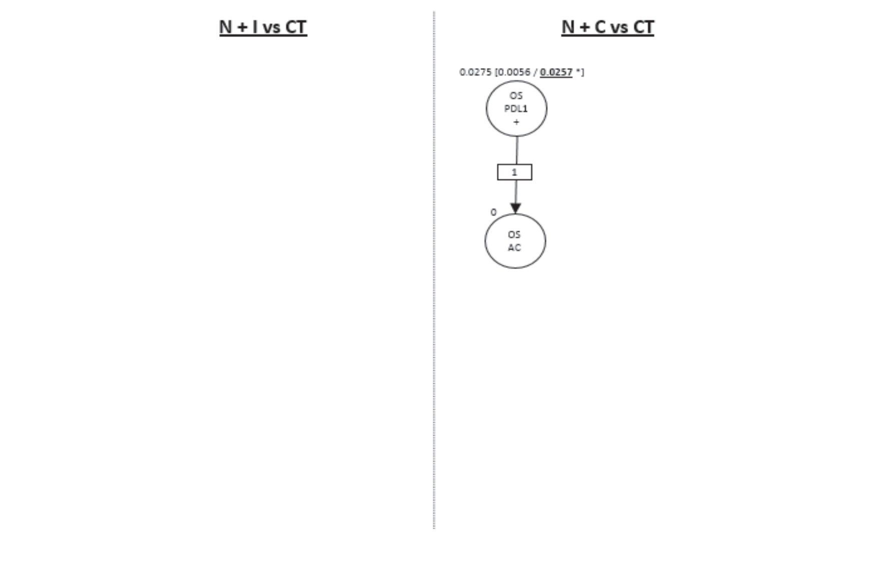
** Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 68% observed information fraction at Time 1

*** Nominal signifiance levels for Time 1 / Time 2, based on Pocock with 65% observed information fraction at Time 1

Step 4: Reject OS in all randomized subjects for N+I vs. CT (shaded vertex) at Time 2. As a result:

- the rejected hypothesis and its related edges are removed;
- its entire significance level is passed to the last testable primary endpoint, OS in PDL1-expressing subjects for N+C vs. CT. Thus latter's local significance level is updated to 0.0275. The alpha spending function (O'Brien-Fleming with 68% information fraction at Time 1) is then applied to this updated alpha level, and it can be tested at 0.0257.

Table 10-4: Step4: Time 2, Rejection of OS AC for N+I vs. CT

Part A: Reject Hypothesis	Part B: Update Graph after Rejection
 N + I vs CT N + C vs CT 0.01 [0.0013 / 0.0096 *] OS PDL1 + 1 0 OS AC 0.0175 [0.0131 / 0.0078 **] OS AC 1 1	 N + I vs CT N + C vs CT 0.0275 [0.0056 / 0.0257 *] OS PDL1 + 1 0 OS AC

* Nominal signifiance levels for Time 1 / Time 2, based on O'Brien-Fleming with 68% observed information fraction at Time 1

** Nominal signifiance levels for Time 1 / Time 2, based on Pocock with 65% observed information fraction at Time 1

Assuming that OS in PDL1-expressing subjects for N+C vs. CT cannot be rejected, the process stops.

APPENDIX 2 MISSING AND PARTIAL RADIOTHERAPY AND SURGERY DATES IMPUTATION ALGORITHMS

Procedures – Imputation Rules.

If reported procedure start date is a full valid date then set start date equal to the date part of procedure start date.

In case of partial date use imputation rules described below:

- If only day is missing then
  - If month and year of procedure match month and year of first dose date then impute as date of first dose;
  - If month and year of procedure don't match month and year of first dose date then impute as first day of that month and year.
- If both day and month are missing, then impute as maximum between 01JAN of the year and date of the first dose;
- If date is completely missing or invalid then leave missing.

Note: Imputation is not applicable to data where start date is not collected (for example "PRIOR RADIOTHERAPY" CRF). Set start date to missing in this case.

If reported end date is a full valid date then set end date equal to the date part of the reported end date.

In case of partial date use imputation rules described below:

- If reported end date is partial then set end date equal to the last possible reported end date based on the partial entered reported end date.
- If reported end date is missing, continuing, unknown or invalid then set end date equal to the most recent database extraction date.

If end date was imputed then compare end date to the death date or last known alive date if subject is not dead. If posterior then end date should be imputed to death date (or last known alive date if subject not dead).

Note: Imputation of partial dates only applies to data entered on "RADIOTHERAPY" CRF page. For other CRF pages in case of partial dates set end date to missing.

Surgeries – Imputation Rules.

If reported surgery date is a full valid date then set start date equal to the date part of surgery date.

In case of partial date, use one of the two imputation rules described below:

A. For data collected on "PRIOR SURGERY RELATED TO CANCER" CRF page:

- If only day is missing then impute as the first day of the month;

- If both day and month are missing then then impute as 01JAN of the year;
- If date is completely missing or invalid then leave missing.

B. For data collected on other CRF pages (deemed to be on-treatment/subsequent surgeries):

- If only day is missing then
  - If month and year of surgery match month and year of first dose date then impute the missing date as the date of first dose;
  - If month and year of surgery don't match month and year of first dose date then impute as first day of that month and year;
- If both day and month are missing then impute as maximum between 01JAN of the year and date of the first dose;
- If date is completely missing or invalid then leave missing.

APPENDIX 3 TIME-TO ONSET AND TIME-TO RESOLUTION DEFINITION AND CONVENTIONS FOR SELECT ADVERSE EVENTS, IMMUNE-MEDIATED ADVERSE EVENTS AND EVENTS OF SPECIAL INTEREST

Time-to onset definition

Time-to onset of AE (any grade) for a specific category is defined as the time between the day of the first dose of study treatment and the onset date of the earliest AE (of any grade) in this category.

The time-to onset of AE (grade 3-5) for a specific category is defined similarly with an onset date corresponding to a grade 3-5 AE.

Time-to onset of drug-related AE (any grade or grade 3-5) for a specific category is defined similarly but restricted to drug-related AE.

Time-to onset for a specific subcategory is defined similarly but restricted to event of this subcategory.

Time-to resolution definition

In order to derive the time-to resolution, overlapping or contiguous AEs within a specific category or subcategory will be collapsed into what will be termed “clustered” AEs. For example, if a subject (without pre-treatment AE) experienced an AE from 1st to 5th January, another AE (with different PT but within same category) from 6th to 11th January and same AE from 10th to 12th January, these will be collapsed into one clustered AE from 1st to 12th January. Table 10-5 is summarizing key derivation steps for each type of clustered AEs.

Time-to resolution of AE (any grade) for a specific category is defined as the longest time from onset to complete resolution or improvement to the grade at baseline among all clustered AEs experienced by the subject in this category per adverse event criteria category. Events which worsened into grade 5 events (death) or have a resolution date equal to the date of death are considered unresolved. If a clustered AE is considered as unresolved, the resolution date will be censored to the last known alive date. Improvement to the grade at baseline implies that all different events in the clustered adverse event should at least have improved to the corresponding (i.e. with same preferred term) baseline grade. This measure is defined only for subjects who experienced at least one AE in the specific category.

The time-to resolution of AE (grade 3-5) for a specific category is defined similarly with an onset date corresponding to a grade 3-5 AE.

Time-to resolution of drug-related AE (any grade or grade 3-5) for a specific category is defined similarly but restricted to drug-related AE.

The time-to resolution of AE (any grade or grade 3-5, drug-related or all) where immune modulating medication was initiated is defined similarly. For data presentation not restricted to IMAE, the additional condition that the subject started an immune modulating medication during the longest AE resolution period will be applied.

Time-to resolution for a specific subcategory is defined similarly but restricted to event of this subcategory.

Table 10-5: Derivation of clustered AE

Type of clustered AE	Derivation
Any grade	Collapse any on-treatment AE from the same category
Drug-related of any grade	Collapse any on-treatment drug-related AE from the same category
Grade 3-5	Collapse any on-treatment AE from the same category. Resolution will be based on the onset date of the earliest grade 3-5 records (if no grade 3-5 record, clustered AE is excluded)
Drug-related of Grade 3-5	Collapse any on-treatment drug-related AE from the same category Resolution will be based on the onset date of the earliest grade 3-5 record (if no Grade 3-5 record, clustered AE is excluded)

The algorithm for collapsing adverse event records is using the following conventions:

For each subject and specified category, the corresponding adverse event records will be collapsed when:

1) Multiple adverse event records have the same onset date.
2) The onset date of an event record is either the same day or 1 day later than the resolution date of a preceding event record (contiguous events).
3) The onset date of an event record is after the onset date and prior to or on the resolution date of a preceding event record (overlapping events).

