Statistical Analysis Plan

Study Code	D419QC00001
Edition Number	4
Date	13Oct2020

A Phase III Randomized, Multicenter, Open-Label, Comparative Study to Determine the Efficacy of Durvalumab or Durvalumab and Tremelimumab in Combination with Platinum-Based Chemotherapy for the First-Line Treatment in Patients with Extensive Disease Small-Cell Lung Cancer (SCLC) (CASPIAN)

Statistical Analysis Plan
Study Code D419QC00001
Edition Number 4
Date 13Oct2020

A Phase III Randomized, Multicenter, Open-Label, Comparative Study to Determine the Efficacy of Durvalumab or Durvalumab and Tremelimumab in Combination with Platinum-Based Chemotherapy for the First-Line Treatment in Patients with Extensive Disease Small-Cell Lung Cancer (SCLC) (CASPIAN)

Study Statistician

PPD (AstraZeneca) Date

A Phase III Randomized, Multicenter, Open-Label, Comparative Study to Determine the Efficacy of Durvalumab or Durvalumab and Tremelimumab in Combination with Platinum-Based Chemotherapy for the First-Line Treatment in Patients with Extensive Disease Small-Cell Lung Cancer (SCLC) (CASPIAN)

Global Product Statistician

PPD (AstraZeneca) Date

TABLE OF CONTENTS PAGE

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation or special term	Explanation
ADA	Anti-drug antibody
AE	Adverse event
APF12	Proportion of patients alive and progression free at 12 months from first dose
Baseline	Refers to the most recent assessment of any variable prior to dosing with study treatment/randomization (as appropriate)
BLQ	Below limit of quantification
BOR	Best objective response
CI	Confidence Interval
CR	Complete Response
CrCl	Creatinine Clearance
CRF/eCRF	Case Report Form (electronic)
CSR	Clinical Study Report
CTC/CTCAE	Common Terminology Criteria for Adverse Event (National Institutes of Health, National Cancer Institute)
CV	Coefficient of variation
DCO	Data Cut-off
DLL3	Delta-like 3 protein
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
EORTC QLQ-C30	European Organization for Research and Treatment of Cancer 30-item core quality of life questionnaire
EORTC QLQ-LC13	European Organization for Research and Treatment of Cancer 13-item lung cancer-specific quality of life questionnaire. Module used as a supplement to EORTC QLQ-C30
EP	Etoposide and platinum-based chemotherapy
EQ-5D	EuroQoL 5-dimension utility index
EQ-5D-3L	EuroQoL 5-dimension, 3-level health state utility index
EQ-5D-5L	EuroQoL 5-dimension, 5-level health state utility index
FA	Final analysis
FAS	Full analysis set
HR	Hazard ratio

Abbreviation or special term	Explanation
HRQoL	Health-related Quality of Life
IA	Interim analysis
IDMC	Independent data monitoring committee
imAE	Immune mediated adverse event
IP	Investigational Product
ITT	Intention to Treat
IVRS	Interactive Voice Response System
MedDRA	Medical Dictionary for Regulatory Activities
MMRM	Mixed Model for Repeated Measures
MRI	Magnetic Resonance Imaging
MTP	Multiple testing procedure
nAb	Neutralizing antibody
NCI	National Cancer Institute
NE	Not evaluable
NSCLC	Non-small cell lung cancer
NTL	Non-target lesions
ORR	Objective Response Rate
OS	Overall Survival
PD	Progressive disease
PD-L1	Programmed death ligand 1
PFS	Progression-free survival
PFS2	Time from randomization to second progression
PK	Pharmacokinetic
PNS	Paraneoplastic syndrome
PR	Partial Response
PRO	Patient Reported Outcome
QoL	Quality of Life
QLQ-LC13	Quality of Life Lung Cancer Module; 13 items self-administered questionnaire from the EORTC for lung cancer
QTcF	QT interval (corrected for heart rate using Fredericia's correction)
irRECIST	Immune-related Response Evaluation Criteria in Solid Tumors
RECIST 1.1	Response Evaluation Criteria In Solid Tumors, Version 1.1

Abbreviation or special term	Explanation
REML	Restricted Maximum Likelihood
SAE	Serious adverse event
SD	Stable disease
TTD	Time to deterioration
TL	Target lesions
TMB	Tumor mutational burden
ULN	Upper limit of normal
WHO	World Health Organization

AMENDMENT HISTORY

Date	Brief description of change
V1.0 CCI	NA – first version
V2.0 CCI	Updated to reflect CSP v3, including changes to primary, secondary and exploratory objectives. Updated to reflect CSP v4, including further changes to primary and secondary objectives, updated MTP and removed BICR. Clarified subsequent anticancer therapy throughout document. Clarified 90 day safety follow-up throughout document, as 90 days following last dose, not following discontinuation. Applied consistent terminology for stratification factor "Planned platinum therapy in Cycle 1". Abbreviations added. Section 1.1 study objectives changed. Section 1.2 detail added to study design. Section 1.3 sample size details updated. Section 2.1 immunogenicity data added to Table 1. Section 2.2 changes to important deviation categories. Section 2.3 added detail about China and Japan analysis plan. Section 3.1 removed text about confirmed responses. Removed section previously numbered 3.1.4 about BICR and irRECIST. Section 3.2 removed "co-primary" terminology. Moved section previously numbered 3.2.1.1 to 3.2.2.1, clarified 2 missed visit rule, corrected example schedule change calculation, removed text on confirmed response & irRECIST. Section 3.2.1.1 removed 7 day window for survival calls. Section 3.2.2.2 added derivation of ORR using investigator assessments, removed visit window for change in tumor size. Removed section previously numbered 3.2.3.1 investigator ORR as not an exploratory endpoint (moved to earlier section). Section 3.3.1 added derivations for EORTC scales, Listed the 5 key symptoms, simplified Table 5. Section 3.3.1.1 minor clarifications. Removed sections previously numbered 3.3.1.2 & 3.3.1.3 (symptom/function improvement rate). Section 3.3.4 added detail on analysis visit window. Section 3.4 clarified 'on treatment' period, and specified FAS for listing & summary of deaths. Section 3.4.1 added OAE section, updated AESI section in line with revised CSP, and added imAE section.

Section 3.4.2 added +1 day to exposure definition, rearranged to clarify logic and add adjustment for etoposide. Added +3 days in duration of delays/interruptions, and sum only positive delays.

Section 3.4.3 RDI to be calculated only for durvalumab and tremelimumab, and not derived separately for the different treatment periods.

Section 3.4.4 Cockcroft-Gault formula added for CrCl, which can be both entered and derived. Added lymphocytes to list of parameters with bi-directional changes.

Added new Section 3.5 Biomarker variables.

Section 3.6 updated PK and immunogenicity analysis.

Section 4.1 revised in line with changes to primary objectives. Specified decimal places for efficacy outputs.

Section 4.1.1 Allow pre-dose scan to be used for RECIST in absence of a pre-randomization scan.

Added numbered section 4.1.3 for imputation rules, including clarifications on causal relationship to durvalumab, and imputation of partial death dates. Removed requirement for "overall" summary.

Section 4.2 table 6 updated in line with changed objectives. Removed formal treatment group comparison for APF6, APF12 and OS18.

Section 4.2.1 and figure 3 updated in line with changed objectives and planned interim analyses. Detail added on alpha recycling. Clarified multiple testing for ePRO endpoints, and changed from Holm to Bonferroni adjustment.

Section 4.2.2.1 clarified primary analysis as separate models for each treatment comparison, added adjusted confidence intervals (x2). Clarified and extended subgroup analysis, adding AJCC Stage and geographic region, changed 'ethnicity' to 'race'. Moved text previously in section 5.1 about "other baseline variables" into this section. Added section describing interaction test for stratification variable.

Moved section previously numbered 4.2.2.2 to 4.2.3.1, removed adjusted confidence intervals. Added K-M plots for sensitivity analyses. Added separate sensitivity analysis for subsequent anticancer therapy. Removed section on disagreements between investigator and central reviews. Removed exploratory irRECIST section.

Added new main section heading 4.2.3 for "Secondary endpoints".

Section 4.2.3.2 clarified that analysis uses a subset of the FAS.

Section 4.2.3.3 & 4.2.3.4 removed formal treatment group comparison for APF6, APF12 and OS18.

Added section 4.2.3.5 describing PK data presentation.

Added section 4.2.3.6 describing immunogenicity data presentation and moved some content from section 3.6.2.

Section 4.2.3.7 clarified, removed symptom improvement rate. Reduced MMRM analysis to the 5 key symptoms only.

Section 4.2.3.9 special data handling section added to describe issue with ePRO data at site PPD.

Date	Brief description of change
	Added new main section heading 4.2.4 for "Exploratory endpoints", including subsection for endpoints to be reported outside the main CSR. Section 4.2.4.1 removed Kaplan Meier plots. Section 4.2.5 added summary of AEs before first dose or after 90 days following last dose of study treatment. Removed some AE summaries. Defined 'most common' cutoff as 2% per treatment group. Simplified denominator for event rates. Changed summary of deaths to include all deaths not just up to 90 days after discontinuation. Added imAE section. Potential Hy's Law summaries and thyrotoxicity only include data from the 'on treatment' period. Added reference to laboratory parameters listed in current CSP. Removed box plots & scatter plots for labs and vital signs. Removed time to subsequent therapy from discontinuation of study treatment. Section 4.2.6 added pathology at diagnosis summary. Changed reference to drug dictionary. Section 4.2.7 added detail in exposure summaries, added dose interruption summaries, corrected text about identifying delays. Section 4.2.8 specified a separate summary of radiotherapy received after discontinuation of treatment. Section 5 updated with details of new planned interim analyses. Section 6 changes removed where no longer discrepant from current protocol. Added rationale for new changes. Section 7 removed various unused references. Appendix A added clarification note. Added Appendix B containing details of alpha spending procedure. Added Appendix C containing details of ePRO data to be excluded.
V3.0 CCI	Abbreviations added. Section 2.1 removed "initially" from Safety analysis set definition. Clarified that analysis sets include patients randomized prior to the end of global recruitment. Clarified PK analysis set includes patients with post-dose *PK* data, and removed review of deviations in relation to PK. Section 3.2.2.1 corrected week 11 to week 12. Section 3.2.3.1 removed 2 missed visit rule for PFS2. Added "other" reason. Section 3.3.1 clarified that definitions for QLQ-LC13 items include side effects. Section 3.4.1 added onset time to definition of "treatment emergent". Moved OAEs to end of section. Section 3.4.3 RDI to use the last day of dosing of the respective treatment. Added new Section 3.4.7 defining concomitant medication. Section 3.6.1 updated handling of BLQ values in PK summaries. Section 3.7 clarified handling of missing discharge dates regarding DCO. Section 4.1.3 added imputation rules for end dates.

Date	Brief description of change
	Section 4.2.1 added detail on alpha spending function for PFS. Section 4.2.2.1 added description of summary of duration of follow up and prematurely censored, and demography for prematurely censored patients for OS. Specified SAS code for applying log-rank test. Section 4.2.3.1 added summary of days between RECIST assessments. Section 4.2.3.3 and 4.2.3.4 updated derivation to use 1 month = 30.4375 days. Section 4.2.3.7 clarified 75% missing data criterion. Section 4.2.8 therapy on same day as discontinuation counts as subsequent therapy. Section 5.1 added detail on alpha spending function for PFS. Section 6 added rationale for adding alpha spending adjustment for PFS. Appendix B added detail on alpha spending function for PFS, and adjusted significance levels rounded down.
V4.0 CCI	Added Section 5.3 containing details of planned long-term safety follow-up analysis.

1. STUDY DETAILS

1.1 Study Objectives

All objectives will be evaluated for all patients, unless otherwise indicated.

1.1.1 Primary objectives

Primary objective:	Outcome measures:
To assess the efficacy of durvalumab + tremelimumab + EP treatment compared with EP and the efficacy of durvalumab + EP treatment compared with EP in terms of OS	OS

EP Etoposide and platinum-based chemotherapy; OS Overall survival.

1.1.2 Secondary objectives

Secondary objectives:	Outcome measures:
To further assess the efficacy of durvalumab + tremelimumab + EP treatment compared with EP and the efficacy of durvalumab + EP treatment compared with EP in terms of PFS, ORR, APF6 (PFS rate at 6 months), APF12 (PFS rate at 12 months), and OS18 (OS rate at 18 months)	PFS, ORR, APF6 and APF12 using site Investigator assessments according to RECIST 1.1 OS18
To assess the efficacy of durvalumab + tremelimumab + EP treatment compared with durvalumab + EP in terms of PFS and OS	PFS using site Investigator assessments according to RECIST 1.1 OS
To assess the PK of durvalumab and durvalumab + tremelimumab	Concentration of durvalumab and tremelimumab in blood and non-compartmental PK parameters, such as peak concentration and trough (as data allow; sparse sampling)
To investigate the immunogenicity of durvalumab and durvalumab + tremelimumab	ADA (confirmatory results: positive or negative; titers [ADA neutralizing antibodies will also be assessed])

Secondary objectives:	Outcome measures:
To assess the effect of the treatment on changes in symptoms and health-related QoL using EORTC QLQ-C30 v3 and QLQ-LC13	EORTC QLQ-C30: symptoms (fatigue, pain, nausea/vomiting, dyspnea, loss of appetite, insomnia, constipation, and diarrhea). Health-related QoL/functioning (physical function, role function, emotional function, cognitive function, social function, and global health status/QoL). EORTC QLQ-LC13: disease-related symptoms (dyspnea, cough, hemoptysis, chest pain, arm/shoulder pain, and other pain). Changes in WHO/ECOG performance status will also be assessed.

ADA Anti-drug antibody; AE Adverse event; APF12 Proportion of patients alive and progression free at 12 months from randomization (ie, PFS rate at 12 months); APF6 Proportion of patients alive and progression free at 6 months from randomization (ie, PFS rate at 6 months); ECOG Eastern Cooperative Oncology Group; EORTC European Organisation for Research and Treatment of Cancer; EP Etoposide and platinum-based chemotherapy; OS Overall survival; OS18 Overall survival at 18 months after randomization; PFS Progression-free survival; 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; QoL Quality of life; RECIST 1.1 Response Evaluation Criteria in Solid Tumors, version 1.1; WHO World Health Organization.

1.1.3 Safety objectives

Safety objective:	Outcome measures:
To assess the safety and tolerability profile of durvalumab and durvalumab + tremelimumab in combination with EP treatment compared with EP	AEs; physical examinations; vital signs including blood pressure and pulse rate; electrocardiograms; and laboratory findings including clinical chemistry, hematology, and urinalysis

AE Adverse event; EP Etoposide and platinum-based chemotherapy.

1.1.4 Exploratory objectives

Exploratory objectives:	Outcome measures:
To further assess the efficacy of durvalumab + tremelimumab + EP treatment compared with EP and to assess the efficacy of durvalumab + EP compared with EP in terms of PFS2	PFS2 using local standard clinical practice[a]
To investigate the relationship between durvalumab PK exposure and clinical outcomes, efficacy, AEs, and/or safety parameters, and biomarkers, if deemed appropriate	A graphical and/or a data modelling approach will be used to analyze durvalumab PK exposure and the relationship with clinical outcomes, efficacy, AEs, and/or safety parameters, as deemed appropriate.
To characterize EP PK when in combination with durvalumab and tremelimumab	Concentration of etoposide, cisplatin or carboplatin in blood

Exploratory objectives:	Outcome measures:
To explore the impact of treatment and disease on health care resource use	Health care resource use will be captured, including inpatient admissions, intensive care unit admissions, and length of stay in hospital
To explore the impact of treatment and disease state on health state utility using the EQ-5D-5L	The EQ-5D-5L health state utility index will be used to derive health state utility based on patient-reported data
To assess AEs by patient self-reporting of specific CTCAE symptoms	Collection of approximately 11 symptoms via the patient-reported outcomes version of the CTCAE (PRO-CTCAE)
To assess patients' overall impression of the change in their health status since the start of study treatment	PGIC item will be collected directly from patients.
To investigate the relationship between a patient's expression of select genes, for example IFN-γ, within the tumor microenvironment and efficacy outcomes with durvalumab ± tremelimumab and EP	Levels of gene expression, for example, IFN-γ, within the tumor microenvironment relative to efficacy outcomes (for example, APF6, APF12, PFS, and OS)
To investigate the relationship between a patient's PD-L1 expression and spatial distribution within the tumor microenvironment and efficacy outcomes with durvalumab ± tremelimumab and EP	Tumoral and/or infiltrating immune cell expression of PD-L1 and spatial distribution within the tumor microenvironment relative to efficacy outcomes (for example, APF6, APF12, PFS, and OS)
To investigate the relationship between a patient's level of DLL3 expression on tumor cells and efficacy outcomes with durvalumab ± tremelimumab and EP	Tumoral expression of DLL3 relative to efficacy outcomes (for example, APF6, APF12, PFS, and OS)
To investigate the relationship between a patient's TMB and/or somatic mutations/genomic alterations and efficacy outcomes with durvalumab ± tremelimumab and EP	Levels of TMB and somatic aberrations in tumor and/or plasma relative to efficacy outcomes (for example, APF6, APF12, PFS, and OS)
To explore potential biomarkers in residual biological samples (eg, tumor and blood), which may influence the progression of cancer (and associated clinical characteristics) and/or prospectively identify patients likely to respond to durvalumab or durvalumab + tremelimumab treatment	Correlation of biomarkers with response to durvalumab or durvalumab + tremelimumab treatment and/or the progression of cancer

Exploratory objectives:	Outcome measures:
To collect and store DNA according to each country's local and ethical procedures for future exploratory research into genes/genetic variation that may influence response (ie, distribution, safety, tolerability, and efficacy) to study treatments and/or susceptibility to disease (optional)	Correlation of polymorphisms with variation in PK, pharmacodynamics, safety, or response parameters observed in patients treated with durvalumab or durvalumab + tremelimumab and/or susceptibility to disease

AE Adverse event; APF12 Proportion of patients alive and progression free at 12 months from randomization (ie, PFS rate at 12 months); APF6 Proportion of patients alive and progression free at 6 months from randomization (ie, PFS rate at 6 months); CTCAE Common Terminology Criteria for Adverse Events; DLL3 Delta-like canonical Notch ligand 3; DNA Deoxyribonucleic acid; EQ-5D-5L EuroQol 5-Dimension, 5-level health state utility index; OS Overall survival; PD-L1 Programmed cell death ligand 1; PFS Progression-free survival; PFS2 Progression-free survival after subsequent anticancer therapy; PGIC Patient's Global Impression of Change; PK Pharmacokinetic(s); TMB Tumor mutational burden

[a] PFS2 will be defined as the time from the date of randomization to the earliest progression event subsequent to that used for the PFS endpoint or death.

