AstraZeneca

Statistical Analysis Plan

Study Code	D419MC00004
Edition Number	5
Date	25 February 2021

A Phase III, Randomized, Multi-Center, Open-Label, Comparative Global Study to Determine the Efficacy of Durvalumab or Durvalumab and Tremelimumab in Combination With Platinum-Based Chemotherapy for First-Line Treatment in Subjects With Metastatic Non-Small-Cell Lung Cancer (NSCLC) (POSEIDON)

TABLE OF CONTENTS PAGE

LIST OF ABBREVIATIONS

Abbreviation or special term	Explanation
ADA	Anti-drug antibodies
AE	Adverse event
AESI	Adverse event of special interest
AJCC	American Joint Committee on Cancer
ALK	Anaplastic lymphoma kinase
ALP	Alkaline phosphatase
ALT	Alanine aminotransferase
APF12	Proportion of subjects alive and progression free at 12 months from randomization
AST	Aspartate aminotransferase
BoR	Best objective response
BICR	Blinded Independent Central Review
bTMB	Blood tumor mutational burden
CFDA	China Food and Drug Authority
CI	Confidence interval
CR	Complete response
CRO	Contract Research Organization
CSP	Clinical study protocol
CSR	Clinical study report
CT	Computed tomography
CTC	Common toxicity criteria
CTCAE	Common Terminology Criteria for Adverse Events
COVID-19	Corona Virus Disease 2019
DBP	Diastolic blood pressure
DCO	Data cut-off
DoR	Duration of response
eCRF	Electronic case report form
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
EGFR	Epidermal growth factor receptor

Abbreviation or special term	Explanation
EORTC	European Organization for Research and Treatment of Cancer
CCI	[redacted]
FAS	Full analysis set
FH	Fleming-Harrington
GGT	Gamma-glutamyl transferase
HR	Hazard ratio
HRQoL	Health-related quality of life
IDMC	Independent Data Monitoring Committee
IHC	Immunohistochemistry
ILD	Interstitial lung disease
imAE	Immune-mediated adverse event
IO	Immuno-oncology
IPD	Important protocol deviation
CCI	[redacted]
ITT	Intent-to-treat
IVRS	Interactive Voice Response System
LD	Longest diameter
MedDRA	Medical Dictionary for Regulatory Activities
MRI	Magnetic resonance imaging
MTP	Multiple testing procedure
NA	Not applicable
NCI	National Cancer Institute
NE	Not evaluable
NED	No evidence of disease
NPH	Non-proportional hazards
NSCLC	Non-small-cell lung cancer
NTL	Non-target lesion
OAE	Other significant adverse event
ORR	Objective response rate
OS	Overall survival
PD	Progressive disease

Abbreviation or special term	Explanation
PD-L1	Programmed cell death ligand 1
PD-L1 TC <1%	PD-L1 expression on less than 1% of tumor cells
PD-L1 TC ≥25%	PD-L1 expression on at least 25% of tumor cells
PD-L1 TC <25%	PD-L1 expression on less than 25% of tumor cells
PD-L1 TC ≥50%	PD-L1 expression on at least 50% of tumor cells
PD-L1 TC <50%	PD-L1 expression on less than 50% of tumor cells
PFS	Progression-free survival
PFS2	Time from randomization to second progression
PH	Proportional hazards
CCI	
PK	Pharmacokinetic
PMDA	Pharmaceutical and Medical Devices Agency
PR	Partial response
PRO	Patient-reported outcome
CCI	
PT	Preferred term
q6w	Every 6 weeks
q8w	Every 8 weeks
QC	Quality control
QLQ-C30 v3	30-item Core Quality of Life Questionnaire, version 3
QLQ-LC13	13-item Lung Cancer Quality of Life Questionnaire
RDI	Relative dose intensity
RECIST 1.1	Response Evaluation Criteria in Solid Tumors, Version 1.1
RMST	Restricted mean survival time
SAE	Serious adverse event
SAP	Statistical analysis plan
SBP	Systolic blood pressure
SD	Stable disease
SI	International system of units
SoC	Standard of Care

Abbreviation or special term	Explanation
SOC	System organ class
T_3	Triiodothyronine
T_4	Thyroxine
TMB	Tumor mutational burden
tTMB	Tissue tumor mutational burden
TEAE	Treatment emergent adverse event
TSH	Thyroid-stimulating hormone
TL	Target lesion
ULN	Upper limit of normal
VAS	Visual analogue scale
WHO	World Health Organization

AMENDMENT HISTORY

Date	Brief description of change
27SEP2017	N/A – First version
14FEB2019	In line with Clinical Study Protocol (CSP) amendments: • Updated objectives in Section 1.1 per CSP v4.0. • Clarified the definition of subjects in China in Section 1.1 and Section 3.5, and the planned number of subjects in China in Section 3.5. • Replaced Figure 1 (overall study design) and Figure 2 (study flow chart) with revised figures in the protocol amendments. The schedule for pemetrexed was changed to either every 3 weeks (q3w) or every 4 weeks (q4w) for Treatment Arm 3. • Sample size was increased from 801 to 1000 and China enrolment was slightly updated in Section 1.3. • Added new analysis sets (PD-L1 TC <25% analysis set, blood-based TMB (bTMB) high analysis set (>=20 mut/Mb, >=16 mut/Mb and >=12 mut/Mb),) and specified their use for outcome variables, and updated PRO data to use FAS only in Section 2.1 and Table 1. • Updated section structure in Section 3.2 to reflect PFS and OS are dual primary variables. • Updated statistical hypotheses and event numbers for PFS and OS analyses (dual primary endpoints), and analysis sets for efficacy and PRO data analysis in Section 4.1. • Updated planned analyses in Table 1312 and wording/analyses in Section 4.2.2 to reflect dual primary and key secondary PFS and OS endpoints and repeat analyses using new analysis sets (PD-L1 TC <25% analysis set, bTMB high analysis set (>=20, >=16 and >=12)). • Updated multiple testing strategy in Section 4.2.1 and interim analysis in Section 5.1 (i.e. event numbers, procedures) to reflect the changes due to dual primary PFS and OS endpoints.. • Updated the defined subgroups for subgroup analyses in Section 4.2.2.

Date	Brief description of change
	• Added bTMB and CCI [redacted] in Section 4.2.6. • Updated Section 5.2 to remove 21 days of follow-up in the condition for initial safety review and add description of safety review for Chinese subjects. Updated Section 6 to remove the points incorporated into protocol amendments since the previous SAP version. Other changes: • Updated immunogenicity analysis set to ADA-evaluable set for consistency with other durvalumab studies. • Updated the definition of 2 missed visits (RECIST assessments) in the PFS derivation in Section 3.2.1.1. • Removed the rule of 2 missed visits in the PFS2 derivation in Section 3.2.2.3 for consistency with other durvalumab studies. • Clarified that for PFS2 derivation only the site investigator reported PFS is considered for initial PFS (not BICR) (Section 3.2.2.3) • Updated slightly QLQ-C30 scale/item abbreviation in Table 6. • Updated individual item numbers used in QLQ-LC13 scale derivation in Table 8. • Updated which items time to deterioration should be derived for in Sections 3.3.1.1 and 3.3.2.1.Updated treatment exposure calculation and defined exposure separately for combination stage and maintenance stage in Section 3.4.5, with treatment exposure summaries updated accordingly in Section 4.2.8. Added additional supportive summaries, updated interaction tests using Gail and Simon approach, and moved exploratory analyses to end of the section in Section 4.2.2.1 (PFS). • Added summary of OS at 12 months, assessment of proportionality assumption, additional supportive summaries, subgroup analyses, effect of covariate analyses and interaction tests using Gail and Simon approach in Section 4.2.2.2 (OS).

Date	Brief description of change
	• Changed word 'stratified' to 'adjusting' for stratification factors in ORR logistic regression models in Section 4.2.2.3 as they will be in the model statement. • Removed analysis of expected duration of response (EDoR) for DoR in Section 4.2.2.4. • Specified outputs of tumor size produced for FAS based on BICR data in Section 4.2.2.7. • Updated denominator of event rate in Section 4.2.4.1 to be over the same time period as the numerator. • Updated summaries of AEs, death and AESIs in Section 4.2.4.1 for consistency with other durvalumab studies. • Updated summaries of thyroid test variables and removed the line plot of potiential Hy's law in Section 4.2.4.2. • In the last paragraph of Section 4.2.4.2, updated selection criteria for plot of liver biochemistry test results over time and individual subject data listing to subjects with potential Hy's law. • Added descriptive summaries of vital signs over time in Section 4.2.4.4.Slightly re-structured sections (Sections 4.2.4 – 4.2.8) so PK, ADA, CCI demographic and other baseline characteristics, and treatment exposure are as parallel sections with safety instead of sub-sections under safety. • PK analysis were updated (Sections 3.4.3 and 4.2.5) • Removed histology type at screening, time from diagnosis to randomization from presentation of disease characteristics at screening, time to subsequent therapy from last dose and plot of post-discontinuation anti-cancer therapy, and updated TNM classification summary to be at diagnosis in Section 4.2.7. • Updated Table 2 with additional two IDPs • Section 4.2.2.8, clarified pain item to be used to control for overall Type I error • Updated section 3.1 to clarify visit response based on investigator's review process

Date	Brief description of change
	• Section 3.2.1 added a note that NE visit is not considered as missing visits for derivation of PFS • Added an imputation technique for missing death date (Section 4.1.3) • Updated the definition of TEAE (section 3.4.1) • Updated the definition of baseline for efficacy and PRO assessments (Section 4.1) • Efron approach replaced Breslow approach to handle ties (Section 4.2.2) • Clarified that CRF data will be used • Clarified that AEs and AEs casually related to treatment consider any CTCAE grade 3 or 4 rather than maximum grade 3 or 4 • Added summary for all deaths occurring on treatment and up to 90 days after last dose of study treatment • In section 3.4.4 and 4.2.8 specified that RDI will be derived and summarized only for immunotherapies
11 March 2019	• Included details of sample size calculations for dual primary endpoints from the protocol and added information on key secondary endpoints, and secondary endpoint of OS between Arm 1 and Arm 3 in bTMB20 high population in Section 1.3 • Simplified the language about PD after missed visits for BoR in Section 3.2.2.5 to be consistent with other durvalumab studies • Added further details about the timing of the analyses in Section 4.2.1 • Updated analysis populations used for BoR in Section 4.2.2.3 to be consistent with Table 1 • Added details on interim analysis in bTMB high population

Date	Brief description of change
	• Section 2.1.10 –immunogenicity analysis set changed to ADA evaluable set for consistency
	• Format of Figure 3 added bTMB to the Figure to make clear it is blood TMB and added footnotes to explain.
	• Section 4.2.6 updated to remove wording on bTMB equivalent cut off
17 Dec 2020	• Sections 2.2, 4.2.2.2, 4.2.4.1 and 4.2.9 updated to add details of additional summaries to assess impact of COVID-19.
	• Section 3.2.1.2 updated to allow OS derivations to use a range of CRF dates, in addition to the date recorded in the SURVIVE module, to obtain the last known date alive for OS censoring in all OS analyses (including the final analysis).
	• Section 4.2.1 clarified by adding 'e.g. 5%' to account for multiple scenarios where the alpha level may differ from 5%, depending on the route taken in the upper levels of the MTP.
25 Feb 2021	• Abbreviations added for FH, IO, PH, NPH, RMST
	• Section 4.2.2.2 Overall survival updated to include overall survival estimates, based on Kaplan-Meier method, at 12, 18, 24 and 36 months to account for longer follow up
	• Section 4.2.2.2 Overall survival updated to include the use of the Grambsch-Themeau non-proportionality test to assess the proportional hazards assumption, and the use of a stratified max-combo test as a sensitivity analysis. The Restricted Mean Survival Time (RMST) of an area-under-the-curve approach (Kaplan-Meier method), pseudovalues approach and/or Royston-Parmar model could be used to determine the difference in means between treatment groups if a lack of proportionality is observed.
	• Section 4.2.2.6 Time from randomization to second progression updated to include the PFS2 rate estimates, based on Kaplan-Meier method, at 12, 18, 24 and 36 months to account for longer follow up

Date	Brief description of change
	• Section 7 References added for updated text in section 4.2.2.2, regarding Grambsch-Themeau non-proportionality test, stratified max-combo test, Restricted Mean Survival Time (RMST) of an area-under-the-curve approach, pseudovalues approach and Royston-Parmar model.

1. STUDY DETAILS

1.1 Study objectives

1.1.1 Primary objectives

Objectives	Outcome measure
• To assess the efficacy of durvalumab monotherapy + SoC chemotherapy compared with SoC chemotherapy alone in terms of PFS and OS in all subjects	• PFS in all subjects using BICR assessments according to RECIST 1.1 • OS in all subjects

Note: Sensitivity analyses of PFS will be performed based on the investigator's assessment according to RECIST 1.1.

BICR Blinded Independent Central Review; PFS Progression-free survival; RECIST 1.1 Response Evaluation Criteria in Solid Tumors, Version 1.1; SoC Standard of care.

1.1.2 Secondary objectives

Objectives	Outcome measures
• To assess the efficacy of durvalumab + tremelimumab combination therapy + SoC chemotherapy compared with SoC chemotherapy alone in terms of PFS and OS	• PFS in all subjects using BICR assessments according to RECIST 1.1 (key secondary objective) • OS in all subjects (key secondary objective)
• To further assess the efficacy of durvalumab + tremelimumab combination therapy + SoC chemotherapy compared with SoC chemotherapy alone in terms of PFS, OS, ORR, BoR, DoR, APF12 and PFS2	• PFS in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25% and subjects with PD-L1 TC <1% using BICR assessments according to RECIST 1.1 • OS in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25% and subjects with PD-L1 TC <1% • ORR, DoR, BoR and APF12 in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25%, subjects with PD-L1 TC <1% and

Objectives	Outcome measures
	all subjects using BICR assessments according to RECIST 1.1
	• PFS2 in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25%, subjects with PD-L1 TC <1% and all subjects using local standard clinical practice
• To further assess the efficacy of durvalumab monotherapy + SoC chemotherapy compared with SoC chemotherapy alone in terms of PFS, OS, ORR, DoR, BoR, APF12 and PFS2	• PFS in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25% and subjects with PD-L1 TC <1% using BICR assessments according to RECIST 1.1
	• OS in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25% and subjects with PD-L1 TC <1%
	• ORR, DoR, BoR, and APF12 in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25% and subjects with PD-L1 TC <1% and all subjects using BICR assessments according to RECIST 1.1
	• PFS2 in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25%, subjects with PD-L1 TC <1% and all subjects using local standard clinical practice
• To assess the efficacy of durvalumab + tremelimumab combination therapy + SoC chemotherapy compared with durvalumab monotherapy + SoC chemotherapy in terms of PFS, OS and ORR	• PFS and ORR in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25%, subjects with PD-L1 TC <1% and all subjects using BICR assessments according to RECIST 1.1

Objectives	Outcome measures
	• OS in subjects with PD-L1 TC <50%, subjects with PD-L1 TC <25%, subjects with PD-L1 TC <1% and all subjects
• To assess the association of tumor mutation burden (TMB) with the efficacy of durvalumab + tremelimumab combination therapy + SoC chemotherapy compared with SoC chemotherapy alone in terms of PFS, OS, ORR, BoR, DoR, APF12 and PFS2	• PFS, ORR, BoR, DoR, APF12in subjects with TMB high using BICR assessments according to RECIST 1.1 • PFS2 in subjects with TMB high using local standard clinical practice • OS in subjects with TMB high
• To assess the association of TMB with the efficacy of durvalumab + tremelimumab combination therapy + SoC chemotherapy compared with durvalumab monotherapy + SoC chemotherapy in terms of PFS, OS, ORR, BoR, DoR, APF12 and PFS2	• PFS, ORR, BoR, DoR, APF12 in subjects with TMB high using BICR assessments according to RECIST 1.1 • PFS2 in subjects with TMB high using local standard clinical practice • OS in subjects with TMB high
• To assess the association of TMB with the efficacy of durvalumab monotherapy + SoC chemotherapy compared with SoC chemotherapy in terms of PFS, OS, ORR, BoR, DoR, APF12 and PFS2	• PFS, ORR, BOR, DoR, APF12 in subjects with TMB high using BICR assessments according to RECIST 1.1 • PFS2 in subjects with TMB high using local standard clinical practice • OS in subjects with TMB high
• To assess the PK of durvalumab + tremelimumab combination therapy + SoC chemotherapy and	• Concentration of durvalumab and tremelimumab

Objectives	Outcome measures
durvalumab monotherapy + SoC chemotherapy	
• To investigate the immunogenicity of durvalumab and tremelimumab	• Presence of ADAs for durvalumab and tremelimumab
• To assess disease-related symptoms and HRQoL in subjects treated with durvalumab + tremelimumab combination therapy + SoC chemotherapy and durvalumab monotherapy + SoC chemotherapy compared with SoC chemotherapy alone using the EORTC QLQ-C30 v3, the QLQ-LC13 module, and WHO/ECOG performance status assessments	• EORTC QLQ-C30 • EORTC QLQ-LC13 • Changes in WHO/ECOG performance status will also be assessed.

ADA Anti-drug antibody; APF12 Proportion of subjects alive and progression free at 12 months from randomization; BICR Blinded Independent Central Review; BoR Best objective response; DoR Duration of response; ECOG Eastern Cooperative Oncology Group; EORTC European Organization for Research and Treatment of Cancer; HRQoL Health-related quality of life; ORR Objective response rate; OS Overall survival; PD-L1 Programmed cell death ligand 1; PFS Progression-free survival; PFS2 Time from randomization to second progression; PK Pharmacokinetic(s); QLQ-C30 v3 30-item Core Quality of Life Questionnaire, version 3; QLQ-LC13 13-item Lung Cancer Quality of Life Questionnaire; RECIST 1.1 Response Evaluation Criteria in Solid Tumors, Version 1.1; SoC Standard of care; TC Tumor cell; TMB Tumor mutational burden; WHO World Health Organization.

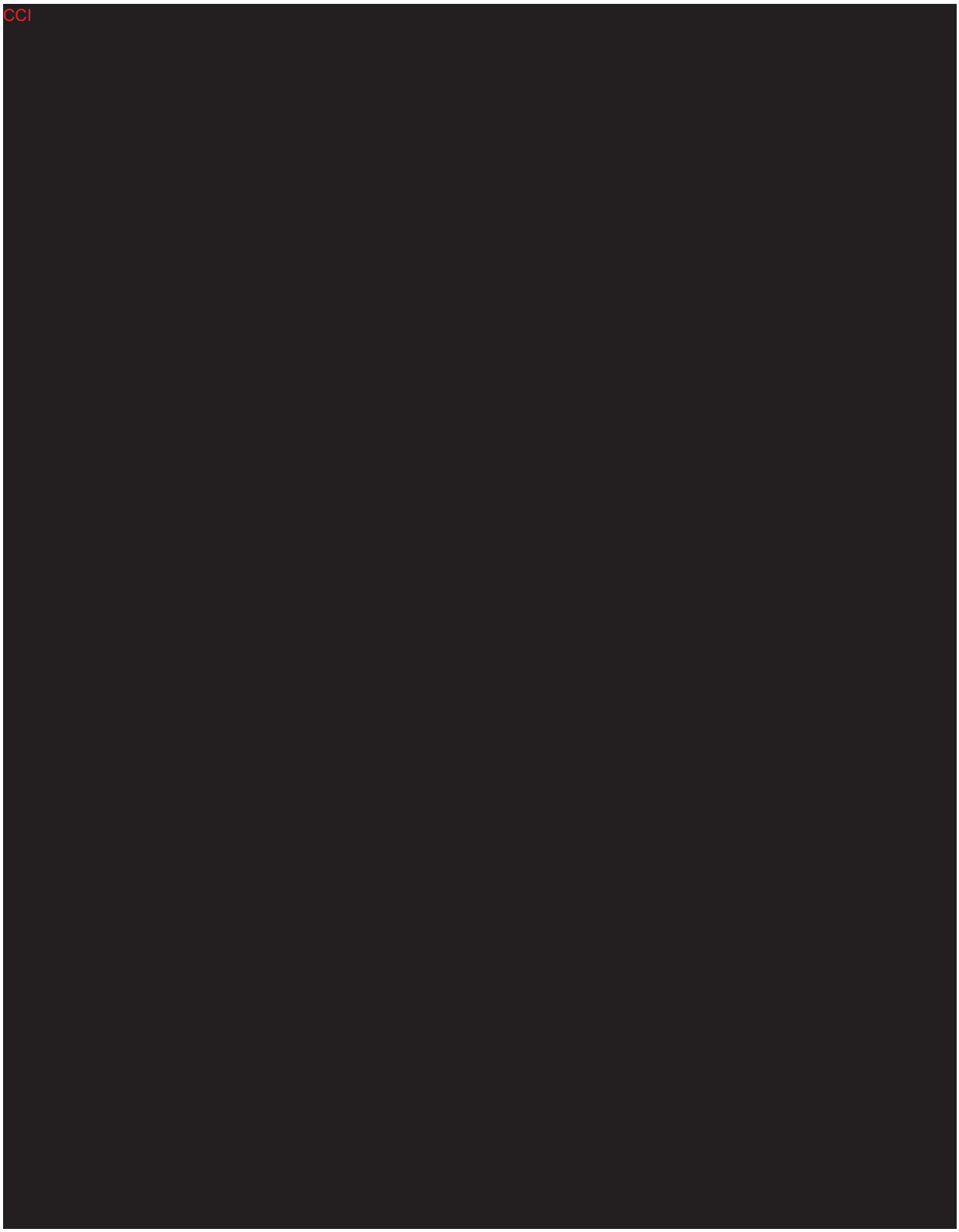
1.1.3 Safety objectives

Objectives	Outcome measures
• To assess the safety and tolerability profile of durvalumab + tremelimumab combination therapy + SoC chemotherapy and durvalumab monotherapy + SoC chemotherapy compared with SoC chemotherapy alone	• AEs, physical examinations, laboratory findings, and vital signs