APPENDIX 4 IMMUNOGENICITY ANALYSIS: BACKGROUND AND RATIONALE

The following summary is from the FDA Guidance for Industry Immunogenicity Assessment for Therapeutic Protein Products and White Paper on Assessment and Reporting of the Clinical Immunogenicity of Therapeutic Proteins and Peptides – Harmonized Terminology and Tactical Recommendations by Shankar et al. The program-level definitions of sample- and subject-level ADA status are based on recommendation from the BMS Immunogenicity Council.

Immune responses to therapeutic protein products may pose problems for both patient safety and product efficacy. Immunologically based adverse events, such as anaphylaxis and infusion reactions, have caused termination of the development of therapeutic protein products or limited the use of otherwise effective therapies. Unwanted immune responses to therapeutic proteins may also neutralize the biological activity of therapeutic proteins and may result in adverse events not only by inhibiting the efficacy of the therapeutic protein product, but by cross-reacting to an endogenous protein counterpart, if present. Because most of the adverse effects resulting from elicitation of an immune response to a therapeutic protein product appear to be mediated by humoral mechanisms, circulating antibody has been the chief criterion for defining an immune response to this class of products.

ADA is defined as biologic drug-reactive antibody, including pre-existing host antibodies that are cross-reactive with the administered biologic drug (baseline ADA). Titer is a quasiquantitative expression of the level of ADA in a sample. By employing a serial dilution-based test method, titer is defined as the reciprocal of the highest dilution of the sample (e.g., dilution of 1/100 = titer of 100). The ADA is also tested, via a cell-based biologic assay or a non cell-based competitive ligand-binding assay for a subpopulation of ADA known as neutralizing antibodies (NAb), which inhibits or reduces the pharmacological activity of the biologic drug molecule regardless of its in vivo clinical relevance. Non-neutralizing ADA (non-NAb) is ADA that binds to the biologic drug molecule but does not inhibits its pharmacological activity.

ADA should be tested using sensitive and valid methods and employing an appropriate strategy for elucidating immunogenicity. Detection of ADA is typically performed in three tiers (screening, confirmatory, and titer) using statistically determined cutpoints and samples testing positive in the ADA assay are analyzed for neutralizing activity, especially in late-stage clinical studies. “Detection” of ADA implies that drug-specific ADA was confirmed. The ‘drug tolerance” of an assay (highest drug concentration that does not interfere in the ADA detection method) is not an absolute value and differs between individuals due to the varying avidities of ADA immune responses. An ADA sampling strategy of collecting samples at times when the least drug concentration is anticipated (trough concentrations) can increase the likelihood of accurate ADA detection.

It is useful to present ADA results from clinical studies as (a) characteristics of the ADA immune response, (b) relationship of ADA with pharmacokinetics (PK) and, when relevant, pharmacodynamics (PD) biomarkers, and (c) relationship of ADA with clinical safety and efficacy.

Clinical consequences of ADA can range from no apparent clinical effect to lack of efficacy (primary treatment failure), loss of efficacy (secondary treatment failure) or heightened effect due to altered exposure to the biologic drug, adverse drug reactions (administration-related systemic or site reactions), and severe adverse drug reactions (anaphylaxis and unique clinical problems associated with cross-reactivity and neutralization of endogenous molecules). Thus it becomes important to examine any associations between ADA or any of its attributes with the various clinical sequelae. The presence of ADA may or may not preclude the administration of drug to ADA-positive subjects because the outcome is dependent upon the magnitude of the impact of ADA on PK and PD. Hence, the relationship of ADA with PK/PD is an important additional consideration, but does not necessarily result in a clinically impactful consequence per se.

Immunogenicity Endpoints

A fundamental metric that informs clinical immunogenicity interpretation is the incidence of ADA in a study or across comparable studies. ADA incidence is defined as the proportion of the study population found to have seroconverted or boosted their pre-existing ADA during the study period.

Terms and Definitions

Validated ADA test methods enable characterization of samples into ADA-positive vs. ADA-negative. To classify the ADA status of a subject using data from an in vitro test method, each sample from the subject is categorized based on the following definitions:

Sample ADA Status**:**

- Baseline ADA-positive sample: ADA is detected in the last sample before initiation of treatment
- Baseline ADA-negative sample: ADA is not detected in the last sample before initiation of treatment
- ADA-positive sample: After initiation of treatment, (1) an ADA detected (positive seroconversion) sample in a subject for whom ADA is not detected at baseline, or (2) an ADA detected sample with ADA titer to be at least 4-fold or greater (≥) than baseline positive titer
- ADA-negative sample: After initiation of treatment, ADA not positive sample relative to baseline

Next, using the sample ADA status, subject ADA status is defined as follows:

Subject ADA Status:

- Baseline ADA-positive subject: A subject with baseline ADA-positive sample
- **ADA-positive subject**: A subject with at least one ADA positive-sample relative to baseline at any time after initiation of treatment

1) *Persistent Positive (PP):* ADA-positive sample at 2 or more consecutive time points, where the first and last ADA-positive samples are at least 16 weeks apart
2) *Not PP-Last Sample Positive:* Not persistent positive with ADA-positive sample at the last sampling time point

3) *Other Positive:* Not persistent positive but some ADA-positive samples with the last sample being negative
4) *Neutralizing Positive:* At least one ADA-positive sample with neutralizing antibodies detected

- **ADA-negative subject:** A subject with no ADA-positive sample after the initiation of treatment.

(Note: 16 weeks was chosen based on a long half-life of IgG4.)

Population for Analyses

Analysis of immunogenicity data will be based on ADA evaluable subjects defined as all treated subjects with baseline and at least 1 post-baseline immunogenicity assessment. Analysis dataset and data listing will include all available ADA samples. However, subject-level ADA status will be defined based on only adequate samples (e.g., excluding 1-hour post-infusion samples when clearly indicated).

APPENDIX 5 ANALYSES OF DATA FROM CHINA

Analysis methods for China

Analyses detailed in SAP will be repeated for the patients randomized in China using the analysis sets described in this appendix if feasible for an individual analysis (eg, see the minimal subject required in Section 7.5.4.3). The analysis methods for the China subpopulation will be the same as for the global population unless otherwise noted. No formal hypothesis testing will be performed to evaluate consistency of China subpopulation. Instead, summary statistics and estimates will be provided to assess the consistency descriptively. All statistical analyses for China will be performed when at least one of the primary endpoints is statistically significant.

No adjustment for multiplicity will be made.

Only unstratified analyses will be performed for efficacy endpoints and the sensitivity analysis of investigating non-proportionality in Kaplan-Meier curves will not be performed.

Definition of analysis sets

Primary China subpopulation

Primary China subpopulation is defined as certified Greater China subgroup. It should contain subjects that are Chinese by race and enrolled from all Mainland China sites, National Medical Products Administration (NMPA) certified Taiwan sites and NMPA certified Hong Kong sites.

Secondary China subpopulation

Secondary China subpopulation is defined as Greater China subgroup. It should contain all subjects that are Chinese by race and enrolled from Mainland China, Taiwan and Hong Kong.

Asian subpopulation

Asian subpopulation is defined as Asian subgroup. It should contain subjects that are Asian by race and enrolled from Asian countries which are defined as the combination of J/K/T and Rest of Asia.