A further objective, to fulfil China Food and Drug Administration (CFDA) requirements, is to evaluate the consistency in efficacy and safety among patients from China for benefit-risk assessment of durvalumab + tremelimumab in combination with EP treatment compared to EP and durvalumab + EP compared to EP.

1.2 Study design

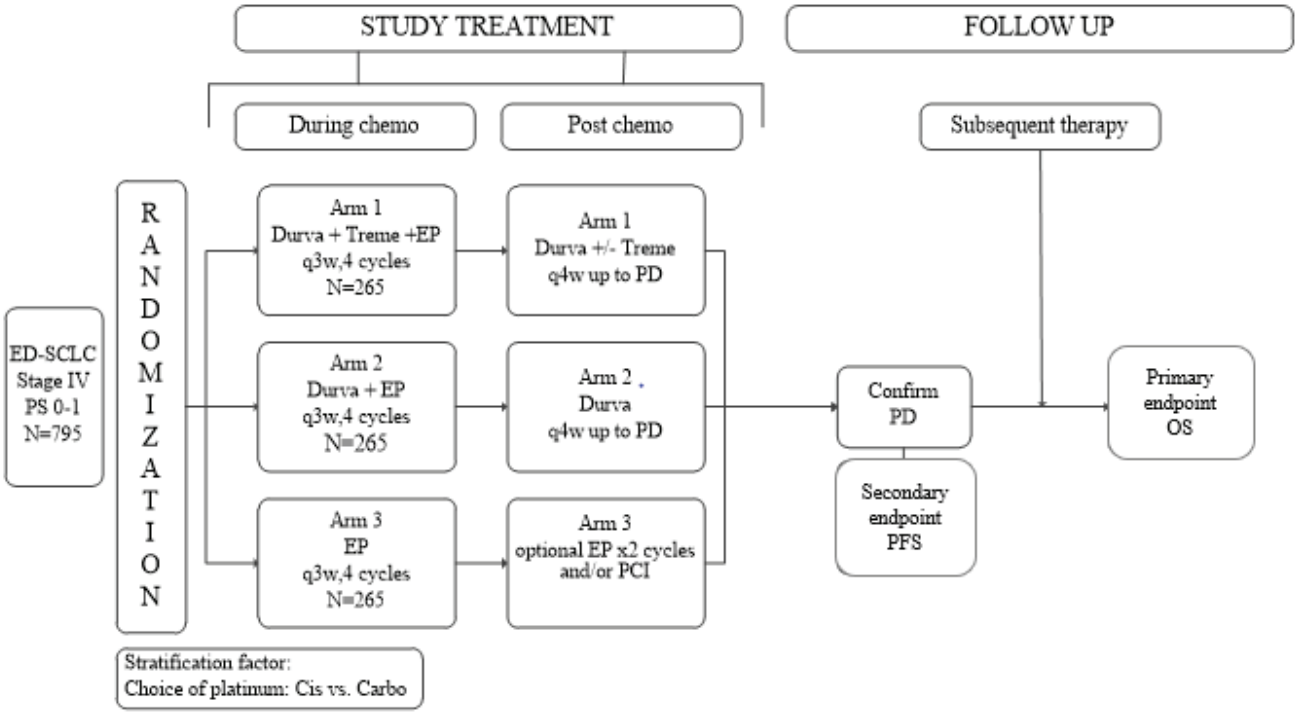
This is a Phase III, randomized, open-label, comparative, multicenter, global study to determine the efficacy and safety of durvalumab + tremelimumab + EP or durvalumab + EP versus EP chemotherapy as first-line treatment in patients with ED SCLC. A schematic diagram of the overall study design is shown in Figure 1, and a detailed study flow chart is shown in Figure 2.

This study will randomize approximately 795 eligible patients at sites worldwide. Once global enrollment of 795 patients is completed, recruitment will continue in China only. A total of up to 189 patients from China will be randomized (see CSP Section 8.6 for detail).

Patients will be randomized in a 1:1:1 ratio in a stratified manner according to the planned platinum-based therapy for Cycle 1 (cisplatin or carboplatin) to receive treatment with durvalumab + tremelimumab + EP (Arm 1), durvalumab + EP (Arm 2), or EP (Arm 3). Cross over will not be permitted as part of this study. Doses and treatment regimens are described in CSP Section 7.2. Assessments will be conducted as indicated in CSP Table 2 and CSP Table 3.

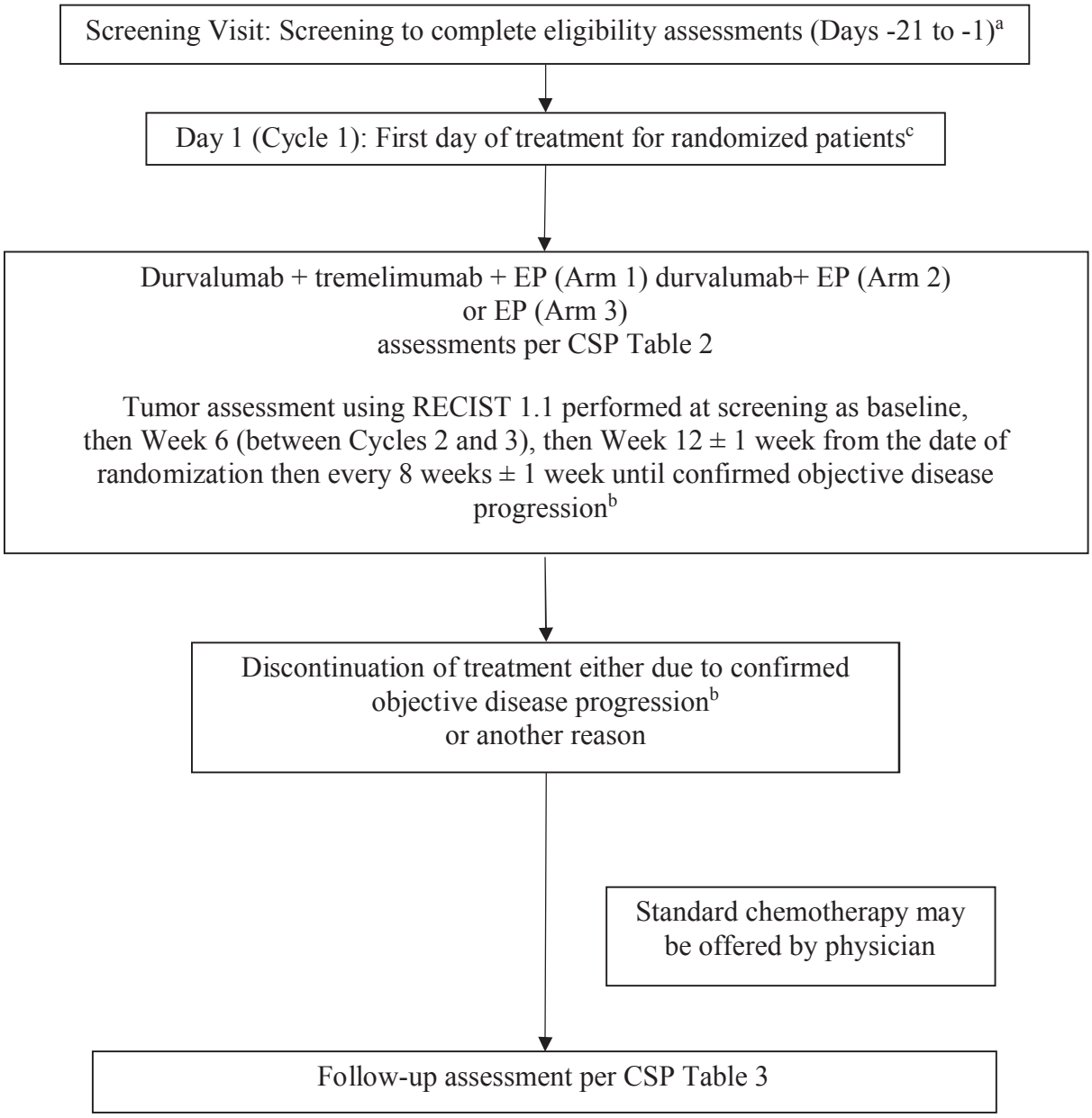
Tumor assessments will be performed at Screening as baseline with follow-ups at Week 6 ±1 week from the date of randomization, at Week 12 ±1 week and then every 8 weeks ±1 week until confirmed objective disease progression (please refer to CSP Appendix F).

Figure 1 Overall study design



Durva durvalumab; Treme tremelimumab; ED Extensive-stage disease; EP Etoposide and platinum-based chemotherapy; OS Overall survival; PCI prophylactic cranial irradiation; SCLC small-cell lung cancer; Treme Tremelimumab.
Note that only one dose of tremelimumab will be administered post chemotherapy in Arm 1 if a patient receives 4 combination doses during chemotherapy, ie, up to 5 durvalumab+tremelimumab combination doses in total – see CSP Section 7.2.

Figure 2 Study flow chart



[a] Informed consent and completion of study procedures and baseline CT/MRI tumor assessment.
[b] In the absence of clinically significant deterioration, a confirmatory scan is always required following the initial demonstration of radiologic PD preferably at the next scheduled imaging visit and no earlier than 4 weeks after the prior scan with radiologic PD by RECIST 1.1. (See CSP Section 5.1 and Appendix F for more information.)
[c] A window of up to 2 days is permitted between randomization and first dose of IP.
EP Etoposide and platinum-based chemotherapy; RECIST 1.1 Response Evaluation Criteria in Solid Tumors, version 1.1.

Independent Data Monitoring Committee (IDMC)

An IDMC comprised of independent experts will be convened to confirm the safety and tolerability of the proposed dose and schedule of durvalumab ± tremelimumab in combination

with platinum based chemotherapy at two early stages of enrolment. A step wise approach will be adopted. The initial safety review will take place when the first 30 patients (10 in each arm) have completed the 1st cycle of treatment and had 21 days of follow up. A second review will take place when an additional 30 patients (10 in each arm) who have completed the 1st cycle of treatment and have had 21 days of follow up. At the time of the 2nd review, it is expected that the initial 30 patients would have had at least 6 weeks of follow up, with some patients receiving much longer. These two reviews will be carried out by the IDMC in an unblinded manner. After review, the IDMC will make a recommendation on whether the study should continue recruitment as planned, or hold recruitment. The IDMC recommendation will be communicated to all sites when available.

In addition, the IDMC will meet approximately every 6 months thereafter to continue safety monitoring.

Details on the IDMC are provided in Section 5.2 and full details of the IDMC procedures and processes can be found in the IDMC Charter.

1.3 Number of patients

The study will randomize approximately 795 eligible patients 1:1:1 to durvalumab + tremelimumab + EP (Arm 1), durvalumab + EP (Arm 2), or EP (Arm 3). The randomization will be stratified based on planned platinum-based therapy in cycle 1 (carboplatin or cisplatin). Once global enrollment achieves 795 randomized patients, recruitment will continue in China only. A total of up to 189 patients from China will be randomized into the study.

The primary objective of this study is to assess the efficacy of durvalumab + tremelimumab + EP treatment compared with EP (Arm 1 vs. 3) and the efficacy of durvalumab + EP treatment compared with EP (Arm 2 vs. 3) in terms of OS. To control for type 1 error, a significance level of 1% will be used for the analysis of Arm 1 vs. 3, and a significance level of 4% will be used for the analysis of Arm 2 vs.3. The study will be considered positive (a success) if either of the OS analysis results are statistically significant. The sizing assumes a 3-month delay in separation of the OS curves between arm 1 vs. arm 3 and between arm 2 vs. arm 3, hence the use of average HRs.

Non-uniform accrual of patients is assumed when estimating the analysis times. The total proportion of patients randomized at time t ($t \leq 15$ months) following the start of the study is assumed to be $(t/15)^2$.

There will be 2 data cut-off timepoints in the study.

The interim analysis of OS will occur when approximately 318 OS events have occurred (60% maturity) in the durvalumab + tremelimumab + EP and EP treatment arms and approximately 318 OS events have occurred (60% maturity) in the durvalumab + EP and EP treatment arms. With a 15-month recruitment period in the global cohort and a minimum follow-up period of approximately 13 months, it is anticipated that this analysis will be performed approximately 28 months after the first patient is randomized.

The data cut-off for the final analysis of OS will occur when approximately 425 OS events have occurred across the durvalumab + tremelimumab + EP and EP treatment arms (80% maturity) and approximately 425 OS events have occurred across the durvalumab + EP and EP treatment arms (80% maturity). If the average true OS HR is 0.69 for both comparisons, the

study will have 89% power to demonstrate a statistically significant difference at the final analysis with a 2-sided 0.93% significance level (for an overall alpha of 1%) for the comparison of durvalumab + tremelimumab + EP versus EP (Arm 1 vs 3), and 96% power to demonstrate a statistically significant difference at a 2-sided 3.57% significance level (for an overall alpha of 4%) for the comparison of durvalumab + EP versus EP (Arm 2 vs 3); this translates to a 4.8-month benefit in median OS over EP (15.7 months vs 10.9 months). The smallest treatment difference that would be statistically significant is an average HR of 0.78 for durvalumab + tremelimumab + EP versus EP and 0.82 for durvalumab + EP versus EP. With a 15-month recruitment period and a minimum follow-up period of 27 months assumed, it is anticipated that this analysis will be performed 42 months after the first patient has been randomized.

A key secondary objective is to assess the efficacy of durvalumab + tremelimumab + EP treatment compared with EP (Arm 1 vs. 3) and the efficacy of durvalumab + EP treatment compared with EP (Arm 2 vs. 3) in terms of PFS. These analyses of PFS will be included in the MTP, as described in Section 4.2.1. If the average true PFS HR is 0.71, the study will have 90% power to demonstrate a statistically significant difference at the 5% level (using a 2-sided test) for the PFS comparisons when approximately 360 PFS events have been observed in the two treatment arms to be compared.

2. ANALYSIS SETS

2.1 Definition of analysis sets

Full analysis set (Intention to treat (ITT))

The full analysis set (FAS) will include all patients randomized prior to the end of global recruitment. The full analysis set will be used for demographics, patient characteristics and efficacy analyses (including PROs). Treatment groups will be compared on the basis of randomized study treatment, regardless of the treatment actually received. Patients who were randomized but did not subsequently go on to receive study treatment are included in the analysis in the treatment group to which they were randomized.

Safety analysis set

The safety analysis set will consist of all patients recruited prior to the end of global recruitment who received at least 1 dose of study treatment. Safety data will be summarized using the safety analysis set, according to the treatment received, that is, erroneously treated patients (eg, those randomized to treatment A but actually given treatment B) will be summarized according to the treatment they actually received.

Pharmacokinetic analysis set

All patients recruited prior to the end of global recruitment who received at least 1 dose of investigational product (IP) per the protocol for whom any post-dose PK data are available will be included in the pharmacokinetic analysis set.

Definitions of the analysis sets for each outcome variable are provided in Table 1.

Table 1 Summary of outcome variables and analysis populations

Outcome variable	Population
Efficacy data	
OS and PFS	Full analysis set (ITT population)
APF6, APF12, PFS2, ORR, OS18, and PROs	Full analysis set (ITT population)
Demography and baseline characteristics	
Baseline and disease characteristics	Full analysis set (ITT population)
Medical/surgical history	Full analysis set (ITT population)
Previous anticancer therapy	Full analysis set (ITT population)
Important deviations	Full analysis set (ITT population)
Concomitant medications/procedures	Full analysis set (ITT population)
Subsequent anticancer therapy	Full analysis set (ITT population)
PK data	PK analysis set
Immunogenicity data	Safety 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

2.2 Protocol deviations

The important protocol deviations will be listed and summarized by randomized treatment group. Deviation 1, below, will lead to exclusion from the safety analysis set. None of the other deviations will lead to patients 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 patients with significant protocol deviations is not planned; however, a 'deviation bias' sensitivity analysis will be performed excluding patients with deviations that may affect the efficacy of the trial therapy if >10% of patients:

- 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.

