Official Protocol Title:	A Randomized, Double-Blind, Phase III Study of Platinum+ Pemetrexed Chemotherapy with or without Pembrolizumab (MK-3475) in First Line Metastatic Non-squamous Non-small Cell Lung Cancer Subjects (KEYNOTE-189)
NCT number:	NCT03950674
Document Date:	16-Nov-2017

Supplemental Statistical Analysis Plan (sSAP)

TABLE OF CONTENTS

1.0 INTRODUCTION

This supplemental statistical analysis plan (sSAP) is a companion document to the protocol. In addition to the information presented in the protocol SAP which provides the principal features of analyses for this trial, this sSAP provides additional statistical analysis details/data derivations and documents modifications or additions to the analysis plan that are not ‘principal’ in nature and result from information that was not available at the time of protocol finalization. Compared to the protocol SAP, this sSAP contains more details on PRO analyses for inclusion in the clinical study report of the protocol.

2.0 SUMMARY OF CHANGES

This sSAP aligns with the protocol amendment v07 for the global study with regard to statistical analysis plan.

3.0 ANALYTICAL AND METHODOLOGICAL DETAILS FOR GLOBAL STUDY

3.1 Statistical Analysis Plan Summary

This section contains a brief summary of the statistical analyses for this trial. Full detail is Sections 3.2-3.12.

Study Design Overview	A Phase 3 Study of Platinum+Pemetrexed Chemotherapy with or without Pembrolizumab (MK-3475) in First Line Metastatic Nonsquamous Non-small Cell Lung Cancer Subjects (KEYNOTE-189)
Treatment Assignment	Approximately 570 subjects will be randomized in a 2:1 ratio to receive pembrolizumab or saline placebo in combination with pemetrexed/platinum. Stratification factors are in Section 5.4 of the protocol. This is a randomized double-blinded study.
Analysis Populations	Efficacy: Intention to Treat (ITT) Safety: All Subjects as Treated (ASaT)
Primary Endpoints/Hypotheses	Progression-free Survival (PFS) per RECIST 1.1 assessed by BICR Overall Survival (OS)
Statistical Methods for Key Efficacy Analyses	The primary hypotheses for PFS and OS will be evaluated by comparing pembrolizumab to saline placebo in combination with pemetrexed/platinum using a stratified Log-rank test. The HR will be estimated using a stratified Cox regression model. Event rates over time will be estimated within each treatment group using the Kaplan-Meier method.
Statistical Methods for Key Safety Analyses	The analysis of safety results will follow a tiered approach. There are no Tier 1 safety parameters in this trial. All safety parameters are considered either Tier 2 or Tier 3. Tier 2 parameters will be assessed via point estimates with 95% confidence intervals provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters. The between-treatment difference will be analyzed using the Miettinen and Nurminen method. In the primary safety comparison, subjects who crossover to pembrolizumab are censored at time of crossover (ie, AEs occurring during treatment with pembrolizumab are excluded for control-arm subjects). An exploratory safety analysis will be conducted for the crossover population including all safety events starting from the date of first dose of pembrolizumab.
Interim Analyses	Two interim analyses are planned in this study. Results will be reviewed by an external data monitoring committee. Details are provided in Section 3.8.