APPENDIX 6 R PROGRAM FOR THE STUDY DESIGN

This appendix includes the program (R-v3.1.3) that was used for the study design (as specified in the protocol).

```
######################################################################
# Program name: CA209648 FINAL DESIGN SAP
# Purpose: Sample size calculation for CA209648 revprot1.0
# Author:[redacted]
# Date: Dec 2016
# Comment: CA209648 design assumes the same OS/PFS for the two experimental
arms (N+C and N+I)
#          This program generates datasets for
#               - both primary endpoints of PFS and OS (generated by separate
calls of the same macros)
#             - the control vs ONE of the two experimental arms
#                    Consequently, the # of randomized subjects, # of events,
etc will be understood
#                   PER COMPARISON (i.e. N+I vs. CT and N+C vs. CT)
# Program Description:
#  I. Set up Macros
#        5 macros: my.summary, Report_Results, Run_iterations, DataCohort,
PrimaryTest
#     Structure:
#     (#my.summary#: customized summary of simulation results)
#     #Report_results#
#        1. calls #Run_Iterations#: to run simulations and collect results in
one dataframe;
#           For each iteration:
#             i. calls #DataCohort#: to generate time to event data for each
cohort
#           ii. pools cohort results and adds random randomization month/day to
each subject
#           iii. calls #PrimaryTest#:
#               a) for each patient, adds censoring;
#               b) runs Cox PH model
#        2. summarizes and structures simulation results for presentation
#  II. Set up calls (for accrual; for OS and PFS)
##############################################################################
#######

######################################################################

library(msm)            # package to access rpexp (to generate piecewise exp
function)
library(ldbounds)     # package to calculate significance level given an alpha
spending function
library(survival)     # coxph, survfit
options(scipen=999)   # to disable scientific format in output

# self-made function for customized summary of simulation results (including
90% CI)
my.summary <- function(x){
  c(min=round(min(x,na.rm=TRUE),3),
    percent=round(quantile(x, probs=0.05,na.rm=TRUE),3),
    median=round(median(x,na.rm=TRUE),3),
    mean=round(mean(x,na.rm=TRUE),3),
    percent=round(quantile(x, probs=0.95,na.rm=TRUE),3),
    max=round(max(x,na.rm=TRUE),3),
    N.A=sum(is.na(x)))
}

##############################################################################
#######
######### I. Set up Macros
##############################################################################
#######

######################################################################
####### DataCohort generate time to event data for each cohort
##### INPUT Parameters
## acc_num:             a vector containing the accrual at each month, can have
zero element
## ctl_mos, trt_mos:    median of the control/experimental group
## trt_cure_os:         % of patients cured in experimental group
## delay_OS:            length of delayed effect (months)
## nprop_trt:           % of patients randomized to the experimental arm
##### OUTPUT
#### A data matrix consist of:
## trt:                 treatment assignment (0 = control, 1 = experimental)
## OS:                  time to event for given endpoint (without censoring)
######################################################################
DataCohort = function(acc_num, ctl_mos, trt_mos, trt_cure_os, delay_OS,
nprop_trt)
{
  DataOutput = NULL
  # calculate parameter for exp distribution: control arm, experimental arm
before and after delay
  ctl_lambda_os = log(2)/ctl_mos
  trt_lambda_os = log(2)/trt_mos
  # set up empty objects
  trt = matrix(0,acc_num,1)
  OS  = matrix(0,acc_num,1)

  ####### Generate vector of treatment assignments (0 = control, 1 =
experimental)
      for (i in 1:acc_num){
        # for each cohort, first half of the vector is experimental, the 2nd
half is control;
        # if acc_num is uneven, the last element's treatment assignment is
determined at random
        # this reflects block structure
        if(i<=floor(nprop_trt*acc_num)){
          trt[i,1] =1
        }
        if(i>ceiling(nprop_trt*acc_num)){
          trt[i,1] =0
        }
        if(i==ceiling(nprop_trt*acc_num) & ceiling(nprop_trt*acc_num)-
floor(nprop_trt*acc_num)>0){
          trt[i,1] =rbinom(1,1,nprop_trt)
        }

  ######    Generate time to event
```

```
        #### control group: exponential distribution
        if(trt[i,1]==0){
            OS[i,1] = rexp(1, rate = ctl_lambda_os)
        }

        #### treatment group
        if(trt[i,1]==1){
          seed1 = runif(1)  # random number to decide of each patient whether
cure or not

          # cured patients: for those who live after delay, very long survival
is generated
          if(seed1 < trt_cure_os){
            OS[i,1]  =  rpexp(1,  rate  =  c(ctl_lambda_os,  log(2)/(10^6)),
t=c(0,delay_OS))
          }

          # non-cure patients: for those who live after delay, exponential
distribution
          if(seed1 >= trt_cure_os){
            OS[i,1]  =  rpexp(1,  rate  =  c(ctl_lambda_os,  trt_lambda_os),
t=c(0,delay_OS))
          }
        }
      }

DataOutput = cbind(trt, OS)
      colnames( DataOutput)=c("trt", "OS")

return(DataOutput)
}


#######################################################################
####### PrimaryTest: 1. for each patient, adds randomization date and censoring;
#                    2. runs Cox PH model
##### INPUT Parameters
## OS_fraction_pd:  Information fraction for IA in the primary subpopulation
(If 1 then no IA)
## event_OS_FA_pd:  Final number of events in the primary subpopulation (defines
end of study)
## DataPool:        Output vector from Run_Iterations, after running DataCohort
for all cohorts, and adding entry dates
##                                                            ("trt",   "OS",
"entry_day","dummy_order","n_percohort1","n_percohort2","n_percohort3","n_per
cohort4", "group_ind","dur_accr","entry_month")
#  Note  that  the  1st  3  parameters  for  PrimaryTest  are  identified  in  the
Report_results call

##### OUTPUT
#### A data matrix consist of:
### results of Cox regression:
# event_OS_FA_sim, event_OS_IA_sim, event_OS_FA_pd,     event_OS_IA_pd:      #
of events at Final and Interim, each for AC and PDL1+, both arms
# event_OS_FA_ctrl,event_OS_IA_ctrl,event_OS_FA_ctrl_pd,event_OS_IA_ctrl_pd: #
of events at Final and Interim, each for AC and PDL1+, control arm
```

```
# event_OS_FA_trt, event_OS_IA_trt, event_OS_FA_trt_pd, event_OS_IA_trt_pd:  #
of events at Final and Interim, each for AC and PDL1+, experimental arm
# OS_IA_time, OS_IA_time_pd, OS_FA_time, OS_FA_time_pd:                    time
of analysis at interim (AC and PDL1+) and at Final (AC and PDL1+), months
# hr_OS_IA, hr_OS_IA_pd, hr_OS_FA, hr_OS_FA_pd:                       simulated
HR at interim (AC and PDL1+) and at Final (AC and PDL1+)
#    pvalue_OS_IA,    pvalue_OS_IA_pd,    pvalue_OS_FA,    pvalue_OS_FA_pd:
simulated p-value at interim (AC and PDL1+) and at Final (AC and PDL1+)
#  kmmedian_IA_ctl,  kmmedian_IA_pd_ctl,  kmmedian_FA_ctl,  kmmedian_FA_pd_ctl:
simulated Kaplan-Meier medians at interim (AC and PDL1+) and at Final (AC and