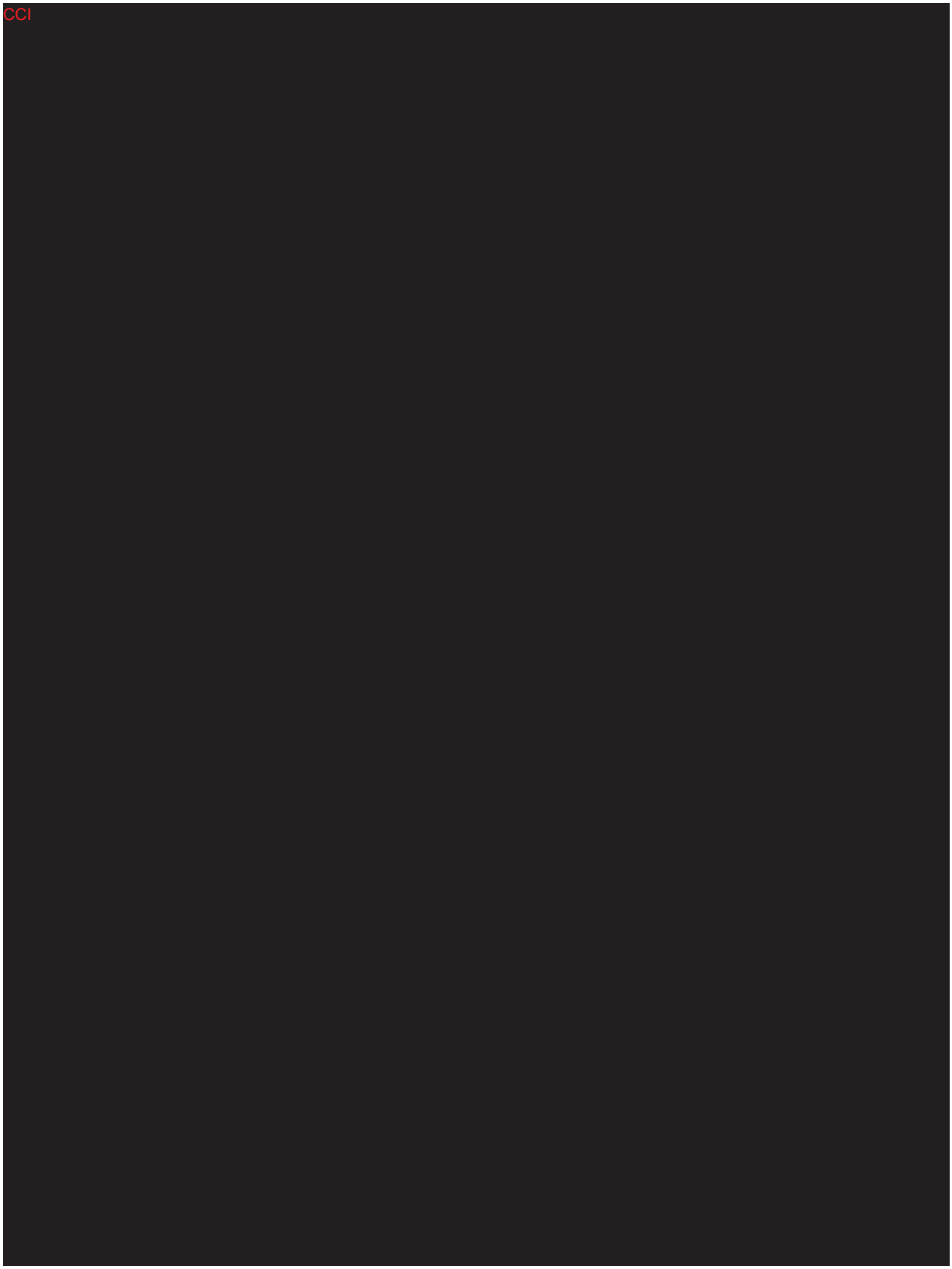
AE Adverse event; SoC Standard of care.

1.1.4 Exploratory objectives

CCI



CCI



A further objective to meet China Food and Drug Authority (CFDA) and Pharmaceutical and Medical Devices Agency (PMDA) requirements is to evaluate consistency in efficacy and safety among subjects in China and in Japan for benefit-risk assessments of durvalumab + tremelimumab combination therapy + Standard of Care (SoC) chemotherapy compared with SoC chemotherapy alone and durvalumab monotherapy + SoC chemotherapy compared with SoC chemotherapy alone.

Subjects in China and in Japan will be defined as subjects who are enrolled in Chinese sites and in Japanese sites, respectively.

1.2 Study design

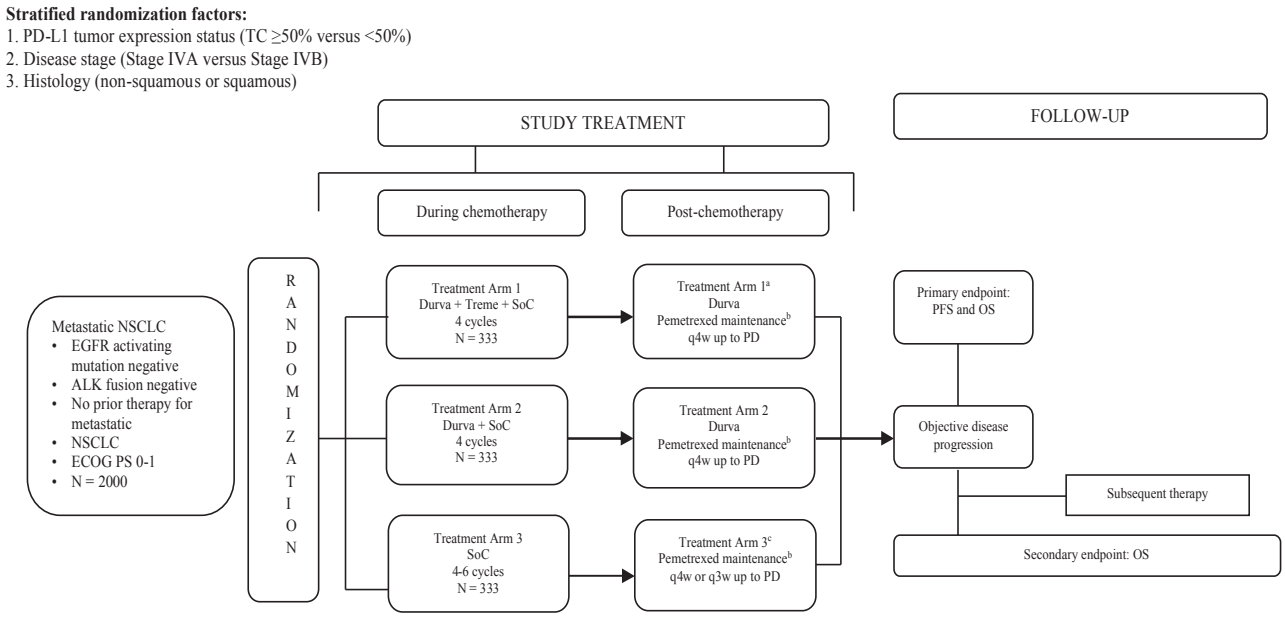
This is a randomized, open-label, multi-center, global, Phase III study to determine the efficacy and safety of durvalumab + tremelimumab combination therapy + SoC chemotherapy or durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone as first-line treatment in subjects with metastatic non-small-cell lung cancer (NSCLC) with

tumors that lack activating epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) fusions. SoC chemotherapy will be one of the following regimens: abraxane + carboplatin, pemetrexed + cisplatin or carboplatin, or gemcitabine + cisplatin or carboplatin.

Tumor evaluation scans will be performed at screening (as baseline) with follow-ups at Week 6±1 week from the date of randomization, at Week 12±1 week from the date of randomization, and then every 8 weeks (q8w) ±1 week until radiological progression. The management of subjects will be based solely upon the results of the tumor evaluation scans conducted by the investigator. The Blinded Independent Central Review (BICR) of all radiologic scans will be performed to derive the progression-free survival (PFS), objective response rate (ORR), duration of response (DoR), best objective response (BoR), and proportion of subjects alive and progression free at 12 months from randomization (APF12) endpoints according to Response Evaluation Criteria in Solid Tumors, Version 1.1 (RECIST 1.1).

A schematic diagram of the overall study design is presented in Figure 1, and a detailed study flow chart is presented in Figure 2.

Figure 1 Overall study design



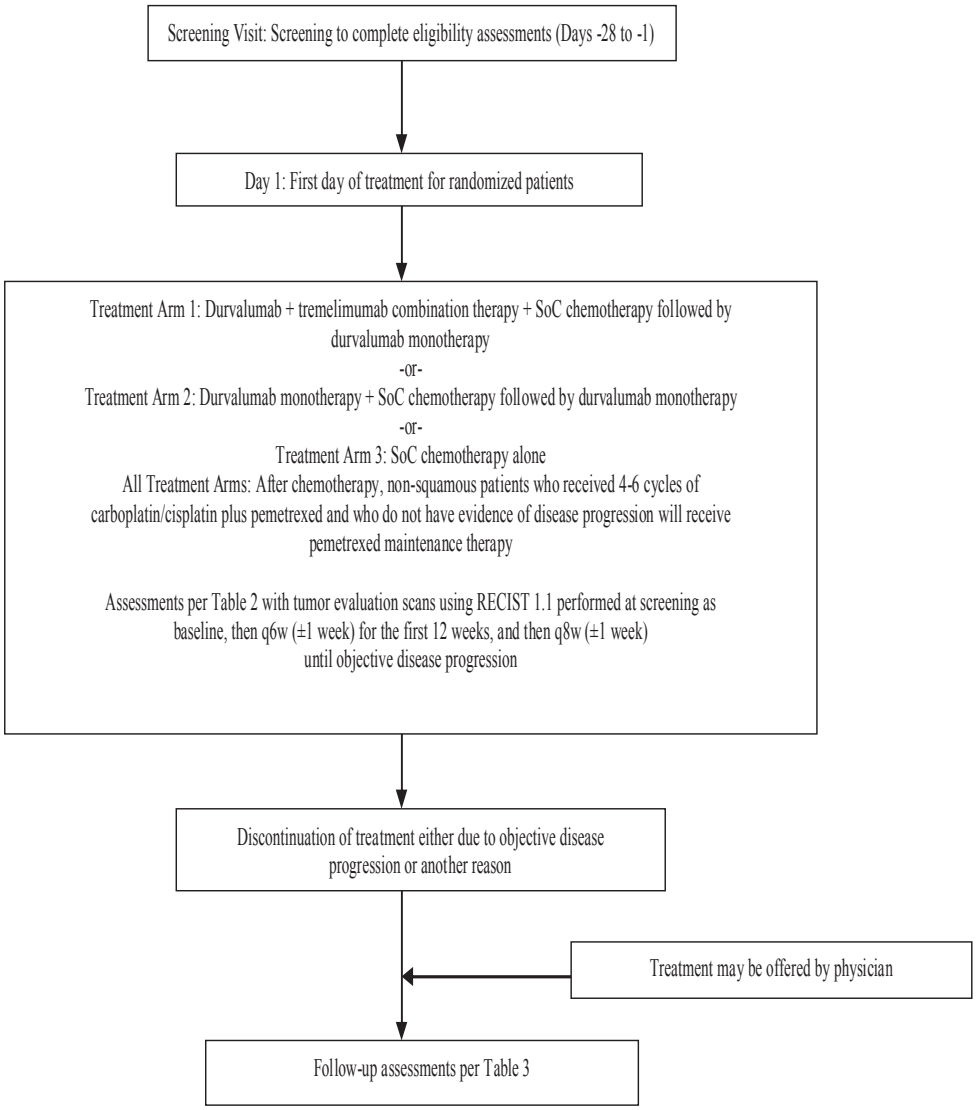
[a] One additional durvalumab + tremelimumab combination therapy dose will be given at Week 16 post-chemotherapy. In the case of dose delay(s), more than 1 durvalumab + tremelimumab combination dose can be given at and after Week 16 post-chemotherapy to ensure that up to 5 combination doses are administered in Treatment Arm 1. If subjects receive fewer than 4 cycles of platinum doublet chemotherapy, the remaining cycles of combined durvalumab/tremelimumab (up to a total of 5) should be given after combination of platinum doublet chemotherapy (with maintenance pemetrexed, if applicable) (see CSP Section 7.2).

[b] Pemetrexed maintenance therapy from Week 12 to clinical progression or radiological progression for non-squamous NSCLC subjects who initially received treatment with pemetrexed and carboplatin/cisplatin, unless contraindicated per the Investigator.

[c] SoC chemotherapy will be given q3w up to 4 doses; extension into Weeks 12 and 15 is as clinically indicated, at the Investigator's discretion.

Durva Durvalumab; ECOG Eastern Cooperative Oncology Group; EGFR Epidermal growth factor receptor; NSCLC Non-small-cell lung cancer; OS Overall survival; PD Progressive disease; PFS Progression-free survival; q4w Every 4 weeks; q3w - Every 3 weeks; SoC Standard of care; TC Tumor cell; Treme Tremelimumab.

Figure 2 Study flow chart



q6w - Every 6 weeks; q8w - Every 8 weeks; RECIST 1.1 - Response Evaluation Criteria in Solid Tumors, Version 1.1; SoC - Standard of care.

1.3 Number of subjects

This study will randomize approximately 1000 eligible subjects in a 1:1:1 ratio to the following treatment arms: durvalumab + tremelimumab combination therapy + standard of care (SoC) chemotherapy (Arm1), durvalumab monotherapy + SoC chemotherapy (Arm2), or SoC chemotherapy alone (Arm3) (approximately 333 patients in each treatment arm) including at least approximately 250 patients in each treatment arm with programmed cell death ligand 1 (PD-L1) expression on less than 50% of tumor cells (PD-L1 TC <50%). Once global enrollment is complete, enrollment may continue in China only. A total of up to 180 subjects, including up to 135 subjects in total with programmed cell death ligand 1 expression on less than 50% of tumor cells (PD-L1 TC <50%), from China will be randomized. Subjects who fulfill all of the inclusion criteria and none of the exclusion criteria will be randomized in a stratified manner according to the following:

- PD-L1 tumor expression status (PD-L1 expression on at least 50% of tumor cells [PD-L1 TC ≥50%] versus PD-L1 TC <50%)
- Disease stage (Stage IVA versus Stage IVB)
- Histology (non-squamous versus squamous)

The study is sized for dual primary endpoints to characterize the PFS and OS benefits of durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone in the intent-to-treat (ITT) population. Sample size and power analysis for the dual primary endpoints and key secondary endpoints are described below.

One interim analysis of PFS will be performed when approximately 80% of the target PFS events have occurred. Three interim analyses of overall survival (OS) will be performed; the first at the time of the interim PFS analysis (approximately 45% of the target OS events), the second at the time of the primary PFS analysis (approximately 61% of the target OS events) and the third when approximately 84% of the target OS events have occurred (information fraction). The alpha will be split between the interim and final analyses using the Lan and DeMets (Karrison et al 2016

Karrison, et al. 2016. “Versatile Tests for Comparing Survival Curves Based on Weighted Log-Rank Statistics.” Stata Journal 16 (3). StataCorp LP: 678–90

Lan and DeMets 1983) spending function that approximates an O’Brien Fleming approach, with the boundaries for the treatment comparison derived based upon the exact number of events at the time of analysis.

The final (primary) PFS analysis for superiority will be performed when the following conditions have been met:

- Approximately 497 BICR PFS events from the global cohort have occurred across the durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone treatment arms (75% maturity)

The final OS analysis for superiority will be performed when the following conditions have been met:

- Approximately 532 OS events have occurred across the durvalumab monotherapy + SOC chemotherapy and SoC chemotherapy alone treatment arms (80% maturity)

Dual primary Endpoints:

Durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone

(PFS in ITT population)

Assuming the true PFS HR is 0.67 and the median PFS in SoC chemotherapy alone arm is 6 months, 497 PFS events from the global cohort (75% maturity) will provide greater than 90% power to demonstrate statistical significance at the 2-sided alpha level of 0.9% (with overall alpha for PFS 1%), allowing for 1 interim analysis conducted at approximately 80% of the target events (information fraction). The smallest treatment difference that is statistically significant will be an HR of 0.79 at final analysis. Assuming a recruitment period of 16 months, this analysis is anticipated to be 25 months from first patient in (FPI).

Durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone

(OS in ITT population)

Assuming the true OS HR is 0.7 and the median OS in SoC arm is 12.9 months, 532 OS events (80% maturity) will provide greater than 90% power to demonstrate statistical significance at the 2-sided alpha level of a 3.3% (with overall alpha for OS 4%), allowing for 3 interim analyses conducted at approximately 45%, 61% and 84% of the target events (information fraction). The smallest treatment difference that is statistically significant will be an HR of 0.83 at final analysis. Assuming a recruitment period of 16 months, this analysis is anticipated to be 46 months from FPI.

Key secondary Endpoints:

Durvalumab + tremelimumab combination therapy + SoC chemotherapy versus SoC chemotherapy alone

(PFS in ITT population)

Assuming the true PFS HR is 0.51 and the median PFS in SoC chemotherapy alone arm is 6 months, 465 PFS events from the global cohort (70% maturity) will provide greater than 90% power to demonstrate statistical significance at the 2-sided alpha level of 0.9% (with overall alpha for PFS 1%), allowing for 1 interim analysis conducted at approximately 80% of the target events (information fraction). The smallest treatment difference that is statistically

significant will be an HR of 0.78. Assuming a recruitment period of 16 months, this analysis is anticipated to be 25 months from FPI.

Durvalumab + tremelimumab combination therapy + SoC chemotherapy versus SoC chemotherapy alone

(OS in ITT population)

Assuming the true OS HR is 0.7 and the median OS in SoC arm is 12.9 months, 532 OS events (80% maturity) will provide greater than 90% power to demonstrate statistical significance at the 2-sided alpha level of a 3.3% (with overall alpha for OS 4%), allowing for 3 interim analyses conducted at approximately 45%, 61% and 84% of the target events (information fraction). The smallest treatment difference that is statistically significant will be an HR of 0.83. Assuming a recruitment period of 16 months, this analysis is anticipated to be 46 months from FPI.

Durvalumab + tremelimumab combination therapy + SoC chemotherapy versus SoC chemotherapy alone

(OS in bTMB20 high population)

It is estimated there will be approximately 70 patients in each treatment arm in the bTMB 20 high analysis set given the prevalence observed in other studies within the development program. Assuming the true OS HR is 0.45 and the median OS in SoC arm is 12.9 months, an estimated 101 OS events (72% maturity) are expected to have occurred at the time of the final OS analysis. With a minimum 101 deaths, the study will provide greater than 90% power to demonstrate statistical significance at the 2-sided alpha level of 4.12% (with overall alpha 5%), allowing for 3 interim analyses conducted at approximately 44%, 60% and 83% of the target events (information fraction), respectively. The smallest treatment difference that could be statistically significant will be a HR of 0.66.

2. ANALYSIS SETS

2.1 Definition of analysis sets

Analysis sets will be applied to each outcome variable as stated in Table 1.

Table 1 Summary of outcome variables and analysis sets

Outcome variable	Analysis set
Efficacy data	
PFS and OS	FAS (ITT)
ORR[a], DoR[b], BoR, APF12, PFS2	FAS (ITT)
PROs and symptom endpoints	FAS (ITT)

PFS, OS, ORR[a], BoR, DoR[b], APF12, and PFS2	PD-L1 TC <50% analysis set, PD-L1 TC <25% analysis set, PD-L1 TC <1% analysis set
PFS, OS, ORR[a], BoR, DoR[b], APF12, and PFS2	bTMB20 high analysis set
PFS, OS, ORR[a], DoR[b]	bTMB16 high analysis set, bTMB12 high analysis set,
PK data	PK analysis set
Safety data	
Exposure	Safety analysis set
AEs	Safety analysis set
Laboratory measurements	Safety analysis set
Vital signs	Safety analysis set
ECGs	Safety analysis set

AE Adverse event; APF12 Proportion of subjects alive and progression free at 12 months from randomization; BoR Best objective response; DoR Duration of response; ECG Electrocardiogram; FAS Full analysis set; ITT Intent-to-treat; ORR Objective response rate; OS Overall survival; PD-L1 Programmed cell death ligand 1; PFS Progression-free survival; PFS2 Time from randomization to second progression; PK Pharmacokinetic(s); PRO Patient-reported outcome; TC Tumor cell; bTMB blood based tumor mutational burden

[a] Subjects who are evaluable for the analysis of ORR are those with measurable disease at baseline.