	Interim analysis 1 (IA1) o Timing: To be performed after target number of PFS events (~370) are observed o Purpose: To demonstrate superiority of pembrolizumab in combination with pemetrexed/platinum in PFS and OS. ORR will be tested after superiority of pembrolizumab in combination with pemetrexed/platinum is demonstrated in PFS and OS. Interim analysis 2 (IA2) o Timing: To be performed after ~468 PFS events are observed o Purpose: To demonstrate superiority of pembrolizumab in combination with pemetrexed/platinum in PFS and OS.
Final Analysis	Final analysis (FA) o Timing: To be performed after target number of deaths (~416) are observed o Purpose: To demonstrate superiority of pembrolizumab in combination with pemetrexed/platinum in OS.
Multiplicity	The overall Type I error rate for each endpoint in the group sequential tests is strictly controlled at 2.5% (one-sided); for both PFS and OS, this is based on the Lan-DeMets O'Brien-Fleming spending function (see Section 3.9). Between the endpoints, the type I error is controlled by the following rollover rule. The total type I error allocated to PFS (0.0095) is subject to rollover to OS if the PFS test is positive. The type I error allocated to OS (0.0155) is subject to rollover to PFS if the OS test is positive. Furthermore, the total type I error (0.025) is subject to rollover to ORR at IA1 if the PFS and OS tests are both positive.
Sample Size and Power	Enrollment of 570 subjects is assumed to occur over 12 months at 2:1 ratio between the experimental and control groups. The actual enrollment is 616 subjects within 13 months. With 370 PFS events at IA1, the study has ~72% power for detecting a PFS HR of 0.7 at 0.0095 (one-sided) and ~84% power for detecting a HR of 0.7 at 0.025 (one-sided). With 468 PFS events at IA2, the study has ~90% power for detecting a HR of 0.7 at 0.0095 (one-sided) and ~96% power for detecting a HR of 0.7 at 0.025 (one-sided). The duration of PFS in the control group is assumed to follow an exponential distribution with a median of 6.5 months based on historical data. The assumed follow-up time after last patient enrolled is 13 months for IA2. An exponential dropout rate of 0.35% per month is assumed. With 242 deaths at IA1, the study has ~37% power for detecting an OS HR of 0.7 at 0.0155 (one-sided) and ~47% power for detecting a HR of 0.7 at 0.025 (one-sided) when the PFS test is significant. With 332 deaths at IA2, the study has ~73% power for detecting a HR of 0.7 at 0.0155 (one-sided) and ~80% power for detecting a HR of 0.7 at 0.025 (one-sided) when the PFS test is significant. With 416 deaths at FA, the study has ~90% power for detecting a HR of 0.7 at 0.0155 (one-sided) and ~93% power for detecting a HR of 0.7 at 0.025 (one-sided) when the PFS test is significant. The duration of OS in the control group is assumed to follow an exponential distribution with a median of 13 months based on historical data. The exponential dropout rate assumed for OS is 0.1% per month.

3.2 Responsibility for Analyses/In-House Blinding

The statistical analysis of the data obtained from this study will be the responsibility of the Clinical Biostatistics department of the SPONSOR.

The SPONSOR will generate the randomized allocation schedule(s) for study treatment assignment for this protocol, and the randomization will be implemented in IVRS.

This trial is double blinded with a crossover phase. At the time of documented progression, subjects will have treatment assignment unblinded and be able to continue therapy in the Crossover Phase, please refer to the protocol Section 2.1 Trial Design for details. In addition, independent central radiologist(s) will perform the central imaging review without knowledge of treatment assignment.

The study team at the Sponsor consisting of clinical, statistical, statistical programming and data management personnel, will be blinded to subject-level PD-L1 biomarker results. An unblinded data management personnel, unblinded Sponsor statistician and unblinded Sponsor statistical programmer will have access to the subject-level PD-L1 results for the purpose of data review and will have no other responsibilities associated with the study. A summary of PD-L1 biomarker prevalence may be provided to the study team at the Sponsor by the IVRS vendor or the unblinded Sponsor statistician.

An external data monitoring committee (eDMC) will be convened to review accumulating safety to provide an opportunity to terminate the study early if there are concerns regarding safety. The eDMC will also review the unblinded efficacy results at the planned interim analysis. The eDMC responsibilities and review schedules will be outlined in the eDMC charter. The recommendation of the eDMC will be communicated to an executive oversight committee of the Sponsor and, in the event of a recommendation to halt the trial early due to safety concerns, to the appropriate regulatory agencies. If the eDMC recommends modifications to the design of the protocol or discontinuation of the study, this executive oversight committee may be unblinded to results at the treatment level in order to act on these recommendations.

Limited numbers of additional SPONSOR personnel may be unblinded, if required, in order to act on the recommendations of the eDMC. The extent to which individuals are unblinded with respect to the results will be documented. Additional logistical details, revisions to the above plan and data monitoring guidance will be provided in the eDMC Charter.

3.3 Hypotheses/Estimation

Objectives and hypotheses of the study are stated in Section 3.0 of the protocol.

3.4 Analysis Endpoints

3.4.1 Efficacy Endpoints

Primary

Progression-free survival - RECIST 1.1 assessed by blinded independent central imaging vendor (BICR)

Progression-free-survival (PFS) is defined as the time from randomization to the first documented disease progression per RECIST 1.1 based on blinded central imaging vendor review or death due to any cause, whichever occurs first. See Section 3.6.1 for definition of censoring.

Overall Survival

Overall Survival (OS) is defined as the time from randomization to death due to any cause. Subjects without documented death at the time of analysis will be censored at the date of last known contact.