PDL1+) - Control arm
#  kmmedian_IA_trt,  kmmedian_IA_pd_trt,  kmmedian_FA_trt,  kmmedian_FA_pd_trt:
simulated Kaplan-Meier medians at interim (AC and PDL1+) and at Final (AC and
PDL1+) - Experimental arm
### simulated results checking that simulation is correct:
# n_ctrl,n_trt,n_ctrl_pd,n_trt_pd:                     simulated # of randomized
subjects for control/experimental, each for AC and PDL1+
#  n_percohort1,n_percohort2,n_percohort3,n_percohort4:  simulated  #  of  total
randomized subjects within each cohort
# simprop_trt:                                         simulated % of randomized
patients in the experimental arm
# dur_accr:                                             duration of accrual (the
same for every iteration - depends on accr_total and N_rand)
####################################################################

PrimaryTest = function(OS_fraction_pd, event_OS_FA_pd, DataPool){

  ##### reading the data sets and generating variables needed for the analysis
  treat = DataPool[,"trt"]
  entry_month = DataPool[,"entry_month"]
  entry_day = DataPool[,"entry_day"]
  OS = DataPool[,"OS"]
  PDL1 = ifelse(DataPool[,"group_ind"]==1 | DataPool[,"group_ind"]==3,1,0)
  group_ind = DataPool[,"group_ind"]
  treat_pd = treat[PDL1==1]
  OS_pd = OS[PDL1==1]

  ##### reading the data sets and generating variables needed only for reporting
  dur_accr=median(DataPool[,"dur_accr"])
  n_percohort1 = median(DataPool[,"n_percohort1"])  #(same for every value)
  n_percohort2 = median(DataPool[,"n_percohort2"])
  n_percohort3 = median(DataPool[,"n_percohort3"])
  n_percohort4 = median(DataPool[,"n_percohort4"])

  ##### calculate which day did the subject joint the study (entry_month and
entry_day was randomly assigned to each subject in Run_Iterations)
  enter = (entry_month - 1) + entry_day

  ##### days up to event/censoring
  OS_day = enter + OS
  OS_day_pd = OS_day[PDL1==1]

  OS_day_sorted = sort(OS_day)
  OS_day_sorted_pd = sort(OS_day_pd)

  ####Determine PDL1+ analysis time
  # FA
  OS_FA_time_pd = OS_day_sorted_pd[event_OS_FA_pd] + 0.01
  # IA
  if(OS_fraction_pd==1){OS_IA_time_pd = OS_FA_time_pd}
  if(OS_fraction_pd                    <1){OS_IA_time_pd                        =
OS_day_sorted_pd[ceiling(event_OS_FA_pd*OS_fraction_pd)] + 0.01}
  # AC analyses are at the time of PDL1+
  OS_IA_time=OS_IA_time_pd
  OS_FA_time=OS_FA_time_pd

  # time from randomization until time of analysis
  follow_OS_IA     = OS_IA_time - enter             # IA in AC
  follow_OS_IA_pd  = OS_IA_time_pd - enter[PDL1==1] # IA in PDL1+
  follow_OS_FA     = OS_FA_time - enter             # FA in AC
  follow_OS_FA_pd  = OS_FA_time_pd - enter[PDL1==1] # FA in PDL1+
  # set up censoring: OS <= follow_up, event, then censor=1 o.w. censor = 0
(censored)
  censor_OS_IA = ( OS <= follow_OS_IA ) + 0                  # IA in AC
  censor_OS_IA_pd = ( OS[PDL1==1] <= follow_OS_IA_pd ) + 0   # IA in PDL1+
  censor_OS_FA = ( OS <= follow_OS_FA ) + 0                  # FA in AC
  censor_OS_FA_pd = ( OS[PDL1==1] <= follow_OS_FA_pd ) + 0   # FA in PDL1+

  # set up vectors for censored survival time
  OS_IA <- OS_FA <- OS                # IA and FA in AC
  OS_IA_pd <- OS_FA_pd <- OS[PDL1==1] # IA and FA in PDL1+
  # truncate time-to-event until analysis time
  OS_IA[which(OS_IA>follow_OS_IA)] = follow_OS_IA[which(OS_IA>follow_OS_IA)]
  OS_IA_pd[which(OS_IA_pd>follow_OS_IA_pd)]                                   =
follow_OS_IA_pd[which(OS_IA_pd>follow_OS_IA_pd)]
  OS_FA[which(OS_FA>follow_OS_FA)] = follow_OS_FA[which(OS_FA>follow_OS_FA)]
  OS_FA_pd[which(OS_FA_pd>follow_OS_FA_pd)]                                   =
follow_OS_FA_pd[which(OS_FA_pd>follow_OS_FA_pd)]

  ## generating KM medians
  # control arm
  kmmedian_IA_ctl      <-      quantile(survfit(Surv(OS_IA,censor_OS_IA)      ~
treat))[[1]][3]
  kmmedian_IA_pd_ctl   <-   quantile(survfit(Surv(OS_IA_pd,censor_OS_IA_pd)   ~
treat_pd))[[1]][3]
  kmmedian_FA_ctl      <-      quantile(survfit(Surv(OS_FA,censor_OS_FA)      ~
treat))[[1]][3]
  kmmedian_FA_pd_ctl   <-   quantile(survfit(Surv(OS_FA_pd,censor_OS_FA_pd)   ~
treat_pd))[[1]][3]
  # experimental arm
  kmmedian_IA_trt      <-      quantile(survfit(Surv(OS_IA,censor_OS_IA)      ~
treat))[[1]][4]
  kmmedian_IA_pd_trt   <-   quantile(survfit(Surv(OS_IA_pd,censor_OS_IA_pd)   ~
treat_pd))[[1]][4]
  kmmedian_FA_trt      <-      quantile(survfit(Surv(OS_FA,censor_OS_FA)      ~
treat))[[1]][4]
  kmmedian_FA_pd_trt   <-   quantile(survfit(Surv(OS_FA_pd,censor_OS_FA_pd)   ~
treat_pd))[[1]][4]

  ###### Cox models with log rank test
  # IA, AC
  survtest_OS_IA <- coxph(Surv(OS_IA,censor_OS_IA) ~ treat)
  event_OS_IA_sim = summary(survtest_OS_IA)$nevent
  hr_OS_IA = summary(survtest_OS_IA)$conf.int[1]
  pvalue_OS_IA =  summary(survtest_OS_IA)$waldtest[3]
  # IA, PDL1+
  survtest_OS_IA_pd <- coxph(Surv(OS_IA_pd,censor_OS_IA_pd) ~ treat_pd)
  event_OS_IA_pd = summary(survtest_OS_IA_pd)$nevent
  hr_OS_IA_pd = summary(survtest_OS_IA_pd)$conf.int[1]
  pvalue_OS_IA_pd =  summary(survtest_OS_IA_pd)$waldtest[3]
  # FA, AC
  survtest_OS_FA <- coxph(Surv(OS_FA,censor_OS_FA) ~ treat)
  event_OS_FA_sim  = summary(survtest_OS_FA)$nevent
  hr_OS_FA  = summary(survtest_OS_FA)$conf.int[1]
  pvalue_OS_FA  =  summary(survtest_OS_FA)$waldtest[3]
  # FA, PDL1+
  survtest_OS_FA_pd <- coxph(Surv(OS_FA_pd,censor_OS_FA_pd) ~ treat_pd)
  event_OS_FA_sim_pd  = summary(survtest_OS_FA_pd)$nevent
  hr_OS_FA_pd  = summary(survtest_OS_FA_pd)$conf.int[1]
  pvalue_OS_FA_pd  =  summary(survtest_OS_FA_pd)$waldtest[3]

  # number of events by arm (for reporting)
  event_OS_FA_ctrl <- length(censor_OS_FA[censor_OS_FA==1 & treat==0])
  event_OS_FA_trt <- length(censor_OS_FA[censor_OS_FA==1 & treat==1])
  event_OS_IA_ctrl <- length(censor_OS_IA[censor_OS_IA==1 & treat==0])
  event_OS_IA_trt <- length(censor_OS_IA[censor_OS_IA==1 & treat==1])
  event_OS_FA_ctrl_pd    <-    length(censor_OS_FA_pd[censor_OS_FA_pd==1    &
treat_pd==0])
  event_OS_FA_trt_pd    <-    length(censor_OS_FA_pd[censor_OS_FA_pd==1    &
treat_pd==1])
  event_OS_IA_ctrl_pd    <-    length(censor_OS_IA_pd[censor_OS_IA_pd==1    &
treat_pd==0])
  event_OS_IA_trt_pd    <-    length(censor_OS_IA_pd[censor_OS_IA_pd==1    &
treat_pd==1])
  # simulated # of randomized patients (to check simulations)
  n_ctrl <- length(censor_OS_FA[treat==0])          # AC, control arm
  n_trt <- length(censor_OS_FA[treat==1])           # AC, experimental arm
  n_ctrl_pd <- length(censor_OS_FA_pd[treat_pd==0])  # PDL1+, control arm
  n_trt_pd <- length(censor_OS_FA_pd[treat_pd==1])   # PDL1+, experimental arm
  # simulated % of cured patients (to check simulations)
  simprop_trt <- n_trt/(n_ctrl+n_trt)

  # generate output vector from PrimaryTest, for the ith iteration
  re=c(event_OS_IA_sim, OS_IA_time, event_OS_IA_pd ,OS_IA_time_pd, OS_FA_time,
OS_FA_time_pd,   event_OS_FA_sim,event_OS_FA_pd,   hr_OS_IA,   pvalue_OS_IA,
hr_OS_IA_pd,        pvalue_OS_IA_pd,        hr_OS_FA,        pvalue_OS_FA,
pvalue_OS_FA_pd,hr_OS_FA_pd,

event_OS_FA_ctrl,event_OS_FA_trt,event_OS_IA_ctrl,event_OS_IA_trt,event_OS_FA
_ctrl_pd,event_OS_FA_trt_pd,event_OS_IA_ctrl_pd,event_OS_IA_trt_pd,
       n_ctrl,n_trt,n_ctrl_pd,n_trt_pd,simprop_trt, dur_accr,