[b] Subjects who are evaluable for the analysis of DoR are those who responded in the ORR analysis.

2.1.1 Full analysis set

The full analysis set (FAS) will include all randomized subjects. Treatment arms will be compared on the basis of randomized study treatment, regardless of the treatment actually received. Subjects who were randomized but did not subsequently go on to receive study treatment are included in the analysis in the treatment arm to which they were randomized. The analysis of data using the FAS therefore follows the principles of Intent-to-treat (ITT).

2.1.2 PD-L1 TC <50% analysis set

The PD-L1 TC <50% analysis set will include the subset of subjects in the FAS whose PD-L1 status is PD-L1 TC <50% as defined by the Ventana SP263 PD-L1 immunohistochemistry (IHC) assay (i.e., <50% PD-L1-membrane expression in tumoral tissue).

2.1.3 PD-L1 TC <25% analysis set

The PD-L1 TC <25% analysis set will include the subset of subjects in the FAS whose PD-L1 status is PD-L1 TC <25% as defined by the Ventana SP263 PD-L1 IHC assay (i.e., <25% PD-L1-membrane expression in tumoral tissue).

2.1.4 PD-L1 TC <1% analysis set

The PD-L1 expression on less than 1% of tumor cells (PD-L1 TC <1%) analysis set will include the subset of subjects in the FAS whose PD-L1 status is PD-L1 TC <1% as defined by the Ventana SP263 PD-L1 IHC assay (i.e., <1% PD-L1-membrane expression in tumoral tissue).

2.1.5 bTMB20 high analysis set

The bTMB20 high analysis set will include the subset of subjects in the FAS whose bTMB>=20 mutations per megabase (mut/Mb).

2.1.6 bTMB16 high analysis set

The bTMB16 high analysis set will include the subset of subjects in the FAS whose bTMB>=16 mut/Mb.

2.1.7 bTMB12 high analysis set

The bTMB12 high analysis set will include the subset of subjects in the FAS whose bTMB>=12 mut/Mb.

2.1.8 Safety analysis set

The safety analysis set will consist of all subjects who received at least one dose of study treatment. Safety data will not be formally analyzed but summarized, according to the treatment received; that is, erroneously treated subjects (e.g., those randomized to treatment A but actually given treatment B) will be summarized according to the treatment they actually received. Subjects in the durvalumab + tremelimumab combination therapy + SoC chemotherapy arm or durvalumab monotherapy + SoC chemotherapy arm who received only SoC chemotherapy and no doses of durvalumab or tremelimumab will be summarized in the SoC chemotherapy alone arm. Decisions on how to assign erroneously treated subjects to treatment groups will be made on a case-by-case basis.

2.1.9 Pharmacokinetic analysis set

All subjects who received at least one dose of study treatment per the protocol for whom any post-dose data are available and who do not violate or deviate from the protocol in ways that would significantly affect the pharmacokinetic (PK) analyses will be included in the PK analysis set. The population will be defined by the Study Physician, Pharmacokineticist, and Statistician prior to any analyses being performed.

2.1.10 ADA-evaluable set

Subjects in the safety analysis set with non-missing baseline ADA sample and at least 1 post-baseline ADA sample will be included in the ADA-evaluable set.

2.2 Violations and deviations

The following general categories will be considered important protocol deviations (IPDs) and will be programmatically derived from the electronic case report form (eCRF) data. These will be listed and discussed in the clinical study report (CSR) as appropriate:

Table 2 Important protocol deviations categories

Deviation	Important protocol deviation categories
1	Subjects randomized but who did not receive study treatment
2	Subjects who deviate from Inclusion criteria 3, 4 or 5 or from Exclusion criteria 5 per the CSP
3	Baseline RECIST scan > 42 days before randomisation
4	No baseline RECIST 1.1 assessment on or before date of randomization
5	Received prohibited concomitant systemic anti-cancer medications (including other anti-cancer agents). Please refer to the CSP section 7.7 for those medications that are detailed as being 'excluded' from permitted use during the study. This will be used as a guiding principle for the physician review of all medications prior to database lock.
6	Subjects randomized who received an incorrect study treatment, i.e. a treatment different to that which they were randomized to. Subjects who receive the wrong treatment at any time will be included in the safety analysis set as described in Section 2.1. During the study, decisions on how to handle errors in treatment dispensing (with regard to continuation/discontinuation of study treatment or, if applicable, analytically) will be made on an individual basis with written instruction from the study team leader and/or statistician.

CSP Clinical study protocol; RECIST 1.1 Response Evaluation Criteria in Solid Tumors, Version 1.1.

The IPDs will be listed and summarized by randomized treatment group. IPDs will also be summarised separately whether they are due to COVID-19 or not due to COVID-19. COVID-19 PDs that are not IPDs will also be listed. Deviation 1 will lead to exclusion from the safety analysis set. None of the other deviations will lead to subjects being excluded from the analysis sets described in Section 2.1 (with the exception of the PK analysis set, if the deviation is considered to impact upon PK). A per-protocol analysis excluding subjects with specific IPDs is not planned; however, a 'deviation bias' sensitivity analysis may be performed on the PFS and OS endpoint excluding subjects with deviations that may affect the efficacy of the trial therapy if > 10% of subjects in either treatment group:

- Did not have the intended disease or indication or
- Did not receive any randomized therapy.

The need for such a sensitivity analysis will be determined following review of the protocol deviations ahead of database lock and will be documented prior to the primary analysis being conducted.

In addition to the programmatic determination of the deviations above, other study deviations captured from the eCRF module for inclusion/exclusion criteria will be tabulated and listed. Any other deviations from monitoring notes or reports will be reported in an appendix to the CSR.

3. PRIMARY AND SECONDARY VARIABLES

3.1 Derivation of RECIST visit responses

For all subjects, the RECIST tumor response data will be used to determine each subject's visit response according to RECIST version 1.1 (see clinical study protocol [CSP] Appendix F). It will also be used to determine if and when a subject has progressed in accordance with RECIST and their BoR to study treatment.

Baseline radiological tumor assessments are to be performed no more than 28 days before the date of randomization and ideally as close as possible to randomization. Tumor assessments are then performed every 6 weeks (q6w), ±1 week, for the first 12 weeks relative to the date of randomization, and every 8 weeks (q8w), ±1 week, thereafter until objective disease progression, as defined in CSP Section 5.1.1, or death (see schedule of assessments for treatment period [CSP Table 2] and follow-up [CSP Table 3]).

If an unscheduled assessment is performed, and the subject has not progressed, every attempt should be made to perform the subsequent assessments at their scheduled visits. This schedule is to be followed in order to minimize any unintentional bias caused by some subjects being assessed at a different frequency than other subjects.

For subjects who discontinue study treatment due to toxicity in the absence of objective disease progression, objective tumor assessments should be continued q6w±1 week for 12 weeks (relative to the date of randomization), then q8w±1 week until objective disease progression or death.

Following objective disease progression, subjects should continue to be followed up for survival status at Months 2, 3, and 4, and then q8w ±2 weeks, as outlined in the follow-up schedules of assessments (CSP Table 3).

BICR according to RECIST 1.1 will be regarded as primary in terms of the efficacy analyses. A sensitivity analysis will be performed using investigator assessment data. CCI [redacted]

From the investigator's review of the imaging scans, the RECIST tumour response data will be used to determine each patient's visit response according to RECIST version 1.1. At each visit, patients will be programmatically assigned a RECIST 1.1 visit response of CR, PR, SD or PD, using the information from target lesions (TLs), non-target lesions (NTLs) and new lesions and depending on the status of their disease compared with baseline and previous assessments. If a patient has had a tumour assessment that cannot be evaluated then the patient will be assigned a visit response of not evaluable (NE), (unless there is evidence of progression in which case the response will be assigned as PD).

Please refer to Table 3 and Table 4 for the definitions of CR, PR, SD and PD.

RECIST outcomes (i.e. PFS, ORR etc.) will be calculated programmatically for the site investigator data (see Section 3.2) from the overall visit responses.

The following sections pertain to site investigator data and the programmatic derivation of visit response.

3.1.1 Investigator RECIST 1.1-based assessments: Target lesions

Measurable disease is defined as having at least one measurable lesion, not previously irradiated prior to randomization, which is ≥ 10 mm in the longest diameter (LD), (except lymph nodes which must have short axis ≥ 15 mm) with computed tomography (CT) or magnetic resonance imaging (MRI) and which is suitable for accurate repeated measurements.

A subject can have a maximum of five measurable lesions recorded at baseline with a maximum of two lesions per organ (representative of all lesions involved and suitable for accurate repeated measurement) and these are referred to as TLs. Lymph nodes, in any location, are collectively considered as a single organ, with a maximum of 2 lymph nodes as TLs. A bilateral or multi-lobular organ is considered as a single organ. If more than one baseline scan is recorded then measurements from the one that is closest and prior to randomization will be used to define the baseline sum of TLs. It may be the case that, on occasion, the largest lesion does not lend itself to reproducible measurement, in which circumstance the next largest lesion, which can be measured reproducibly, should be selected.

All other lesions (or sites of disease) not recorded as TLs should be identified as NTLs at baseline. Measurements are not required for these lesions, but their status should be followed at subsequent visits.

Only subjects with measurable target disease at baseline should be included in the study. However, if subjects who do not have measurable disease at entry (i.e. no TLs), but have non-measurable disease, are enrolled in the study, evaluation of overall visit responses will be based on the overall NTL assessment and the absence/presence of new lesions (see Section 3.1.3 for further details). If a subject does not have measurable disease at baseline then the TL visit response will be not applicable (NA).

If no TLs and no NTLs are recorded at a visit, both the TL and NTL visit response will be recorded as NA and the overall visit response will be no evidence of disease (NED). If a new lesion is observed then the overall visit response will be PD.

Table 3 TL visit responses (RECIST 1.1)

Visit responses	Description
Complete response (CR)	Disappearance of all TLs. Any pathological lymph nodes selected as TLs must have a reduction in short axis to < 10mm.
Partial response (PR)	At least a 30% decrease in the sum of diameters of TLs, taking as reference the baseline sum of diameters as long as criteria for PD are not met.
Progressive disease (PD)	A ≥ 20% increase in the sum of diameters of TLs and an absolute increase of ≥ 5mm, taking as reference the smallest sum of diameters since treatment started including the baseline sum of diameters.
Stable disease (SD)	Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD.
Not evaluable (NE)	Only relevant if any of the TLs at follow-up were not assessed or not evaluable (e.g. missing anatomy) or had a lesion intervention at this visit. Note: If the sum of diameters meets the PD criteria, PD overrides NE as a TL response.
Not applicable (NA)	No TLs are recorded at baseline.

Rounding of TL data

For calculation of PD and PR for TLs percentage changes from baseline and previous minimum should be rounded to one decimal place before assigning a TL response. For example 19.95% should be rounded to 20.0% but 19.94% should be rounded to 19.9%

Missing TL data

For a visit to be evaluable, all TL measurements should be recorded. However, a visit response of PD should still be assigned if any of the following occurred

- A new lesion is recorded
- A NTL visit response of PD is recorded
- The sum of TLs is sufficiently increased to result in a 20% increase, and an absolute increase of ≥ 5mm, from nadir even assuming the non-recorded TLs have disappeared

Note: the nadir can only be taken from assessments where all the TLs had a LD recorded.

If there is at least one TL assessment of NE and a visit response of PD cannot be assigned, the visit response is NE.

Lymph nodes

For lymph nodes, if the size reduces to < 10mm then these are considered non-pathological. However, a size will still be given and this size should still be used to determine the TL visit response as normal. In the special case where all lymph nodes are < 10mm and all other TLs are 0mm then although the sum may be > 0mm the calculation of TL response should be over-written as a CR.

TL visit responses subsequent to CR

A CR can only be followed by CR, PD or NE. If a CR has occurred then the following rules at the subsequent visits must be applied:

- Step 1: If all lesions meet the CR criteria (i.e. 0mm or < 10mm for lymph nodes) then response will be set to CR irrespective of whether the criteria for PD of TL is also met i.e. if a lymph node LD increases by 20% but remains < 10mm.

- Step 2: If some lesion measurements are missing but all other lesions meet the CR criteria (i.e. 0mm or < 10mm for lymph nodes) then response will be set to NE irrespective of whether, when referencing the sum of TL diameters, the criteria for PD are also met.

- Step 3: If not all lesions meet the CR criteria and the sum of lesions meets the criteria for PD then response will be set to PD

- Step 4: If after steps 1 – 3 a response can still not be determined the response will be set to remain as CR

TL too big to measure

If a TL becomes too big to measure this should be indicated in the database and a size ('x') above which it cannot be accurately measured should be recorded. If using a value of x in the calculation of TL response would not give an overall visit response of PD, then this will be flagged and reviewed by the study team blinded to treatment assignment. It is expected that a visit response of PD will remain in the vast majority of cases.

TL too small to measure

If a TL becomes too small to measure then this will be indicated as such on the eCRF and a value of 5mm will be entered into the database and used in TL calculations. However a smaller value may be used if the radiologist has not indicated 'too small to measure' on the case report form and has entered a smaller value that can be reliably measured. If a TL

response of PD results then this will be reviewed by the study team blinded to treatment assignment.

Irradiated lesions/lesion intervention

Any TL (including lymph nodes), which has had intervention during the study (for example, irradiation / palliative surgery / embolization), should be handled in the following way and once a lesion has had intervention then it should be treated as having intervention for the remainder of the study noting that an intervention will most likely shrink the size of tumors:

- Step 1: the diameters of the TLs (including the lesions that have had intervention) will be summed and the calculation will be performed in the usual manner. If the visit response is PD, this will remain as a valid response category.

- Step 2: If there was no evidence of progression after step 1, treat the lesion diameter (for those lesions with intervention) as missing and if ≤ 1/3 of the TLs have missing measurements then scale up as described in the 'Scaling' section below. If the scaling results in a visit response of PD then the subject would be assigned a TL response of PD.

- Step 3: If, after both steps, PD has not been assigned, then, if appropriate (i.e. if ≤ 1/3 of the TLs have missing measurements), the scaled sum of diameters calculated in step 2 should be used, and PR or SD then assigned as the visit response. Subjects with intervention are evaluable for CR as long as all non-intervened lesions are 0 (or <10mm for lymph nodes) and the lesions that have been subject to intervention have a value of 0 (or <10mm for lymph nodes) recorded. If scaling up is not appropriate due to too few non-missing measurements then the visit response will be set as NE.

At subsequent visits, the above steps will be repeated to determine the TL and overall visit response. When calculating the previous minimum, lesions with intervention should be treated as missing and scaled up (as per step 2 above).

Scaling (applicable only for irradiated lesions/lesion intervention)

If > 1/3 of TL measurements are missing (because of intervention) then the TL response will be NE, unless the sum of diameters of non-missing TL would result in PD (i.e. if using a value of 0 for missing lesions, the sum of diameters has still increased by 20% or more compared to nadir and the sum of TLs has increased by ≥5mm from nadir).

If ≤ 1/3 of the TL measurements are missing (because of intervention) then the results will be scaled up (based on the sizes at the nadir visit) to give an estimated sum of diameters and this will be used in calculations; this is equivalent to comparing the visit sum of diameters of the non-missing lesions to the nadir sum of diameters excluding the lesions with missing measurements.

Example of scaling

Lesion	Longest diameter at nadir visit	Longest diameter at follow-up visit
1	7.2	7.1
2	6.7	6.4
3	4.3	4.0
4	8.6	8.5
5	2.5	Intervention
Sum	**29.3**	**26**

Lesion 5 is missing at the follow-up visit.

The sum of lesions 1-4 at the follow-up is 26 cm. The sum of the corresponding lesions at the nadir visit is 26.8 cm.

Scale up as follows to give an estimated TL sum of 28.4 cm:

(26 / 26.8) * 29.3 = 28.4 cm

CR will not be allowed as a TL response for visits where there is missing data. Only PR, SD or PD (or NE) could be assigned as the TL visit response in these cases. However, for visits with ≤ 1/3 lesion assessments not recorded, the scaled up sum of TLs diameters will be included when defining the nadir value for the assessment of progression.

Lesions that split in two

If a TL splits in two, then the LDs of the split lesions should be summed and reported as the LD for the lesion that split.

Lesions that merge

If two TLs merge, then the LD of the merged lesion should be recorded for one of the TL sizes and the other TL size should be recorded as 0cm.

Change in method of assessment of TLs

CT, MRI and clinical examination are the only methods of assessment that can be used within a trial, with CT and MRI being the preferred methods and clinical examination only used in special cases. If a change in method of assessment occurs (between CT and MRI) this will be considered acceptable and no adjustment within the programming is needed.

If a change in method involves clinical examination (e.g. CT changes to clinical examination or vice versa), any affected lesions should be treated as missing.

3.1.2 Investigator RECIST 1.1-based assessments: Non-target lesions and new lesions

At each visit, the investigator should record an overall assessment of the NTL response. This section provides the definitions of the criteria used to determine and record overall response for NTL at the investigational site at each visit.

NTL response will be derived based on the investigator's overall assessment of NTLs as follows:

Table 4 NTL visit responses

Visit responses	Description
Complete response (CR)	Disappearance of all NTLs present at baseline with all lymph nodes non-pathological in size (<10 mm short axis).
Progressive disease (PD)	Unequivocal progression of existing NTLs. Unequivocal progression may be due to an important progression in one lesion only or in several lesions. In all cases, the progression MUST be clinically significant for the physician to consider changing (or stopping) therapy.
Non-CR/Non-PD	Persistence of one or more NTLs with no evidence of progression.
Not evaluable (NE)	Only relevant when one or some of the NTLs were not assessed and, in the investigator's opinion, they are not able to provide an evaluable overall NTL assessment at this visit. Note: For subjects without TLs at baseline, this is relevant if any of the NTLs were not assessed at this visit and the progression criteria have not been met.
Not applicable (NA)	Only relevant if there are no NTLs at baseline.

To achieve 'unequivocal progression' on the basis of NTLs, there must be an overall level of substantial worsening in non-target disease such that, even in the presence of SD or PR in TLs, the overall tumor burden has increased sufficiently to merit a determination of disease progression. A modest 'increase' in the size of one or more NTLs is usually not sufficient to qualify for unequivocal progression status.

Details of any new lesions will also be recorded with the date of assessment. The presence of one or more new lesions is assessed as progression.

A lesion identified at a follow up assessment in an anatomical location that was not scanned at baseline is considered a new lesion and will indicate disease progression.

The finding of a new lesion should be unequivocal: i.e. not attributable to differences in scanning technique, change in imaging modality or findings thought to represent something other than tumor.

If a new lesion is equivocal, for example because of its small size, the treatment and tumor assessments should be continued until the previously new lesion has been assessed as unequivocal and then the progression date should be declared using the date of the initial scan when the new lesion first appeared.

New lesions will be identified via a Yes/No tick box. The absence and presence of new lesions at each visit should be listed alongside the TL and NTL visit responses.

A new lesion indicates progression so the overall visit response will be PD irrespective of the TL and NTL response.

If the question 'Any new lesions since baseline' has not been answered with Yes or No and the new lesion details are blank this is not evidence that no new lesions are present, but should not overtly affect the derivation. This scenario (i.e. whereby new lesion response is NE), should only occur in exceptional cases, as missing data for the new lesion field should always be queried.

Symptomatic progression is not a descriptor for progression of NTLs: it is a reason for stopping study therapy and will not be included in any assessment of NTLs.

Subjects with 'symptomatic progression' requiring discontinuation of treatment without objective evidence of disease progression at that time should continue to undergo tumor assessments where possible until objective disease progression is observed or death occurs (whichever comes first).

3.1.3 Investigator RECIST 1.1-based assessments: Overall visit response

Table 5 defines how the previously defined TL and NTL visit responses will be combined with new lesion information to give an overall visit response.

Table 5 Overall visit responses

Target lesions	Non-target lesions	New lesions	Overall visit response
CR	CR or NA	No (or NE)	CR
CR	Non-CR/Non-PD or NE	No (or NE)	PR
PR	Non-PD or NE or NA	No (or NE)	PR
SD	Non-PD or NE or NA	No (or NE)	SD
PD	Any	Any	PD
Any	PD	Any	PD
Any	Any	Yes	PD
NE	Non-PD or NE or NA	No (or NE)	NE

NA	CR	No (or NE)	CR
NA	Non-CR/Non-PD	No (or NE)	SD
NA	NE	No (or NE)	NE
NA	NA	No (or NE)	NED

CR Complete response, NA Not applicable (only relevant if there were no NTLs at baseline), NE Not evaluable, NED No evidence of disease, PD Progressive disease, PR Partial response, SD Stable disease.