Secondary

Objective response rate (ORR) – RECIST 1.1 assessed by BICR

Overall Response Rate (ORR) is defined as the proportion of subjects who have a complete response (CR) or a partial response (PR). Responses are based on confirmed assessments by BICR review per RECIST 1.1.

Response Duration (DOR) – RECIST 1.1 assessed by BICR

For subjects who demonstrated confirmed CR or PR, DOR is defined as the time from first documented evidence of CR or PR until disease progression or death. Response duration for subjects who have not progressed or died at the time of analysis will be censored at the date of their last tumor assessment.

Exploratory

PFS – RECIST 1.1 assessed by BICR, OS and ORR – RECIST 1.1 assessed by BICR in subjects with different PD-L1 expression levels.

PFS, ORR and DOR assessed by the investigator using RECISIT 1.1.

PFS, ORR and DOR per investigator assessed irRECIST response criteria.

Patient-reported outcomes (PROs) while on treatment and post-discontinuation will be examined. The PRO endpoints include EORTC QLQ-C30, EORTC QLQ-LC13 and EuroQoL(EQ)-5D.

PFS and OS following crossover to pembrolizumab in subjects initially treated with saline placebo in combination with chemotherapy.

3.4.2 Safety Endpoints

Safety measurements are described in Section 4.2.3.3 of the protocol.

3.5 Analysis Population

3.5.1 Efficacy Analysis Population

The Intention-to-Treat (ITT) population will serve as the population for primary efficacy analysis. All randomized subjects will be included in this population. Subjects will be included in the treatment group to which they are randomized.

If an unexpectedly large number of randomized subjects are not treated, analyses may be performed using the Full Analysis Set (FAS), including all randomized subjects who received at least 1 dose of study treatment and did not have a major protocol violation.

Details on the approach to handling missing data are provided in Section 3.6 Statistical Methods.

3.5.2 Safety Analysis Populations

The All Subjects as Treated (ASaT) population will be used for the analysis of safety data in this study. The ASaT population consists of all randomized subjects who received at least one dose of study treatment. Subjects will be included in the treatment group corresponding to the study treatment they actually received for the analysis of safety data using the ASaT population. For most subjects this will be the treatment group to which they are randomized. Subjects who take incorrect study treatment for the entire treatment period will be included in the treatment group corresponding to the study treatment actually received. Any subject who receives the incorrect study medication for one cycle but receives the correct treatment for all other cycles will be analyzed according to the correct treatment group and a narrative will be provided for any events that occur during the cycle for which the subject is incorrectly dosed.

At least one laboratory or vital sign measurement obtained subsequent to at least one dose of study treatment is required for inclusion in the analysis of each specific parameter. To assess change from baseline, a baseline measurement is also required.

Details on the approach to handling missing data for safety analyses are provided in Section 3.6 Statistical Methods.

3.6 Statistical Methods

3.6.1 Statistical Methods for Efficacy Analyses

This section describes the statistical methods that address the primary and secondary objectives. Methods for biomarker exploratory objectives 4) and 8) and pharmacokinetic objective 7) will be documented in separate SAPs.

Efficacy results that will be deemed to be statistically significant after consideration of the Type I error control strategy are described in Section 3.9, Multiplicity. Nominal p-values will be computed for other efficacy analyses, but should be interpreted with caution due to potential issues of multiplicity.

All statistical tests, unless otherwise specified, will be stratified for treatment and stratification factors.

3.6.1.1 Progression-Free Survival (PFS)

The non-parametric Kaplan-Meier method will be used to estimate the PFS curve in each treatment group. The treatment difference in PFS will be assessed by the stratified log-rank test (based on the stratification factors defined in Section 5.4 of the protocol). A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (i.e., hazard ratio) between the treatment arms. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported. The stratification factors used for randomization (See Section 5.4 of the protocol) will be applied to both the stratified log-rank test and the stratified Cox model.

Since disease progression is assessed periodically, progressive disease (PD) can occur any time in the time interval between the last assessment where PD was not documented and the assessment when PD is documented. For the primary analysis, for the subjects who have PD, the true date of disease progression will be approximated by the date of the first assessment at which PD is objectively documented per RECIST 1.1 by central imaging vendor, regardless of discontinuation of study drug. Death is always considered as a confirmed PD event. Sensitivity analyses will be performed for comparison of PFS based on investigator's assessment.