       n_percohort1,n_percohort2,n_percohort3,n_percohort4,
       kmmedian_IA_ctl,        kmmedian_IA_pd_ctl,        kmmedian_FA_ctl,
kmmedian_FA_pd_ctl,  kmmedian_IA_trt,  kmmedian_IA_pd_trt,  kmmedian_FA_trt,
kmmedian_FA_pd_trt)
  names(re) = c("event_OS_IA", "OS_IA_time", "event_OS_IA_pd","OS_IA_time_pd",
"OS_FA_time",      "OS_FA_time_pd",      "event_OS_FA_sim","event_OS_FA_pd",
"hr_OS_IA",  "pvalue_OS_IA",  "hr_OS_IA_pd",  "pvalue_OS_IA_pd",  "hr_OS_FA",
"pvalue_OS_FA", "pvalue_OS_FA_pd", "hr_OS_FA_pd",

"event_OS_FA_ctrl","event_OS_FA_trt","event_OS_IA_ctrl","event_OS_IA_trt","ev
```

```
ent_OS_FA_ctrl_pd","event_OS_FA_trt_pd","event_OS_IA_ctrl_pd","event_OS_IA_tr
t_pd",
                "n_ctrl","n_trt","n_ctrl_pd","n_trt_pd","simprop_trt",
"dur_accr",
                "n_percohort1","n_percohort2","n_percohort3","n_percohort4",
                "kmmedian_IA_ctl",  "kmmedian_IA_pd_ctl",  "kmmedian_FA_ctl",
"kmmedian_FA_pd_ctl",        "kmmedian_IA_trt",        "kmmedian_IA_pd_trt",
"kmmedian_FA_trt", "kmmedian_FA_pd_trt")
  return(re)
}

####################################################################
####### Run_Iterations: for each iteration, creates time to event datasets and
result datasets from Cox PH analyses
##### INPUT Parameters
# Note that all parameters are identified in the Report_results call
## N_sim:           Number of datasets (iterations) to be simulated
## event3:         Final number of events in the primary subpopulation (defines
end of study)
## random_seed:     seed used for random generation
## OS_fraction_pd:  Information fraction for IA in the primary subpopulation
(If 1 then no IA)
## acc_num:         accrual pattern
## N_rand:          # of randomized subjects (will be less than sum(acc_num))
##### OUTPUT
#### A data matrix generating all output variables from PrimaryTest, but one
row for each iteration
####################################################################

Run_Iterations = function(N_sim, event3, random_seed, OS_fraction_pd, acc_num,
N_rand){

  set.seed(random_seed)
  result = NULL
  result_list = list(result)
  # read in parameter values for Datacohort
  group1          =                    list(prop_group1,          ctl_mos_1,
trt_mos_1,trt_cure_os_1,delay_OS_1,nprop_trt_1)
  group2          =                    list(prop_group2,          ctl_mos_2,
trt_mos_2,trt_cure_os_2,delay_OS_2,nprop_trt_2)
  group3          =                    list(prop_group3,          ctl_mos_3,
trt_mos_3,trt_cure_os_3,delay_OS_3,nprop_trt_3)
  group4          =                    list(prop_group4,          ctl_mos_4,
trt_mos_4,trt_cure_os_4,delay_OS_4,nprop_trt_4)

  # create empty month vector for accrual period (length = # of randomized
patients)
  entry_month <- matrix(0,N_rand,1)
  n = length(acc_num)                  # # of months in accrual period
  # fill entry_month with values:
  # each month value (1, ..., n) will be included as many times as many patients
are randomized during that period
  for (month in 1:n){
    Mn = acc_num[month]         # # of patients randomized in this month
    # generate Mn_cum (= cumulative accrual until beginning of given month)
    if(month==1){Mn_cum=0}
    if(month>1){Mn_cum = sum(acc_num[1:(month-1)])}
    # Case#1: all subjects are to be recruited within the given month
    if(Mn_cum+Mn < N_rand){
    if(Mn > 0){entry_month[(Mn_cum+1):min((Mn_cum+Mn),N_rand),1] = month}
    }
    # Case#2  (last  month  of  the  accrual):  not  necessarily  all  subjects  from
acc_num are needed anymore
    if(Mn_cum+Mn >= N_rand & Mn_cum < N_rand){
    if(Mn > 0){entry_month[(Mn_cum+1):min((Mn_cum+Mn),N_rand),1] = month
               dur_accr = month}
    }
  }

  #################################################
  ### SIMULATIONS BEGIN ###########################
  #################################################
  for (NN in 1:N_sim){

    # empty vectors
    entry_day <- dummy_order <- group_ind <- matrix(0,N_rand,1)
    #  determine  #  of  subjects  per  cohort:  random  sample  from  multinomial
distribution
    n_percohort    <-    rmultinom(n=1,    size    =    N_rand,    prob    =
c(group1[[1]],group2[[1]],group3[[1]],group4[[1]]))
    # retrieve cohort sizes for reporting purposes
    n_percohort1 <- n_percohort[[1]]
    n_percohort2 <- n_percohort[[2]]
    n_percohort3 <- n_percohort[[3]]
    n_percohort4 <- n_percohort[[4]]
    # create variable that indicates cohort assignment for each patient
    group_ind[1:n_percohort[[1]],1] = 1

if(n_percohort[[2]]>0){group_ind[(n_percohort[[1]]+1):(n_percohort[[1]]+n_per
cohort[[2]]),1] = 2}

if(n_percohort[[3]]>0){group_ind[(n_percohort[[1]]+n_percohort[[2]]+1):(n_per
cohort[[1]]+n_percohort[[2]]+n_percohort[[3]]),1] = 3}

if(n_percohort[[4]]>0){group_ind[(n_percohort[[1]]+n_percohort[[2]]+n_percoho
rt[[3]]+1):(n_percohort[[1]]+n_percohort[[2]]+n_percohort[[3]]+n_percohort[[4
]]),1] = 4}

    # run DataCohort for each cohort
      Data_group1                                                            =
DataCohort(n_percohort[[1]],group1[[2]],group1[[3]],group1[[4]],group1[[5]],g
roup1[[6]])
      Data_group2                                                            =
DataCohort(n_percohort[[2]],group2[[2]],group2[[3]],group2[[4]],group2[[5]],g
roup2[[6]])
      Data_group3                                                            =
DataCohort(n_percohort[[3]],group3[[2]],group3[[3]],group3[[4]],group3[[5]],g
roup3[[6]])
      Data_group4                                                            =
DataCohort(n_percohort[[4]],group4[[2]],group4[[3]],group4[[4]],group4[[5]],g
roup4[[6]])
    # pool data
      Data = rbind(Data_group1, Data_group2, Data_group3, Data_group4)
    # ADD RANDOMIZATION DATE (entry_month, entry_day)
    for (k in 1:N_rand){
      entry_day[k,1] = round(runif(1))  # entry day is random within a month
      dummy_order[k,1] = runif(1)               # this will be used to assign
each patient to a random randomization month
    }
    Data                                                                   <-
cbind(Data,entry_day,dummy_order,n_percohort1,n_percohort2,n_percohort3,n_per
cohort4, group_ind, dur_accr)
    Data <- Data[order(dummy_order),]
    Data  <-  cbind(Data,entry_month)      #added  entry_month  to  the  randomly
ordered dataset

    colnames(Data)=c("trt",                                               "OS",
"entry_day","dummy_order","n_percohort1","n_percohort2","n_percohort3","n_per
cohort4", "group_ind","dur_accr","entry_month")

    # run Cox regression on the pooled cohorts
    test=PrimaryTest(event_OS_FA_pd=event3,     OS_fraction_pd=OS_fraction_pd,
DataPool=Data)

    result_list[[1]] = rbind(result_list[[1]], test)

    if(NN %% 250 == 0) {
      print("Generation of samples is ongoing, latest iteration number:", quote
= FALSE)
      print(NN)
      }

  }
  #################################################
  ### SIMULATIONS END ##############################
  #################################################

  return(result_list)
}

#######################################################################
####### Report_results: determines significance/power; summarizes and outputs
all results