3.1.4 Blinded Independent Central Review of RECIST 1.1-based assessments

A planned BICR of radiological images will be carried out using RECIST 1.1. All radiological scans for all subjects (including those at unscheduled visits, or outside visit windows) will be collected on an ongoing basis and sent to an AstraZeneca appointed Contract Research Organization (CRO) for quality control (QC) and storage. The imaging scans will be reviewed by two independent radiologists using RECIST 1.1 and will be adjudicated, if required (i.e. two reviewers review the scans and adjudication is performed by a separate reviewer in case of a disagreement). For each subject, the BICR will define the overall visit response (i.e. the response obtained overall at each visit by assessing TLs, NTLs and new lesions) data and no programmatic derivation of visit response is necessary. (For subjects with TLs at baseline: CR, PR, SD, PD, NE; for subjects with NTLs only: CR, SD [Non-CR/Non-PD], PD, NE; for subjects with no evidence of disease at baseline [NED] evaluation of overall visit response at each visit will be based on absence/presence of new lesions: PD, NED, NE). If a subject has had a tumor assessment that cannot be evaluated then the subject will be assigned a visit response of NE (unless there is evidence of progression in which case the response will be assigned as PD). RECIST assessments/scans contributing towards a particular visit may be performed on different dates and for the central review the date of progression for each reviewer will be provided based on the earliest of the scan dates of the component that triggered the progression.

Adjudication is triggered by any difference in overall timepoint response between the two primary reviewers. If adjudication is performed, the reviewer that the adjudicator agreed with will be selected as a single reviewer (note in the case of more than one review period, the latest adjudicator decision will be used). In the absence of adjudication, the records for all visits for a single reviewer will be used. The reviewer selected in the absence of adjudication will be the reviewer who read the baseline scan first. The records from the single selected reviewer will be used to report all BICR RECIST information including dates of progression, visit response, censoring and changes in TL dimensions. Endpoints (of ORR, PFS and DoR) will be derived programmatically from this information.

Results of this independent review will not be communicated to investigators and the management of subjects will be based solely upon the results of the RECIST 1.1 assessment conducted by the investigator.

A BICR of all subjects will be performed for the final database lock for PFS, which will cover all of the scans up to the data cut-off (DCO).

Further details of the BICR will be documented in the Imaging Charter.

3.2 Outcome Variables

The analysis of PFS, ORR, DoR, BoR, and APF12 will be based on BICR tumor assessments according to RECIST 1.1. Overall survival (OS) will be evaluated from all-cause mortality. Additionally, time from randomization to second progression (PFS2) will be defined by local clinical practice.

A sensitivity analysis of PFS will be performed using the investigator tumor assessments. CCI [redacted]

3.2.1 Dual primary variables

The dual primary endpoints are PFS using BICR assessments according to RECIST 1.1 and OS, comparing durvalumab monotherapy + SoC chemotherapy with SoC chemotherapy alone in the FAS.

3.2.1.1 Progression-free survival

PFS is defined as the time from randomization until the date of objective disease progression or death (by any cause in the absence of progression) regardless of whether the subject withdraws from randomized therapy or receives another anti-cancer therapy prior to progression (i.e. date of PFS event or censoring – date of randomization + 1). Subjects who have not progressed or died at the time of analysis will be censored at the time of the latest date of assessment from their last evaluable RECIST 1.1 assessment. However, if the subject progresses or dies after two or more missed visits, the subject will be censored at the time of the latest evaluable RECIST 1.1 assessment prior to the two missed visits (Note: NE visit is not considered as missed visit).

Given the scheduled visit assessment scheme (i.e. six-weekly for the first 12 weeks then eight-weekly thereafter) the definition of 2 missed visits will change. If the previous RECIST assessment is less than study day 36 (i.e. week 5) then two missing visits will equate to 14 weeks since the previous RECIST assessment, allowing for early and late visits (i.e., 2 x 6 weeks + 1 week for an early assessment + 1 week for a late assessment = 14 weeks). If the two missed visits occur over the period when the scheduled frequency of RECIST assessments changes from six-weekly to eight-weekly this will equate to 16 weeks (i.e., take the average of 6 and 8 weeks which gives 7 weeks and then apply same rationale hence 2 x 7 weeks + 1 week for an early assessment + 1 week for a late assessment = 16 weeks). The time period for the previous RECIST assessment will be from study days 36 to 78 (i.e. week 5 to week 11). From week 11 onwards (when the scheduling changes to eight-weekly assessments), two

missing visits will equate to 18 weeks (i.e. 2 * 8 weeks + 1 week for an early assessment + 1 week for a late assessment = 18 weeks).

If the subject has no evaluable visits or does not have baseline data they will be censored at Day 1 unless they die within two visits of baseline (12 weeks plus 1 week allowing for a late assessment within the visit window), in which case the date of death is used when deriving PFS.

The PFS time will always be derived based on scan/assessment dates, not visit dates.

RECIST assessments/scans contributing towards a particular visit may be performed on different dates. The following rules will be applied:

- For BICR assessments, the date of progression will be determined based on the earliest of the scan dates of the component that triggered the progression for the adjudicated reviewer selecting PD or of the reviewer who read baseline first if there is no adjudication for BICR data.

- For investigational assessments, the date of progression will be determined based on the earliest of the dates of the component that triggered the progression.

- For both BICR and investigational assessments, when censoring a subject for PFS the subject will be censored at the latest of the dates contributing to a particular overall visit assessment.

Note: for TLs only the latest scan date is recorded out of all scans performed at that assessment for the TLs and similarly for NTLs only the latest scan date is recorded out of all scans performed at that assessment for the NTLs.

3.2.1.2 Overall survival

OS is defined as the time from the date of randomization until death due to any cause regardless of whether the subject withdraws from randomized therapy or receives another anti-cancer therapy (i.e. date of death or censoring – date of randomization + 1). Any subject not known to have died at the time of analysis will be censored based on the last recorded date on which the subject was known to be alive (SUR_DAT, recorded within the SURVIVE module of the eCRF).

Note, survival calls will be made in the week following the date of DCO for the analysis, and if subjects are confirmed to be alive or if the death date is post the DCO date these subjects will be censored at the date of DCO. The status of ongoing, withdrawn (from the study) and "lost to follow-up" subjects at the time of the final OS analysis should be obtained by the site personnel by checking the subject's notes, hospital records, contacting the subject's general practitioner and checking publicly-available death registries. In the event that the subject has actively withdrawn consent to the processing of their personal data, the vital status of the

subject can be obtained by site personnel from publicly available resources where it is possible to do so under applicable local laws.

Note, for OS analyses, it may be necessary to use all relevant eCRF fields to determine the last recorded date on which the subject was known to be alive. The last date for each individual subject is defined as the latest among the following dates recorded on the eCRFs:

- AE start and stop dates
- Admission and discharge dates of hospitalization
- Study treatment date
- End of treatment date
- Laboratory test dates
- Date of vital signs
- Disease assessment dates on RECIST eCRF
- Start and stop dates of alternative anti-cancer treatment
- Date last known alive on survival status eCRF
- End of study date

3.2.2 Secondary variables

3.2.2.1 Objective response rate

ORR (per RECIST 1.1 using BICR assessments) is defined as the percentage of subjects with at least one visit response of CR or PR, with the denominator defined as subset of all randomized subjects with measurable disease at baseline per BICR. ORR will also be obtained, using the same algorithm, for the RECIST 1.1 site investigator tumor data, based on a subset of all randomized subjects with measurable disease at baseline per the site investigator.

Data obtained up until progression, or last evaluable assessment in the absence of progression, will be included in the assessment of ORR. Subjects who discontinue randomized treatment without progression, receive a subsequent anti-cancer therapy (note that for this analysis radiotherapy is not considered a subsequent anti-cancer therapy) and then respond will not be included as responders in the ORR.

CCI

3.2.2.2 Duration of response

DoR (per RECIST 1.1 using BICR assessments) will be defined as the time from the date of first documented response until date of documented progression or death in the absence of disease progression (i.e. date of PFS event or censoring – date of first response + 1). The end of response should coincide with the date of progression or death from any cause used for the RECIST 1.1 PFS endpoint. The time of the initial response will be defined as the latest of the dates contributing towards the first visit response of PR or CR.

If a subject does not progress following a response, then their DoR will use the PFS censoring time.

DoR will not be defined for those subjects who do not have a documented response.

3.2.2.3 Time from randomization to second progression

PFS2 is defined as the time from the date of randomization to the earliest of the progression event (subsequent to that used for the primary variable PFS i.e. PFS event as reported by the site investigator) or death (i.e. date of PFS2 event or censoring – date of randomization + 1). The date of the first progression will be programmatically determined from investigator-assessed data (see Section 3.2.1.1 for details). The date of second progression will be recorded by the investigator and defined according to local standard clinical practice and may involve any of: objective radiological imaging, symptomatic progression, or death. The date of the PFS2 assessment and investigator opinion of progression status (progressed or non-progressed) at each assessment will be recorded in the eCRF. Second progression status will be reviewed (q6w for the first 12 weeks relative to the date of randomization and q8w thereafter) following the progression event used for the primary variable PFS (the first progression) and status recorded. Subjects alive and for whom a second disease progression has not been observed should be censored at the last time known to be alive and without a second disease progression, i.e. censored at the latest of the PFS or PFS2 assessment date if the subject has not had a second progression or death.

3.2.2.4 Proportion of subjects alive and progression-free at 12 months

APF12 will be defined as the Kaplan-Meier estimate of PFS (per RECIST 1.1 using BICR assessments) at 12 months.

3.2.2.5 Best objective response

BoR is calculated based on the overall visit responses from each RECIST 1.1 assessment, described in CSP Appendix F. It is the best response a subject has had following randomization, but prior to starting any subsequent cancer therapy and up to and including RECIST 1.1 progression or the last evaluable assessment in the absence of RECIST 1.1 progression, as determined by BICR. Categorization of BoR will be based on RECIST using the following response categories: CR, PR, SD, PD and NE.

For determination of a best response of SD, the earliest of the dates contributing towards a particular overall visit assessment will be used. SD should be recorded at least 6 weeks minus 1 week, i.e. at least 35 days (to allow for an early assessment within the assessment window), after randomization. For CR/PR, the initial overall visit assessment that showed a response will use the latest of the dates contributing towards a particular overall visit assessment.

BoR will be determined programmatically based on RECIST 1.1 from the overall visit response using all BICR data up until the first progression event. It will also be determined programmatically based on RECIST using all site investigator data up until the first progression event.

The denominator will be consistent with those used in the ORR analysis.

For subjects whose progression event is death, BoR will be calculated based upon all evaluable RECIST assessments prior to death.

For subjects who die with no evaluable RECIST 1.1 assessments, if the death occurs ≤91 days (i.e. 12 weeks plus 1 week allowing for a late assessment within the visit window) after randomization, then BoR will be assigned to the PD category. For subjects who die with no evaluable RECIST 1.1 assessments, if the death occurs >91 days after the date of randomization then BoR will be assigned to the NE category.

Progression events that have been censored due to them being more than two missed visits after the last assessment will not contribute to the BoR derivation.

A subject will be classified as a responder if the RECIST criteria for a CR or PR are satisfied at any time following randomization, prior to RECIST progression and prior to starting any subsequent cancer therapy.

3.3 Patient-reported outcome variables

Patient-reported outcome (PRO) questionnaires, a secondary endpoint of interest, will be assessed using the European Organization for Research and Treatment of Cancer (EORTC) 30-item Core Quality of Life Questionnaire (QLQ-C30) with the 13-item Lung Cancer Quality of Life Questionnaire (QLQ-LC13) module (health-related quality of life [HRQoL] and lung cancer specific symptoms), CCI [redacted] All items/questionnaires will be scored according to published scoring guidelines or the developer's guidelines, if published guidelines are not available.

3.3.1 EORTC QLQ-C30

The EORTC QLQ-C30 consists of 30 questions that can be combined to produce 5 functional scales (physical, role, cognitive, emotional, and social), 3 symptom scales (fatigue, pain, and nausea/vomiting), 6 individual items including 5 symptoms (dyspnea, insomnia, appetite loss,

constipation, and diarrhea) and an item assessing financial difficulties), and a 2-item global measure of health status. The EORTC QLQ-C30 will be scored according to the EORTC QLQ-C30 Scoring Manual (Fayers et al 2001). An outcome variable consisting of a score from 0 to 100 will be derived for each of the symptom scales/symptom items, each of the functional scales, and the global health status scale in the EORTC QLQ-C30 according to the EORTC QLQ-C30 Scoring Manual. Higher scores on the global health status and functioning scales indicate better health status/function, but higher scores on symptom scales and individual symptom items represent greater symptom severity. The EORTC QLQ-C30 functional and symptom scales, individual symptom items and global health status are derived as follows.

1. Calculate the average of the items that contribute to the scale or take the value of an individual item, i.e. the raw score (RS):

   $$RS = (I_1 + I_2 + \ldots + I_n) / n,$$

   where $I_1 + I_2 + \ldots + I_n$ are the items included in a scale and n is the number of items in a scale.

2. Use a linear transformation to standardize the raw score, so that scores range from 0 to 100, where a higher score represents a higher ("better") level of functioning, or a higher ("worse") level of symptoms.

   Functional scales: $Score = (1 - [RS - 1] / range) * 100$

   Symptom scales/items; global health status: $Score = ([RS - 1] / range) * 100,$

   where *range* is the difference between the maximum and the minimum possible value of RS.

The number of items and item range for each scale/item are displayed in Table 6.

Table 6 Scoring the EORTC QLQ-C30

Scale/ item	Scale/ item abbreviation	Number of items (n)	Item range	Item numbers
Global health status/ QoL	QL	2	6	29, 30
Functional scales				
Physical	PF	5	3	1-5
Role	RF	2	3	6, 7
Cognitive	CF	2	3	20, 25
Emotional	EF	4	3	21-24

Scale/ item	Scale/ item abbreviation	Number of items (n)	Item range	Item numbers
Social	SF	2	3	26, 27
Symptom scales				
Fatigue	FA	3	3	10, 12, 18
Pain	PA	2	3	9, 19
Nausea/ vomiting	NV	2	3	14, 15
Symptom items				
Dyspnea	DY	1	3	8
Insomnia	SL	1	3	11
Appetite loss	AP	1	3	13
Constipation	CO	1	3	16
Diarrhea	DI	1	3	17
Financial difficulties	FI	1	3	28

For each subscale, if <50% of the subscale items are missing, then the subscale score will be divided by the number of non-missing items and multiplied by the total number of items on the subscales (Fayers et al 2001). If at least 50% of the items are missing, then that subscale will be treated as missing. Missing single items are treated as missing. The reason for any missing questionnaire will be identified and recorded. If there is evidence that the missing data are systematic, missing values will be handled to ensure that any possible bias is minimized.

Changes in score compared with baseline will be evaluated as described below.

Clinically meaningful changes

A minimum clinically meaningful change is defined as an absolute change in the score from baseline of ≥ 10/≤ -10 for scales/items from the EORTC QLQ-C30 (Osoba et al 1998). At each post-baseline assessment, the change in symptoms/functioning from baseline will be categorized as improvement, no change or deterioration as shown in Table 7.

Table 7 EORTC QLQ-C30 Clinically Meaningful Changes

Score	Change from baseline	Assessment period response
	≥ 10	Improvement
	< 10, > -10	No change

EORTC QLQ-C30 global health status/ QoL and functional scales	≤ -10	Deterioration
EORTC QLQ-C30 symptom scales and items	≤ -10	Improvement
	> -10, < 10	No change
	≥ 10	Deterioration

3.3.1.1 Time to HRQoL/symptom deterioration

Time to deterioration will be derived for all QLQ-C30 items (see Table 6), except financial difficulties.

Time to deterioration will be defined as the time from the date of randomization until the date of the first clinically meaningful deterioration that is confirmed at a subsequent visit or death (by any cause) in the absence of a clinically meaningful deterioration, regardless of whether the subject withdraws from study treatment or receives another anti-cancer therapy prior to HRQoL/symptom deterioration. Subjects with a single deterioration and with no further assessments will be treated as deteriorated in the analysis.

Death will be included as an event only if the death occurs within 2 visits of the last PRO assessment where the HRQoL/symptom change could be evaluated.

Subjects whose HRQoL/symptom (as measured by EORTC QLQ-C30) has not shown a clinically meaningful deterioration and who are alive at the time of the analysis will be censored at the time of their last PRO assessment where the HRQoL/symptom could be evaluated. Also, if the HRQoL/symptom deteriorates after 2 or more missed PRO assessment visits or if the subject dies after 2 or more missed PRO assessment visits, the subject will be censored at the time of the last PRO assessment where HRQoL/symptom could be evaluated. If a subject has no evaluable visits or does not have baseline data they will be censored at date of randomization.

The population for analysis of time to global health status/HRQoL or function deterioration will include a subset of the FAS population who have baseline scores ≥10; the population for analysis of time to symptom deterioration will include a subset of the FAS population who have baseline scores ≤90.

3.3.1.2 Symptom improvement rate

The symptom improvement rate will be defined as the number (%) of subjects with 2 consecutive assessments at least 14 days apart that show a clinically meaningful improvement in that symptom from baseline. The denominator will consist of a subset of the FAS who have a baseline symptom score ≥10. Symptom improvement rate will be derived for the 3 symptom scales and the 5 individual symptom items.

3.3.1.3 HRQoL/function improvement rate

The HRQoL/function improvement rate (hereafter function improvement rate) will be defined as the number (%) of subjects with 2 consecutive assessments at least 14 days apart that show a clinically meaningful improvement in that function scale from baseline. The denominator will consist of a subset of the FAS who have a baseline HRQoL/function score ≤90. Function improvement rate will be derived for the 5 functional scales and the global health status/HRQoL scale.

3.3.2 Lung cancer module (EORTC QLQ-LC13)

The QLQ-LC13 is a lung cancer specific module from the EORTC for lung cancer comprising 13 questions to assess lung cancer symptoms (cough, hemoptysis, dyspnea, and site-specific pain), treatment-related symptoms (sore mouth, dysphagia, peripheral neuropathy, and alopecia), and pain medication. The scoring approach for the QLQ-LC13 is identical in principle to that for the symptom scales/items of the QLQ-C30 (see Section 3.3.1), using Table 8.

Table 8 Scoring the EORTC QLQ-LC13

Scale/ item	Scale/ item abbreviation	Number of items (n)	Item range	Item numbers
Lung cancer symptoms				
Cough	LCCO	1	3	31
Hemoptysis	LCHA	1	3	32
Dyspnea	LCDY	3	3	33-35
Site-specific pain				
Pain in chest	LCPC	1	3	40
Pain in arm/shoulder	LCPA	1	3	41
Pain in other parts	LCPO	1	3	42
Treatment-related symptoms				
Sore mouth	LCSM	1	3	36
Dysphagia	LCDS	1	3	37
Peripheral neuropathy	LCPN	1	3	38
Alopecia	LCHR	1	3	39

The dyspnea scale will only be used if all 3 items have been scored; otherwise the items are treated as single-item measures.

Changes in score compared with baseline will be evaluated as described below.

Clinically meaningful changes

A minimum clinically meaningful change is defined as an absolute change in the score from baseline of ≥ 10/≤ -10 for scales/items from the EORTC QLQ-LC13 (Osoba et al 1998). At each post-baseline assessment, the change in symptoms from baseline will be categorized as improvement, no change or deterioration as shown in Table 9.

Table 9 EORTC QLQ-LC13 Clinically Meaningful Changes

Change from baseline	Assessment period response
≤ -10	Improvement
> -10, < 10	No change
≥ 10	Deterioration

3.3.2.1 Time to symptom deterioration

Time to deterioration will be derived for all items in the lung cancer symptoms scale (see Table 8) in the QLQ-LC13.

Time to symptom deterioration will be defined as the time from the date of randomization until the date of the first clinically meaningful symptom deterioration that is confirmed at a subsequent visit or death (by any cause) in the absence of a clinically meaningful symptom deterioration, regardless of whether the subject withdraws from study treatment or receives another anti-cancer therapy prior to symptom deterioration. Subjects with single deterioration and no further assessments will be treated as deteriorated in the analysis.

Death will be included as an event only if the death occurs within 2 visits of the last PRO assessment where the symptom change could be evaluated.

Subjects whose symptoms (as measured by QLQ-LC13) have not shown a clinically meaningful deterioration and who are alive at the time of the analysis will be censored at the time of their last PRO assessment where the symptom could be evaluated. Also, if symptoms deteriorate after 2 or more missed PRO assessment visits or the subject dies after 2 or more missed PRO assessment visits, the subject will be censored at the time of the last PRO assessment where the symptom could be evaluated. If a subject has no evaluable visits or does not have baseline data they will be censored at date of randomization.

The population for analysis of time to symptom deterioration will include a subset of the FAS population who have baseline scores ≤90.

3.3.2.2 Symptom improvement rate

The symptom improvement rate will be defined as the number (%) of subjects with 2 consecutive assessments at least 14 days apart that show a clinically meaningful improvement in that symptom from baseline. The denominator will consist of a subset of the FAS who have a baseline symptom score ≥10.

3.3.3 CCI

CCI

CCI

3.3.4 CCI

CCI

3.3.5 CCI

CCI

3.3.6 PRO Compliance

Summary measures of overall compliance and compliance over time will be derived for EORTC QLQ-C30 and QLQ-LC13, and for CCI respectively. These will be based upon:

- Received questionnaire = a questionnaire that has been received and has a completion date and at least one individual item completed.
- Expected questionnaire = a questionnaire that is expected to be completed at a scheduled assessment time, e.g. a questionnaire from a subject who has not withdrawn from the study at the scheduled assessment time but excluding subjects in countries with no available translation. For subjects that have progressed, the latest of progression and safety follow-up will be used to assess whether the subject is still under PRO follow-up at the specified assessment time. Date of study

discontinuation will be mapped to the nearest visit date to define the number of expected forms.