Eligibility criteria deviations are deviations from the protocol inclusion and exclusion criteria. Post-entry deviations are deviations from the protocol that occurred after the patient was assigned to the study.

The following general categories will be considered important deviations and be listed and discussed in the Clinical Study Report (CSR) as appropriate for the study. If a 'deviation bias' sensitivity analysis is conducted, then patients with these deviations will be excluded from the sensitivity analysis:

- Deviation 1: Patients randomized but who did not receive study treatment.
- Deviation 2: Patients who deviate from key entry criteria as per the CSP. These are inclusion criteria 3, 5, 7 and 9 and exclusion criteria 7, 11, 17.
- Deviation 3: Baseline RECIST scan >42 days before randomization.
- Deviation 4: No baseline RECIST 1.1 assessment on or before randomization.
- Deviation 5: Received prohibited concomitant medications (including other anticancer agents). Please refer to the Clinical Study Protocol (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 to identify those likely have an impact on efficacy.
- Deviation 6: Patients randomized who received treatment other than that to which they were randomized.

In addition to the programmatic determination of the deviations above (based on the clinical database), monitoring notes or summaries will be reviewed to determine any important post entry deviations that are not identifiable via programming, and to check that those identified via programming are correctly classified. The final classification of deviations will be made at the blinded data review meeting (BDRM) prior to database lock or data freeze. Decisions made at the BDRM will be documented and approved by AstraZeneca prior to analysis.

Patients 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.

2.3 China and Japan cohorts

The China cohort is defined as all patients from sites in China accredited by CFDA and enrolled into the study prior to the last patient last visit (LPLV) of the global cohort. The China cohort will include approximately 189 randomized patients. The global cohort consists of all patients randomized by the documented date of last patient randomized of the global cohort. Once global enrollment is complete, recruitment across all sites except for those from China will be closed, and the recruitment of patients in China will continue.

Hence, a patient randomized in the China cohort prior to the last patient randomized in the global cohort will be included in both the global FAS and China FAS. A patient randomized in the China cohort after the last patient was randomized in the global cohort will be included only in the China FAS.

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 the Chinese and Asian populations is required to facilitate the benefit-risk assessment for Chinese patients. Thus, the efficacy and safety data in the China cohort will be summarized and analyzed separately where the same endpoint definitions (as described in CSP Section 8.4) and the same methods for statistical analyses (as detailed in CSP Section 8.5) are applied.

The China FAS will include all patients randomized in the China cohort and will be used for all China only efficacy analyses.

The China safety analysis set will consist of all patients randomized in the China cohort who receive at least one dose of study treatment.

Efficacy analyses for the China cohort will be performed when the OS data from the patients in this cohort are of similar maturity to those of the global cohort where significant clinical efficacy is established in the global cohort, eg, if OS efficacy is established at the primary analysis, a similar maturity to this will be used for the consistency evaluation.

All statistical analyses will be considered exploratory and only performed if sufficient numbers of events or patients are available (eg, ≥20 OS events) unless specified otherwise, descriptive statistics only will be presented. No adjustment for multiplicity will be made and the procedure for hierarchical testing detailed in study CSP Section 8.5 will not be followed. OS and PFS efficacy evaluation for the China cohort will be performed once, respectively.

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

Japan cohort consists of all patients 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 Japan cohort is required to facilitate the benefit-risk assessment for Japanese patients.

3. PRIMARY AND SECONDARY VARIABLES

3.1 Derivation of RECIST 1.1 Visit Responses

For all patients, the RECIST version 1.1 (see further details in Appendix F in the CSP) tumor response data will be used to determine each patient's visit response. It will also be used to determine if and when a patient has progressed and also their best objective response.

The baseline assessment should be performed no more than 28 days before randomization and ideally as close as possible to the start of study treatment. Efficacy for all patients will be assessed on images collected q6w ± 1w for the first 12 weeks relative to the date of randomization, and q8w±1w thereafter until confirmed objective disease progression or off-study (refer to the study plans in CSP: Table 2 [treatment period] and Table 3 [follow-up]). If an unscheduled assessment is performed, and the patient has not progressed, every attempt should be made to perform the subsequent assessments at his or her next regulatory imaging visit.

For patients who discontinue study treatment for reasons other than the confirmed objective RECIST disease progression, objective tumor assessments per the scheduled assessments should be performed q6w ± 1w for the first 12 weeks (relative to the date of randomization), and then q8w ± 1w thereafter until confirmed objective disease progression or until death (whichever comes first) as defined in CSP Table 3.

Categorization of objective tumor response assessment will be based on the RECIST 1.1 criteria of response: complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD). Target lesion progression will be calculated in comparison to when the tumor burden was at a minimum (ie, smallest sum of diameters previously recorded on study). In the absence of progression, tumor response (CR or PR) and SD will be calculated in comparison to the baseline tumor measurements obtained before starting treatment.

Following confirmed progression, patients should continue to be followed up for survival every 2 months as outlined in the study plan (Table 3 in the CSP).

Subsequent anticancer therapy information will be collected at the timepoints indicated in Table 3 in the CSP.

3.1.1 Investigator RECIST 1.1-based assessments: Target lesions (TLs)

At each visit, patients will be programmatically assigned a RECIST 1.1 visit response of CR, PR, SD, or PD depending on the status of their disease compared with baseline and previous assessments. Baseline will be assessed within the 28 days prior to randomization. If a patient has had a tumor 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).

Measurable disease is defined as having at least 1 measurable lesion, not previously irradiated, which is ≥10 mm in the longest diameter (except lymph nodes which must have short axis ≥15 mm) with CT or MRI and which is suitable for accurate repeated measurements.

A patient can have a maximum of 5 measurable lesions recorded at baseline with a maximum of 2 lesions per organ (representative of all lesions involved suitable for accurate repeated measurement) and these are referred to as target lesions (TLs). If more than 1 baseline scan is recorded then measurements from the scan that is closest and prior to the date of randomization (or prior to the start of study treatment in the absence of a pre-randomization scan) 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 TL should be identified as non-target lesions (NTL) at baseline. Measurements are not required for these lesions, but their status should be followed at subsequent visits.

Measurable disease (ie, at least 1 TL) is one of the entry criteria for the study. However, if a patient with non-measurable disease is enrolled in the study, the evaluation of overall visit responses will be based on the overall NTL assessment and the absence/presence of new lesions. If a patient does not have measurable disease at baseline then the TL visit response will be not applicable (NA).

Table 2 TL visit responses

Visit Reponses	Description
Complete Response (CR)	Disappearance of all TLs. Any pathological lymph nodes selected as TLs must have a reduction in short axis to <10 mm.
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)	At least a 20% increase in the sum of diameters of TLs, taking as reference the smallest sum on study (this includes the baseline sum if that is the smallest on study). In addition to the relative increase of 20%, the sum must also demonstrate an absolute increase of at least 5 mm.
Stable disease (SD)	Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD.
Not Evaluable (NE)	Only relevant in certain situations (ie, if any of the target lesions were not assessed or not evaluable or had a lesion intervention at this visit; and scaling up could not be performed for lesions with interventions). Note: If the sum of diameters meets the progressive disease criteria, progressive disease overrides not evaluable as a target lesion 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 1 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 ≥5 mm, 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 lesion diameter recorded.

Lymph nodes

For lymph nodes, if the size reduces to <10 mm 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 <10 mm and all other TLs are 0 mm then although the sum may be >0 mm 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 (ie, 0 mm or <10 mm for lymph nodes) then response will be set to CR irrespective of whether the criteria for PD of TL is also met (ie, if a lymph node LD increases by 20% but remains <10 mm).
- Step 2: If some lesion measurements are missing but all other lesions meet the CR criteria (ie, 0 mm or <10 mm 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, a value of 5 mm will be entered into the database and used in TL calculations, unless the radiologist has indicated and 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

Previously irradiated lesions (ie, lesion irradiated prior to entry into the study) should be recorded as NTLs and should not form part of the TL assessment.

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, a scaled sum of diameters will be calculated (as long as ≤1/3 of the TLs have missing measurements), treating the lesion with intervention as missing, and PR or SD then assigned as the visit response. Patients with intervention are evaluable for CR as long as all non-intervened lesions are 0 (or <10 mm for lymph nodes) and the lesions that have been subject to intervention also have a value of 0 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 where appropriate (as per step 2 above).

Scaling (applicable only for irradiated lesions/lesion intervention)

If >1/3 of TL measurements are treated as missing (because of intervention) then TL response will be NE, unless the sum of diameters of non-missing TL would result in PD (ie, 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 ≥5 mm from nadir).

If ≤1/3 of the TL measurements are treated as 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 (mm)	Longest diameter at follow-up visit (mm)
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 mm. The sum of the corresponding lesions at nadir visit is 26.8 mm.

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

$$\frac{26}{26.8} \times 29.3 = 28.4mm$$

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 2, then the LDs of the split lesions should be summed and reported as the LD for the lesion that split.

Lesions that merge

If 2 TLs merge, then the LD of the merged lesion should be recorded for 1 of the TL sizes and the other TL size should be recorded as 0 mm.

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 (eg, 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 (NTLs) and new lesions

At each visit, an overall assessment of the NTL response should be recorded by the investigator. 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 3 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 1 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 1 or more NTLs with no evidence of progression.
Not Evaluable (NE)	Only relevant when 1 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 patients 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 1 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 1 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: ie, not attributable to differences in scanning technique, change in imaging modality or findings thought to represent something other than tumor.

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.

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.

Patients 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.

3.1.3 Investigator RECIST 1.1-based assessments: Overall visit response

Table 4 defines how the previously defined TL and NTL visit responses will be combined with new lesion information to give an overall visit response.

Table 4 Overall Visit Responses

Target lesions	Non-target lesions	New Lesion	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

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

3.2 Outcome Variables

The analysis of the primary endpoints will evaluate OS from all-cause mortality. The analyses of the secondary efficacy endpoints (PFS, ORR, APF6 and APF12) will be based on site Investigator tumor assessments according to RECIST 1.1. In addition, time from randomization to second progression (PFS2) will be based on second progression as defined by local clinical practice.

All RECIST assessments, whether scheduled or unscheduled, will be included in the calculations. This is also regardless of whether a patient discontinues study treatment or receives another anticancer therapy.

3.2.1 Primary endpoints

The primary endpoints of this study are OS (Arm 1 vs. 3) and OS (Arm 2 vs. 3)

3.2.1.1 Overall survival

Overall survival (OS) is defined as the time from the date of randomization until death due to any cause (ie, date of death or censoring – date of randomization + 1). Any patient not known to have died at the time of analysis will be censored based on the last recorded date on which the patient was known to be alive (SUR_DAT, recorded within the SURVIVE module of the electronic Case Report Form [eCRF]).

Note: Survival calls will be made in the week following the date of Data Cut-Off (DCO) for the analysis; if patients are confirmed to be alive or if the death date is post the DCO date these patients will be censored at the date of DCO. The status of ongoing, withdrawn (from the study) and "lost to follow-up" patients at the time of the final OS analysis should be obtained by the site personnel by checking the patient's notes, hospital records, contacting the patient's general practitioner and checking publicly-available death registries. In the event that the patient has actively withdrawn consent to the processing of their personal data, the vital status of the patient can be obtained by site personnel from publicly-available resources where it is possible to do so under applicable local laws.

Note that for any OS analysis performed prior to the final OS analysis, in the absence of survival calls being made, it may be necessary to use all relevant CRF fields to determine the last recorded date on which the patient was known to be alive for those patients still on treatment (since the SURVIVE module is only completed for patients off treatment if a survival sweep is not performed).

3.2.2 Secondary endpoints

3.2.2.1 Progression-free survival

Progression-free survival (PFS) (per RECIST 1.1 using site Investigator assessments) will be defined as the time from the date of randomization until the date of objective disease progression or death (by any cause in the absence of progression) regardless of whether the patient withdraws from randomized therapy or receives another anticancer therapy prior to progression (ie, date of PFS event or censoring – date of randomization + 1). Patients 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 patient progresses or dies after 2 or more missed visits, the patient will be censored at the time of the latest evaluable RECIST 1.1 assessment prior to the 2 missed visits (Note: NE visit is not considered a missed visit). If the patient has no evaluable visits or does not have baseline data, he or she will be censored at date of randomization unless the patient died within 2 visits from randomization (2 x 6 weeks for tumor assessments + 2 x 7 days for permitted visit window).

Given the scheduled visit assessment scheme (ie, q6w ± 1w for the first 12 weeks then q8w ± 1w thereafter) the definition of 2 missed visits will change. If there is no RECIST assessment during the period when the scheduled frequency of RECIST assessments is q6w (ie, 2 * 6 + 1 week for an early assessment +1 week for a late assessment), this will be considered as 2 missed visits. If the 2 missed visits occur over the period when the scheduled frequency of RECIST assessments changes from q6w to q8w this will equate to 16 weeks (ie, 6 weeks + 8 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 (ie, Week 5 to Week 11). From Week 12 onwards (when the scheduling changes to q8w), 2 missing visits will equate to 18 weeks (ie, 2 x 8 weeks + 1 week for an early assessment + 1 week for a late assessment = 18 weeks). If a patient has no evaluable visit or does not have baseline data, he/she will be censored at date of randomization unless they die within 2 visits from randomization date (2 x 6 weeks + 1 week for a late assessment).

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

RECIST 1.1 assessments/scans contributing towards a particular visit may be performed on different dates. The following rules will be applied:

- The date of progression will be determined based on the earliest of the RECIST 1.1 assessment/scan dates of the component that indicates progression.
- When censoring a patient for PFS, the patient will be censored at the latest of the dates contributing to a particular overall visit assessment

Note: For target lesions, only the latest scan date is recorded out of all scans performed at that assessment for the target lesions, and similarly for non-target lesions, only the latest scan date is recorded out of all scans performed at that assessment for the non-target lesions.

3.2.2.2 Objective response rate

ORR (per RECIST 1.1 using site Investigator assessments) is defined as the number (%) of patients with at least 1 visit response of CR or PR. If any patients do not have measurable disease at baseline then the analysis of ORR will exclude these patients, so that the denominator is a subset of the FAS who have measurable disease at baseline.

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

Best objective response

Best objective response (BOR) is calculated based on the overall visit responses from each RECIST assessment, described in Appendix F in the CSP. It is the best response a patient has had following randomization but prior to starting any subsequent anticancer therapy (palliative or adjuvant radiotherapy excluded) up until RECIST 1.1 progression or the last evaluable assessment in the absence of RECIST 1.1 progression, as determined by site Investigator.

Categorization of BOR will be based on RECIST 1.1 (Appendix F in Protocol) using the following response categories: CR, PR, SD, PD, and NE.

BOR will be determined programmatically based on RECIST 1.1 using all site Investigator assessments up until the first progression event, the start of any subsequent anticancer therapy (palliative or adjuvant radiotherapy excluded) or the last evaluable assessment in the absence of progression. For patients whose progression event is death, BOR will be calculated based upon all evaluable RECIST 1.1 assessments prior to death.

For determination of a best response of SD, the earliest of the dates contributing toward a particular overall visit assessment will be used. SD should be recorded at least 6 weeks minus 1 week, ie, at least 35 days after randomization to allow for an early assessment within the assessment window (ie, study day 36). For CR/PR, the initial overall visit assessment which showed a response will use the latest of the dates contributing towards a particular overall visit assessment.

For patients who die with no evaluable RECIST 1.1 assessments, if the death occurs ≤91 days (ie, 2x6 weeks +7 days) after randomization, then BOR will be assigned to the progression (PD) category. For patients 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 after 2 or more missed visits from their last evaluable assessment will not contribute to the BOR derivation.

Change in tumor size

For supportive purposes, the percentage change from baseline in tumor size will be derived at each scheduled tumor assessment visit (ie, Week 6, Week 12, etc. ... hereafter referred to as

Week X for convenience). Best percentage change from baseline in tumor size will also be derived as the biggest decrease or the smallest increase in tumor size from baseline.

This is based on RECIST 1.1 target lesion measurements taken at baseline and subsequently. Assessments closest to the protocolled visit day will be selected. Tumor size is defined as the sum of the longest diameters of the target lesions for the site Investigator data based upon RECIST 1.1 assessments. Target lesions are measurable tumor lesions. Baseline for RECIST 1.1 is defined to be the last evaluable assessment prior to randomization. The change in target lesion tumor size at Week X will be obtained for each patient by taking the difference between the sum of the target lesions at Week X and the sum of the target lesions at baseline. To obtain the percentage change in target lesion tumor size at Week X the change in target lesion tumor size is divided by the sum of the target lesions at baseline and multiplied by 100 (ie, [Week X – baseline] / baseline * 100). More details on target lesions and measurements can be found in Section 3.1.

The above derivations will be programmed for the site Investigator data based upon RECIST 1.1 assessments.

3.2.2.3 Proportion of patients alive and progression free at 6 and 12 months

The proportion of patients alive and progression free at 6 and 12 months (APF6 and APF12) will be defined as the Kaplan-Meier estimate of PFS (per RECIST 1.1 as assessed using site Investigator assessments) at 6 and 12 months, respectively.