##### INPUT Parameters
# same input parameters as for Run_Iterations; in addition:
## alpha_OS:        Significance level for the analysis based on all randomized
(secondary), overall across timepoints
## alpha_OS_pd:    Significance level for the analysis based on PDL1+ (primary),
overall across timepoints
##### OUTPUT
#### A list including the following elements:
### Elements to display simulation parameters (input):
## sim.par:              Simulation Parameters in function Report_results
## sim.par2:             Simulation Parameters by Cohort
### Elements to check that simulation produces result as expected (summary
statistics of generated parameters across iterations):
## result_nrand:         Checking Simulation: # of Randomized Subjects
## result_nrandcoh:      Checking Simulation: % Randomized by Cohort
### Elements displaying simulation results, presented in the protocol (summary
statistics of generated parameters across iterations):
```

```
## powerOS_DFS:              SIMULATED POWER (Reported in Protocol Table 8.1-
1,8.1-2)
## result_hr:          Simulated HR (Protocol Table 8.1-1,8.1-2: 'Hypothesized
Overall HR')
## max_hr_result:        Maximum Significant HR (Protocol Table 8.1-1,8.1-2:
'Critical HR')
## min_obs_median_result: Minimum Significant Difference in Medians, months
(Protocol Table 8.1-1,8.1-2: 'Minimal Difference in Median')
## result_alpha:         Alpha Levels Used in the Simulated Datasets  (Protocol
Table 8.1-1,8.1-2: 'Significance Level' for IA and FA)
## result_event:           # of Simulated Events (Protocol Table 8.1-1,8.1-2:
'Projected # of Events')
## result_kmmedian:        Simulated KM Medians, months (Protocol Table 8.1-
1,8.1-2: 'Hypothesized Median...')
## result_dur:              Enrolment Period and Projected Time of Analyses,
months  (Protocol Table 8.1-1: 'Projected Time of LPLV (from FPFV)')
####################################################################

Report_results = function(N_sim, seed, alpha_OS, alpha_OS_pd, event_OS_FA_pd,
OS_fraction_pd, accr_pattern, N_rand){

  all_result     =     Run_Iterations(N_sim=N_sim,     event3=event_OS_FA_pd,
random_seed=seed,     OS_fraction_pd=OS_fraction_pd,     acc_num=accr_pattern,
N_rand=N_rand)

  #SIGNIFICANCE
  alpha_OS_IA   <-   alpha_OS_FA   <-   alpha_OS_IA_pd   <-   alpha_OS_FA_pd   <-
rep(0,N_sim)
if(OS_fraction_pd<1){
  bds_nivo_pd<-bounds(t=c(OS_fraction_pd,1),                      iuse=c(1,1),
alpha=c(alpha_OS_pd/2,alpha_OS_pd/2))
  alpha_OS_IA_pd <- (1-pnorm(bds_nivo_pd[[8]][1]))*2
  alpha_OS_FA_pd <- (1-pnorm(bds_nivo_pd[[8]][2]))*2
}
if(OS_fraction_pd==1){
  alpha_OS_IA_pd <- alpha_OS_pd
  alpha_OS_FA_pd <- alpha_OS_pd
}
OS_fraction2=(all_result[[1]][,"event_OS_IA"]/all_result[[1]][,"event_OS_FA_s
im"])
for(m in 1:N_sim){
  if(OS_fraction2[m]<1){
    bds_nivo<-bounds(t=c(OS_fraction2[m],1),                      iuse=c(1,1),
alpha=c(alpha_OS/2,alpha_OS/2))
    alpha_OS_IA[m] <- (1-pnorm(bds_nivo[[8]][1]))*2
    alpha_OS_FA[m] <- (1-pnorm(bds_nivo[[8]][2]))*2
  }
  if(OS_fraction2[m]>=1){
    alpha_OS_IA[m] <- alpha_OS
    alpha_OS_FA[m] <- alpha_OS
  }
}

  test_OS_IA = all_result[[1]][,"pvalue_OS_IA"] < alpha_OS_IA
  test_OS_IA_pd = all_result[[1]][,"pvalue_OS_IA_pd"] < alpha_OS_IA_pd
  test_OS_FA = all_result[[1]][,"pvalue_OS_FA"] < alpha_OS_FA
  test_OS_FA_pd = all_result[[1]][,"pvalue_OS_FA_pd"] < alpha_OS_FA_pd
```

```
#### calculate the power
power_OS_IA = sum(test_OS_IA) / length(test_OS_IA)
power_OS_IA_pd = sum(test_OS_IA_pd) / length(test_OS_IA_pd)
power_OS_FA = sum(test_OS_FA) / length(test_OS_FA)
power_OS_FA_pd = sum(test_OS_FA_pd) / length(test_OS_FA_pd)

#### events
event_OS_IA_sim= summary(all_result[[1]][,"event_OS_IA"])
event_OS_IA_pd= summary(all_result[[1]][,"event_OS_IA_pd"])
event_OS_FA_pd_sim= summary(all_result[[1]][,"event_OS_FA_pd"])
event_OS_FA_sim= summary(all_result[[1]][,"event_OS_FA_sim"])
####events by arm
event_OS_IA_ctrl= summary(all_result[[1]][,"event_OS_IA_ctrl"])
event_OS_FA_ctrl= summary(all_result[[1]][,"event_OS_FA_ctrl"])
event_OS_IA_trt= summary(all_result[[1]][,"event_OS_IA_trt"])
event_OS_FA_trt= summary(all_result[[1]][,"event_OS_FA_trt"])
event_OS_IA_ctrl_pd= summary(all_result[[1]][,"event_OS_IA_ctrl_pd"])
event_OS_FA_ctrl_pd= summary(all_result[[1]][,"event_OS_FA_ctrl_pd"])
event_OS_IA_trt_pd= summary(all_result[[1]][,"event_OS_IA_trt_pd"])
event_OS_FA_trt_pd= summary(all_result[[1]][,"event_OS_FA_trt_pd"])

#### N by arm
n_ctrl= summary(all_result[[1]][,"n_ctrl"])
n_trt= summary(all_result[[1]][,"n_trt"])
n_ctrl_pd= summary(all_result[[1]][,"n_ctrl_pd"])
n_trt_pd= summary(all_result[[1]][,"n_trt_pd"])
simprop_trt= summary(all_result[[1]][,"simprop_trt"])
# to check sample size per cohort
p_pergroup1 <- summary((all_result[[1]][,"n_percohort1"])/N_rand)
p_pergroup2 <- summary((all_result[[1]][,"n_percohort2"])/N_rand)
p_pergroup3 <- summary((all_result[[1]][,"n_percohort3"])/N_rand)
p_pergroup4 <- summary((all_result[[1]][,"n_percohort4"])/N_rand)

# medians
kmmedian_IA_ctl = my.summary(all_result[[1]][,"kmmedian_IA_ctl"])
kmmedian_IA_pd_ctl = my.summary(all_result[[1]][,"kmmedian_IA_pd_ctl"])
kmmedian_FA_ctl = my.summary(all_result[[1]][,"kmmedian_FA_ctl"])
kmmedian_FA_pd_ctl = my.summary(all_result[[1]][,"kmmedian_FA_pd_ctl"])
kmmedian_IA_trt = my.summary(all_result[[1]][,"kmmedian_IA_trt"])
kmmedian_IA_pd_trt = my.summary(all_result[[1]][,"kmmedian_IA_pd_trt"])
kmmedian_FA_trt = my.summary(all_result[[1]][,"kmmedian_FA_trt"])
kmmedian_FA_pd_trt = my.summary(all_result[[1]][,"kmmedian_FA_pd_trt"])

#### enrolment duration  and time of analyses
dur_accr = my.summary(all_result[[1]][,"dur_accr"])
month_OS_IA = my.summary(all_result[[1]][,"OS_IA_time"])
month_OS_IA_pd = my.summary(all_result[[1]][,"OS_IA_time_pd"])
month_OS_FA = my.summary(all_result[[1]][,"OS_FA_time"])
month_OS_FA_pd = my.summary(all_result[[1]][,"OS_FA_time_pd"])

####  output HR at each test
hr_OS_IA = all_result[[1]][,"hr_OS_IA"]
hr_OS_IA_pd = all_result[[1]][,"hr_OS_IA_pd"]
hr_OS_FA = all_result[[1]][,"hr_OS_FA"]
hr_OS_FA_pd = all_result[[1]][,"hr_OS_FA_pd"]
```

```
  #### find the largest HR s.t. cox gives significant result
  max_hr_OS_IA = max(hr_OS_IA[test_OS_IA])
  max_hr_OS_IA_pd = max(hr_OS_IA_pd[test_OS_IA_pd])
  max_hr_OS_FA = max(hr_OS_FA[test_OS_FA])
  max_hr_OS_FA_pd = max(hr_OS_FA_pd[test_OS_FA_pd])

  min_obs_mOS_IA                                                        =
median(all_result[[1]][,"kmmedian_IA_ctl"])*(1/max_hr_OS_IA - 1)
  min_obs_mOS_IA_pd                                                     =
median(all_result[[1]][,"kmmedian_IA_pd_ctl"])*(1/max_hr_OS_IA_pd - 1)
  min_obs_mOS_FA                                                        =
median(all_result[[1]][,"kmmedian_FA_ctl"])*(1/max_hr_OS_FA - 1)
  min_obs_mOS_FA_pd                                                     =
median(all_result[[1]][,"kmmedian_FA_pd_ctl"])*(1/max_hr_OS_FA_pd - 1)

  max_hr_result     =     c(max_hr_OS_IA,     max_hr_OS_FA,max_hr_OS_IA_pd,
max_hr_OS_FA_pd)
  names(max_hr_result)  =  c("Interim,  AC",  "Final,  AC",  "Interim,  PDL1+",
"Final, PDL1+")

  min_obs_median_result = c(min_obs_mOS_IA, min_obs_mOS_FA, min_obs_mOS_IA_pd,
min_obs_mOS_FA_pd)
  names(min_obs_median_result)  =  c("Interim,  AC",  "Final,  AC",  "Interim,
PDL1+", "Final, PDL1+")