- Evaluable questionnaire = a questionnaire with a completion date and at least one subscale that is non-missing.

- Overall PRO compliance rate is defined as: Total number of evaluable questionnaires across all time points, divided by total number of questionnaires expected to be received across all time points multiplied by 100.

- Compliance over time will be calculated separately for each visit, including baseline, as the number of evaluable questionnaires at the time point (as defined above), divided by number of expected questionnaires multiplied by 100.

- Similarly, the evaluability rate over time will be calculated overall and separately for each visit, including baseline, as the number of evaluable questionnaires (per definition above), divided by the number of received questionnaires multiplied by 100.

The number of subjects with received, expected and evaluable questionnaires, as well as compliance rate and evaluability rate will be summarized at each scheduled assessment time point and overall.

3.4 Safety variables

Data from all cycles of treatment will be combined in the presentation of safety data.

3.4.1 Adverse events

Adverse events (AEs) (both in terms of Medical Dictionary for Regulatory Activities [MedDRA] preferred terms [PTs] and CTCAE grade) will be listed individually by subject.

Any AE occurring before the start of study treatment will be included in the data listings but will not be included in the summary tables of AEs. Any AE occurring within 90 days of the last dose in any treatment arm may be included in the AE summaries, but the majority of those summaries will omit those AEs observed after a subject has received subsequent anti-cancer therapy (including radiotherapy, with the exception of palliative radiotherapy) following discontinuation of study treatment (whichever occurs first). Any events in the period that occur after a subject has received further anti-cancer therapy (following discontinuation of study treatment) will be flagged in the data listings.

3.4.1.1 Adverse events of special interest

An adverse event of special interest (AESI) is one of the scientific and medical interest specific to understanding the study treatment and may require close monitoring and rapid communication by the investigator to the Sponsor. An AESI may be serious or non-serious.

The rapid reporting of AESIs allows ongoing surveillance of these events in order to characterize and understand them in association with the use of study treatment.

AESIs for durvalumab ± tremelimumab include, but are not limited to, events with a potential inflammatory or immune-mediated mechanism and which may require more frequent monitoring and/or interventions such as steroids, immunosuppressants, and/or hormone replacement therapy. These AESIs are being closely monitored in clinical studies with durvalumab monotherapy and combination therapy. An immune-mediated adverse event (imAE) is defined as an AESI that is associated with drug exposure and is consistent with an immune-mediated mechanism of action and where there is no clear alternate etiology. Serologic, immunologic, and histologic (biopsy) data, as appropriate, should be used to support an imAE diagnosis. Appropriate efforts should be made to rule out neoplastic, infectious, metabolic, toxin, or other etiologic causes of the imAE.

AESIs observed with durvalumab ± tremelimumab include, but are not limited to, the following:

- Diarrhea/colitis
- Pneumonitis/interstitial lung disease (ILD)
- Alanine aminotransferase (ALT)/aspartate aminotransferase (AST) increases/hepatitis/hepatotoxicity
- Neuropathy/neuromuscular toxicity (eg, Guillain-Barré, myasthenia gravis)
- Endocrinopathies (ie, events of hypophysitis, hypopituitarism, adrenal insufficiency, diabetes insipidus, hyper- and hypothyroidism, type I diabetes mellitus)
- Rash/dermatitis
- Nephritis/blood creatinine increases
- Pancreatitis (or laboratory results suggestive of pancreatitis [eg, increased serum lipase, increased serum amylase])
- Other inflammatory responses that are rare with a potential immune-mediated aetiology include, but are not limited to, myocarditis, pericarditis, uveitis, vasculitis, non-infectious meningitis, and non-infectious encephalitis

In addition, infusion-related reactions and hypersensitivity/anaphylactic reactions with a different underlying pharmacological etiology are also considered AESIs.

3.4.1.2 Other significant adverse events

During the evaluation of the AE data, an AstraZeneca medically qualified expert will review the list of AEs that were not reported as serious adverse events (SAEs) and AEs leading to discontinuation. Based on the expert's judgment, significant AEs of particular clinical importance may, after consultation with the Global Subject Safety Physician, be considered as other significant adverse events (OAEs) and reported as such in the CSR. A similar review of laboratory/vital signs/electrocardiogram (ECG) data will be performed for identification of OAEs. Examples of these are marked hematological and other laboratory abnormalities and certain events that lead to intervention (other than those already classified as serious) or significant additional treatment.

3.4.2 Other safety variables

3.4.2.1 Laboratory assessments

Laboratory data reported in this study has been collected in local laboratories, and local reference ranges have been used for the primary interpretation of laboratory data at the local laboratories. However, project reference ranges will be used throughout the study for reporting purposes. A list of laboratory variables to be included in the outputs is provided in CSP Tables 4 (clinical chemistry), 5 (hematology) and 6 (urinalysis).

Change from baseline in hematology and clinical chemistry variables will be calculated for each post-dose visit on treatment. Common toxicity criteria (CTC) grades will be defined at each visit according to the CTC grade criteria using local or project ranges as required, after conversion of lab result to corresponding preferred units. The following parameters have CTC grades defined for both high and low values: potassium, sodium, magnesium, glucose, calcium and corrected calcium. For these parameters high and low CTC grades will be calculated.

Corrected calcium will be derived during creation of the reporting database using the following formula:

Corrected calcium (mmol/L) = Total calcium (mmol/L) + ([40 – albumin (g/L)] * 0.02)

Creatinine clearance will be derived during creation of the reporting database according to the Cockroft-Gault formula:

Creatinine clearance (mL/min) = ([140 – age at randomization] * weight (kg) [* 0.85 if subject is female]) / (72 * serum creatinine (mg/dL))

Absolute values will be compared to the project reference ranges and classified as low (below the lower limit of reference range), normal (within reference range, upper and lower limit included) and high (above upper limit of reference range).

The maximum or minimum on-treatment value (depending on the direction of an adverse effect) will be defined for each laboratory parameter as the maximum (or minimum) post-dose value at any time on treatment (defined as between the start of study treatment and up to and including the earlier of 90 days following the date of last dose of study treatment or the date of initiation of the first subsequent anti-cancer therapy).

The denominator used in laboratory summaries will only include evaluable subjects, i.e. those who had sufficient data to have the possibility of an abnormality.

For example:

- If a CTCAE criterion involves a change from baseline, evaluable subjects would have both a pre-dose and at least 1 post-dose value recorded
- If a CTCAE criterion does not consider changes from baseline, evaluable subjects need only have 1 post-dose value recorded.

3.4.2.2 Electrocardiogram

All ECG data will be listed.

The QTcF will be derived during creation of the reporting database using the reported ECG values (RR and QT).

QTcF = QT/RR^ (1/3) where RR is in seconds.

3.4.2.3 Vital signs

Vital signs data collected up to and including 28 days (or weight collected at 28 days, and 2, 3 and 6 months) following the date of last dose of study treatment will be used for reporting. All vital signs data will be listed.

3.4.3 Pharmacokinetic and immunogenicity variables

3.4.3.1 Pharmacokinetic analysis

Individual concentrations of durvalumab and tremelimumab will be listed by visit. Summary statistics of durvalumab and tremelimumab concentrations will be calculated and tabulated by visit. Peak and trough concentrations will be determined as data allow. Individual concentrations of EP treatments will also be listed by time point. Samples below the lower limit of quantification will be treated as missing in the analyses.

3.4.3.2 Immunogenicity analysis

Immunogenicity results will be analyzed descriptively by summarizing the number and percentage of subjects who develop detectable anti-drug antibodies (ADAs) against durvalumab and tremelimumab. The immunogenicity titer and presence of neutralizing ADAs will be reported for samples confirmed positive for the presence of ADAs. CCI

CCI [redacted]

3.4.4 CCI [redacted]

CCI [redacted]

3.4.5 Treatment exposure

Exposure will be defined for each molecule as follows:

- Total (or intended) exposure = the earliest of (date of last dose of study drug +xx days, death date or DCO) – first dose date + 1day.

Where xx = 20 if the last dose occurs during chemotherapy, and 27 if the last dose occurs during the post-chemotherapy period, for Abraxane xx = 6, for Gemcitabine xx = 6 (if last infusion/dose was first dose of cycle) and xx = 13 days (if last infusion/dose was second dose of cycle).

Actual exposure (calculated for durvalumab and tremelimumab only)

- Actual exposure is defined as above, but excluding total duration of dose interruptions and cycle delays.

Calculation of duration of dose delays/interruptions (for actual exposure):

> Duration of dose delays/interruptions = Sum of positive values of [Date of the dose - Date of previous dose – (xx+3) days] where xx is 21 for the chemotherapy period, and 28 for the post-chemotherapy period.

Dose reductions are not permitted per Section 6.9.1 of the CSP for the immunotherapy agents (durvalumab, tremelimumab). The actual exposure calculation makes no adjustment for any dose reductions that may have occurred.

:

Exposure of each study treatment will also be calculated separately for the combination stage and maintenance stage, with the cut-off date between the two stages defined as follows:

Table 10 Cut-off date to define combination stage

Treatment Arm	SoC chemotherapy	Cut-off date
Durva + Treme + SoC (Treatment Arm 1)	Abraxane + Carboplatin, or Gemcitabine + Cisplatin, or Gemcitabine + Carboplatin	Maximum of the dates subject discontinues all SoC chemotherapies
	Pemetrexed + Carboplatin, or Pemetrexed + Cisplatin	Maximum of the dates subject discontinues SoC chemotherapy except Pemetrexed
Durva + SoC (Treatment Arm 2)	Abraxane + Carboplatin, or Gemcitabine + Cisplatin, or Gemcitabine + Carboplatin	Maximum of the dates subject discontinues all SoC chemotherapies
	Pemetrexed + Carboplatin, or Pemetrexed + Cisplatin	Maximum of the dates subject discontinues SoC chemotherapy except Pemetrexed
SoC (Treatment Arm 3)	Abraxane + Carboplatin, or Gemcitabine + Cisplatin, or Gemcitabine + Carboplatin	Maximum of the dates subject discontinues all SoC chemotherapies

Treatment Arm	SoC chemotherapy	Cut-off date
	Pemetrexed + Carboplatin, or Pemetrexed + Cisplatin	Maximum of the dates subject discontinues SoC chemotherapy except Pemetrexed

Note: the dates subject discontinues study treatment are taken from DOSDISC modules.

Exposure in the combination stage will be calculated using the above formula with last infusion/dose date and first infusion/dose date as the last infusion/dose date and the first infusion/dose date on or prior to the cut-off date in Table 10 .

Exposure in maintenance stage (only applicable for durvalumab/tremelumumab and pemetrexed treatments) will be calculated using the above formula with last infusion/dose date and first infusion/dose date as the last infusion/dose date and the first infusion/dose date after the cut-off date in Table 10.

Exposure will also be measured by the number of cycles received. For chemotherapy, a cycle corresponds to a period of 21 days (during chemotherapy)/ 28 days (post-chemotherapy). For immunotherapy, a cycle corresponds to one dose of treatment. If a cycle is prolonged due to toxicity, this should still be counted as one cycle. A cycle will be counted if treatment is started, even if the full dose is not delivered. The number of cycles will also be calculated separately for combination stage and maintenance stage based on the cut-off date in Table 10.

Subjects who permanently discontinue during a dose interruption or delay

If a decision is made to permanently discontinue study treatment in-between cycles or during a dose interruption or delay, then the date of last administration of study medication will be used in the programming.

3.4.6 Dose intensity

Relative dose intensity (RDI) is the percentage of the actual dose delivered relative to the intended dose through to treatment discontinuation. RDI will be defined for Durvalumab and Tremelimumab (RDI is not calculated for chemotherapy treatments) as follows:

- RDI = 100% * d/D, where d is the actual cumulative dose delivered up to the actual last day of dosing and D is the intended cumulative dose up to the or the actual last day of dosing. D is the total dose that would be delivered, if there were no modification to dose or schedule. When accounting for the calculation of intended cumulative dose 3 days should be added to the date of last dose to reflect the protocol allowed window for dosing.

When deriving actual dose administered the volume before and after infusion will also be considered.

3.5 China and Japan cohort

It is planned to randomize up to 180 subjects from China, including up to 135 subjects with PD-L1 TC <50%, before global enrollment completion. This is to ensure adequate Chinese subject participation to satisfy CFDA requirements to evaluate consistency in safety and efficacy in Chinese subjects. If China enrollment cannot complete prior to the end of global recruitment, recruitment across all sites will be closed except for those sites in mainland China, where recruitment of subjects will continue.

The China cohort consists of all subjects from sites in China and enrolled prior to the last subject last visit of the global cohort. Per CFDA guidance, in addition to the evaluation of the global cohort data for primary, secondary and safety objectives, evaluation of consistency in efficacy and safety in Chinese and Asian populations is required to facilitate the benefit-risk assessment for Chinese subjects.

Details of the China cohort and Asian population analyses, including the vendor to perform the analyses, will be specified in the China supplementary statistical analysis plan (SAP), which is to be finalized before the global cohort data locks for analysis.

Japan cohort consists of all subjects from Japan sites. In addition to the evaluation of the global cohort data for primary, secondary and safety objectives, evaluation of consistency in efficacy and safety in the Japan cohort is required to facilitate the benefit-risk assessment for Japanese subjects.

4. ANALYSIS METHODS

4.1 General principles

The formal statistical analysis will be performed to test the following main hypotheses:

- H0: No difference between durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone
- H1: Difference between durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone

The dual primary endpoints are PFS and OS (durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone) in the ITT population (with PFS using BICR assessments per RECIST 1.1).

The final (primary) PFS analysis for superiority will be performed when the following conditions have been met:

- Approximately 497 BICR PFS events from the global cohort have occurred across durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone treatment arms (75% maturity)

The final OS analysis for superiority will be performed when the following conditions have been met:

- Approximately 532 OS events have occurred across durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone treatment arms (80% maturity)

The following general principles will apply:

- Descriptive statistics will be used for all variables, as appropriate, and will be presented by treatment arm. Continuous variables will be summarized by the number of observations, mean, standard deviation, median, upper and lower quartiles (where appropriate), minimum, and maximum. For log-transformed data it is more appropriate to present geometric mean, coefficient of variation (CV), median, minimum and maximum. Categorical variables will be summarized by frequency counts and percentages for each category.

- Unless otherwise stated, percentages will be calculated out of the population total for the corresponding treatment arm.

- For continuous data, the mean and median will be rounded to 1 additional decimal place compared to the original data. The standard deviation will be rounded to 2 additional decimal places compared to the original data. Minimum and maximum will be displayed with the same accuracy as the original data.

- For categorical data, percentages will be rounded to 1 decimal place.

- SAS® version 9.2 or higher will be used for all analyses.

For safety variables, baseline will be the last value obtained prior to the first dose of study treatment. For laboratory data, any assessments made on Day 1 will be considered pre-dose. If two visits are equally eligible to assess subject status at baseline (e.g., screening and baseline assessments are both on the same date prior to first dose with no washout or other intervention in the screening period), the average can be taken as a baseline value. For non-numeric laboratory tests (i.e. for urinalysis parameters), where taking an average is not possible, the best value would be taken as baseline as this is the most conservative approach. In the scenario where there are two assessments on Day 1 and assessment time is recorded for only one of those records, that record would be selected as baseline.

In general, for efficacy and PRO endpoints the last observed measurement prior to randomisation will be considered the baseline measurement. However, if an evaluable

assessment is only available after randomisation but before the first dose of randomised treatment then this assessment will be used as baseline.

For assessments on the day of first dose where time is not captured, a nominal pre-dose indicator, if available, will serve as sufficient evidence that the assessment occurred prior to first dose. Assessments on the day of the first dose where neither time nor a nominal pre-dose indicator are captured will be considered prior to the first dose if such procedures are required by the protocol to be conducted before the first dose.

All data collected will be listed. Efficacy data will be summarized and analyzed by treatment arm based on the FAS, the PD-L1 TC <50% analysis set, the PD-L1 TC <25% analysis set, the PD-L1 TC <1% analysis set, bTMB20 high analysis set, bTMB16 high analysis set and bTMB12 high analysis set. PRO data will be summarized and analyzed by treatment arm based on the FAS. PK data will be summarized and analyzed based on the PK analysis set. Safety and treatment exposure data will be summarized using the safety analysis set. Study population and demography data will be summarized based upon the FAS (unless otherwise stated).

4.1.1 Common derivations

For safety data, study day will be calculated from the date of first dose of study treatment and will be used to show start/stop day of assessments and events.

If the date of the event is on or after the date of first dose then:

- Study day = (date of event – date of first dose) + 1.

If the date of the event is prior to the date of first dose then:

- Study day = (date of event – date of first dose).

In the situation where the event date is partial or missing, study day, and any corresponding durations will appear missing in the listings.

In all summaries change from baseline for quantitative measurements will be calculated as:

- Post-treatment value – value at baseline

The percentage change from baseline will be calculated as:

- (Post-baseline value – value at baseline) x 100 / value at baseline

4.1.2 Visit windowing

The following conventions will be applied when windowing visits:

The time windows should be exhaustive so that data recorded at any time point, including unscheduled visit data, has the potential to be summarized. Inclusion within the time window should be based on the actual date and not the target date of the visit.

The window for the visits following baseline will be constructed in such a way that the upper limit of the interval falls half way (midpoint) between the two visits (the lower limit of the first post-baseline visit will be Day 2, and thereafter half way between that visit and the previous visit). If an even number of days exists between two consecutive visits then the upper limit will be taken as the midpoint value minus 1 day. For example, visit windows for vital signs are constructed as in Table 11.

Table 11 10 Vital signs visit windows

Visit	Window	Target day
Week 1	2 - 14	8
Week 3	15 - 25	22
Week 4	26 - 35	29
Week 6	36 – 46	43
Week 7	47 – 56	50
Week 9	57 – 67	64
Week 10	68 – 77	71
Week 12	78 – 98	85
Week 16	99 – 126	113
Every 4 weeks	(Target day - 14) – (target day + 13)	Previous target day + 28

If there is more than one value per subject within a time window then the closest value to the scheduled visit (target) date should be summarized, or in the event that two values are equidistant from the target date, the earlier value should be summarized. If there are two values recorded on the same day and the parameter is CTCAE gradeable then the record with the highest toxicity grade should be used. Alternatively, if there are two records recorded on the same day and the toxicity grade is the same (or is not calculated for the parameter) then the average of the two records should be used. Listings will display all values (including unscheduled values) contributing to a time point for a subject, but will flag the windowed values wherever feasible. Note that in summaries of extreme values, or last value on treatment, all post-baseline values collected are used, including those collected at unscheduled visits regardless of whether or not the value is closest to the target date.

4.1.3 Missing values

Missing safety data will generally not be imputed. However, safety assessment values of the form of "< x" (i.e. below the lower limit of quantification) or "> x" (i.e. above the upper limit of quantification) will be imputed as "x" in the calculation of summary statistics but displayed as "< x" or "> x" in the listings. Additionally, AEs that have missing causality (after data querying) will be assumed to be related to study treatment.

Similarly, missing efficacy data will generally not be imputed, unless otherwise specified in the relevant sections below.

During the creation of the reporting database, partial AE or medication start dates will be imputed as follows:

- If only day is missing: impute day as the first day of the month (unless month and year are the same as month and year of first dose of study treatment, in which case day should be imputed as date of first dose);
- If day and month are missing: impute day and month as the first day of the year (unless year is the same as year of first dose of study treatment, in which case day and month should be imputed as date of first dose);
- If the start date is missing, then the analysis start date will not be imputed.

Partial AE or medication end dates will be imputed as follows:

- If only day is missing: impute day as the earlier of either the DCO or the last day of the month;
- If day and month are missing: impute day and month as the earlier of either the DCO or the last day of the year;
- If the end date is missing, then the analysis end date will not be imputed.

If either both the start and end date of an AE or medication are missing, or the start date of an AE or medication is missing, but the end date is complete or imputed and on or after the date of first dose, then the AE or medication is considered treatment emergent or concomitant.

If a patient is known to have died where only a partial death date is available then the date of death will be imputed as the latest of the last date known to be alive +1 from the database and the death date using the available information provided:

a. For Missing day only – using the 1st of the month
b. For Missing day and Month – using the 1st of January

If there is evidence of death but the date is entirely missing, it will be treated as missing, i.e. censored at the last known alive date.

4.2 Analysis methods

Results of all statistical analyses will be presented using appropriately sized confidence intervals (CIs) and 2-sided p-values, unless otherwise stated.

Table 12 details which endpoints are to be subjected to formal statistical analysis, together with pre-planned sensitivity analyses, making it clear which analysis is regarded as primary for that endpoint.