3.2.2.4 Proportion of patients alive at 18 months

The proportion of patient alive at 18 months (OS18) will be defined as the Kaplan-Meier estimate of OS at 18 months.

3.2.3 Exploratory endpoints

3.2.3.1 Time from randomization to second progression (PFS2)

Time from randomization to second progression (PFS2) is defined as the time from the date of randomization to the earliest of the progression events subsequent to that used for the PFS endpoint or death (ie, 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 for details). The date of second progression will be recorded by the Investigator in the eCRF (PFS2 module) and defined according to local standard clinical practice and may involve any of the following: objective radiological imaging, symptomatic progression, other, or death. Patients 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, that is, censored at the latest of the PFS or PFS2 assessment date if the patient has not had a second progression or death. Patients who died as a first PFS are then censored for PFS2 at the date of death. Patients who have a first PFS event and then die subsequently will have their PFS2 event at date of death. Patients without any first PFS event will be censored at their last available scan.

Time from randomization to first subsequent therapy or death

As a supportive summary, time to first subsequent therapy or death (TFST) is defined as the time from the date of randomization to the earlier of start date of the first subsequent anticancer therapy after discontinuation of randomized treatment, or death (i.e. date of first subsequent anticancer therapy/death or censoring – date of randomization + 1). Any patient not known to have had a first subsequent anticancer therapy will be censored at the last date that the patient was known not to have received a first subsequent anticancer therapy (obtained from the TTSCAPRX form). If a patient terminated the study for reason other than death before first subsequent therapy, these patients will be censored at the earliest of their last known to be alive and termination dates. Patients not receiving randomized treatment would have TFST calculated in the same way, i.e. time from date of randomization to the subsequent therapy or death.

3.3 Patient-reported outcome (PRO) variables

Patient reported outcome (PRO) questionnaires will be assessed using the EORTC QLQ-C30 v3 (core questionnaire) with the QLQ-LC13 (lung cancer module), patient-reported outcomes version of the CTCAE (PRO-CTCAE), Patient's Global Impression of Change (PGIC), and the EuroQol 5-Dimensions, 5-level health state utility index (EQ-5D-5L). All items/questionnaires will be scored according to published scoring guidelines or the developer's guidelines, if published guidelines are not available. All PRO analyses will be based on the FAS.

3.3.1 Symptoms and HRQL

The EORTC QLQ-C30 v3 questionnaire consists of 30 questions that can be combined to produce 5 functional scales (physical, role, cognitive, emotional, and social), a global health/quality of life (QoL) scale, 3 symptom scales (fatigue, pain, and nausea/vomiting), 6 individual items (dyspnea, insomnia, appetite loss, constipation, diarrhea and financial difficulties. The EORTC QLQ-C30 will be scored according to the EORTC QLQ-C30 scoring manual (Fayers et al 2001). The EORTC QLQ-LC13 is a lung cancer specific module comprising 13 questions to assess lung cancer symptoms (cough, hemoptysis, dyspnea and site-specific pain), treatment related side-effects (sore mouth, dysphagia, peripheral neuropathy and alopecia) and pain medication. With the exception of a multi-item scale for dyspnea, all are single items.

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/QoL scale in the EORTC QLQ-C30 and for each of the symptom/side-effect scales/items in the EORTC QLQ-LC13 according to the EORTC QLQ-C30 Scoring Manual and EORTC QLQ-LC13 instructions, respectively.

Higher scores on the global health status/QoL and functioning scales indicate better health status/function, but higher scores on symptom scales/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 5.

Table 5 Scoring the EORTC QLQ-C30

Scale/ item	Scale/ item abbreviation	Number of items (n)	Item range	Item numbers
Global health status/ QoL	QL2	2	6	29, 30
Functional scales				
Physical	PF2	5	3	1-5
Role	RF2	2	3	6, 7
Cognitive	CF	2	3	20, 25
Emotional	EF	4	3	21-24
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

Baseline will be defined as described in Section 4.1.1.

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. The dyspnea scale (LC13) is only used if all 3 items have been scored; otherwise the items are treated as single-item measures. 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.

The primary PRO measures will be subject-reported lung cancer symptoms assessed using the EORTC QLQ-LC13 and EORTC QLQ-C30, namely:

- Dyspnea (multi-item scale based on three questions: "Were you short of breath when you rested; walked; climbed stairs?")
- Cough: one item ("How much did you cough?")
- Chest pain: one item ("Have you had pain in your chest?")
- Fatigue (multi-item scale based on three questions: "Did you need rest; Have you felt weak; Were you tired?")
- Appetite loss: one item ("Have you lacked appetite?")

Definition of clinically meaningful changes

Changes in score compared with baseline will be evaluated. A minimum clinically meaningful change is defined as an absolute change in the score from baseline of ≥10 for scales/items from the EORTC QLQ-C30 (Osoba et al 1998). For example, a clinically meaningful improvement in physical function (as assessed by EORTC QLQ-C30) is defined as an increase in the score from baseline of ≥10, whereas a clinically meaningful deterioration is defined as a decrease in the score from baseline of ≥10. 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 6.

Table 6 Mean change and visit response in symptoms and health-related quality of life

Score	Change from baseline	Visit response
EORTC QLQ-C30 symptom scales/items and QLQ-LC13 symptom/side-effect scales/items	≥+10	Deterioration
	≤-10	Improvement
	Otherwise	No change
	≥+10	Improvement

Score	Change from baseline	Visit response
EORTC QLQ-C30 functional scales and global health status/QoL	≤-10	Deterioration
	Otherwise	No change

For the visit level summaries of Improvement/Deterioration/No change then all patients with a baseline and post-baseline score will be included thus the denominator may differ from the time to deterioration endpoints derived below.

3.3.1.1 Time to HRQoL/function deterioration

Time to deterioration (TTD) will be analyzed for all function domains and Global health status/Qol (C30). TTD will be defined as the time from the date of randomization until the date of the first clinically meaningful deterioration (a decrease in the score from baseline of ≥10) that is confirmed at a subsequent visit or death (by any cause) in the absence of a clinically meaningful deterioration, regardless of whether the patient withdraws from study treatment or receives another anticancer therapy prior to Global health status/QoL or function deterioration (ie, date of Global health status QoL/function deterioration event or censoring-date of randomization + 1). Patients with a 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 health status/QoL/function change could be evaluated.

Patients whose HRQoL/function (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/function could be evaluated. Also, if HRQoL/function deteriorates after 2 or more missed PRO assessment visits (using the same definitions for 2 missed visits as used in the 'Time to Symptom deterioration' derivation in Section 3.3.1.2) or the patient dies after 2 or more missed PRO assessment visits, the patient will be censored at the time of the last PRO assessment where HRQoL/function could be evaluated. If the patient has no evaluable visits or does not have baseline data they will be censored at 1 day unless they die within 2 visits of baseline (16 weeks plus 1 week allowing for a late assessment within the visit window).

The population for analysis of time to HRQoL/function deterioration will include a subset of the FAS who have baseline scores ≥10. In this analysis, RECIST 1.1 progression will not be considered as HRQoL/function deterioration and data will not be affected by RECIST 1.1 progression.

3.3.1.2 Time to symptom deterioration

For each of the symptom scales/items in the QLQ-LC13 and QLQ-C30, time to deterioration will be defined as the time from the date of randomization until the date of the first clinically meaningful symptom deterioration (an increase in the score from baseline of ≥10) 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 patient withdraws from study treatment or receives another anticancer therapy prior to symptom deterioration (ie, date of

symptom deterioration event or censoring – date of randomization + 1). Patients with a 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.

Patients whose symptoms (as measured by QLQ-LC13 or QLQ-C30) 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 patient dies after 2 or more missed PRO assessment visits, the patient will be censored at the time of the last PRO assessment where the symptom could be evaluated. If the patient has no evaluable visits or does not have baseline data they will be censored at 1 day unless they die within 2 visits of baseline (48 days (ie, 2 x 3 weeks x 7 days) plus 2 x 3 days allowing for a late assessment within the visit window.

Given the scheduled PRO assessment scheme (ie, every cycle for first 4 cycles, where 1 cycle is a 3-week period), 2 missed visits will equate to 48 days (see above).

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

In this analysis, RECIST 1.1 progression will not be considered as symptom deterioration and data will not be affected by RECIST 1.1 progression.

3.3.2 Patient's Global Impression of Change (PGIC)

The PGIC is a global measure indicating the degree of change in overall health status since start of treatment, as evaluated by the patient. Patients will be asked to evaluate "Since the start of the treatment I have received in this study, my overall health is" and response options are a 7-point Likert response scale with the following categories: Very much improved; Much improved; Minimally improved; No change; Minimally worse; Much worse and; Very much worse (Dworkin et al., 2008).

Efficacy data related to health status will be presented by the subgroups of PGIC:

- Improved (values of 1 or 2)
- No change (values of 3, 4 or 5)
- Worse (values of 6 or 7)

3.3.3 Health state utility (EQ-5D-5L)

The EQ-5D is a standardized measure of health status developed by the EuroQol Group in order to provide a simple, generic measure of health for clinical and economic appraisal. Applicable to a wide range of health conditions and treatments, it provides a simple descriptive profile and a single index value for health status that can be used in the clinical and economic evaluation of health care.

The EQ-5D-5L index comprises 5 dimensions of health (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression). For each dimension, respondents select which statement best describes their health on that day from a possible 5 options of increasing levels of severity (no problems, slight problems, moderate problems, severe problems, and unable to/extreme problems). A unique EQ-5D health state is referred to by a 5-digit code allowing for a total of 3125 health states. For example, state 11111 indicates no problems on any of the 5 dimensions. These data will be converted into a weighted health state index by applying scores from EQ-5D value sets elicited from general population samples (the base case will be the United Kingdom valuation set, with other country value sets applied in scenario analyses). Where values sets are not available, the EQ-5D-5L to EQ-5D-3L crosswalk will be applied (Oemar and Jansen 2013). In addition to the descriptive system, respondents also assess their health on the day of assessment on a visual analogue scale, ranging from 0 (worst imaginable health) to 100 (best imaginable health). This score is reported separately.

The evaluable population will comprise a subset of the FAS who have a baseline EQ-5D-5L assessment.

3.3.4 PRO Compliance Rates

Summary measures of overall compliance and compliance over time will be derived for the EORTC-QLQ-C30, LC13 and EQ-5D-5L respectively. These will be based upon:

- Received questionnaire = a questionnaire that has been received and has a completion date and at least 1 individual item completed.

- Expected questionnaire = a questionnaire that is expected to be completed at a scheduled assessment time eg, a questionnaire from a patient who has not withdrawn from the study at the scheduled assessment time but excluding patients in countries with no available translation. For patients that have progressed, the latest of progression and safety follow-up will be used to assess whether the patient is still under HRQoL 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 1 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.

- Overall patient compliance rate is defined for each randomized treatment group as: Total number of patients with an evaluable baseline and at least 1 evaluable follow-up questionnaire (as defined above), divided by the total number of patients expected to have completed at least a baseline questionnaire multiplied by 100.

Compliance over time will be calculated separately for each visit, including baseline, as the number of patients with an evaluable questionnaire (as defined above) within the analysis visit window (as defined in Section 4.1.2), divided by number of patients still expected to complete

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

3.4 Safety

Safety and tolerability will be assessed in terms of adverse events (AEs) (including serious adverse events [SAEs]), deaths, laboratory data, vital signs, and exposure. These will be collected for all patients.

Safety data will be summarized from the 'On treatment' period, unless otherwise specified.

'On treatment' will be defined as assessments between date/time of the first dose and 90 days following last dose of study treatment ('day 28' visit following last dose of study treatment for vital signs assessments), or up to and including the date of initiation of the first subsequent therapy (whichever occurs first). Palliative or adjuvant radiotherapy is not considered a subsequent anticancer therapy.

The safety analysis set will be used for reporting of all safety data except summary table and listing of deaths that will be produced on the FAS.

3.4.1 Adverse events (AEs)

AEs and SAEs will be collected throughout the study, from date of first dose until 90 days after the last dose of IP (durvalumab or tremelimumab or EP). The Medical Dictionary for Regulatory Activities (MedDRA) will be used to code the AEs, using the latest version. AEs will be graded according to the National Cancer Institute Common Terminology Criteria for AEs (NCI CTCAE Version 4.03). A treatment emergent adverse event (TEAE) is an AE with an onset date/time, or a pre-existing AE worsening, during the 'on-treatment' period as defined above. For Arm1 and Arm2, in the unlikely event of the components being administered separately then date of first dose/last dose will be considered as the earliest/latest dosing date of either component.

AEs of special interest

An adverse event of special interest (AESI) is one of scientific and medical interest specific to understanding of the IP 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 this investigational product.

AESIs for durvalumab ± tremelimumab include, but are not limited to, events with a potential inflammatory or immune-mediated mechanisms 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.

Some clinical concepts (including some selected individual preferred terms and higher level terms) have been considered AESI observed with durvalumab ± tremelimumab. These AESIs have been identified as:

- Diarrhea/Colitis and intestinal perforation
- Pneumonitis/ILD
- Hepatitis/transaminase increases
- Endocrinopathies (ie, events of hypophysitis/hypopituitarism, adrenal insufficiency, hyper- and hypothyroidism and type I diabetes mellitus)
- Rash/Dermatitis
- Nephritis/Blood creatinine increases
- Pancreatitis/serum lipase and amylase increases
- Myocarditis
- Myositis/Polymyositis
- Neuropathy/neuromuscular toxicity (eg, Guillain-Barré, and myasthenia gravis)
- Other inflammatory responses that are rare/less frequent with a potential immune mediated etiology include, but are not limited to, pericarditis, sarcoidosis, uveitis and other events involving the eye, skin, hematological and rheumatological events.
- In addition, infusion-related reactions and hypersensitivity/anaphylactic reactions with a different underlying pharmacological etiology are also considered AESIs.

Other categories may be added or existing terms may be merged as necessary. An AstraZeneca medically qualified expert after consultation with the Global Patient Safety Physician has reviewed the AEs of interest and identified which preferred terms contribute to each AESI. A further review will take place prior to Database lock (DBL) to ensure any further terms not already included are captured within the categories.

Immune-mediated Adverse Events (imAE)

To fully characterize the AESI (excluding AESI group Infusion related/ Hypersensitivity/ Anaphylactic reactions) during which systemic steroids, endocrine therapy, or other immunosuppressants were administered), the Sponsor will classify AESIs as immune-mediated AEs (imAEs) or not imAEs. Further details are provided in an imAE Charter.

Other significant adverse events (OAE)

During the evaluation of the AE data, an AstraZeneca medically qualified expert will review the list of AEs that were not reported as SAEs and 'Discontinuation of Investigational Product due to Adverse Events' (DAEs). Based on the expert's judgement, significant adverse events of particular clinical importance may, after consultation with the Global Patient Safety Physician, be considered other significant adverse events (OAEs) and reported as such in the CSR. A similar review of laboratory/vital signs/ECG data will be performed for identification of OAEs.

Examples of these are marked haematological 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 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 the chemotherapy period, and 27 if the last dose occurs during the post-chemotherapy period. For etoposide, xx = 18 and 25 days respectively, if last dose is administered on day 3 of the cycle, with appropriate adjustment if dosing is stopped on day 1 or 2.

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 will also be measured by the number of cycles received. If a cycle (period of 3 weeks during chemotherapy and 4 weeks thereafter) is prolonged due to toxicity, this should still be counted as 1 cycle. A cycle will be counted if treatment is started even if the full dose is not delivered.

For EP exposure in Arm 3, two additional q3w cycles (on Week 12 and Week 15) can be given if clinically indicated, at the investigator's discretion before the patient enter Follow-up. These will be combined with other previous administrations considering the same rules as during the Chemotherapy period.

Patients who permanently discontinue during a dose delay

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

3.4.3 Dose intensity

Relative dose intensity (RDI) is the percentage of the actual dose intensity delivered relative to the intended dose intensity through to treatment discontinuation.

RDI will be defined as follows for durvalumab and tremelimumab (RDI is not calculated for EP treatments):

- RDI = 100% * d/D, where d is the actual cumulative dose delivered up to the actual last day of dosing of the respective treatment (durvalumab or tremelimumab) and D is the intended cumulative dose up to the actual last day of dosing of the respective treatment. 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 may be added to the date of last dose to reflect the protocol allowed window for dosing.

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

3.4.4 Laboratory data

Laboratory data will be collected throughout the study, from screening up to the follow-up visits as described in the CSP. Blood and urine samples for determination of hematology, clinical chemistry, and urinalysis will be collected as described in Section 5.2.1 of the CSP. For the definition of baseline and the derivation of post-baseline visit values considering visit window and how to handle multiple records, derivation rules as described in Section 4.1.2 below will be used.

Change from baseline in hematology and clinical chemistry variables will be calculated for each post-dose visit on treatment. 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 results to corresponding preferred units. The following parameters have CTC grades defined for both high and low values: lymphocyte count, potassium, sodium, magnesium, glucose and corrected calcium, so high and low CTC grades will be calculated.

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

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

Calculated creatinine clearance (CrCl) will be derived in the reporting database using the Cockcroft-Gault formula:

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

Plasma creatinine may be used as a substitute where no serum creatinine measurement is available. At some visits, CrCl is also calculated at site.