  sim.par   =   c(N_sim,   seed,     alpha_OS,   alpha_OS_pd,event_OS_FA_pd,
OS_fraction_pd, N_rand)
  names(sim.par)       =       c("N_sim",       "seed",       "alpha_OS",
"alpha_OS_pd","event_OS_FA_pd", "OS_fraction_pd", "Total_Rand")

  sim.par2                                                              =
matrix(c(prop_group1,prop_group2,prop_group3,prop_group4,nprop_trt_1,nprop_tr
t_2, nprop_trt_3, nprop_trt_4,
                      ctl_mos_1, ctl_mos_2,ctl_mos_3,ctl_mos_4,
                      trt_mos_1,trt_mos_2,trt_mos_3,trt_mos_4,
                      delay_OS_1, delay_OS_2,delay_OS_3, delay_OS_4,
                      trt_cure_os_1,trt_cure_os_2,trt_cure_os_3, trt_cure_os_4
  ),ncol=4, byrow=TRUE)
  colnames(sim.par2)=c(label1, label2, label3, label4)
  rownames(sim.par2)=c("% Total Rand by Cohort","% Rand to Experimental Arm",
"Median  in  Control,  months","Median  in  Experimental,  months",  "Delay,
months","% Cure in Experimental")

  result_nrand = rbind(n_ctrl, n_trt, n_ctrl_pd, n_trt_pd, simprop_trt)
  rownames(result_nrand) = c("# in Control","# in Experimental","# in Control,
PDL1+","# in Experimental, PDL1+","%Rand to Experimental Arm")

  result_nrandcoh = rbind(p_pergroup1,p_pergroup2,p_pergroup3,p_pergroup4)
  rownames(result_nrandcoh)   =   c("%Rand   Cohort1","%Rand   Cohort2","%Rand
Cohort3","%Rand Cohort4")

  result_event  =  rbind(event_OS_FA_sim,  event_OS_FA_ctrl,  event_OS_FA_trt,
event_OS_IA_sim,   event_OS_IA_ctrl,   event_OS_IA_trt,   event_OS_FA_pd_sim,
event_OS_FA_ctrl_pd, event_OS_FA_trt_pd, event_OS_IA_pd, event_OS_IA_ctrl_pd,
event_OS_IA_trt_pd)
  rownames(result_event)  =  c("Final,  AC,  both  arms","Final,  AC,  Control",
"Final,   AC,   Experimental","Interim,   Ac,   both   arms","Interim,   AC,
Control","Interim, Ac, Experimental", "Final, PDL1+, both arms", "Final, PDL1+,
Control", "Final, PDL1+, Experimental", "Interim, PDL1+, both arms", "Interim,
PDL1+, Control", "Interim, PDL1+, Experimental")

result_dur = rbind(dur_accr, month_OS_IA, month_OS_FA, month_OS_IA_pd,
month_OS_FA_pd)
rownames(result_dur) = c("Accrual Period", "Analysis IA, AC", "Analysis FA,
AC", "Analysis IA, PDL1+", "Analysis FA, PDL1+")

result_hr = rbind(
summary(hr_OS_IA),summary(hr_OS_FA),summary(hr_OS_IA_pd),summary(hr_OS_FA_pd)
)
rownames(result_hr) = c("Interim All Randomized","Final All
Randomized","Interim PDL1+","Final PDL1+")

result_kmmedian = rbind(kmmedian_IA_ctl, kmmedian_IA_pd_ctl, kmmedian_FA_ctl,
kmmedian_FA_pd_ctl, kmmedian_IA_trt, kmmedian_IA_pd_trt, kmmedian_FA_trt,
kmmedian_FA_pd_trt)
rownames(result_kmmedian) = rbind("Control All Randomized, IA", "Control PDL1+,
IA", "Control All Randomized, FA", "Control PDL1+, FA", "Experimental All Rand,
IA", "Experimental PDL1+, IA", "Experimental All Rand, FA", "Experimental PDL1+,
FA")

  result_alpha = rbind(summary(alpha_OS_IA), summary(alpha_OS_FA),
summary(alpha_OS_IA_pd), summary(alpha_OS_FA_pd))
  rownames(result_alpha) = c("alpha IA - AC", "alpha FA - AC", "alpha IA-PDL1+",
"alpha FA-PDL1+")

  powerOS_DFS <- matrix(c(power_OS_IA_pd, power_OS_FA_pd, power_OS_IA,
power_OS_FA),
                        byrow=TRUE, ncol=2, nrow=2)
  rownames(powerOS_DFS) <- c("PDL1+","AC")
  colnames(powerOS_DFS) <- c("Interim", "Final")

rownames(all_result)=NULL
  result_list = list(sim.par, sim.par2, result_nrand, result_nrandcoh,
powerOS_DFS, result_hr, max_hr_result, min_obs_median_result, result_alpha,
result_event, result_kmmedian, result_dur)
  names(result_list) = c("Simulation Parameters in function
Report_results","Simulation Parameters by Cohort", "Checking Simulation: # of
Randomized Subjects", "Checking Simulation: % Randomized by Cohort","SIMULATED
POWER (Reported in Protocol Table 8.1-1,8.1-2)", "Simulated HR (Protocol Table
8.1-1,8.1-2: 'Hypothesized Overall HR')","Maximum Significant HR (Protocol
Table 8.1-1,8.1-2: 'Critical HR')", "Minimum Significant Difference in Medians,
months (Protocol Table 8.1-1,8.1-2: 'Minimal Difference in Median')", "Alpha
Levels Used in the Simulated Datasets (Protocol Table 8.1-1,8.1-2:
'Significance Level' for IA and FA)", "# of Simulated Events (Protocol Table
8.1-1,8.1-2: 'Projected # of Events')", "Simulated KM Medians, months (Protocol
Table 8.1-1,8.1-2: 'Hypothesized Median...')", "Enrolment Period and Projected
Time of Analyses, months (Protocol Table 8.1-1: 'Projected Time of LPLV (from
FPFV)')")

  return(result_list)
}
```

```
##############################################################################
#######
######### II. Set up calls for different endpoints
##############################################################################
#######

##############################################################################
#######
#############################ACCRUAL RATE
#note that sum(accr_total) > N_rand; accr_total is to be truncated as needed
(see Run_Iterations)
#TOTAL accrual rate for the 3 arms:
accr_total = c(3,  3,  6,     6,    9,    9,    12,    12,    18,    27,    36,
rep(45,100))
#accrual rate for 2 arms:
accr_total2 = round(accr_total/3*2)

#program assumes that we stratify within these groups
label1="Western PDL1+"
label2="Western PDL1-"
label3="Asia PDL1+"
label4="Asia PDL1-"
prop_group1=0.2*0.5
prop_group2=0.2*0.5
prop_group3=0.8*0.5
prop_group4=0.8*0.5
#randomization ratio
nprop_trt_1=nprop_trt_2=nprop_trt_3=nprop_trt_4=0.5

##############################################################################
#######
################OS#######################################
# OS median for control
ctl_mos_1=6    #West PDL1+
ctl_mos_2=6       #West PDL1-
ctl_mos_3=10    #Asia PDL1+
ctl_mos_4=10       #Asia PDL1-
# OS median for experimental, after delay (without cure)
trt_mos_1=ctl_mos_1/0.65    #West PDL1+
trt_mos_2=ctl_mos_2/0.85
trt_mos_3=ctl_mos_3/0.65    #Asia PDL1+
trt_mos_4=ctl_mos_4/0.85
# OS delay
delay_OS_1=delay_OS_3=3   #PDL1+
delay_OS_2=delay_OS_4=4   #PDL1-
# OS cure in experimental arm
trt_cure_os_1=trt_cure_os_3=0.15   #PDL1+
trt_cure_os_2=trt_cure_os_4=0.1   #PDL1-

####################

OS_RESULTS <- Report_results(N_sim=10000, seed=13484,
                              accr_pattern=accr_total2,N_rand=626,
                              alpha_OS=0.01,
                              event_OS_FA_pd=250,OS_fraction_pd=0.7,
alpha_OS_pd=0.01)
OS_RESULTS
```

```
##############################################################################
#######
################PFS#######################################
# PFS median for control
ctl_mos_1=4    #West PDL1+
ctl_mos_2=4       #West PDL1-
ctl_mos_3=4    #Asia PDL1+
ctl_mos_4=4       #Asia PDL1-
# PFS median for experimental, after delay
trt_mos_1=ctl_mos_1/0.55    #West PDL1+
trt_mos_2=ctl_mos_2/0.75
trt_mos_3=ctl_mos_3/0.55    #Asia PDL1+
trt_mos_4=ctl_mos_4/0.75
# PFS delay
delay_OS_1=delay_OS_3=1   #PDL1+
delay_OS_2=delay_OS_4=2   #PDL1-
# PFS cure in experimental arm
trt_cure_os_1=trt_cure_os_3=0   #PDL1+
trt_cure_os_2=trt_cure_os_4=0   #PDL1-

####################

PFS_RESULTS <- Report_results(N_sim=10000, seed=13484,
                              accr_pattern=accr_total2,N_rand=626,
                              alpha_OS=0.015,
                              event_OS_FA_pd=250,OS_fraction_pd=1,
alpha_OS_pd=0.015)
PFS_RESULTS
```