Table 12 Pre-planned statistical and sensitivity analyses to be conducted

Endpoints analyzed	Notes
Progression-free survival	Stratified log-rank tests for: • Dual primary analysis using BICR RECIST 1.1 assessments: – Durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone for the ITT population • Key secondary analysis using BICR RECIST 1.1 assessments: – Durvalumab + tremelimumab combination therapy + SoC chemotherapy and SoC chemotherapy alone for the ITT population • Other secondary analyses using BICR RECIST 1.1 assessments: – Durvalumab + tremelimumab combination therapy + SoC chemotherapy and SoC chemotherapy alone for PD-L1 TC <50% population (stratified only for disease stage and histology), PD-L1 TC <25% population (stratified only for disease stage and histology), PD-L1 TC <1% population (stratified only for disease stage and histology) and bTMB high population (>=20, >=16 and >=12) – Durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone for PD-L1 TC <50% population (stratified only for disease stage and histology), PD-L1 TC <25% population (stratified only for disease stage and histology), PD-L1 TC <1% population (stratified only for disease stage and histology) and bTMB high population (>=20, >=16 and >=12) • Sensitivity analyses using investigator assessments (RECIST 1.1) • CCI [redacted]

Endpoints analyzed	Notes
Overall survival	Stratified log-rank tests for: • Dual primary analysis – Durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone for the ITT population • Key secondary analysis – Durvalumab + tremelimumab combination therapy + SoC chemotherapy and SoC chemotherapy alone for the ITT population • Other secondary analyses – Durvalumab + tremelimumab combination therapy + SoC chemotherapy and SoC chemotherapy alone for PD-L1 <50% population (stratified only for disease stage and histology), PD-L1 TC <25% population (stratified only for disease stage and histology), PD-L1 TC <1% population (stratified only for disease stage and histology) and bTMB high population (>=20, >=16 and >=12) – Durvalumab monotherapy + SoC chemotherapy and SoC chemotherapy alone for PD-L1 <50% population (stratified only for disease stage and histology), PD-L1 TC <25% population (stratified only for disease stage and histology), PD-L1 TC <1% population (stratified only for disease stage and histology) and bTMB high population (>=20, >=16 and >=12)
Objective response rate	Logistic regression for: • Secondary analysis for the ITT, PD-L1 <50%, PD-L1 <25%, PD-L1 <1% and bTMB high population (>=20, >=16 and >=12) using BICR RECIST 1.1 assessments • Sensitivity analysis for the ITT, PD-L1 <50%, PD-L1 <25%, PD-L1 <1% and bTMB high population (>=20, >=16 and >=12) using investigator RECIST 1.1 assessments • CCI [redacted]
Duration of response	Kaplan-Meier estimates for: • Secondary analysis using BICR assessments (RECIST 1.1)
Proportion of subjects alive and progression free at 12 months	Kaplan Meier estimates of progression-free survival at 12 months
Time from randomization to second progression	Stratified log-rank test

Endpoints analyzed	Notes
Time to deterioration (EORTC QLQ-C30 and QLQ-LC13 endpoints)	Stratified log-rank test
Symptom improvement rates (EORTC QLQ-C30 and QLQ-LC13 endpoints)	Logistic regression

4.2.1 Multiple testing strategy

In order to strongly control the type I error at 5% (2-sided), a multiple testing procedure (MTP) with gatekeeping strategy will be used across the dual primary endpoints and the secondary endpoints included in MTP.

- The dual primary endpoints: PFS and OS (durvalumab monotherapy +SoC chemotherapy versus SoC chemotherapy alone) in the ITT population (with PFS using BICR assessments per RECIST 1.1).
- The key secondary endpoints: PFS and OS (durvalumab + tremelimumab combination therapy + SoC chemotherapy and SoC chemotherapy alone) in the ITT population (with PFS using BICR assessments per RECIST 1.1).
- The secondary endpoints: OS (durvalumab + tremelimumab combination therapy + SoC chemotherapy and SoC chemotherapy alone) in the bTMB20 high population, bTMB16 high population, bTMB12 high population

Hypotheses will be tested using a MTP with an alpha-exhaustive recycling strategy (Burman et al 2009). With this approach, hypotheses will be tested in a pre-defined order as outlined in Figure 3. According to alpha (test mass) splitting and alpha recycling, if the higher level hypothesis in the MTP is rejected for superiority, the next lower level hypothesis will then be tested. The test mass that becomes available after each rejected hypothesis is recycled to lower level hypotheses not yet rejected. This testing procedure stops when the entire test mass is allocated to non-rejected hypotheses. Implementation of this pre-defined ordered testing procedure, including recycling, will strongly control type I error at 5% (2-sided), among all the dual primary endpoints and the secondary endpoints included in MTP.

The overall 5% type 1 error will be first split between the dual primary endpoints of PFS and OS so an alpha level of 1% will be allocated to the primary PFS analysis and 4% will be allocated to the primary OS analysis. If PFS primary endpoint is significant, the 1% alpha level will be recycled to the key secondary PFS comparison of Arm 1 vs Arm 3. If OS primary endpoint is significant, the 4% alpha level will be recycled to the key secondary OS comparison of Arm 1 vs Arm 3. If PFS comparison of Arm 1 vs Arm 3 is significant, the 1% alpha level will be recycled to OS comparison of Arm 1 vs Arm 3. If OS comparison of Arm 1

vs Arm 3 is significant, the 4% alpha level will be recycled to PFS comparison of Arm 1 vs Arm 3.

If both PFS and OS comparisons of Arm 1 vs Arm 3 are significant, the available alpha level (e.g. 5%) will be recycled to OS comparison of Arm 1 vs Arm 3 in bTMB20 high population. If the comparison is significant, the alpha level (e.g. 5%) will be recycled to OS comparison of Arm 1 vs Arm 3 in bTMB16 high population. If the comparison is significant, the alpha level (e.g. 5%) will be recycled to OS comparison of Arm 1 vs Arm 3 in bTMB12 high population.

The primary and the key secondary PFS endpoints are tested at 2 time points, 1 interim analysis and 1 final analysis. The primary and the key secondary OS endpoints are tested at 4 time points: 3 interim analyses and 1 final analysis. The secondary endpoints of OS in the bTMB20 high population, bTMB16 high population, and bTMB12 high population are tested at 4 timepoints, with 3 interim analyses and 1 final analysis. The tests including the interim and the final analyses that are for the same comparison/endpoint (i.e., shown in 1 rectangle box in Figure 3) will be considered as 1 test family. As long as 1 test in the family can be rejected, the family is rejected. Thus, the assigned total alpha to the family will be recycled to next MTP level.

If the interim or final analyses indicate superiority in the tested hypothesis, then subsequent analyses of endpoints will be performed hierarchically in accordance with the MTP strategy. The alpha level allocated to the interim or final analyses will be determined by the Lan DeMets (Karrison et al 2016
Karrison, et al. 2016. “Versatile Tests for Comparing Survival Curves Based on Weighted Log-Rank Statistics.” Stata Journal 16 (3). StataCorp LP: 678–90

Lan and DeMets 1983) spending function that approximates an O’Brien Fleming approach, where the alpha level applied at the interim depends upon the proportion of information available at time of the analysis. A separate Lan DeMets (O’Brien Fleming) spending function will be used to determine the alpha level at the interim and final analyses for the next lower level of hypothesis testing under MTP based on the information proportion of that lower level of hypothesis.

If the interim results do not meet the criterion of stopping for superiority for a given hypothesis, then follow-up will continue until the next target number of events for that comparison has been observed, following which the hypothesis will be re-tested. If the hypothesis is then rejected, subsequent testing will continue hierarchically in accordance with the MTP strategy.

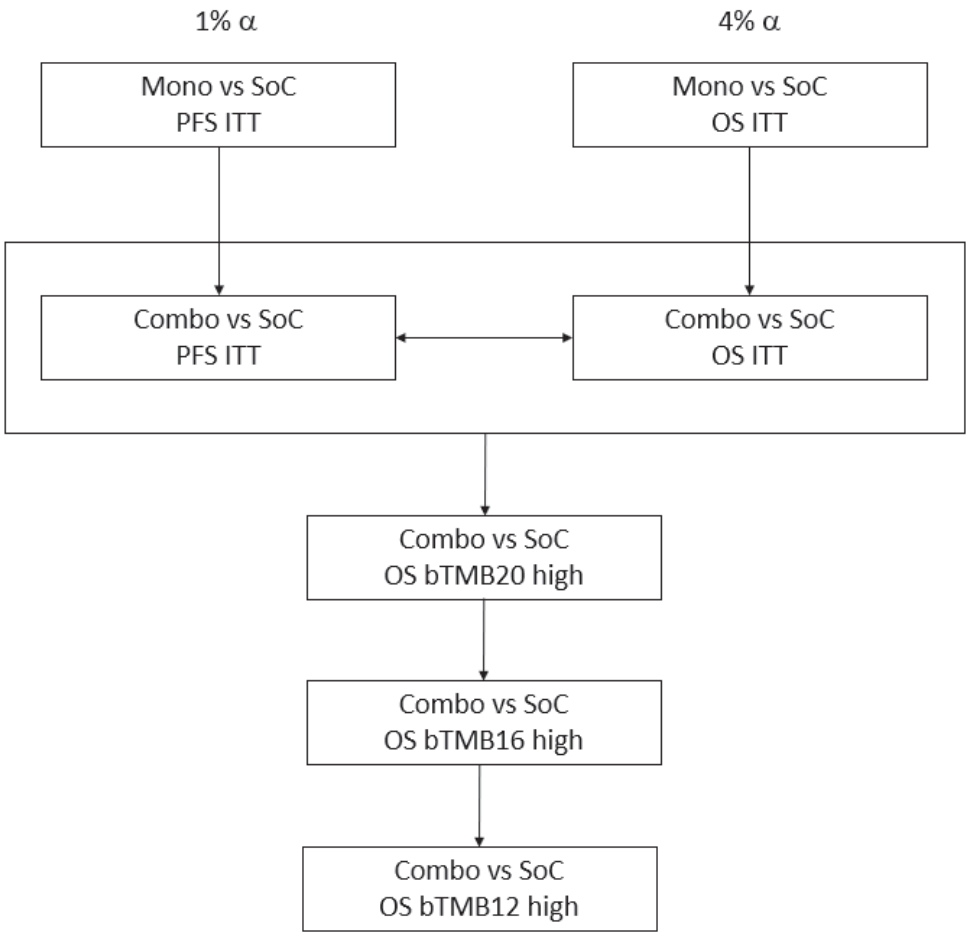
Assuming that 1% alpha level is available to a PFS hypothesis testing, if exactly 80% of the final target PFS events are available at the time of the interim, the 2-sided alpha level to be applied for the interim and final analyses of PFS would be 0.0034 and 0.009, respectively.

Assuming that 4% alpha level is available to a OS hypothesis testing, and if exactly 45%, 61% and 84% of the OS events required at the time of the final OS analysis are available at the times of the interim analyses, the 2-sided alpha level to be applied for the first interim, second interim, third interim and final analyses of OS would be 0.001, 0.0055, 0.0204 and 0.033 respectively. Refer Section 5.1 for details of interim analysis.

If the interim results do not meet the criterion of stopping for superiority for a given hypothesis, then follow-up will continue until the final target number of events for that comparison has been observed, following which the hypothesis will be re-tested. If the hypothesis is then rejected, subsequent testing will continue hierarchically. The above testing procedure will ensure strong control of the family-wise error rate (Glimm et al, 2010).

If statistical significance is achieved at one of the planned interim analyses, then that will be considered the final analysis for that endpoint. Further analyses for that endpoint may occur depending on the needs for long term follow-up with more mature data. If statistical significance is achieved at the interim analysis for PFS for the comparison of Arm 2 and 3, and statistical significance is not achieved at the first interim analysis for OS in ITT population for the comparison of Arm 2 and 3, the second interim analysis for OS in ITT population for the comparison of Arm 2 and 3 will occur when approximately 328 OS events have been observed across Arms 2 and 3 in the ITT population. Similar principle will apply for testing of subsequent endpoints in the multiple testing procedure.

Figure 3 Multiple testing procedures for controlling the type 1 error rate



Combo durvalumab + tremelimumab combination therapy + SoC chemotherapy; ITT Intent-to-treat; Mono durvalumab monotherapy + SoC chemotherapy; OS Overall survival; PFS Progression-free survival; SoC Standard of care; vs versus; bTMB20 high blood tumor mutational burden>=20 mut/Mb; bTMB16 high blood tumor mutational burden >=16 mut/Mb; bTMB12 high blood tumor mutational burden >=12 mut/Mb.

4.2.2 Analysis of the primary and secondary endpoints

4.2.2.1 Progression-free survival

The dual primary PFS analysis will be based on the BICR tumor assessments according to RECIST 1.1. The full analysis set will be used. The analysis will use a stratified log-rank test adjusting for PD-L1 tumor expression (PD-L1 ≥50% versus PD-L1 <50%), histology (squamous versus non-squamous), and disease stage (Stage IVA and Stage IVB) for generation of the p-value. The covariates in the statistical modelling will be based on the

values entered into interactive voice response system (IVRS) at randomization, even if it is subsequently discovered that these values were incorrect.

The hazard ratio (HR) and its CI will be estimated from a stratified Cox proportional hazards model (with ties = Efron and PD-L1 tumor expression (PD-L1≥ 50% versus PD-L1 <50%), histology (squamous versus non-squamous), and disease stage (Stage IVA and Stage IVB) included in the STRATA statement) and the CI calculated using a profile likelihood approach.

Key secondary PFS analysis will be performed using the same methodology as for the dual primary PFS analysis described above.

Kaplan-Meier plots of PFS will be presented by treatment arm and PD-L1 tumor status and TMB subgroup, where appropriate. Summaries of the number and percentage of subjects experiencing a PFS event and the type of event (RECIST 1.1 or death) will be provided along with median PFS for each treatment.

The assumption of proportionality will be assessed. Proportional hazards will be tested firstly by examining plots of complementary log-log (event times) versus log (time) and, if these raise concerns, by fitting a time dependent covariate (adding a treatment-by-time or treatment-by-ln[time] interaction term) to assess the extent to which this represents random variation. If a lack of proportionality is evident, the variation in treatment effect can be described by presenting piecewise HR calculated over distinct time-periods. In such circumstances, the HR can still be meaningfully interpreted as an average HR over time unless there is extensive crossing of the survival curves. If lack of proportionality is found this may be a result of a treatment-by-covariate interaction, which will be investigated. In addition, the Kaplan-Meier curve along with landmark analyses (e.g., one year PFS rate) will also help in understanding the treatment benefit.

The treatment status at progression of subjects at the time of analysis will be summarized. This will include the number (%) of subjects who were on treatment at the time of progression, the number (%) of subjects who discontinued study treatment prior to progression, the number (%) of subjects who have not progressed and were on treatment or discontinued treatment. This will also provide distribution of number of days prior to progression for the subjects who have discontinued treatment.

Sensitivity analyses

The following sensitivity analyses will be performed for the treatment comparisons of the dual primary and key secondary endpoints based on the FAS:

- A sensitivity analysis will be performed to assess possible evaluation-time bias that may be introduced if scans are not performed at the protocol-scheduled time points. The midpoint between the time of progression and the previous evaluable RECIST assessment (using the final date of the assessment) will be analyzed using a stratified log-rank test. Note that midpoint values resulting in non-integer values

should be rounded down. For subjects whose death was treated as PFS event, the date of death will be used to derive the PFS time used in the analysis. This approach has been shown to be robust even in highly asymmetric assessment schedules (Sun and Chen 2010). To support this analysis, the mean of subject-level average inter-assessment times will be tabulated for each treatment. This approach will use BICR tumor assessments.

- Attrition bias will be assessed by repeating the dual primary/key secondary PFS analysis except that the actual PFS event times, rather than the censored times, of subjects who progressed or died in the absence of progression immediately following two or more non-evaluable tumor assessments will be included. In addition, and within the same sensitivity analysis, subjects who take subsequent therapy (note that for this analysis radiotherapy is not considered a subsequent anti-cancer therapy) prior to their last evaluable RECIST assessment or progression or death will be censored at their last evaluable assessment prior to taking the subsequent therapy. This analysis will be supported by a Kaplan-Meier plot of the time to censoring using the PFS data from the dual primary/key secondary analysis and where the censoring indicator of the PFS analysis is reversed.

- Ascertainment bias will be assessed by analyzing the site investigator data. The stratified log-rank test will be repeated on the programmatically derived PFS using the site investigator data based upon RECIST 1.1. The HR and CI will be presented.

  If there is an important discrepancy between the dual primary/key secondary analysis using the BICR assessments and this sensitivity analysis using investigator assessments, the proportion of subjects with site but no central confirmation of progression will be summarized; such subjects have the potential to introduce bias in the central review due to informative censoring. An approach that imputes an event at the next visit in the central review analysis may help inform the most likely HR value (Fehrenbacher et al 2016, Fleischer et al 2011), but only if an important discrepancy exists.

  Disagreements between investigator and central reviews of RECIST 1.1 progression will be presented for each treatment group. The summary will include the early discrepancy rate which is the frequency of central review declared progressions before the investigator review as a proportion of all central review progressions and the late discrepancy rate which is the frequency of central review declared progressions after the investigator review as a proportion of all discrepancies.

- An additional sensitivity analysis will be performed with the covariates used in the statistical model derived from eCRF data rather than using the values from IVRS.

A forest plot illustrating the HR and 95% CI will be provided for each treatment comparison to compare the dual primary/key secondary analysis and corresponding sensitivity analyses of progression free survival.

No adjustment to the significance level for testing of the sensitivity analyses will be made since all these analyses will be considered supportive of the analysis of PFS.

Additional supportive summaries

In addition, the number of subjects prematurely censored will be summarized by treatment arm together with baseline prognostic factors of the prematurely censored subjects. A subject would be defined as prematurely censored if they had not progressed (or died in the absence of progression) and the latest scan prior to DCO was more than one scheduled tumor assessment interval plus 2 weeks (8 weeks if time from randomization to DCO for that subject is 12 weeks or less, and 10 weeks otherwise) prior to the DCO date.

Additionally, summary statistics will be given for the number of days from censoring to DCO and from last tumor assessment to DCO for all censored subjects.

A summary of the duration of follow-up will be summarized using median time from randomization to date of censoring (date last known to have not progressed) in censored (not progressed) subjects only, presented by treatment group.

Additionally, summary statistics for the number of weeks between the time of progression and the last evaluable RECIST assessment prior to progression will be presented for each treatment group.

Summaries of the number and percentage of subjects who miss two or more consecutive RECIST assessments will be presented for each treatment group.

All of the collected RECIST 1.1 data will be listed for all randomized subjects. In addition, a summary of new lesions (i.e. sites of new lesions) will be produced.

A summary table of first subsequent cancer therapies relative to progression by treatment arm will be provided.

Other secondary analyses

The dual primary/key secondary PFS analysis will be repeated for the following endpoints, using BICR assessments according to RECIST 1.1:

- Comparison of durvalumab + tremelimumab combination therapy + SoC chemotherapy with SoC chemotherapy alone and durvalumab monotherapy + SoC chemotherapy with SoC chemotherapy alone:

  - PFS in subjects with PD-L1 TC <50% (stratified only for disease stage and histology)
  - PFS in subjects with PD-L1 TC <25% (stratified only for disease stage and histology)
  - PFS in subjects with PD-L1 TC <1% (stratified only for disease stage and histology)
  - PFS in subjects with bTMB high (>=20, >=16 and >=12)

- Comparison of durvalumab + tremelimumab combination therapy + SoC chemotherapy with durvalumab monotherapy + SoC chemotherapy:
  - PFS in all subjects (full analysis set)
  - PFS in subjects with PD-L1 TC <50% (stratified only for disease stage and histology)
  - PFS in subjects with PD-L1 TC <25% (stratified only for disease stage and histology)
  - PFS in subjects with PD-L1 TC <1% (stratified only for disease stage and histology)
  - PFS in subjects with bTMB high (>=20, >=16 and >=12)

Subgroup analyses

Subgroup analyses will be conducted comparing PFS (per RECIST 1.1 using BICR assessments) between the treatments concerned in the dual primary and key secondary endpoints in the following subgroups of the FAS (but not limited to):

- Sex (male versus female)
- Age at randomization (<65 versus ≥65 years of age)
  - This will be determined from the date of birth (BRTHDAT in the DM module) and date of randomization (IERNDDAT in the IE1 module) on the eCRF at screening. Subjects with a partial date of birth (i.e. for those countries where year of birth only is given) will have an assumed date of birth of 1st Jan [given year]). Subjects with a missing age value will be included using the mean age (overall FAS) and categorized accordingly.
- PD-L1 status (PD-L1 TC ≥50% versus TC <50%)

- Histology (squamous versus non-squamous)
- Chemotherapy (abraxane doublet versus pemetrexed doublet versus gemcitabine doublet)
- Smoking (current smoker, former smoker, never smoker)
- Race (Asian versus non-Asian)
  - This will be determined from the response to "Race" (DEM module) on the eCRF at screening.
- PD-L1 using cut points of 1% and 25% tumor expression (<1% versus ≥1%, and <25% versus ≥25%)
- bTMB (>=20 vs <20)
- bTMB (>=16 vs <16)
- bTMB (>=12 vs <12)
- Performance status (normal activity [PSTAT=0] versus restricted activity [PSTAT=1])
  - This will be determined from the response to "Performance status" (PSTAT module) on the eCRF at screening. Subjects with a missing performance status will be included in the 'restricted activity' category.
- Brain metastasis (Yes [DISSITES=1 and LOCADMET=3] versus No)
  - This will be derived from the responses to the "Site of local/metastatic disease" and "Metastatic/locally advanced" (DISEXT module) on the eCRF at screening.
- Disease stage (stage IVA [AJCC_STG=41] versus stage IVB [AJCC_STG=42])
  - This will be determined from the response to "Stage/AJCC stage" (PATHGEN module) on the eCRF at screening.