Absolute values will be compared to the project reference range and classified as low (below range), normal (within range or on limits of range) and high (above range).

Local reference ranges will be used for the primary interpretation of laboratory data at the local laboratory. Project reference ranges will be used throughout for reporting purposes. The denominator used in laboratory summaries of CTC grades will only include evaluable patients, in other words those who had sufficient data to have the possibility of an abnormality.

For example:

- If a CTCAE criterion involves a change from baseline, evaluable patients would have both a pre-dose and at least 1 post-dose value recorded

- If a CTCAE criterion does not consider changes from baseline, to be evaluable the patient need only have 1 post dose-value recorded.

3.4.5 ECGs

ECG data will be collected at Screening and as clinically indicated through the study. Triplicate ECGs will be obtained in case of clinically significant ECG abnormalities.

At each time point the Investigator's assessment of the ECG will be collected locally. The QT interval corrected for heart rate using Fredericia's correction (QTcF) will be entered by the site using the following formula:

QTcF (ms) = QT/RR^(1/3) where QT and RR are in seconds

For triplicate ECGs, the mean of the 3 ECG assessments collected at the site will be used to determine the value at that time point. The Investigator's opinion on clinical significance of any abnormalities will be recorded for each ECG.

3.4.6 Vital signs

Vital signs data obtained up until the 28 days from date of last dose of study treatment will be used for reporting. Change from baseline in vital signs variables will be calculated for each post-baseline visit on treatment. For derivation of post-baseline visit values considering visit window and to handle multiple records, derivation rules as described in Section 4.1.2 below will be used.

The denominator in vital signs data should include only those patients with recorded data.

3.4.7 Concomitant medication

Any medications taken by the patient at any time between the date of the first dose (including the date of the first dose) of study treatment up to the date of last dose of study treatment + 90

days in the study will be considered as concomitant medication. Any medication that started prior to the first dose of the study treatment and ended after the first dose or is ongoing will be considered as both prior and concomitant medication.

Allowed and disallowed concomitant medications will be presented by ATC classification and generic term.

3.5 Biomarker variables

Data collected on PD-L1 expression will contain the dates when each sample was sectioned and stained. A flag will be derived in the reporting database to identify cases that were outside the recommended cut slide stability, i.e. stained date – sectioned date > 90 days.

3.6 Pharmacokinetic and Immunogenicity variables

3.6.1 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 (BLQ) will be treated as missing in the summary for a time point if less than or equal to half of the results are BLQ; when more than half of the results are BLQ, the mean, median, minimum, and geometric mean are presented as BLQ and the SD, CV, and geometric CV are presented as N/A.

3.6.2 Immunogenicity analysis

Serum samples for ADA assessments will be conducted utilizing a tiered approach (screen, confirm, titer), and ADA data will be collected at scheduled visits shown in the CSP. ADA result from each sample will be reported as either positive or negative. If the sample is positive, the ADA titer will be reported as well. In addition, the presence of neutralizing antibody (nAb) will be tested for all ADA positive samples using a ligand-binding assay. The nAb results will be reported as positive or negative. A patient is defined as being ADA-positive if a positive ADA result is available at any time, including baseline and all post-baseline measurements; otherwise ADA negative.

The following ADA categories for durvalumab (Arm 1 and Arm 2) and tremelimumab (Arm 1) will be determined.

- ADA positive at any visit (at baseline and/or post-baseline
- ADA positive post-baseline and positive at baseline
- ADA positive post-baseline and not detected at baseline (treatment-induced ADA)
- ADA not detected post-baseline and positive at baseline
- Baseline ADA titer that was boosted by ≥4-fold following drug administration (treatment-boosted ADA)
- Persistently positive ADA
- Transiently positive ADA
- nAb positive at any visit (at baseline and/or post-baseline)

3.7 Health Resource Use

Health resource use outcome variables include the following:

- Types of hospitalization
- Length of stay of people admitted
  - into hospital for at least 1 overnight stay (end date > start date)
  - to intensive care/high dependency units
- Primary sign or symptom the patient present with

The length of hospital stay will be calculated as the difference between the date of hospital discharge (or death date) and the start date of hospitalization or start of study drug if the start of study drug is after start date of hospitalization (length of hospital stay = end date of hospitalization – start date of hospitalization + 1). Patients with missing discharge dates will be calculated as the difference between the last day with available data (up to DCO, where applicable) and the start date of hospitalization.

Sum of total duration of hospital stay will be considered for analysis if >20% of patients who were hospitalized were admitted to hospital more than 1 time during study period.

The length of ICU stay will be calculated in the using the same method as detailed above for the length of hospital stay.

4. ANALYSIS METHODS

4.1 General principles

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

- H0: No difference between durvalumab + tremelimumab + EP and EP
- H1: Difference between durvalumab + tremelimumab + EP and EP
- H0: No difference between durvalumab + EP and EP
- H1: Difference between durvalumab + EP and EP

Formal statistical analysis will also be performed to test the secondary hypotheses for PFS:

- H0: No difference between durvalumab + tremelimumab + EP and EP
- H1: Difference between durvalumab + tremelimumab + EP and EP
- H0: No difference between durvalumab + EP and EP

- H1: Difference between durvalumab + EP and EP

The 2 primary endpoints are OS (Arm 1 vs. 3) and OS (Arm 2 vs. 3). The study has been sized to characterize the OS benefit of durvalumab + tremelimumab + EP versus EP and OS benefit of durvalumab + EP versus EP.

The general principles as mentioned below will be followed throughout the study:

- Descriptive statistics will be used for all variables, as appropriate, and will be presented by treatment group. Continuous variables will be summarized by the number of observations, mean, standard deviation, median, 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 analysis set total for the corresponding treatment group.

- 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 efficacy analysis summaries, hazard ratios will be rounded to 2 decimal places, with CIs to 3 decimal places. Median OS or PFS and associated CIs will be displayed to 1 decimal place. P-values will be rounded to 4 decimal places, except for those below 0.00005, which will be displayed as '<0.0001'.

- For PK data the geometric mean and coefficient of variation (CV) will be presented to 4 significant figures (sf), minimum and maximum will be presented to 3 sf and n will be presented as an integer.

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

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

4.1.1 Baseline

In general, for efficacy and PRO endpoints the last observed measurement prior to randomization will be considered the baseline measurement. However, if an evaluable assessment is only available after randomization but before the first dose of randomized treatment then this assessment will be used as baseline. For safety endpoints the last observation before the first dose of study treatment will be considered the baseline measurement unless otherwise specified. 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.

4.1.2 Visit window for safety and PRO assessments

Time windows will be defined for any presentations that summarize values by visit. The following conventions will apply:

- The time windows will be exhaustive so that data recorded at any time point has the potential to be summarized. Inclusion within the time window will be based on the actual date and not the intended date of the visit.

- All unscheduled visit data have the potential to be included in the summaries.

- The window for the visits following baseline will be constructed in such a way that the upper limit of the interval falls half way between the two scheduled visits (the lower limit of the first post-baseline visit will be Day 2). 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.

  The following algorithm will be considered:

  1. Min_X = Max_(X-1) +1

  2. Diff_X = Day (Assessment_(X+1) – Day (Assessment_X)

  3. Max_X = Day (Assessment_X) +

     (i) (Diff_X/2) -1 if Diff_X is an even number

     (ii) Integer part of (Diff_X/2) if Diff_X is an odd number

  An example is provided in Appendix A (Section 8).

- The visit windows for all patients will be based on the visit schedule in Table 2 of the protocol, even for patients who have switched to a different schedule due to treatment discontinuation.

- For summaries showing the maximum or minimum values, the maximum/minimum value recorded on treatment will be used (regardless of where it falls in an interval).

- Listings will display all values contributing to a time point for a patient.

- For visit based summaries:

  – If data are recorded in both plasma and serum from the same sample, the average of the two records will be used.

  – If there is more than one value per patient within a time window then the closest value to the scheduled visit date will be used, or the earlier in the event the values are equidistant from the nominal visit date. If there are two values recorded on the same day and the parameter is CTCAE gradable then the

record with the highest toxicity grade will 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 will be used. The listings will highlight the value for that patient that went into the summary table, wherever feasible.

- To prevent very large tables or plots being produced that contain many cells with meaningless data, for each treatment group visit data will only be summarized if the number of observations is greater than the minimum of 20 and > 1/3 of patients dosed.

- For summaries at a patient level, all values will be included, regardless of whether they appear in a corresponding visit based summary, when deriving a patient level statistic such as a maximum.

- For safety assessments, baseline will be defined as the last non-missing measurement prior to the first dose of study treatment. For laboratory data, any assessments made on the day of the first dose will be considered pre-dose. If there are two visits equally eligible to assess patient status at baseline (e.g., screening and baseline assessments both on the same date prior to first dose with no washout or other intervention in the screening period) with assessment time missing, the average can be taken as a baseline value. For non-numeric laboratory tests (i.e. some of the urinalysis parameters) where taking an average is not possible then the best value would be taken as baseline as this is the most conservative. In the scenario where there are two assessments on the day of first dose, one with time recorded and the other without time recorded, the one with time recorded would be selected as baseline.

Where safety data are summarized over time, study day will be calculated in relation to date of first treatment.

4.1.3 Imputation rules

Missing safety data will generally not be imputed. However, safety assessment values of the form of "<x" (ie, below the lower limit of quantification) or >x (ie, 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, adverse events that have missing causality (after data querying) will be assumed to be related to study drug, with the exception of causality assessments that are not applicable for the patient, e.g. causal relation to durvalumab for patients in Arm 3.

Age at randomization will be derived for analysis purpose from the date of birth (DEM module) and the randomization date (IE1 module) on the eCRF at screening as: year (randomization date) – year (date of birth), -1 if "day and month" of the randomization date is before "day and month" of the date of birth. Patients with a partial date of birth (ie, for those countries where year of birth only is given due to local regulatory constraints) will have an assumed date of birth of 1st Jan ([given year]). For patients with a missing age, the mean age (overall FAS) will be imputed.

Handling of missing/incomplete dates

The original incomplete or missing dates will be presented in the listings.

- Adverse events: all AEs will be considered as treatment-emergent unless the opposite can be clearly stated. Imputation will be done only in the context of identifying TEAEs.

- Concomitant medications: all medications will be considered as concomitant unless the opposite can be clearly stated.

In practice, for adverse events and concomitant medications, original incomplete or missing start dates will be imputed as below:

- Missing day: Impute the 1st of the month unless month is same as month of first dose of study drug then impute first dose date

- Missing day and month: impute 1st January unless year is the same as first dose date then impute first dose date

- Completely missing – impute first dose date unless the end date suggests it could have started prior to this in which case impute the 1st January of the same year as the end date.

When imputing a start date, ensure that the new imputed date is sensible ie, is prior to the end date of the AE or med.

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.

Imputed dates will not be used for deriving duration of events or treatment.

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:

- For missing day only – using the 1st of the month

- 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.

No other imputation will be made.

The following other general principles will also apply:

- All data collected will be listed.
- Efficacy and PRO data will be summarized and analyzed 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 on the safety analysis set. Study population and demography data will be summarized based upon the FAS.
- All outputs will be summarized by treatment group, unless otherwise specified.
- Safety data will be summarized at baseline and for the 'on-treatment' period.

4.2 Analysis methods

Results of all statistical analyses will be presented using a 95% Confidence Interval (CI) and 2-sided p-value, unless otherwise stated.

The following table (Table 7) 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 7 Pre-planned statistical and sensitivity analyses to be conducted

Endpoint analyzed	Notes
Overall survival	• Primary analysis using a stratified log-rank test adjusted for planned platinum therapy during Cycle 1, for: – durvalumab + tremelimumab + EP vs EP – durvalumab + EP vs EP • Secondary analysis (same method as primary analysis) – durvalumab + tremelimumab + EP vs durvalumab + EP • Sensitivity analysis using a Kaplan-Meier plot of time to censoring where the censoring indicator of the primary analysis is reversed – attrition bias – durvalumab + tremelimumab + EP vs EP – durvalumab + EP vs EP • Subgroup analysis using an unstratified Cox model (same comparisons as for the sensitivity analysis above) • Effect of covariates on HR estimate using a stratified Cox model (same comparisons as for the sensitivity analysis above)

Endpoint analyzed	Notes
Progression-free survival	• Secondary analysis using site Investigator RECIST 1.1 assessments, using a stratified log-rank test adjusted for planned platinum therapy during Cycle 1 – durvalumab + tremelimumab + EP vs EP – durvalumab + EP vs EP – durvalumab + tremelimumab + EP vs durvalumab + EP • Sensitivity analysis (durvalumab ± tremelimumab + EP vs EP) using site Investigator RECIST 1.1 assessments: – Interval censored analysis – evaluation time bias (log-rank test) – Analysis using alternative censoring rules – attrition bias (same method as for OS) – Subsequent anticancer therapy • Subgroup analysis (unstratified) using site Investigator RECIST 1.1 assessments (same method as for OS) • Effect of covariates on HR estimate using a stratified Cox model (same method as for OS)
Objective response rate	Logistic regression for durvalumab ± tremelimumab +EP vs EP
Proportion of patients alive and progression free at 6 and 12 months	Kaplan-Meier estimates with CI using log-log transformation
Proportion of patients alive at 18 months	Kaplan-Meier estimates with CI using log-log transformation
Time from randomization to second progression	Stratified log-rank test for durvalumab ± tremelimumab +EP vs EP adjusting for platinum therapy planned in Cycle 1
Change from baseline in PRO symptoms	Average change from baseline using a Mixed model for repeated measures – durvalumab + tremelimumab+ EP versus EP – durvalumab + EP versus EP
Time to symptom deterioration (EORTC QLQ-C30 and QLQ-LC13 endpoints)	Stratified log-rank test for: – durvalumab + tremelimumab+ EP versus EP – durvalumab + EP versus EP

EORTC European Organisation for Research and Treatment of Cancer; EP Etoposide and platinum-based chemotherapy; 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.

4.2.1 Multiple testing strategy

The multiple testing procedure (Figure 3) will define which significance levels should be applied to the interpretation of the raw p-values.

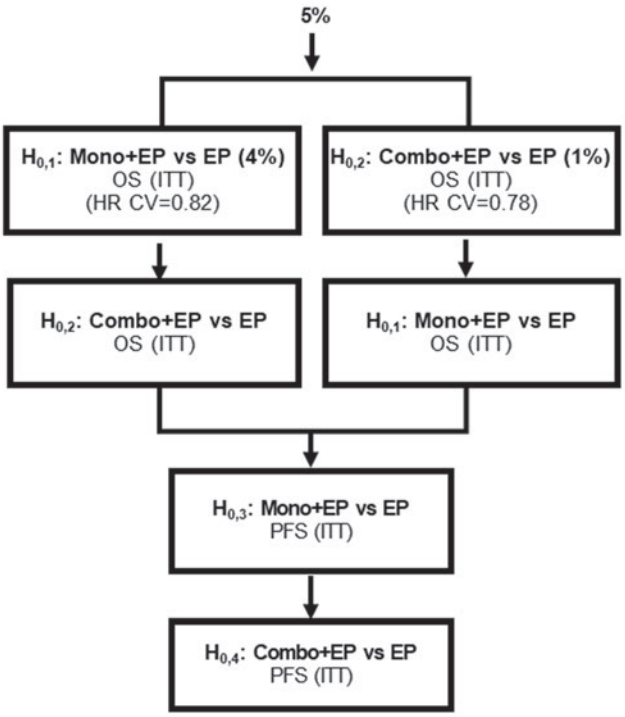
In order to strongly control the type I error at 5% 2-sided, a multiple testing procedure (MTP) with gatekeeping strategy will be used across the 2 primary endpoints of OS (Arm 1 vs. 3) and OS (Arm 2 vs. 3) and the 2 key secondary endpoints of PFS (Arm 1 vs. 3) and PFS (Arm 2 vs. 3). If the higher-level hypothesis in the MTP is rejected for superiority, the following hypothesis will then be tested as shown in Figure 3.

Hypotheses will be tested using a multiple testing procedure 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, the test mass that becomes available after each rejected hypothesis is recycled to secondary hypotheses not yet rejected.

Note given OS is tested at multiple timepoints (i.e., interim and final analyses), the OS tests for the same comparison (i.e., shown in 1 box in the MTP) will be considered as 1 test family. As long as one test in the family can be rejected, the family is rejected thus the assigned total alpha to the family can be recycled to next MTP level. 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 key hypotheses.

Figure 3 shows the multiple testing framework.

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



Note: Alpha recycling between Mono+EP vs EP and Combo+EP vs EP OS comparisons
Mono+EP vs EP = comparison of durvalumb+EP vs EP
Combo+EP vs EP = comparison of durvalumab + tremelimumab + EP vs EP

OS Overall survival; PFS Progression-free survival; EP Etoposide and platinum-based chemotherapy; HR Hazard ratio; CV Critical value.

The testing procedure is hierarchical in that it starts with testing the 2 primary endpoints as outlined in Figure 3. The overall 5% type 1 error will be initially split between the primary endpoints: an alpha level of 4% will be allocated to the analysis of OS (Arm 2 vs. 3), and an alpha level of 1% will be allocated to the analysis of OS (Arm 1 vs. 3).

Implementation of this pre-defined ordered testing procedure, including recycling, will strongly control type I error at 5% (2-sided), among all key hypotheses (Glimm et al 2009

). Below is a detailed procedure.