APPENDIX 7 CHANGE OF THE TRIGGER FOR PFS FA/OS IA

Per the original protocol, the planned interim analysis (PFS final analysis and OS interim analysis) was to be triggered when 136 PFS events per blinded independent central review committee (BICR) are observed among the PD-L1 expressing subjects in the chemotherapy arm (Arm C). The primary PFS endpoint is based on the primary PFS definition (i.e. PFS events truncated at subsequent therapy as defined in Section 8.3.1). PFS event tracking is conducted by an independent external statistical group (Axio, Inc) which supports statistical analyses and generates reports for review by an independent Data Monitoring Committee. Event tracking commenced in July 2020 and per tracking plan, BMS has been receiving information on the number of PFS events per primary and secondary (PFS events not truncated by subsequent therapy, sensitivity analysis in Section 7.5.2.2) definitions. BMS remains blinded to the number of PFS events in the experimental arms.

Based on the latest event tracking report (report is stored in eTMF Section 01.03.03 Committee report) utilizing the September 16 2020 BICR data transfer (data is stored in unix folder “/gbs/prod/clin/data/ca/209/648/stable/blinded/level2”), a total of 102 PFS events per BICR were observed by primary definition among the PD-L1 expressing subjects in Arm C. The PFS events per secondary definition was 135 among the PD-L1 expressing subjects in Arm C and upon further investigation, it could be determined that the target of 136 PFS events per BICR for triggering the interim analysis (IA) per protocol is unlikely to be reached for the reasons specified below, acknowledging that data review is ongoing.

- Under the 1:1:1 randomization ratio and based on the assumption of 50% (from Protocol Section 3.1) and 48.3% (% of patients in the control arm in Table 11.2-1 in the ONO-4538-24 (CA209473) Primary Clinical Study Report (CSR)[19]) for prevalence of Tumor PD-L1 expression ≥ 1% in advanced ESCC, it is estimated that approximately 158 PD-L1 expressing subjects would be randomized to the chemotherapy arm.

- Assuming 158 PD-L1 expressing subjects randomized to the chemotherapy arm and considering the number of 102 PFS events observed so far, there would be an additional 158-102=56 subjects for whom a PFS event may still be expected per primary definition. However, per secondary definition which accounts for disease progression after subsequent therapy, a total of 135 PFS events per BICR were observed. Therefore, it is evident that for 135-102=33 subjects, PFS events per primary definition should not be expected.
- Event tracking in all randomized subjects across the three treatment arms indicated there are a total of 94 subjects are censored per primary definition due to withdrawal of consent, lost to follow-up, no scans or without PD/death after subsequent therapies. Assuming equal subject distribution across treatment arms and PD-L1 expression subgroups, it is estimated that 15 of these subjects would represent PD-L1 expressing subjects in the chemotherapy arm, thus reducing the potential pool of subjects for whom a PFS event could be expected even further.
- Therefore, among the 56 subjects without observed PFS events (per primary definition), only 56-33-15=8 subjects may contribute a PFS event per primary definition, whereas an additional 34 PFS events are needed to trigger the IA.

Based on the reasons stated above, the IA will be performed when 136 PFS events are observed or when a 12-month minimum follow-up (defined as the time from the date the last patient was randomized to the clinical cutoff date) is reached if the target number of events is unlikely to be reached among the PD-L1 expressing subjects in the chemotherapy arm**.** The table below provides the power based on different numbers of PFS events among PD-L1+ subjects in chemo arm using the same protocol assumption.

PFS events among PD-L1+ in chemo	Total number of PFS events for one comparison (i.e. for N+I vs CT; and for N+C vs CT)	Power
136	250	90%
121	218	85%
110	195	80%
100	175	75%

13 REFERENCES

[1] Dolan P. Modeling valuations for EuroQol health states. Medical Care 1997;35: 1095-1108

[2] The EuroQol Group: EuroQol: A new facility for the measurement of health-related quality of life—The EuroQol Group. Health Policy 16:199-208, 1990.

[3] Pickard AS, Neary MP, Cella D. Estimation of minimally important differences in EQ-5D utility and VAS scores in cancer. Health Qual Life Outcomes 2007;5:70.

[4] Schoenfeld, David A., and Anastasios A. Tsiatis. "A modified log rank test for highly stratified data." Biometrika 74.1 (1987): 167-175.

[5] Kernan, Walter N., et al. "Stratified randomization for clinical trials." Journal of clinical epidemiology 52.1 (1999): 19-26.

[6] De Stavola, B. L., and D. R. Cox. "On the consequences of overstratification." Biometrika 95.4 (2008): 992-996.

[7] Feng, Changyong, Hongyue Wang, and Xin M. Tu. "Power loss of stratified log-rank test in homogeneous samples." Journal of Quality and Reliability Engineering 2010 (2010).

[8] Brookmeyer R. and Crowley J. A confidence interval for the median survival time. Biometrics 38:29-41, 1982

[9] Klein, J. P. and Moeschberger, M. L. (1997), Survival Analysis: Techniques for Censored and Truncated Data, New York: Springer-Verlag.

[10] Greenwood, M. The errors of sampling of the survivorship tables, Reports on Public Health and Statistical Subjects, 33, Appendix 1, HMSO, London, 1926

[11] Kalbfleisch, J. D. and Prentice, R. L. (1980), The Statistical Analysis of Failure Time Data, New York: John Wiley & Sons.

[12] Statistical methodology in the pharmaceutical sciences / edited by Berry DA, Chapter 13 Categorical Data analysis p. 415 and 417 ff., Marcel Dekker, 1990

[13] Willi MAURER and Frank BRETZ: Multiple Testing in Group Sequential Trials Using Graphical Approaches. American Statistical Association Statistics in Biopharmaceutical Research November 2013, Vol.5, No.4

[14] Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. Statistics in Medicine 2009; 28: 586-604.

[15] Clopper, CJ and Pearson, ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404-413, 1934

[16] Clopper, CJ and Pearson, ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26: 404-413, 1934

[17] Adverse Event Domain Requirements Specification. Bristol-Myers Squibb Co. PRI. Version 2.4.0. July 25, 2016.

[18] Non-Study Medication Domain Requirements Specification. Bristol-Myers Squibb, Co. PRI. Version 2.9 July 25, 2016

[19] Final ONO-4538-24 (CA209473) Clinical Study Report: Phase III Study, a Multicenter, Open-Label, Randomized Study in Patients with Esophageal Cancer. Ono Pharmaceuticals, 2019. Document Control No. 930141648.