The subgroup analyses will be based on values recorded on the eCRF (except the chemotherapy which will be based on the interactive voice response system (IVRS)).

Other baseline variables may also be assessed if there is clinical justification or an imbalance is observed between the treatment arms. The purpose of the subgroup analyses is to assess the consistency of treatment effect across expected prognostic and/or predictive factors.

No adjustment to the significance level for testing of the subgroup analyses will be made since all these analyses will be considered supportive of the analysis of PFS.

For each subgroup level of a factor, the HR and 95% CI will be calculated from an un-stratified Cox proportional hazards model with treatment as only covariate. The Cox models will be fitted using SAS® PROC PHREG with the Efron method to control for ties, using the by statement to obtain a HR and 95% CI for each subgroup level separately.

These HRs and associated two-sided 95% CIs will be summarized and presented on a forest plot for each treatment comparison, along with the results of the overall dual primary/key secondary analysis.

Unless there is a marked difference between the results of the statistical analyses of the PFS from the BICR data and that of the site investigator tumor data, these subgroup analyses will only be performed on the PFS endpoint using the BICR data.

If there are too few events available for a meaningful analysis of a particular subgroup (it is not considered appropriate to present analyses where there are less than 20 events in a subgroup), the relationship between that subgroup and PFS will not be formally analyzed. In this case, only descriptive summaries will be provided.

Effect of covariates on HR estimate

Cox proportional hazards modelling will be employed to assess the effect of pre-specified covariates on the HR estimate for the treatment comparisons of the dual primary and key secondary endpoints based on the FAS. Before embarking on more detailed modelling, an initial model will be constructed containing treatment and the stratification factors alone to ensure that any output from the Cox modelling is likely to be consistent with the results of the stratified log-rank test.

The results from the initial model and the model containing additional covariates will be presented.

Additional covariates for this model will be age at randomization, sex, smoking status and race.

The model will include the effect regardless of whether the inclusion of effect significantly improves the fit of the model providing there is enough data to make them meaningful.

Consistency of treatment effect between subgroups

Interactions between treatment and stratification factor will be tested to rule out any qualitative interaction using the approach of Gail and Simon (Gail and Simon 1985). This test will be performed separately for the treatment comparisons of the dual primary and key secondary endpoints based on the FAS.

Exploratory analyses

CCI

4.2.2.2 Overall survival

OS will be analyzed using stratified log-rank tests, using the same methodology as described for the PFS endpoints. The effects of durvalumab + tremelimumab combination therapy + SoC chemotherapy versus SoC chemotherapy alone and durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone will be estimated by the HR together with its corresponding CI and p-value, in the full analysis set, in subjects with PD-L1 TC <50% (stratified only for disease stage and histology), in subjects with PD-L1 TC <25% (stratified only for disease stage and histology), in subjects with PD-L1 TC <1% (stratified only for disease stage and histology) and in subjects with bTMB high (>=20, >=16 and >=12). In addition, the effect of durvalumab + tremelimumab combination therapy + SoC chemotherapy versus durvalumab monotherapy + SoC chemotherapy will also be estimated in the same subject populations. Kaplan-Meier plots will be presented by treatment arm and PD-L1 tumor status and bTMB subgroup, where appropriate. Summaries of the number and percentage of subjects who have died, those still in survival follow-up, those lost to follow-up, and those who have withdrawn consent will be provided along with the median OS for each treatment. Survival rate at for example 12, 18, 24 and 36 months, based on Kaplan-Meier method, will also be presented along with its 95% CI.

The boundaries (i.e., adjusted alpha levels) for the treatment comparison at the interim and final analyses for OS will be derived based upon the exact number of OS events using the Lan and DeMets approach that approximates the O'Brien Fleming spending function.

The assumption of proportionality will be assessed in the same way as for PFS, whereby proportional hazards will be tested firstly by examining plots of complementary log-log (event times) versus log (time) and, if these raise concerns, by fitting a time dependent covariate (adding a treatment-by-time or treatment-by-ln(time) interaction term) to assess the extent to which this represents random variation. For overall survival the Grambsch-Themeau non-proportionality test may also be used to check non-proportional hazards (NPH) violation (Grambsch and Therneau 1994).

A three-component stratified max-combo test will be used as a sensitivity analysis with the same stratification factors as the primary analysis. The max-combo test is an adaptive procedure, optimizing test statistics as the maximum of the log rank test ($FH^{0,0}$) and the selected Fleming-Harrington (FH) weighted log-rank tests (Fleming and Harrington 1991) ($FH^{0,1}$ and $FH^{1,1}$) i.e. Zmax = max{ $FH^{0,0}$, $FH^{0,1}$ and $FH^{1,1}$}, with multiplicity adjustment based on asymptotic multivariate distribution (Karrison et al. 2016). The Fleming-Harrington tests of $FH^{0,1}$ and $FH^{1,1}$ assign less weights to early events and are more powerful in the scenario of delayed effect, and the log-rank test is optimal in the scenario of proportional

hazards (PH) (Schoenfeld 1981). Under proportional hazards (PH), the power loss from Zmax is minimal (Lin et al. 2020). As a result, the Zmax statistic is a robust test with consideration of possible delayed effect scenarios (NPH) and proportional hazards scenario and is recommended by the Cross-Pharma Non-proportional Hazard Working Group when NPH are expected (Lin et al. 2020).

The variation in treatment effect can be described by presenting piecewise HR calculated over distinct time-periods, for example 0-6m, 6-12m, 12-18m etc. Under NPH, the HR from the primary analysis can still be meaningfully interpreted as an average HR over time unless there is extensive crossing of the survival curves. In addition, the survival rates at clinically meaningful time points are particularly useful in interpretation of treatment benefit. The Restricted Mean Survival Time (RMST) of an area-under-the-curve approach (Kaplan-Meier method), may also be analysed over a specified time period, for example up until the maximum event time in either treatment arm, with standard error, for each treatment group, along with the estimate of difference in means between treatment groups, confidence interval and p-value. In addition, pseudovalues approach (Andersen et al. 2004) and Royston-Parmar model (Royston and Parmar 2011, 2013) may also be used while controlling stratification factors.

If lack of proportionality is found, this may be a result of delayed effect in immuno-oncology (IO) agents and/or a treatment-by-covariate interaction, which will be investigated.

CCI

Sensitivity analysis and additional supportive summaries

A sensitivity analysis for OS will examine the censoring patterns to rule out attrition bias with regards to the treatment comparisons of the dual primary and key secondary endpoints, achieved by a Kaplan-Meier plot of time to censoring where the censoring indicator of OS is reversed.

A sensitivity analysis may be conducted to assess for the potential impact of COVID-19 deaths on OS. This will be assessed by repeating the OS analysis except that any patient who

had a death with primary/secondary cause as COVID-19 infection, or a COVID-19 infection reported as a fatal AE, will be censored at their COVID infection death date.

The number of subjects prematurely censored will be summarized by treatment arm. A subject would be defined as prematurely censored if their survival status was not defined at the DCO.

In addition, duration of follow-up will be summarized using medians and other descriptive statistics:

- In censored subjects who are alive at DCO only: Time from randomization to date of censoring (date last known to be alive) by treatment arm.
- In all subjects: Time from randomization to the date of death (i.e. overall survival) or to the date of censoring (date last known to be alive) for censored subjects regardless of treatment arm.

Subgroup analyses and forest plots will be generated comparing OS between treatments of the dual primary and key secondary endpoints in the same way as previously specified for PFS.

No adjustment to the significance level for testing will be made since all these subgroup analyses will be considered supportive of the analysis of OS.

The effect of covariates upon the HR estimate and the consistency of treatment effect between subgroups will be analyzed for OS with regards to the treatment comparisons of the dual primary and key secondary endpoints, using the same methods as those described for PFS.

4.2.2.3 Objective response rate

The ORR will be based on the BICR tumor assessments according to RECIST 1.1. The ORR will be compared between durvalumab + tremelimumab combination therapy + SoC chemotherapy versus SoC chemotherapy alone and durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone using logistic regression models adjusting for the same factors as the PFS endpoints. The results of the analysis will be presented in terms of an odds ratio (an odds ratio greater than 1 will favor the experimental arms) together with its associated profile likelihood 95% CI (e.g. using the option 'LRCI' in SAS procedure GENMOD) and p-value (based on twice the change in log-likelihood resulting from the addition of a treatment factor to the model). This analysis will be performed in the FAS, in subjects with PD-L1 TC <50% (adjusting only for disease stage and histology), in subjects with PD-L1 TC <25% (adjusting only for disease stage and histology), in subjects with PD-L1 TC <1% (adjusting only for disease stage and histology) and in subjects with bTMB high (>=20, >=16 and >=12). In addition, the effect of durvalumab + tremelimumab combination therapy + SoC chemotherapy versus durvalumab monotherapy + SoC chemotherapy will also be estimated in the same subject populations.

If there are not enough responses for a meaningful analysis using logistic regression then a Fisher's exact test using mid p-values will be presented.

The mid-p-value modification of the Fisher's exact test amounts to subtracting half of the probability of the observed table from Fisher's p-value.

Fisher's exact test mid p-value = 2-sided p-value – (table probability / 2)

This analysis of ORR in all subject populations will be repeated using the results of the programmatically derived ORR using the site investigator tumor data based upon RECIST 1.1 as a sensitivity analysis to confirm the results of the primary analysis derived from the eCRFs.

CCI

Summaries will be produced that present the number and percentage of subjects with a tumor response (CR/PR), based upon the number of subjects with measurable disease at baseline per BICR/investigator as appropriate. Overall visit response data will be listed for all subjects (ie, the FAS). For each treatment arm, BoR will be summarized by n (%) for each category (CR, PR, SD, PD, and NE), in the FAS, in subjects with PD-L1 TC <50%, in subjects with PD-L1 TC <25%, in subjects with PD-L1 TC <1% and in subjects with bTMB high (>=20). No formal statistical analyses are planned for BoR.

4.2.2.4 Duration of response

Descriptive data will be provided for the DoR in responding subjects by treatment arm, including the associated Kaplan-Meier curves (without any formal comparison of treatment arms or p-value attached). This analysis will be performed in the FAS, in subjects with PD-L1 TC <50%, in subjects with PD-L1 TC <25%, in subjects with PD-L1 TC <1% and in subjects with bTMB high (>=20, >=16 and >=12).

4.2.2.5 Proportion of subjects alive and progression free at 12 months

The APF12 (where 12 months equates to study day 366) will be summarized (using the Kaplan-Meier curve) and presented by treatment arm.

This analysis may be performed in the FAS, in subjects with PD-L1 TC <50%, in subjects with PD-L1 TC <25%, in subjects with PD-L1 TC <1% and in subjects with bTMB20 high population.

4.2.2.6 Time from randomization to second progression

Time from randomization to second progression or death (PFS2) will be analyzed using stratified log-rank tests, using the same methodology as described for the PFS endpoints. The HR for the treatment effect together with its 95% CI will be presented. Medians and Kaplan-Meier plots will be presented to support the analysis. The PFS2 rate at 12, 18, 24 and 36 months, based on Kaplan-Meier method, will also be presented along with its 95% CI.

In addition, a Kaplan-Meier plot of the time to censoring, similar to that of the sensitivity analysis of the PFS endpoints, will be produced where the censoring indicator of PFS2 is reversed.

The number and percentage of subjects experiencing a PFS2 event and the type of progression (symptomatic progression, objective radiological progression or other) will also be summarized by treatment arm. PFS2 will be summarized by treatment arm.

This analysis will be performed in the FAS, in subjects with PD-L1 TC <50%, in subjects with PD-L1 TC <25%, in subjects with PD-L1 TC <1% and in subjects with bTMB20 high population.

4.2.2.7 Change in TL tumor size

The absolute values and percentage change in TL tumor size from baseline will be summarized using descriptive statistics and presented at each time point and by randomized treatment group. The best change in TL tumor size from baseline (where best change in TL size is the maximum reduction from baseline or the minimum increase from baseline in the absence of a reduction) and the proportion of subjects with any reduction and with a reduction of more than 10% will also be presented by randomized treatment group.

Tumor size may also be presented graphically using waterfall plots for each treatment group, to present each subject's best percentage change in TL tumor size as a separate bar, with the bars ordered from the largest increase to the largest decrease. Reference lines at the +20% and –30% change in TL tumor size level will be added to the plots, which correspond with the definitions of progression and 'partial' response respectively. The scale in these plots will be fixed to be from -100 to 100 to avoid presenting extreme values. Values that are capped as a result of this restriction to the scale are marked with '#'. Values are ordered in descending order with the imputations due to death appearing first followed by a gap followed by all other subjects.

The number and percentage of subjects with scaled TLs and new lesions will also be presented.

The above outputs will be programmed for the BICR data based upon RECIST assessments for subjects in the FAS.

4.2.2.8 Patient-reported outcomes

For PRO symptoms and HRQoL endpoints, the overall type I error (5% [2-sided]) will be controlled across the 5 primary PRO measures of cough, dyspnea, and pain in chest as

assessed by the EORTC QLQ-LC13 and fatigue and appetite loss as assessed by the EORTC QLQ-C30 using the Bonferroni-Holm procedure (

Grambsch and Therneau 1994
Grambsch, PM. and Therneau TM. 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81:515–26, https://doi.org/10.1093/biomet/81.3.515.

Holm 1979).

The physical functioning and overall health status domains of the EORTC QLQ-C30 are furthermore pre-specified endpoints of interest.

All PRO analyses will be based on the FAS (ITT population), unless otherwise stated.

EORTC QLQ-C30

Time to deterioration will be analyzed using stratified log-rank tests, using the same methodology as described for the dual primary and key secondary PFS endpoints.

The effect of durvalumab + tremelimumab combination therapy + SoC chemotherapy versus SoC chemotherapy alone and durvalumab monotherapy + SoC chemotherapy versus SoC chemotherapy alone will be estimated by the HR together with its corresponding CI and p-value. Kaplan-Meier plots will be presented by treatment arm. Summaries of the number and percentage of subjects who have an event as well as who were censored will be provided along with the medians for each treatment. A forest plot by scale/item will also be produced for each treatment comparison.

Symptom improvement rate for each of the 3 symptom scales and the 5 individual symptom items and QoL/function improvement rate for each of the 5 function scales (physical, role, emotional, cognitive, and social) and global health status/QoL will be analyzed using logistic regression, similar to the analysis of ORR. A forest plot of the odds ratios by scale/item will also be produced for each treatment comparison.

Summaries of the original and the change from baseline values of each symptom scale/item, the global HRQoL score, and each functional domain will be reported by visit for each treatment arm. Summaries of the number and percentage of subjects in each response category at each visit for each ordinal item will also be produced for each treatment arm. In addition, compliance with the QLQ-C30 will be presented for each visit and overall.

EORTC QLQ-LC13

Time to deterioration will be analyzed using stratified log-rank tests, using the same methodology as described for the dual primary and key secondary PFS endpoints.

The effect of experimental arms versus SoC arm will be estimated by the HR together with its corresponding CI and p-value. Kaplan-Meier plots will be presented by treatment arm.

Summaries of the number and percentage of subjects who have an event as well as who were censored will be provided along with the medians for each treatment. A forest plot by scale/item will also be produced for each treatment comparison.

Symptom improvement rate for each of the 6 individual-symptom items will be analyzed using logistic regression, similar to the analysis of ORR. A forest plot by scale/item will also be produced for each treatment comparison.

Summaries of the original and the change from baseline values of each symptom (dyspnea, cough, hemoptysis, chest pain, arm/shoulder pain, and/or other pain) and each treatment-related symptom (sore mouth, dysphagia, peripheral neuropathy, and alopecia) will be reported by visit for each treatment arm. Summaries of the number and percentage of subjects in each response category at each visit for each ordinal-symptom item will also be produced for each treatment arm. In addition, compliance with the QLQ-LC13 will be presented for each visit and overall.

CCI

4.2.3 CCI

CCI

CCI

4.2.4 Safety

Safety and tolerability data will be presented by treatment arm and unless otherwise specified, using the safety population. No formal statistical analyses will be performed on the safety variables.

Data from all cycles of treatment will be combined in the presentation of safety data. AEs (both in terms of MedDRA PTs and CTCAE grade) will be listed individually by subject. The number of subjects experiencing each AE will be summarized by treatment arm and CTCAE grade. Additionally, data presentations of the rate of AEs per person-years at risk may be produced.

Other safety data will be assessed in terms of laboratory variables (including clinical chemistry and hematology), vital signs, and ECGs. At the end of the study, appropriate summaries of all safety data will be produced, as defined in the following sections. Additional safety summaries may need to be produced to aid interpretation of the safety data.

4.2.4.1 Adverse events

All AEs, both in terms of current MedDRA PT and CTCAE grade, will be summarized descriptively by count (n) and percentage (%) for each treatment group. The latest MedDRA dictionary version will be used for coding. The majority of the AE summaries, unless stated otherwise, will be based on treatment emergent AEs (TEAEs). TEAEs are defined as AEs with an onset date on or after the date of first dose or pre-treatment AEs that increase in severity on or after the date of first dose and up to and including the earlier of 90 days following the date of last dose of study treatment or the date of initiation of the first subsequent anti-cancer therapy (including radiotherapy, with the exception of palliative radiotherapy) following discontinuation of study treatment (whichever occurs first). Excluding AEs after initiation of subsequent therapy will more accurately depict AEs attributable to study treatment only, as AEs observed more than 90 days following the date of last dose of study treatment are likely to be attributable to subsequent therapy. Any pre-treatment AEs (i.e. AEs starting before the date of first dose of study treatment) that do not increase in severity after the first dose will be included in the AE listings, but will not be included in the summary tables (unless otherwise stated).

To assess the longer term toxicity profile, an AE summary (by system organ class [SOC], PT and maximum reported CTCAE grade) will also be produced containing AEs starting or increasing in severity up to and including 90 days following the date of last dose of study treatment, but without taking subsequent systemic anti-cancer therapy into account.

A selection of AE summaries (by SOC and PT) may also be produced containing AEs that started or increased in severity after the initiation of the first subsequent anti-cancer therapy following discontinuation of study treatment up to and including 90 days following the date of

last dose of study treatment (i.e. summarizing those AEs experienced by subjects taking subsequent therapy during the AE collection follow-up window post discontinuation of study treatment). These outputs will only be produced if the number of AEs observed warrant the inclusion of such outputs for interpretational purposes. Any events that started prior to dosing or that started more than 90 days following the date of last dose of study treatment will be presented in a separate summary (by SOC and PT).

An overall summary of AEs in the below categories will be presented at subject level.

The following summaries present subject incidence (frequencies and percentages), counting each subject only once within each SOC and PT:

- All AEs
- All AEs causally related to study treatment (as determined by the reporting investigator)
- AEs with CTCAE grade 3 or 4
- AEs causally related to study treatment with CTCAE grade 3 or 4
- AEs with outcome of death
- AEs with outcome of death, causally related to study treatment (as determined by the reporting investigator)
- All SAEs
- All SAEs, causally related to study treatment (as determined by the reporting investigator)
- AEs leading to discontinuation of study treatment
- AEs leading to discontinuation of study treatment, causally related to study treatment (as determined by the reporting investigator)
- AEs leading to hospitalization
- AEs leading to dose delay/interruption
- AEs leading to dose reduction of chemotherapy treatment
- Infusion reaction AEs (as determined by the reporting investigator)

Summaries of other significant AEs may be produced. For example a summary and listing of COVID-19 infections may be produced.

In addition, truncated AE tables of most common AEs by PT, and of most common AEs of CTCAE grade 3 or 4 by SOC and PT will be produced, including those events that occurred in at least 5% of subjects in any treatment group. This cut-off may be modified after review of the data. When applying the cut-off (i.e. x%), the raw percentage, without prior rounding applied, should be compared to the cut-off (i.e. an AE with frequency of 4.9% will not appear if the cut-off is 5%).

Each AE event rate (per 100 patient years) will also be summarized by PT within each SOC for all AEs. For each PT, the event rate is defined as the number of subjects with that AE divided by the total number of years at risk. The denominator is calculated as the number of days from first dose of study treatment to the date of last dose of study treatment, summed over all subjects, divided by 365.25.

There will also be summaries of causally related AEs with CTCAE grade 3 or 4 by PT and causally related SAEs by PT at the subject level and a summary of AEs by SOC, PT and maximum reported CTCAE grade presenting also subject incidence.

Fluctuations observed in CTCAE grades during study will be listed for those AEs which are CTCAE ≥ 3.

In addition, all reported AEs will be listed along with the date of onset, date of resolution (if AE is resolved) and investigator's assessment of severity and relationship to study treatment. Listings will present all AEs, SAEs, AEs leading to discontinuation of study treatment and AEs requiring treatment with steroids, endocrine therapy or other immunosuppressive agents.