Test OS (Arm 2 vs. 3) at 4% significance level and OS (Arm 1 vs. 3) at 1%, respectively.

A. If neither of the 2 tests is statistically significant, accept $H_{0,1}$ and $H_{0,2}$, and stop procedure
B. If $H_{0,1}$ is not statistically significant at 4%, but $H_{0,2}$ is statistically significant at 1% level, reject $H_{0,2}$ and recycle the 1% to $H_{0,1}$, and retest $H_{0,1}$ at 5% level.
    a) If $H_{0,1}$ is not statistically significant at 5% level, accept $H_{0,1}$ and stop.
    b) If $H_{0,1}$ is statistically significant at 5% level, reject $H_{0,1}$. Then test $H_{0,3}$ at 5%.
        i. If $H_{0,3}$ is not statistically significant at 5% level, accept $H_{0,3}$ and stop the procedure.
        ii. If $H_{0,3}$ is statistically significant at 5% level, reject $H_{0,3}$ and test $H_{0,4}$ at 5%.
            - If $H_{0,4}$ is statistically significant at 5% level, reject $H_{0,4}$ and stop the procedure.
            - Otherwise, accept $H_{0,4}$ and stop the procedure.
C. If $H_{0,2}$ is not statistically significant at 1%, but $H_{0,1}$ is statistically significant at 4% level, reject $H_{0,1}$ and recycle the 4% to $H_{0,2}$, and retest $H_{0,2}$ at 5% level.
    a) If $H_{0,2}$ is not statistically significant at 5% level, accept $H_{0,2}$ and stop.
    b) If $H_{0,2}$ is statistically significant at 5% level, reject $H_{0,2}$. Then test $H_{0,3}$ at 5%. The rest is the same as in B(b) i. and ii.

D. If both tests are statistically significant, reject $H_{0,1}$ and $H_{0,2}$, and test $H_{0,3}$ at 5%. The rest is the same as in B(b) i. and ii.

For the OS endpoint, there is one interim analysis (IA) planned, and the alpha level will be controlled at the interim and final analysis timepoints by using the Lan DeMets (Lan and De Mets, 1983) spending function that approximates an O'Brien Fleming approach. The O'Brien Fleming boundaries for the OS interim and final analyses will be adjusted depending on the alpha used for the OS endpoint. A separate Lan DeMets (O'Brien Fleming) spending function accounting for an interim and final analysis will also be applied to PFS endpoints in the MTP. A detailed calculation of alpha allocation at IA and final analysis (FA) is provided in Appendix B.

In addition, durvalumab + tremelimumab + EP will be compared with durvalumab + EP for OS and PFS. This comparison is not included in the MTP.

The secondary PRO endpoints described below are not part of the main multiple testing procedure.

For the secondary PRO endpoints, the overall type I error (5% 2-sided) will be controlled across the 10 hypothesis tests of 2 treatment comparisons in the MMRM analysis for each of the 5 key PRO measures of cough, dyspnea and chest pain as assessed by the EORTC QLQ-LC13; fatigue and appetite loss as assessed by the EORTC QLQ-C30. Each of the 5 endpoints will be analyzed using a Bonferroni-adjusted 1% significance level and 99% confidence intervals, with a hierarchical testing procedure. For each endpoint, first Arm 2 vs 3 will be tested at the 1% level, and if significant, Arm 1 vs 3 will then be tested.

4.2.2 Primary endpoints

4.2.2.1 Overall survival

The OS primary endpoint will be analyzed, separately for each treatment comparison, using a stratified log-rank test (CCI [redacted]) adjusting for platinum therapy planned in Cycle 1 (carboplatin or cisplatin, as entered into IVRS at randomization). The effect of durvalumab +/- tremelimumab + EP versus EP treatment will be estimated by the HR together with its CI and p-value. The HR and its 95% and ([1-adjusted alpha] × 100%) CIs (with adjustments both without and with alpha recycling – see Appendix B) will be estimated from the Cox proportional hazards model (Cox 1972) (with ties=Efron and platinum therapy planned in Cycle 1 included in the STRATA statement) and the CI calculated using a profile likelihood approach.

Kaplan-Meier plot of OS will be presented by treatment arm. Summaries of the number and percentage of patients who have died, those still in survival follow-up, those lost to follow-up, those who have withdrawn consent and those censored for any other reason will be provided along with the median OS and its 95% CI 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 will 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 treatment-by-covariate interactions, which will be investigated.

Secondary analysis

A **secondary analysis** of OS will be performed to compare durvalumab + tremelimumab + EP versus durvalumab + EP. This analysis will be performed using the same methodology as for the primary endpoint described above. This will not be included in the multiple testing strategy.

Sensitivity analysis and additional supportive summaries

A **sensitivity analysis** for OS will examine the censoring patterns to rule out attrition bias, which is achieved by a Kaplan-Meier plot of time to censoring, where the censoring indicator of OS is reversed.

The number of subjects prematurely censored for OS, and baseline prognostic factors of the prematurely censored patients, 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:

- 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 patients) summarized in all patients regardless of treatment arm, as well as by treatment arm;
- Time from randomization to the date of censoring in censored patients only, presented by treatment arm.

Subgroup analyses

A **subgroup analysis** will be conducted comparing OS between treatments in the following subgroups of the FAS:

- Planned platinum therapy (carboplatin vs. cisplatin) – stratification factor
- Age at randomization (<65 vs. ≥ 65 years of age)
- Gender (male vs. female)
- AJCC Stage (III vs. IV)
- Performance status at baseline (normal vs. abnormal [considering all status except "Normal activity" as abnormal])
- Smoking status at screening (smoker [current or former] vs. non-smoker [never smoked]). This will be determined from the response to 'Nicotine Use Occurrence' (SU module) on the eCRF at screening. A patient is categorized as smoker if there exists a record in SU_NIC with any of the following options for "What type of substance was used?" 'Cigarettes', 'Cigarillos', 'Cigars', 'Pipe Tobacco', 'Tobacco for Smoking'.
- CNS metastasis at baseline (Y [metastatic site = 'Brain/CNS' or 'Other CNS'] vs. N)
- Race (Asian vs. non-Asian)
- Geographical region (Asia vs. Europe vs. North and South America)

The subgroup analyses for the stratification factor will be based on the values entered into the IVRS, all other factors will be based on values recorded on the eCRF as indicated above. Patients with missing data for a subgroup variable will be excluded from the analysis for that subgroup only.

For each subgroup, the HR and 95% CI will be calculated from an unstratified Cox proportional hazards model with treatment as the only covariate. These will be calculated for durvalumab ± tremelimumab + EP vs EP. The Cox models will be fitted using SAS® PROC PHREG with the Efron method to control for ties, using the BY statement to obtain HR and 95% CI for each subgroup level separately.

These hazard ratios and associated two-sided 95% CIs will be summarised and presented on a forest plot, along with the results of the overall primary analysis.

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 OS will not be formally analyzed. In this case, only descriptive summaries will be provided.

Other baseline variables may also be assessed if there is a clinical justification or if 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 factors. No adjustment to the significance level for testing will be made since all these analyses will be considered supportive of the primary analysis of OS.

Effect of covariates on HR estimate

Cox proportional hazards modeling will be employed to assess the effect of the stratification factor and covariates (listed above) on the HR estimate. A model will be constructed, containing treatment and the stratification factor alone, to ensure that any output from the Cox modeling is likely to be consistent with the results of the stratified log-rank test. The result from the initial model and the model containing additional covariates will be presented.

This analysis evaluates the treatment effect adjusting for any potential imbalances in baseline prognostic factors.

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

Only the stratification factor (planned platinum therapy) will be tested formally for interaction with treatment, using a likelihood ratio test. Any quantitative interactions identified using this procedure will then be tested to rule out any qualitative interaction using the approach of Gail and Simon (Gail and Simon 1985).

Impact of switching (crossover outside of this study) to immunotherapies (or other potentially active investigational agents) on OS analyses

Exploratory analyses of OS adjusting for the impact of subsequent immunotherapy or other investigational treatment may be performed, if a sufficient proportion of patients switch. Methods such as Rank Preserving Structural Failure Time (Robins and Tsiatis 1991), Inverse Probability of Censoring Weighting (Robins 1993) and other methods in development will be explored. The decision to adjust and the final choice of methods will be based on a blinded

review of the data and the plausibility of the underlying assumptions. Baseline and time-dependent characteristics will be explored, and summaries of baseline characteristics will be summarized by treatment group, splitting between those that have and haven't switched at the time of the analyses. Further detail will be provided in the Payer Analysis Plan. These analyses are intended to support reimbursement appraisals.

4.2.3 Secondary endpoints

4.2.3.1 Progression-free survival

The secondary endpoint PFS, based on site Investigator data, will be analyzed using a stratified log-rank test adjusting for platinum therapy planned in Cycle 1 (carboplatin or cisplatin). The effects of durvalumab + tremelimumab + EP versus EP treatment and durvalumab + EP versus EP treatment will be estimated by the HR together with its 95% CI and p-value for the FAS (see Section 4.2.2.1 for details). Analysis of PFS will also be performed to compare durvalumab + tremelimumab + EP versus durvalumab + EP. This will use the same methodology as described above, but will not be included in the multiple testing strategy.

Kaplan-Meier plot of PFS will be presented by treatment arm. Summaries of the number and percentage of patients experiencing a PFS event, the type of event (RECIST 1.1 or death) and the number and percentage of censored patients and detailed reason for censoring will be provided along with median PFS and its 95% CI for each treatment.

The assumption of proportionality will be assessed in the same way as for OS.

The treatment status at progression of patients at the time of analysis will be summarised. This will include the number (%) of patients who were on treatment at the time of progression, the number (%) of patients who discontinued study treatment prior to progression, the number (%) of patients 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 patients who have discontinued treatment.

Additional supportive summaries/graphs

In addition, the number of patients prematurely censored will be summarized by treatment group together with baseline prognostic factors of the prematurely censored patients. A patient 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 tumour assessment interval plus 2 weeks (i.e. 6+2 weeks if time period between randomization and DCO for that patient is 13 weeks or less; 8+2 weeks otherwise) prior to the DCO date.

Additionally, summary statistics will be given for the number of days from censoring to data cut-off for all censored patients.

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

Additionally, summary statistics for the number of weeks between the time of progression and the last evaluable RECIST 1.1 assessment prior to progression will be presented for each treatment group.

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

Summaries will be presented of the per-patient mean number of days between RECIST assessments, by treatment group.

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

Sensitivity analyses

- Evaluation-Time bias

A **sensitivity analysis** will be performed on the key secondary PFS analyses of durvalumab ± tremelimumab + EP versus EP treatment, 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 1.1 assessment will be analyzed using a log-rank test. For patients whose death was treated as a PFS event, the date of death will be used to derive the PFS time used in the analysis. Kaplan-Meier plots will be presented by treatment group. This approach has been shown to be robust to even highly asymmetric assessment schedules (Sun and Chen 2010).

- Attrition bias

Attrition bias will be assessed by repeating the key secondary PFS analyses except that the actual PFS event times, rather than the censored times, of patients who progressed or died in the absence of progression immediately following 2 or more non-evaluable tumor assessments will be included. This analysis will be supported by a Kaplan-Meier plot of the time to censoring where the censoring indicator of the PFS analysis is reversed.

- Subsequent anticancer therapy

The key PFS analyses will also be repeated with the adjustment that patients who take subsequent anticancer therapy (palliative/adjuvant radiotherapy excluded) prior to progression or death will be censored at their last evaluable assessment prior to taking the subsequent therapy.

Subgroup analyses

Subgroup analyses will be conducted for PFS comparing durvalumab ± tremelimumab + EP vs EP alone in the same way as previously specified for OS.

Effect of covariates on HR estimate

The same analysis as for OS will be done (see Section 4.2.2.1)

A forest plot illustrating the hazard ratio and 95% confidence interval will be provided to compare the secondary and subgroup/sensitivity analyses of progression free survival.

4.2.3.2 Objective response rate

The ORR will be based on the programmatically derived RECIST 1.1 using the site Investigator tumor data. The ORR will be compared between durvalumab + tremelimumab + EP vs EP using logistic regression models adjusting for the same stratification factor as the primary endpoint. The results of the analysis will be presented in terms of an odds ratio (an odds ratio greater than 1 will favor durvalumab + tremelimumab + EP) together with its associated profile likelihood 95% CI (CCI [redacted]) 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 a subset of the FAS, including all patients with measurable disease at baseline.

The same analysis will be performed to compare durvalumab +EP vs EP.

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 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)

Summaries will be produced that present the number and percentage of patients with a tumor response (CR/PR). Overall visit response data will be listed for all patients (ie, the FAS).

BOR

For each treatment group, BOR will be summarized by n (%) for each category (CR, PR, SD, PD, and NE). No formal statistical analyses are planned for BOR.

Change in tumor size

The absolute values and percentage changes in target lesion tumor size from baseline will be summarized using descriptive statistics and presented at each time point for each treatment group. The best change in target lesion tumor size from baseline, (where best change in target lesion size is the maximum reduction from baseline or the minimum increase from baseline in the absence of a reduction) will also be summarized and presented for each treatment group.

Tumor size will also be presented graphically using waterfall plots for each treatment group, to present each subject's best percentage change in 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 tumor size levels will be added to the plots, which correspond with the definitions of progression and 'partial' response respectively. On each of the waterfall plots the planned platinum therapy in Cycle 1 (carboplatin or cisplatin) for each patient will be indicated. Additional waterfall plots showing percentage change in tumor size at specific timepoints may be produced if it is felt that these are warranted to provide greater clarity.

The above outputs will be programmed for the site Investigator RECIST 1.1 assessments.

4.2.3.3 Proportion of patients alive and progression free at 6 and 12 months

The APF6 and APF12 (where 1 month equates to 30.4375 days) will be summarized (using the Kaplan-Meier curve) and presented by treatment group along with confidence intervals using the log-log transformation.

4.2.3.4 Proportion of patients alive at 18 months

The proportion of patients alive at 18 months (ie, OS18, using 18x30.4375 days) will be summarized (using the Kaplan-Meier curve) and presented by treatment group along with confidence intervals using the log-log transformation.

4.2.3.5 PK data

Pharmacokinetic concentration data of durvalumab and tremelimumab will be listed for each patient and each dosing day, and a summary by visit and timepoint provided for all evaluable patients in the PK analysis set. Peak and trough concentrations will be assessed from the visit summaries.

4.2.3.6 Immunogenicity analysis

A summary of the number and percentage of patients who develop detectable ADA to MEDI4736 or tremelimumab by ADA categories (Section 3.5.3) in different treatment arms will be presented. The denominator for percentage calculations will use the number of ADA-evaluable patients, defined as patients in the safety analysis set who have a non-missing baseline ADA and at least 1 non-missing post-baseline ADA result, in the treatment arm. ADA prevalence (ADA-positive at any visit) and ADA incidence (sum of treatment-induced ADA and treatment-boosted ADA) for each treatment group will be calculated. Immunogenicity results will be listed for all patients in the safety analysis set regardless of ADA-evaluable status. ADA titer and nAb data will be presented for samples confirmed positive for the presence of ADA to durvalumab and/or tremelimumab. The effect of ADA on PK, safety and efficacy will be examined by descriptive summaries if data allow. AEs in ADA positive patients by ADA positive category will be listed.

4.2.3.7 Patient reported outcomes

Treatment efficacy will be evaluated primarily on what patients and clinicians consider the primary symptoms of lung cancer (listed in Section 3.3.1). The assessments of cough, dyspnea (breathlessness) and chest pain as assessed by the EORTC QLQ-LC13 and fatigue and appetite loss from EORTC QLQ-C30 will be used as key secondary efficacy endpoints.

EORTC QLQ-C30 and QLQ-LC13

Summaries of compliance over time by timepoint and overall will be reported.

Time to deterioration will be summarized using the same methodology as described for the primary OS endpoint. Kaplan-Meier plots will be presented by treatment group. Summaries

of the number and percentage of patients who have an event as well as who were censored will be provided along with the medians for each treatment.

Summaries of original and change from baseline values of each symptom scale/item, the global health status/QoL score, and each functional domain (for QLQ-C30) and of each symptom/treatment-related side effect (for QLQ-LC13) will be reported by visit for each treatment group. Graphical presentations may also be produced as appropriate.

Summaries of the number and percentage of patients in each response category at each visit for each ordinal item (in terms of the proportion of patients in the categories of improvement, no change, and deterioration) will also be produced for each treatment group.

Mixed models repeated measures of change from baseline in PRO symptoms

In addition to the time to deterioration endpoints listed above the following longitudinal endpoints are of interest: QLQ-C30 fatigue and appetite loss; QLQ-LC13 dyspnea, cough and chest pain. These are not part of the main multiple testing procedure and are considered a separate set of PRO endpoints. A Bonferroni adjustment to the significance level will be applied to the tests described below to control the overall Type I error at the 5% level as described in Section 4.2.1.

Change from baseline in the primary PRO symptom scores (see Section 3.3.1) will be analyzed separately for each treatment comparison using a mixed model for repeated measures (MMRM) analysis with use of all data from baseline up to PD or 12 months. The analysis will be to compare the average treatment effect from the point of randomization until PD or 12 months (whichever is earlier) excluding visits with excessive missing data (defined as >75% missing data).