In order to evaluate the robustness of the PFS endpoint per RECIST 1.1 by central imaging vendor, we will perform two sensitivity analyses with a different set of censoring rules. The first sensitivity analysis is the same as the primary analysis except that it censors at the last disease assessment without PD when PD or death is documented after more than one missed disease assessment. The second sensitivity analysis is the same as the primary analysis except that it considers discontinuation of treatment or initiation of an anticancer treatment subsequent to discontinuation of study-specified treatments, whichever occurs later, to be a PD event for subjects without documented PD or death. The censoring rules for primary and sensitivity analyses are summarized in Table 1. In case there is an imbalance between the treatment groups on disease assessment schedules or censoring patterns, we will perform an additional PFS sensitivity analysis using time from randomization to scheduled tumor assessment time instead of actual tumor assessment.

Table 1 Censoring rules for Primary and Sensitivity Analyses of PFS

Situation	Primary Analysis	Sensitivity Analysis 1	Sensitivity Analysis 2
No PD and no death; new anticancer treatment is not initiated	Censored at last disease assessment	Censored at last disease assessment	Censored at last disease assessment if still on study therapy; progressed at treatment discontinuation otherwise
No PD and no death; new anticancer treatment is initiated	Censored at last disease assessment before new anticancer treatment	Censored at last disease assessment before new anticancer treatment	Progressed at date of new anticancer treatment
No PD and no death; ≥ 2 consecutive missed disease assessments	Censored at last disease assessment	Censored at last disease assessment prior to ≥2 consecutive missed visits	Censored at last disease assessment
PD or death documented after ≤ 1 missed disease assessment	Progressed at date of documented PD or death	Progressed at date of documented PD or death	Progressed at date of documented PD or death

PD or death documented at any time after ≥ 2 consecutive missed disease assessments	Progressed at date of documented PD or death	Censored at last disease assessment prior to the ≥ 2 consecutive missed disease assessment	Progressed at date of documented PD or death

In case the proportional hazards assumption is not valid, Restricted Mean Survival Time (RMST) method may be conducted for PFS to account for the possible non-proportional hazards effect.

In case of potential gross imbalance in baseline prognostic factors in the ITT population with TPS≥50% (due to lack of stratification according to TPS≥50% vs. TPS<50%), a sensitivity PFS analysis may be performed using the multivariate Cox regression to adjust for those imbalanced baseline prognostic factors.

3.6.1.2 Overall Survival (OS)

The non-parametric Kaplan-Meier method will be used to estimate the survival curves. The treatment difference in survival will be assessed by the stratified log-rank test (based on the stratification factor defined in section 5.4 of the protocol). A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (i.e., the hazard ratio). The hazard ratio and its 95% confidence interval from the stratified Cox model with a single treatment covariate will be reported. The stratification factors used for randomization (See Section 5.4 of the protocol) will be applied to both the stratified log-rank test and the stratified Cox model.

Since subjects in the control arm are allowed to switch to the pembrolizumab treatment after progressive disease, adjustment for the effect of crossover on OS may be performed based on recognized methods, e.g., a two-stage method or the Rank Preserving Structural Failure Time (RPSFT) model proposed by Robins and Tsiatis , based on an examination of the appropriateness of the data to the assumptions required by the methods.

In case of potential gross imbalance in baseline prognostic factors in the ITT population with TPS≥50% (due to lack of stratification according to TPS≥50% vs. TPS<50%), a sensitivity OS analysis may be performed using the multivariate Cox regression to adjust for those imbalanced baseline prognostic factors.

3.6.1.3 Objective response rate (ORR) and Duration of Response (DOR)

Stratified Miettinen and Nurminen’s method will be used for comparison of the ORR between two treatment groups. The difference in ORR and its 95% confidence interval from the stratified Miettinen and Nurminen’s method with strata weighting by sample size with a single treatment covariate will be reported. The stratification factors used for randomization (See Section 5.4 of the protocol) will be applied to the analysis.

If sample size permits, duration of response duration will be summarized descriptively using Kaplan-Meier medians and quartiles. Only the subset of patients who show a complete response or partial response will be included in this analysis.

For each DOR analysis, a corresponding summary of the reasons responding subjects are censored will also be provided. Responding subjects who are alive, have not progressed, have not initiated new anti-cancer treatment, have not been determined to be lost to follow-up, and have had a disease assessment within ~5 months of the data cutoff date are considered ongoing responders at the time of analysis. If a subject meets multiple criteria for censoring, the censoring criterion that occurs earliest will be applied.