Deaths

Two summaries of all deaths will be provided presenting subject incidence by treatment arm in the FAS for the following categories:

- Total number of deaths (regardless of the date of death)
- Death related to disease under investigation only, as determined by investigator (regardless of the date of death)
- Death related to disease under investigation and an AE with outcome of death
    - AE onset prior to subsequent therapy, defined as an AE with onset date (or pre-treatment AEs that increase in severity) on or after the date of first dose and up to and including the earlier of 90 days following the date of last dose of study treatment or the date of initiation of the first subsequent therapy

  - AE onset after start of subsequent therapy, defined as an AE with onset date more than 90 days following the date of last dose of study treatment or after the date of initiation of the first subsequent therapy if this is earlier

- AE with outcome of death only
  - AE onset prior to subsequent therapy, defined as above
  - AE onset after start of subsequent therapy, defined as above
- Death not due to either disease progression or an AE with a start date whilst on treatment or within the safety follow-up period (i.e. 90 days following the date of last dose of study treatment) (*)
- Subjects with unknown reason for death
- Other deaths

This summary will be repeated for all deaths up until 90 days following last dose of study treatment. Hence the category marked (*) will only appear in the first summary.

Listings of all deaths, and all AEs with outcome of death will also be produced.

Adverse events of special interest

The list of PTs used to identify AESIs (as defined in Section 3.4.1.1) will be finalized prior to database lock (DBL) and documented in the Study Master File. Grouped summary tables of certain MedDRA PTs will be produced and may also show the individual PTs which constitute each AESI grouping. Groupings will be based on PTs provided by the medical team prior to DBL, and a listing of the PTs in each grouping will be provided.

Summaries of the above-mentioned grouped AE categories will include number (%) of subjects who have:

- Any AESI, presented by grouped term and PT, including the number of events
- Any AESI, presented by grouped term and maximum reported CTCAE grade
- Any AESI, presented by grouped term and outcome
- Any AESI causally related to study treatment (as determined by the reporting investigator)
- Any AESI leading to discontinuation of study treatment
- Any AESI requiring concomitant treatment

- Any AESI requiring concomitant medication use of steroids
- Any AESI requiring concomitant medication use of high dose steroids
- Any AESI requiring concomitant endocrine treatment
- Any AESI requiring treatment with other immunosuppressive agents

A summary of total duration (days) of AESIs, including the median and range of total duration, will be provided for events which have an end date.

Additionally, time to onset of first AESI will be presented using summary statistics, as well as AESIs by grouped term, PT and first onset of AESI.

Summary of long term tolerability

To assess long term tolerability, provided that there are a sufficient number of subjects with events to warrant it, prevalence plots, life table plots and cumulative incidence plots may be presented for each of the AESI grouped terms and any other events considered important after review of the safety data, provided there are ≥ 10 events.

A prevalence plot provides information on the extent to which the events may be an ongoing burden to subjects. The prevalence at time t after first dose of study treatment is calculated as the number of subjects experiencing the event divided by the number of subjects receiving study treatment or in safety follow-up at time t; generally, t is categorized by each day after dosing. The prevalence is plotted over time split by treatment arm. Multiple occurrences of the same event are considered for each subject, but a subject is only counted in the numerator whilst they are experiencing one of the occurrences of the event. These plots will only be produced for AESIs that have ≥10 events.

A life table plot can be used to describe the time to onset of the event and specifically when subjects are at most risk of first experiencing the event. The hazard, or in other words, the probability of having an AE in a specified time period (e.g. 0-1 months, 1-3 months, 3-6 months, etc.) given that the subject reaches that time period without having an event is plotted for each time period split by treatment. These plots will only be produced for AESIs that have ≥10 events.

A cumulative incidence plot is a plot of the raw cumulative incidence and cumulative incidence function over time with the treatment groups presented on separate plots. The raw cumulative incidence is the actual probability that a subject will have experienced their first occurrence of the event by a given time point. The cumulative incidence function estimates the cumulative incidence if the data cut-off had not been imposed and all subjects had completed safety follow-up (Pintilie 2006). These plots will only be produced for AESIs that have ≥10 events.

4.2.4.2 Laboratory assessments

Data obtained between the start of study treatment and up to and including the earlier of 90 days following the date of last dose of study treatment or the date of initiation of the first subsequent anti-cancer therapy, defined as on treatment, will be used for the reporting of laboratory assessments. Excluding laboratory data after initiation of subsequent therapy will more accurately depict laboratory toxicities attributable to study treatment only, as toxicities observed more than 90 days following the date of last dose of study treatment are likely to be attributable to subsequent therapy.

To assess the longer term toxicity profile, summaries of laboratory data may also be produced containing data collected up to and including 90 days following the date of last dose of study treatment, but without taking subsequent anti-cancer therapy into account.

A selection of summaries of laboratory data may also be produced containing data obtained after the initiation of the first subsequent anti-cancer therapy following discontinuation of study treatment up to and including 90 days following the date of last dose of study treatment (i.e. summarizing laboratory data for subjects taking subsequent therapy during the safety collection follow-up window post discontinuation of study treatment). These outputs will only be produced if the number of laboratory toxicities observed warrant the inclusion of such outputs for interpretational purposes. Any data collected after 90 days following the date of last dose of study treatment will not be summarized.

Data summaries will be provided in international system of units (SI).

The following summaries will be provided for laboratory data:

- Absolute value and change from baseline at each scheduled assessment time (for quantitative measurements)
- Shift tables in hematology and clinical chemistry parameters from baseline to maximum CTCAE grade on treatment, indicating hyper- and hypo-directionality of change for electrolytes:
  - Hematology: hemoglobin, leukocytes, lymphocytes (absolute count), neutrophils (absolute count), platelets
  - Clinical chemistry: ALT, AST, alkaline phosphatase (ALP), amylase, bicarbonate, creatinine, gamma-glutamyl transferase (GGT), lipase, total bilirubin, total protein, magnesium (high/low), sodium (high/low), potassium (high/low), calcium (high/low), corrected calcium (high/low), glucose (high/low)
- Incidence of CTCAE grade changes, presenting subjects who had a shift of at least two grades from baseline, and subjects who changed to grade 3 or 4 since baseline

- Shift table in urinalysis parameters (bilirubin, blood, glucose, ketones, protein) from baseline to maximum value on treatment
- Shift table in creatinine clearance from baseline to minimum value on treatment
- Shift tables in thyroid test variables (TSH, T3, T4) from baseline to maximum and minimum value on treatment
- Scatter plots (shift plots) of baseline to maximum value / minimum value (as appropriate) on treatment, including or excluding outliers, may be produced for certain parameters if warranted after data review

Liver enzyme elevations and potential Hy's law

To capture all elevated liver enzymes and potential Hy's law cases, the following summaries will be produced:

- Incidence of elevated ALT, AST, and total bilirubin during the on treatment period in the following categories:
  - ALT, AST, or either ALT or AST: ≥ 3x – ≤ 5x, > 5x – ≤8x, > 8x – ≤ 10x, >10x – ≤ 20x and >20x the upper limit of normal (ULN)
  - Total bilirubin: ≥2x – ≤3x, >3x – ≤5x, >5x ULN
  - Potential Hy's law: (ALT or AST ≥3x ULN) and total bilirubin ≥2x ULN, where the onset date of the ALT or AST elevation should be prior to or on the date of the total Bilirubin elevation
- Scatter plots of ALT and AST (horizontal axis) versus total bilirubin (vertical axis) by treatment group with reference lines at 3x ULN for ALT and AST and 2x ULN for total bilirubin
- Narratives will be provided in the CSR for subjects who have ALT ≥ 3x ULN plus total bilirubin ≥ 2x ULN or AST ≥ 3x ULN plus total bilirubin ≥ 2x ULN at any assessment

Individual subject data, presenting all assessments for subjects with potential Hy's law, will be listed.

4.2.4.3 Electrocardiogram

Overall evaluation of ECG is collected as clinically indicated throughout the treatment period, in terms of normal or abnormal, and the relevance of the abnormality is termed as "clinically significant" or "not clinically significant". ECG data will be listed only.

4.2.4.4 Vital signs

Box plots for absolute values and change from baseline by week may be presented for certain vital signs parameters if warranted after data review.

Vital signs (systolic blood pressure, diastolic blood pressure, pulse rate, temperature, respiratory rate and weight) will be summarized over time in terms of absolute values and change from baseline at each scheduled measurement by actual treatment group.

Vital signs data will be listed.

4.2.5 Pharmacokinetic and immunogenicity data

Pharmacokinetic data

PK concentration data will be listed for each subject and each dosing day, and a summary will be provided for all evaluable subjects in the PK analysis population.

Immunogenicity analysis

Immunogenicity results will be listed by subject, and a summary of the number and percentage of subjects who develop detectable anti-durvalumab and anti-tremelimumab antibodies will be provided. The immunogenicity titer and neutralizing ADA data will be listed for samples confirmed positive for the presence of anti-durvalumab and anti-tremelimumab antibodies.

The effects of immunogenicity on PK, pharmacodynamics, efficacy, and safety will be evaluated if data allow.

CCI

4.2.6 CCI

CCI

4.2.7 Demographics and other baseline characteristics

Disposition, demographic data and other baseline characteristics will be presented for subjects in the FAS (unless otherwise specified below). No statistical testing will be carried out for demographic or other baseline characteristics. The following data will be summarized:

- Subject disposition

- Important protocol deviations
- Inclusion in analysis sets
- Demographics: age at randomization (years), age at randomization group (years) (≥18 - <50, ≥50 - <65, ≥65 - <75, ≥75), sex, race, ethnic group and country
- Subject characteristics at baseline: height (cm), weight (kg), weight group (kg) (<70, 70 to 90, >90), body mass index (BMI) (kg/m^2), BMI group (kg/m^2) (<25, 25 – 30, >30)
- Stratification factors recorded at randomization by IVRS and on eCRF
- Subject recruitment by region, country and center
- Previous disease-related treatment modalities
- Number of regimens of previous chemotherapy at baseline
- Medical (past and current) and surgical history
- Disease characteristics at screening: World Health Organization (WHO)/Eastern Cooperative Oncology Group（ECOG) performance status, American Joint Committee on Cancer (AJCC) staging, KRAS mutation status, best response to previous cancer therapy, overall disease classification
  - Partial dates of first diagnosis will be imputed as the first day of the month (in case of missing day only) or the first day of the year (in case of missing day and month) in the derivation of time from diagnosis to randomization.
- Disease characteristics at diagnosis: AJCC staging and histology type
- Extent of disease at baseline
- TNM classification at diagnosis
- Disallowed and allowed concomitant medication use (coded using the latest WHO drug dictionary version)
- Smoking status (never, current, former)
- Post-discontinuation disease-related anti-cancer therapy
- WHO/ECOG performance status over time

In addition, listings will present disposition for discontinued subjects, IPDs and baseline weight, height and BMI for the FAS, as well as subjects excluded from analyses.

4.2.8 Treatment exposure

The following summaries related to study treatment will be produced for the safety analysis set by actual treatment received:

- Total exposure of immunotherapies, chemotherapies and overall for each treatment group, and for chemotherapies separately for combination stage and maintenance stage
- Actual exposure of immunotherapies for each treatment group
- Total number of immunotherapy and chemotherapy cycles and infusions received for each treatment group, and for chemotherapies separately for combination stage and maintenance stage
- Reasons for dose delays or infusion interruptions of immunotherapies and reasons for dose delays/interruptions or dose reductions of chemotherapies. Dose interruptions will be based on investigator initiated dosing decisions.
- Number of subjects with dose delays or infusion interruptions of immunotherapies and dose delays/interruptions or dose reductions of chemotherapies.
- Cumulative exposure of durvalumab, tremelimumab and chemotherapies over time
- RDI for the immunotherapies
- Exposure of durvalumab, tremelimumab and chemotherapies over time will be plotted

For subjects ongoing on study treatment at the time of the PFS and OS analyses, the DCO date will be used to calculate exposure.

Administration of study treatment will also be listed for all subjects in the safety analysis set. Total treatment duration (in weeks) of any treatment and cumulative dose of each treatment will be included in the disposition listing for all discontinued subjects.

4.2.9 Coronavirus Disease 2019 (COVID-19)

A summary of COVID-19 study disruptions will be created presenting number of subjects with impacted visits, study drug or concomitant medications. A listing of all patients affected by a COVID-19 related study disruption, by unique subject number identifier, will be generated along with the description of how the individual's participation was altered.

Additional analyses might be conducted to investigate the impact of COVID-19 on study endpoints.

5. INTERIM ANALYSES

5.1 Interim efficacy analysis

Interim safety monitoring will be conducted by an IDMC. Interim analysis for efficacy will be performed by IDMC as described below:

One interim analysis of PFS will be performed when approximately 80% of the target PFS events have occurred across Arms 2 and 3. Three interim analyses of OS will be performed; the first at the time of the interim PFS analysis (approximately 45% of the target OS events in Arms 2 and 3), the second at the time of the primary PFS analysis (approximately 61% of the target OS events in Arms 2 and 3) and the third when approximately 84% of the target OS events have occurred in Arms 2 and 3. The interim analyses will be performed for the analyses specified in MTP, refer to Section 4.2.1. It is expected that global recruitment will have completed prior to the results of the interim analyses being available.

The Lan DeMets spending function that approximates an O'Brien Fleming approach will be used to account for multiplicity introduced by including the one interim analysis for superiority (Karrison et al 2016
Karrison, et al. 2016. "Versatile Tests for Comparing Survival Curves Based on Weighted Log-Rank Statistics." Stata Journal 16 (3). StataCorp LP: 678–90

Lan and DeMets 1983). The boundaries for the treatment comparison will be derived based upon the exact number of events at the time of analyses.

The alpha level that is spent at the interim and final analyses for the PFS analyses will be calculated using the Lan DeMets spending function separately. If exactly 80% of the target events are available at the time of each interim comparison (368 out of 465 events occurred in Arm1 and 3, ITT population; 397 out of 497 events occurred in Arm 2 and 3, ITT population), with overall 2-sided alpha levels of 0.01, the 2-sided alpha level to be applied for the interim and final analyses of each PFS would be 0.0034 and 0.009, respectively.

The alpha level that is spent at the interim and final analyses for the OS analyses will be calculated using the Lan DeMets spending function separately. If exactly 45% of the target events are available at the time of the first interim comparison (243 out of 532 events occurred in Arm 1 and 3 or Arm 2 and 3, ITT population and if exactly 61% of the target events are available at the time of the second interim comparison (328 out of 532 events occurred in Arm 1 and 3 or Arm 2 and 3, ITT population), and if exactly 84% of the target events are available at the time of the third interim comparison (449 out of 532 events occurred in Arm 1 and 3 or Arm 2 and 3, ITT population), with overall 2-sided alpha levels of 0.04, the 2-sided alpha

level to be applied for the first interim, second interim, third interim and final analyses of each OS analysis would be 0.001, 0.0055, 0.0204 and 0.033 respectively.

The alpha level that is spent at the interim and final analyses for OS in Arms 1 and 3 in bTMB20 high population will be calculated using the Lan DeMets spending function that approximates an O'Brien Fleming approach separately for each treatment comparison and/or population. The actual prevalence of bTMB20 high population may be different to that assumed for the sample size calculation (N=70 per arm). Therefore, the sample size may shift accordingly. However, in that case, the final analysis will be planned at approximately 72% maturity and the alpha levels for the 3 interims and the final analysis will be calculated based on the assumption that approximately 72% of the actual randomized bTMB20 high patients in Arms 1 and 3 will have a death event at the final OS analysis. It is planned that analyses will have similar proportion of events as in the ITT population and be conducted at approximately 45%, 61% and 84% of the number of target events.

The alpha level that is spent at the interim and final analyses for OS in bTMB16 high population and bTMB12 high population will be calculated using the Lan DeMets spending function that approximates an O'Brien Fleming approach separately. It will be assumed the maturity in Arms 1 and 3 will be similar as for the bTMB20 high population (72%) at the final analyses and the alpha allocation at the interim analyses will be computed based on this assumption. Interim analyses will be conducted at the same time as the OS ITT population.

If the interim analyses indicate superiority, subsequent analyses of other endpoints will be performed in accordance with the hierarchical multiple testing strategy described in Section 4.2.1.

5.2 Independent Data Monitoring Committee

This study will use an external Independent Data Monitoring Committee (IDMC) to assess ongoing safety analyses as well as the interim efficacy analysis. The initial safety review will take place when the first 30 subjects (10 in each arm) have completed the first cycle of treatment. A second review will add an additional 30 subjects (10 in each arm) who have completed the first cycle of treatment, making a total of 60 subjects. At the time of the second review, it is expected that the initial 30 subjects would have had at least 6 weeks of follow-up, with some subjects having longer follow-up. The IDMC will meet at least every 6 months thereafter. An additional safety review for Japanese subjects will take place when the first 3 subjects in each treatment arm in Japan have completed the first cycle of treatment. An additional safety review for Chinese subjects will take place when the first 10 subjects in each treatment arm in China have completed the first cycle of treatment. Following each meeting, the IDMC will report to the sponsor and may recommend changes in the conduct of the study.

This committee will be composed of therapeutic area experts and biostatisticians, who are not employed by AstraZeneca and are free from conflict of interest.

Following the reviews, the IDMC will recommend whether the study should continue unchanged, be stopped, or be modified in any way. Once the IDMC has reached a recommendation, a report will be provided to AstraZeneca. The report will include the recommendation and any potential protocol amendments and will not contain any unblinding information.

The final decision to modify or stop the study will sit with the sponsor. The sponsor or IDMC may call additional meetings if at any time there is concern about the safety of the study.

Full details of the IDMC procedures and processes can be found in the IDMC Charter. The safety of all AstraZeneca/MedImmune clinical studies is closely monitored on an ongoing basis by AstraZeneca/MedImmune representatives in consultation with the Subject Safety Department. Issues identified will be addressed; this could involve, for instance, amendments to the CSP and letters to investigators.

6. CHANGES OF ANALYSIS FROM PROTOCOL

The following changes of analysis from protocol are based on CSP v4.0, dated 25 September 2018:

- The analysis of expected duration of response (EDoR) was not a required analysis, so not included for DoR endpoints in the SAP. This is consistent with other durvalumab studies.

- The analysis of comparison of APF12 between treatment arms is removed to be consistent with other durvalumab studies.

7. REFERENCES

Andersen et al 2004
Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Anal. 2004 Dec;10(4):335-50.

CCI

Breslow 1974
Breslow, NE. Covariance Analysis of Censored Survival Data. Biometrics. 1974; 30:89–99

Burman et al 2009
Burman CF, Sonesson C, Guilbaud O. A recycling framework for the construction of Bonferroni-based multiple tests. Stat Med 2009;28:739-61.

Fayers et al 2001
Fayers PM, Aaronson NK, Bjordal K, Groenvold M, Curran D, Bottomley A; EORTC Quality of Life Group. The EORTC QLQ-C30 Scoring Manual. 3rd ed. Brussels: European Organisation for Research and Treatment of Cancer: 2001.

Fehrenbacher et al 2016
Fehrenbacher L, Spira A, Ballinger M, Kowanetz M, Vansteenkiste J, Mazieres J, et al. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, Phase 2, randomised, controlled trial. Lancet 2016; published online 09 March 2016.

Fleischer et al 2011
Fleischer F, Gaschler-Markefski B, Bluhmki E. How is retrospective independent review influenced by investigator-introduced informative censoring: a quantitative approach. Stat Med 2011;30(29):3373-86

Fleming and Harrington 1991
Fleming TR, Harrington DP. Counting Processes and Survival Analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley and Sons Inc. 1991;New York.

Gail and Simon 1985
Gail M, Simon R. Tests for qualitative interactions between treatment effects and patient subsets. Biometrics 1985;41(2):361–72.

Glimm et al 2010
Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. Stat Med 2010;29:219–228.

Grambsch and Therneau 1994
Grambsch, PM. and Therneau TM. 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81:515–26, https://doi.org/10.1093/biomet/81.3.515.

Holm 1979
Holm S. A simple sequentially rejective multiple test procedure. Scand J Statistics 1979;6:65-70.

Karrison et al 2016
Karrison, et al. 2016. "Versatile Tests for Comparing Survival Curves Based on Weighted Log-Rank Statistics." Stata Journal 16 (3). StataCorp LP: 678–90

Lan and DeMets 1983
Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983; 70 (3):659-663

Osoba et al 1998
Osoba D, Rodrigues G, Myles J, Zee B, Pater J. Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 1998;16(1):139-44.

Pintilie 2006
Competing risks: A practical perspective. Wiley, 2006.

Royston and Parmar 2011
Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med. 2011 Aug 30;30(19):2409-21.

Royston and Parmar 2013
Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013 Dec 7;13:152.

Sun and Chen 2010
Sun X, Chen C. Comparison of Finkelstein's Method with the conventional approach for interval-censored data analysis. Stat Biopharm Res 2010;2(1):97-108.

SIGNATURE PAGE

This is a representation of an electronic record that was signed electronically and this page is the manifestation of the electronic signature

Document Name: d419mc00004-sap-ed-5		
Document Title:	Statistical Analysis Plan Edition 5	
Document ID:	Doc ID-004027785	
Version Label:	4.0 CURRENT LATEST APPROVED	
Server Date **(dd-MMM-yyyy HH:mm 'UTC'Z)**	**Signed by**	**Meaning of Signature**
08-Mar-2021 15:15 UTC	PPD	Content Approval
08-Mar-2021 13:00 UTC		Author Approval

Notes: (1) Document details as stored in ANGEL, an AstraZeneca document management system.