It is acknowledged that patients will discontinue treatment at different timepoints during the study and that this is an important time with regards to symptoms and HRQoL data collection. To account for this, and in order to include the discontinuation and follow-up assessments, a generic assessment time variable will be derived for each subject in order that the average treatment effect can be analyzed using the above method. Each visit will be assigned a sequential number. The time from randomization to each of these will be derived in order to select only those assessment times occurring within the first 12 months of randomization or until PD.

As an example, say a patient X has data collected at the first 4 scheduled assessments of a 3-weekly schedule for 4 cycles and 4-weekly schedule for later cycles and then discontinues treatment, whilst patient Y discontinues treatment after the first scheduled assessment, the first 6 generic assessment times would be as follows:

Generic visit	Study Day	
	Patient X	Patient Y
Baseline	Baseline	Baseline

1	22	22
2	43	35 (discontinuation)
3	64	64
4	85	85
5	110 (discontinuation)	113
6	141	141

The MMRM model will include treatment, age at randomization (<65 vs ≥65 years of age), sex (male vs female), smoking history (smoker vs non-smoker), visit and the interaction between treatment and visit interaction as fixed factors, baseline as a covariate and further adjusted for the interaction between baseline and visit. Restricted maximum likelihood (REML) estimation will be used. An overall adjusted mean estimate will be derived that will estimate the average treatment effect over visits giving each visit equal weight. For this overall treatment comparison, adjusted mean estimates per treatment group and corresponding 95% CIs will be presented along with an estimate of the treatment difference, 95% CI and p-value.

An unstructured covariance matrix will be used to model the within-subject error and the Kenward-Roger approximation will be used to estimate the degrees of freedom. If the fit of the unstructured covariance structure fails to converge, the following covariance structures will be tried in order until convergence is reached: toeplitz with heterogeneity, autoregressive with heterogeneity, toeplitz, autoregressive and compound symmetry.

Multiple imputation techniques for missing values may be considered to explore the robustness of any treatment effect.

These analyses will be performed in the FAS.

4.2.3.8 WHO/ECOG performance status

WHO/ECOG performance status data will be summarized over time for the FAS.

Data obtained up until 90 days following last dose of study treatment will be used for reporting. Visits containing data from ≥ 20 patients or >1/3 of patients within any treatment group will be presented.

4.2.3.9 Special data handling

Due to an unresolvable data issue (Appendix C), all ePRO data of PPD from site PPD will be excluded from all summary tables, figures and listings. Formal definition of study analysis sets: full analysis set (ITT population), safety analysis set and PK analysis set remain the same. A footnote will mention the exclusion of the data from the impacted tables, figures and listings.

4.2.4 Exploratory endpoints

4.2.4.1 Time from randomization to second progression

PFS2 will be analyzed using the same methodology as described for the secondary PFS endpoint but including 95% CI for HR, for durvalumab ± tremelimumab +EP vs EP. Summaries of the number and percentage of patients who have an event as well as who were censored will be provided along with the medians and 95% CI for each treatment.

This analysis will be performed in FAS.

Time from randomization to first subsequent therapy or death

For supportive purposes, the time to the start of subsequent therapy (palliative/adjuvant radiotherapy excluded) or death will be analyzed using the same methodology and model. The HR for the treatment effect together with its 95% CI will be presented. No multiplicity adjustment will be applied as these are viewed as supportive endpoints.

A summary table of first subsequent therapies by treatment group will be provided.

4.2.4.2 Healthcare resource use

The potential impact the disease and treatment have on health care resource use will be analyzed for the purposes of submissions to payers. Descriptive statistics (as appropriate, including means, median, ranges or frequencies, and percentages) will be provided for each arm on the different types of hospital admissions, the length of stay of people admitted into hospital for at least 1 overnight stay, and length of stay of people admitted to intensive care/high dependency units, as well as the primary sign or symptom the patient presents with. To support submissions to payers, additional analyses may be undertaken, and these will be outlined in a separate payer analysis plan.

4.2.4.3 Patient reported outcomes

PRO-CTCAE

Data from the PRO-CTCAE will be summarized using the FAS. The number (%) of patients with each level of response for each CTCAE item at baseline and over time will be summarized. Further summaries to explore the data (ie, the severity of symptoms) may be produced.

Patients' Global Impression of Change

PGIC data, considered as categorical data, will be presented using summaries and descriptive statistics by visit and overall based on the FAS.

EQ-5D-5L

Summaries of compliance over time by timepoint and overall will be reported.

Descriptive statistics will be reported for health state utility values and the visual analogue scale by visit, as well as the change in these scores from baseline. To support future economic

evaluations, additional appropriate analyses may be undertaken (eg, mean health state utility pre- and post-treatment, and pre- and post-progression) and will be outlined in the payer analysis plan.

4.2.4.4 PK data

Individual PK concentrations of EP treatments will be listed for each patient by time point.

4.2.4.5 Other exploratory endpoints

Remaining exploratory analysis described in the CSP may be reported outside the main CSR, including PK/PD, biomarkers (PD-L1, TMB, IFN-γ, DLL3), DNA. A separate SAP will be prepared where necessary.

4.2.5 Safety data

Safety and tolerability data will be presented using the safety analysis set. Safety data will be summarized only. No formal statistical analyses will be performed on the safety data.

Data from all cycles of treatment will be combined in the presentation of safety data. AEs (both in terms of MedDRA preferred terms and CTCAE grade) will be listed individually by patient. The number of patients experiencing each AE will be summarized by treatment group and CTCAE grade.

Other safety data will be assessed in terms of physical examination, clinical chemistry, hematology, vital signs, and ECGs. Exposure to durvalumab + tremelimumab combination therapy, durvalumab monotherapy, and EP will be summarized. Time on study, dose delays/interruptions in durvalumab, tremelimumab and EP treatments, and EP dose reductions will also be summarized.

The following sections describe the planned safety summaries. However, additional safety tables may be required to aid interpretation of the safety data.

Adverse Events

All AEs, both in terms of current MedDRA preferred term and CTCAE grade, will be summarized descriptively by count (n) and percentage (%) for each treatment group. The majority of the AE summaries, unless stated otherwise, will be based on TEAEs. Any AE occurring before study treatment (ie, before the administration of the first dose) and without worsening after initiation of study treatment will be included in the AE listings, but will not be included in the summary tables (unless otherwise stated). However, any AE occurring before the administration of the first dose on Study Day 1 that increases in severity after the first dose will be regarded as treatment emergent and thus will be included in the majority of summary tables.

AEs observed up until 90 days following last dose of study treatment (durvalumab, tremelimumab or EP) or until the initiation of the first subsequent anticancer therapy (excluding palliative/adjuvant radiotherapy) following discontinuation of treatment (whichever occurs first) will be used for reporting of all of the AE summary tables. This will

more accurately depict AEs attributable to study treatment only as a number of AEs up to 90 days following last dose of study treatment are likely to be attributable to subsequent therapy.

However, to assess the longer term toxicity profile, all of the AE summaries may also be produced containing AEs observed up until 90 days following last dose of study treatment (ie, without taking subsequent anticancer therapy into account).

A summary table with AE which started prior to first dose or after 90 days following last dose of study treatment will be produced.

A selection of AE summaries may also be produced containing AEs (by system organ class and preferred term) observed from the initiation of the first subsequent anticancer therapy following discontinuation of study treatment until 90 days following last dose of study treatment (ie, summarizing those AEs experienced by patients 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.

Frequencies and percentages of patients reporting each preferred term will be presented (ie, multiple events per patient will not be accounted for).

Summary information (the number and percent of patients by system organ class and preferred term separated by treatment group) will be tabulated for:

- All AEs (*)
- All AEs causally related to any study treatment (as determined by the reporting investigator) (*)
- All AEs by any reported CTCAE grade
- Any CTCAE grade 3 or 4 (*)
- Any CTCAE grade 3 or 4 causally related to any study treatment (as determined by the reporting investigator) (*)
- AEs with outcome of death (*)
- AEs with outcome of death causally related to any study treatment (as determined by the reporting investigator) (*)
- AEs by outcome
- All SAEs (*)
- All SAEs causally related to any study treatment (as determined by the reporting investigator) (*)
- SAEs leading to discontinuation of any study treatment

- SAEs leading to discontinuation of any study treatment, causally related to any study treatment (as determined by the reporting investigator)
- AEs leading to discontinuation of any study treatment (*)
- AEs leading to discontinuation of any study treatment, causally related to any study treatment (as determined by the reporting investigator) (*)
- AEs leading to discontinuation of Etoposide
- AEs leading to discontinuation of Carboplatin/Cisplatin
- AEs leading to dose delay/interruption of any study treatment (*)
- Other significant AEs
- Immune mediated AEs (as determined by the reporting investigator) (*)
- Infusion reaction AEs (as determined by the reporting investigator) (*)

An overall summary table will include the number and percentage of patients in each selected category (*). In addition, a truncated AE table of most common AEs and another table showing most common AEs with any CTCAE grade 3 or 4, showing all events that occur in at least 2% of patients in any treatment group will be summarized by preferred term, by decreasing frequency. This cut-off may be modified after review of the data. When applying a cut-off (ie, x %), the raw percentage should be compared to the cut-off, no rounding should be applied first (ie, an AE with frequency of 1.9% will not appear if a cut-off is 2%).

Each AE event rate (per 100 patient years) will also be summarized by preferred term within each system organ class for the output summarizing all AEs. For each preferred term, the event rate is defined as the number of patients with that AE divided by the total drug exposure of patients at risk of AE. The denominator is calculated as the total over each patient of days from first dose to the last day of study treatment.

Fluctuations observed in CTCAE grades during study will be listed for those AEs which are CTCAE $\geq$3 at any time.

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 drug. Any events in this period that occur after a patient has received further therapy for cancer (following discontinuation of study treatment) will be flagged in the data listings.

Deaths

A summary of all deaths will be provided with number and percentage of patients, categorized as:

- Total number of deaths

- Death related to disease under investigation ONLY, as determined by investigator
- Death related to disease under investigation, as determined by the investigator, and with AE with outcome of death (with further subcategories for AEs with onset/worsening date prior to, and AEs with onset date after initiation of subsequent anticancer therapy)
- AE with outcome of death ONLY(with further subcategories for AEs with onset/worsening date prior to, and AEs with onset date after initiation of subsequent anticancer therapy)
- Death occurred more than 90 days after the date of last dose of study treatment, and unrelated to AE or disease under investigation (*)
- Unknown reason for death
- Other deaths.

This summary will be produced twice; firstly for all deaths, and secondly for all deaths up until 90 days following last dose of study treatment. Hence the category maked (*) will only appear in the first summary.

Adverse events of special interest

Preferred terms used to identify adverse events of special interest (as defined in Section 3.4.1) will be listed before DBL and documented in the Trial Master File. Groupings will be based on preferred terms provided by the medical team prior to DBL, and a listing of the preferred terms in each grouping will be provided. Summary tables of AESI categories will be produced and will also show the individual preferred terms which constitute each AESI grouping.

Summaries of these grouped AE categories will include number (%) of patients who have:

- At least 1 AESI
- At least 1 AESI by CTCAE grade
- At least 1 AESI presented by outcome
- At least 1 AESI causally related to any study treatment (as determined by the reporting investigator) by CTCAE grade
- At least 1 AESI leading to discontinuation of any study treatment

An overall AESI summary will be presented, including the number and percentage of patients in each of these categories.

A summary of total duration (in days) of AESI will be provided for events which have an end date and this may be supported by summaries of ongoing AESIs at death and, separately, at

data cut-off. Summary statistics showing the time to onset of the first AESI may also be presented as appropriate.

Additionally, there will be several summaries of AESIs requiring concomitant treatment, and particularly the relationship of AESIs to the use of immunosuppressive agents (ie, depicting which AESI triggered immunosuppressive use) and, separately, to the use of immunosuppressive agents at high doses.

Infection Adverse events

Infection AEs will be summarized by pooled terms and PTs in two ways: (1) using MedDRA HLGT/HLT pooled terms (2) Custom pooled terms. The following summaries will be reported for both HLGT/HLT pooled terms and custom pooled terms and PTs:

- Infection AEs (including event rate)
- Infection AEs by CTCAE grade
- Serious Infection AEs
- Infection AEs presented by outcome
- Infection AEs of CTCAE grade 3 or 4
- Infection AEs with outcome of death
- Infection AEs leading to discontinuation of any study treatment
- Infection AEs leading to dose delay/interruption of any study treatment

Overall Infection AE summaries will be presented, including the number and percentage of patients in each of these categories.

Immune-mediated Adverse events (imAEs)

The imAEs (as classified by the Sponsor) will be summarized in the same manner as for the summaries for AESI described above. See further details in the imAE Charter with respect to derivation rules.

Summary of long term tolerability

To assess long term tolerability, provided that there are a sufficient number of patients 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 patients. The prevalence at time t after first dose of study treatment is calculated as the number of patients experiencing the event divided by the number of patients receiving

study treatment or in safety follow-up at time t; generally, t is categorized by each day after dosing. The prevalence will be plotted over time and presented for each treatment group separately. Multiple occurrences of the same event are considered for each patient but a patient is only counted in the numerator whilst they are experiencing 1 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 patients 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 (eg, 0-1 months, 1-3 months, 3-6 months, etc.) given that the patient reaches that time period without having an event is plotted for each time period.

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 patient 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 patients had completed safety follow-up (Pintilie 2006).

Laboratory assessments

Summaries of laboratory assessments will include the variables listed in Tables 4, 5 and 6 of the CSP. Calculated creatinine clearance values reported by the site (e.g. at the screening visit) will be summarized separately from programmatically derived CrCl.

Data obtained up until the 90 days following last dose of study treatment or until the initiation of the first subsequent therapy following discontinuation of treatment (whichever occurs first) will be used for reporting. This will more accurately depict laboratory toxicities attributable to study treatment only as a number of toxicities up to 90 days following last dose of study treatment are likely to be attributable to subsequent therapy.

A small selection of summaries of laboratory data may also be produced containing data from initiation of the first subsequent therapy following discontinuation of study treatment until 90 days following last dose of study treatment (ie, summarizing the laboratory data collected on patients taking subsequent therapy during the safety collection follow-up window post last dose 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 post 90 days after last study treatment dose will not be summarized.

Data summaries will be provided in International System (SI) of units.

For continuous laboratory assessments, absolute value and change from baseline will be summarized using descriptive statistics at each scheduled assessment time by treatment group.

Shift tables for laboratory values by worst CTC grade on-treatment will be produced, and for specific parameters separate shift tables indicating directionality of change (i.e. change to low

or high) will be produced. The laboratory parameters for which CTC grade shift outputs will be produced are:

- Hematology: hemoglobin, leukocytes, lymphocytes (absolute count, low and high), neutrophils (absolute count), platelets
- Clinical chemistry: ALT, AST, ALP, total bilirubin, albumin, magnesium (low and high), sodium (low and high), potassium (low and high), corrected calcium (low and high), glucose (low and high), creatinine

For the parameters with no CTCAE grading that are listed in the CSP, shift tables from baseline to minimum and maximum values on-treatment will be provided.

Additional summaries will include a shift table for urinalysis (bilirubin, blood, glucose, ketones, protein) comparing baseline value to maximum on-treatment value.

Liver Enzyme Elevations and Hy's law

The following summaries will include the number (%) of patients who have:

- Elevated ALT, AST, and Total bilirubin during the study
  - ALT ≥3x – ≤5x, >5x – ≤8x, >8x - ≤10x, >10x - ≤20x, and >20x Upper Limit of Normal (ULN)
  - AST ≥3x – ≤5x, >5x – ≤8x, >8x - ≤10x, >10x - ≤20x, and >20x ULN
  - Total bilirubin ≥2x - ≤3x, >3x - ≤5x, and >5x ULN
  - ALT or AST ≥3x - ≤5x, >5x - ≤8x, >8x - ≤10x, >10x - ≤20x, and >20x ULN
  - ALT or AST ≥3x ULN and total bilirubin ≥2x ULN (Potential Hy's law): The onset date of ALT or AST elevation should be prior to or on the date of total bilirubin elevation
- Narratives will be provided in the CSR for patients who have ALT ≥3x ULN plus total bilirubin ≥2x ULN or AST ≥3x ULN plus total bilirubin ≥2x ULN at any visit.

Data from start dose up until the 90 days following last dose of study treatment or until the initiation of the first subsequent therapy following discontinuation of treatment (whichever occurs first) will be included.

Liver biochemistry test results over time for patients with elevated ALT or AST (ie, ≥3x ULN), and elevated total bilirubin (ie, ≥2x ULN) (at any time) will be plotted. Individual patient data where ALT or AST (ie, ≥3x ULN) plus total bilirubin (ie, ≥2x ULN) are elevated at any time will be listed also.

Plots of ALT and AST vs. total bilirubin by treatment group will also be produced with reference lines at 3x ULN for ALT, AST, and 2x ULN for Total bilirubin. In each plot, Total bilirubin will be in the vertical axis.

Assessment of Thyrotoxicity

The analysis of thyroid function tests will be based on data up to 90 days after the last dose of study medication or date of initiation of subsequent therapy (whichever occurs first).

Absolute value and change from baseline will be summarized using descriptive statistics at each scheduled assessment time.

Shift tables showing baseline to maximum and baseline to minimum values will be produced for TSH, free T3 and freeT4.

Vital signs

Vital signs data obtained up until the 28-day safety follow-up visit will be included in the summary tables.

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

Physical examination

All individual physical examination data will be listed only.

Other Safety Data

Data from positive pregnancy tests will be listed only.

4.2.6 Demographic and baseline characteristics data

The following will be summarized for all patients in the FAS (unless otherwise specified):

- Patient disposition (including screening failures and reason for screening failure) – all patients
- Important protocol deviations
- Inclusion in analysis sets – all patients
- Demographics (derived age, age group [<50, ≥50 - <65, ≥65 - <75 and ≥75 years], sex, race and ethnicity)
- Patient characteristics at baseline (height, weight, weight group [<70, 70-90, >90 kg], body mass index [BMI] and BMI group)
- Patient recruitment by region, country and center
- Previous treatment modalities
- Disease characteristics at baseline*/ diagnosis (WHO performance status*, primary tumor location, histology type and AJCC stage)
- Extent of disease at baseline
- Pathology at diagnosis with primary tumor, regional lymph nodes and distant metastases TNM classification
- Relevant medical history (past and current)
- Relevant surgical history
- Disallowed concomitant medications
- Allowed concomitant medications
- Nicotine use, categorized (non-smoker, ex-smoker, current smoker) and number of pack years for type of each substance used
- Stratification factor as per the Interactive Voice Response System (IVRS) and corresponding eCRF data describing actual platinum therapy received in cycle 1.

The medications will be coded following AZ standard drug dictionary / WHO Drug dictionary as applicable.

4.2.7 Treatment exposure

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

- Number of infusions received.
- Total exposure of each molecule and overall.
- Actual exposure of durvalumab and tremelimumab.
- Total number of cycles of each treatment received (i.e., cycles at least 1 dose).
- Number of patients receiving ≥1, ≥2, ≥3, ≥4, ≥5, ≥6 cycles of each molecule.
- Number of patients who received the planned starting dose
- Number of, and reasons for, dose delays and interruptions.
- Number of, and reasons for dose reductions of EP combination.
- RDI (relative dose intensity) of durvalumab and tremelimumab.
- Cumulative exposure over time (≥0, ≥3, ≥6, ≥9, ≥12, ≥16 etc…) for each molecule, considering that EP combination can be given for 2 additional q3w cycles on Week 12 and Week 15 for patients in Arm 3.

Numbers of, and reasons for, dose delays and interruptions will be derived from the exposure eCRF items Action Taken = 'Dose interrupted' and Treatment cycle delayed = 'Yes'.

Dose reductions will be identified based on Exposure eCRF item Action Taken='Dose reduced'.

For etoposide, dose delay, dose interruption and dose reduction descriptions are based on the first administration within a cycle.

In these summaries, the molecules involved in EP combination will be presented separately.

For patients on study treatment at the time of the interim OS analysis, the DCO date will be used to calculate exposure.

4.2.8 Subsequent Therapy

Subsequent therapies (chemotherapy or immunotherapy) received after discontinuation of study treatment will have summaries produced by treatment group. A separate summary will be produced for radiotherapy received after discontinuation of study treatment. Therapy received on the same day as discontinuation of study treatment will be considered to be subsequent therapy.

5. INTERIM AND FOLLOW-UP ANALYSES

5.1 Interim Analysis

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

An OS interim analysis is planned for durvalumab + tremelimumab + EP versus EP as well as for durvalumab + EP versus EP. This analysis will be performed by an IDMC.

The interim analysis of OS will occur when approximately 318 OS events have occurred (60% maturity) in the durvalumab + tremelimumab + EP and EP treatment arms and approximately 318 OS events have occurred (60% maturity) in the durvalumab + EP and EP treatment arms.

The Lan DeMets spending function that approximates an O'Brien Fleming approach will be used to account for multiplicity introduced by including an interim analysis for superiority (Lan and DeMets 1983). The alpha level allocated to OS test family will be controlled at the interim and primary time points. The alpha level applied at the interim depends upon the proportion of information available, that is, the actual observed number of deaths and percent of maturity for OS at the time of interim analysis. A separate Lan DeMets (O'Brien Fleming) spending function accounting for an interim and final analysis will also be applied to PFS endpoints in the MTP, and the alpha level at the interim will depend upon the proportion of PFS information available, that is, the actual observed number of progression events and percent of maturity for PFS at the time of interim analysis. A detailed calculation of alpha allocation at IA and FA is provided in Appendix B.

The criterion for superiority is a statistically significant improvement in OS at the interim analysis.

- If the OS results indicate superiority, then analyses of other endpoints will be performed as appropriate, and the results of these analyses will form the basis for submissions for regulatory approval.

- Whether or not the OS IA results indicate superiority, patients will continue to be followed for survival until approximately 425 OS events have occurred across the durvalumab + tremelimumab + EP and EP treatment arms (80% maturity) and approximately 425 OS events have occurred across the durvalumab + EP and EP treatment arms (80% maturity) when the final analysis will be performed.

5.2 Independent Data Monitoring Committee

This study will use an external IDMC to assess ongoing safety analyses, and to perform the formal interim efficacy analysis. The committee will review the safety data after the first 30 patients have been randomized and had 21 days of follow-up, and again after the next 30 patients have been randomized and had 21 days of follow-up. The IDMC will then meet at least every 6 months thereafter. 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/MedImmune and do not have any major 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/MedImmune. 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 Patient Safety Department. Issues identified will be addressed; this could involve, for instance, amendments to the clinical study protocol and letters to investigators.

In addition, safety review of Japanese and Chinese patients will take place to support their entry into the global study as well as supporting their registration:

- For Japanese patients: when the first 12-15 patients have been randomized and have completed the 1st cycle of treatment and had 21 days of follow-up,

- For Chinese patients: when the first 30 patients have been randomized and have completed the 1st cycle of treatment and had 21 days of follow-up.

The recommendations from the IDMC will not reveal the results of the analyses but will take the form of "Continue/Modify/Recommend early submission/Stop."

Details of the IDMC plan and communication process is provided in the IDMC Charter.

5.3 Long-term follow-up Analysis

Following the reporting of the planned interim analysis, and prior to the final analysis DBL, a CSP amendment was implemented for limited assessments to be collected on ongoing patients for assessment of long-term follow-up. Assessments include survival, SAEs, performance status, hospitalizations and exposure.

The analysis, to be reported in a CSR addendum, is planned to be based on a DCO approximately 1 year after the Global final analysis DCO, to achieve a median survival follow-up (in censored patients) of approximately 3 years. It will include all 3 treatment arms, and will constitute an updated summary of:

- the primary OS analysis (including subgroup analysis). This analysis will not be subject to formal statistical testing, but p-values and 95% confidence intervals will be presented.

- key safety outputs, including durvalumab exposure, subsequent anticancer therapies, SAEs, and deaths.

- WHO/ECOG performance status and hospital resource use.

Updated disposition and important protocol deviations will also be summarized, and an additional listing of protocol deviations due to the COVID-19 pandemic will be generated. Depending on the extent of any impact, summaries of data relating to impact of COVID-19 on study conduct (in particular missed visits, delayed or discontinued IP, and other protocol deviations) may be generated.

6. CHANGES OF ANALYSIS FROM PROTOCOL

Table 8 Details of changes from the clinical study protocol

Analysis in Protocol (5.0 29Nov2018)	Analysis in SAP	Rationale for change
For the OS endpoint, there is 1 IA planned, and the alpha level will be controlled at the interim and primary analysis timepoints by using the Lan DeMets (Lan and De Mets, 1983) spending function that approximates an O'Brien Fleming approach.	A separate Lan DeMets (O'Brien Fleming) spending function accounting for an interim and final analysis will also be applied to PFS endpoints in the MTP	To ensure strong control of type I error.
Subgroup analysis	Changed "ethnicity" to "race"	Considered to be a more important subgroup for exploratory subgroup analysis.
For AEs, on treatment (or treatment-emergent AEs) will be defined as any AEs that started after dosing or prior to dosing and which worsens following exposure to the treatment.	A treatment emergent adverse event (TEAE) is an AE with an onset date, or a pre-existing AE worsening, during the 'on-treatment' period as defined above.	Clarification of safety follow-up ("On treatment") period applicable to AEs.
Change from baseline in cough, hemoptysis, dyspnea and chest pain as assessed by the EORTC QLQ - LC13 and insomnia, fatigue and appetite loss as assessed by the EORTC QLQ -C30 will be the primary analysis and assessment of PRO outcome measures	The assessments of cough, dyspnea (breathlessness) and chest pain as assessed by the EORTC QLQ-LC13 and fatigue and appetite loss from EORTC QLQ-C30 will be used as key secondary efficacy endpoints	The importance of hemoptysis and insomnia symptoms has reduced since protocol development, so these are removed from the MMRM analysis
PK data	PK analysis simplified	Due to sparse PK sampling, non-compartmental parameters are not expected.

7. REFERENCES

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.

Cox 1972
Cox DR. Regression models and life-tables. J Royal Stat Society 1972;Series B 34(2):187-220.

Dworkin et al., 2008
Interpreting the Clinical Importance of Treatment Outcomes in Chronic Pain Clinical Trials: IMMPACT Recommendations. The Journal of Pain, Vol 9, No 2 (February), 2008: pp 105-121

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.

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 2009
Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. Stat Med 2009;29:219-28.

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

Oemar and Janssen 2013
Oemar M, Oppe M. EQ-5D-3L User Guide: Basic information on how to use the EQ-5D-3L instrument V5.0 (October 2013). Available from: URL: http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-3L_UserGuide_2013_v5.0_October_2013.pdf. Accessed 07 January 2014.

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.

Robins 1993
Robins JM. Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers. Proceedings of the Biopharmaceutical Section, American Statistical Association 1993; 24-33.

Robins and Tsiatis 1991
Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics-Theory and Methods 1991; 20(8):2609-31.

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.

8. APPENDIX

Appendix A Visit windows

For example, the visit windows for vital signs data (with assessments performed on Day 1 and Day 8 for each of the four 3-week cycles during the chemotherapy period, and Day 1 of each cycle during the post-chemotherapy period up to PD) are:

C1D8: Day 8, visit window 2 – 14

C2D1: Day 22, visit window 15 – 25

C2D8: Day 29, visit window 26 – 35

C3D1: Day 43, visit window 36 – 46

C3D8: Day 50, visit window 47 – 56

C4D1: Day 64, visit window 57 – 67

C4D8: Day 71, visit window 68 – 77

C5D1: Day 85, visit window 78 – 98

C6D1: Day 113, visit window 99 – 126

C7D1: Day 141, visit window 127 – 154

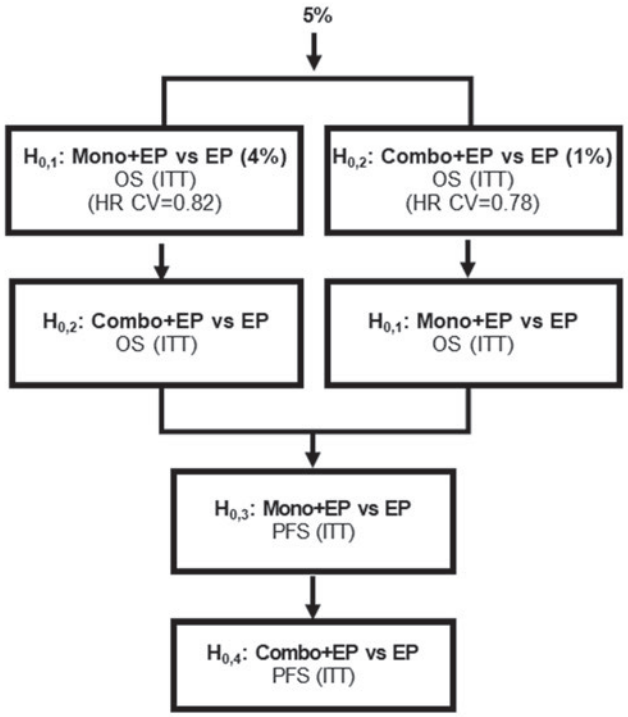
Note: The data included in these windows might not necessarily come from the cycle implied by the nominal visit label, e.g. for patients who have discontinued treatment. Also, due to the differing assessment schedules the visit windows will be different for the different endpoints.

Appendix B Alpha spending

INTRODUCTION

Per protocol (MTP attached below), the mono + EP vs. EP OS test has an overall alpha 4%, and the combo + EP vs. EP OS test has an overall alpha 1%. Both tests require 425 events in the full analysis set for the final analysis. There is one IA for OS, at 75% of target events.

For OS we use the Lan DeMets spending function that approximates an O'Brien Fleming approach to calculate the alpha split for IA and final analysis. The alpha calculation should use the actual observed number of events based on the DBL data. For PFS endpoints in the MTP we use a separate Lan DeMets (O'Brien Fleming) spending function to calculate the alpha split for IA and final analysis of PFS. The alpha calculation should use the actual observed number of PFS events based on the DBL data.



Note: Alpha recycling between Mono+EP vs EP and Combo+EP vs EP OS comparisons
Mono+EP vs EP = comparison of durvalumb+EP vs EP
Combo+EP vs EP = comparison of durvalumab + tremelimumab + EP vs EP

INSTRUCTIONS FOR IA

Alpha Level Calculation:

The alpha calculation process is summarized below:

Step 1: For initial testing of these 2 multiple primary OS analyses, use the corresponding number of events to calculate the alpha split independently. The information fraction for

combo + EP vs. EP for OS test is: # events from combo + EP and EP arms /425 for IA and 1 for final, and the alpha level is 1%. The information fraction for Mono + EP vs. EP is: # events from Mono + EP and EP arms /425 for IA and 1 for final, and the alpha level is 4%.
Step 2: For the 2nd level MTP, which is an alpha recycling step, repeat Step 1 with the same information fraction but an alpha level 5%.
Step 3: For the 3rd level MTP test which is testing PFS for Mono + EP vs. EP, the overall alpha can be 0% or 5%, depending on the primary test results. The information fraction for Mono + EP vs. EP is: # PFS events from Mono + EP and EP arms /477 for IA and 1 for final.
Step 4: For the 4th level MTP test which is testing PFS for combo + EP vs. EP, the overall alpha can be 0% or 5%, depending on the prior MTP test results. The information fraction for combo + EP vs. EP for PFS test is: # events from combo + EP and EP arms /477 for IA and 1 for final.

Software: EAST or R gsDesign or PROC SEQDESIGN in SAS can be used to calculate the adjusted alpha level.

Documentation: EAST screenshots to be documented (including parameter setting and efficacy boundaries calculation) if EAST is used.

Applications: the corresponding alpha-adjusted CI and 95% CI will be reported for the OS and PFS endpoints.

Confidence Interval of HR:

All the OS tests included in the MTP will summarize the corresponding HR CIs at the assigned alpha level, i.e., 1 − 2-sided alpha level **without** alpha recycling, 1 − 2-sided alpha level **with** alpha recycling, and 1 − 2-sided 5% level (i.e., 95% CI). All the tests that are not included in the MTP will summarize the 95% CI. For example, under protocol assumption:

- For the 2 primary OS endpoints, if the alpha level calculated for IA from above Step 1 is 0.23% for combo + EP vs. EP and 1.43% for mono + EP vs. EP, thus a 99.77% (1-0.23%) CI for combo + EP vs. EP and a 98.57% CI for mono + EP vs. EP will be provided. Using the alpha levels calculated from Step 2, the corresponding CI for both comparisons is a 98.09% CI at IA (dependant on the information fraction for each comparison). For both comparisons, a 95% CI will also be calculated.
- The tests included in the 3rd and 4th level of MTP will have 2 possible scenarios according to the co-primary endpoints results, 0% or 5%. If the alpha level calculated for IA from above Step 3 or 4 is 4.26%, a 95.74% CI will be reported, in addition to 95% CIs.
- For the OS/PFS tests that are not included in MTP, only 95% CI will be reported.

CCI [redacted]

CCI

[redacted]

Appendix C Note to file site PPD ePRO data

AstraZeneca MedImmune A member of the AstraZeneca Group	**Study Number: D419QC00001**	PPD
	Country (if applicable): PPD	
	Study Site No (if applicable): PPD	

Title/Reference: PIN Confidentiality Issue related to ePRO at site PPD that resulted in exclusion of data from analysis CCI

Summary:

PPD

PPD

Action: Due to quality concerns and different versions of the situation that were provided by SMM and investigators, it was agreed by Quality and Global Study Team that data from all subjects from this site will be excluded from the analysis. It was impossible to confirm whether this issue affected only two subjects or more as it was not clear whether investigator knew PIN or not.

Impacted subjects:

PPD

PPD

PPD

Global Study Leader

31 July 2018

Full Name, Position

Date

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: d419qc00001-sap-ed-4		
Document Title:	Statistical Analysis Plan Edition 4	
Document ID:	CCI [redacted]	
Version Label:	4.0 CURRENT LATEST APPROVED	
Server Date (dd-MMM-yyyy HH:mm 'UTC'Z)	**Signed by**	**Meaning of Signature**
15-Oct-2020 09:20 UTC	PPD [redacted]	Qualified Person Approval
15-Oct-2020 09:18 UTC	PPD [redacted]	Author Approval

CCI [redacted]