Table 2 Censoring Rules for DOR

Situation	Date of Progression or Censoring	Outcome
No progression nor death, no new anti-cancer therapy initiated	Last adequate disease assessment	Censor (non-event)
No progression nor death, new anti-cancer therapy initiated	Last adequate disease assessment before new anti-cancer therapy initiated	Censor (non-event)
Death or progression after ≥ 2 consecutive missed disease assessments	Last adequate disease assessment prior to ≥ 2 missed adequate disease assessments	Censor (non-event)
Death or progression after ≤ 1 missed disease assessments	PD or death	End of response (Event)
A missed disease assessment includes any assessment that is not obtained or is considered inadequate for evaluation of response.		

3.6.1.4 Exploratory Analyses of Efficacy

An exploratory analysis of PFS2, defined as the time from randomization to subsequent disease progression after initiation of new anti-cancer therapy, or death from any cause, whichever first, may be carried out. Subjects alive and for whom a PFS event has not been observed should be censored at the last time the subject was known alive and without disease progression.

3.6.2 Statistical considerations for patient-reported outcomes (PRO)

The patient-reported outcomes are exploratory objectives in KEYNOTE 189, and thus no formal hypotheses were formulated. This sSAP will focus on PRO endpoints as measured by the EORTC QLQ-C30, EORTC QLQ-LC13 and EuroQol-5D Visual Analogue Scale (EQ-5D VAS).

The global health status/quality of life scale from EORTC QLQ-C30 and the composite endpoint of cough (LC13-Q1), chest pain (LC13-Q10) and dyspnea (QC30 Q8) will be key PRO endpoints. The time to deterioration (TTD) of this composite endpoint has been used as the key PRO endpoint in advanced NSCLC clinical trials [2, 3, 4]. The time to true deterioration of the composite endpoint is defined in as the time to first onset of 10 or more decrease from baseline with confirmation under right-censoring rule. The time to true deterioration will be used in the analysis.

Supportive analyses will include all QLQ-C30 and QLQ-LC13 sub-scales/items, and the EQ-5D VAS. The QLQ-C30 includes five functional dimensions (physical, role, emotional, cognitive, and social), three symptom scales (fatigue, nausea/vomiting, and pain), and six single item measures (dyspnea, sleep disturbance, appetite loss, constipation, diarrhea, and financial difficulties). The QLQ-LC13 comprises 13 questions consisting of one multi-item scale (dyspnea) and nine single items that assess the specific symptoms (dyspnea, cough, hemoptysis, and site specific pain), side effects (sore mouth, dysphagia, neuropathy, and alopecia), and pain medication use of lung cancer patients. Supportive analysis will also include alternative approaches to TTD in addition to estimating the effect of disease progression on HRQoL.

Combined with EORTC QLQ-C30, the EORTC QLQ-LC13 measures lung cancer associated symptoms (cough, hemoptysis, dyspnea, and site specific pain), and treatment related symptoms (sore mouth, dysphagia, peripheral neuropathy, and alopecia).

EQ-5D is another set of endpoints as a measure of health outcome. The analyses in this SAP focus on the EQ-5D VAS within the EQ-5D.

Results for the EORTC QLQC30 and EORTC QLQ-LC13.will be reported for the overall study population and the biomarker positive strata (>=1%) as described in the exploratory PRO objective in the study protocol.

Key PRO Endpoints:

1. The mean score changes from baseline to weeks 12 and 21 as measured by the EORTC QLQ-C30 global health status/quality of life scale.

2. Time to True Deterioration in the composite endpoint of cough (LC13-Q1), chest pain (LC13-Q10) and dyspnea (LC30-Q8).

Supportive PRO Endpoints:

The mean score changes, and the number and proportions of deterioration/stable/improvement from baseline to weeks 12 and 21. Specifically:

- EORTC QLQ-C30 global health status/quality of life scale (the number and proportions of deterioration/stable/improvement only).

- Each EORTC QLQ-C30 functional subscale: physical functioning, role functioning, emotional functioning, cognitive functioning, social functioning.

- Each EORTC QLQ-C30 symptom subscale score: Fatigue, Nausea and vomiting, Pain, Dyspnoea, Insomnia, Appetite loss, Constipation, Diarrhea.

- Each EORTCQLQ-LC13 item: pain, coughing, sore mouth, dysphagia, peripheral neuropathy, alopecia, and haemoptysis (mean score changes from baseline only).

- The EORTC QLQ-LC13 dyspnea multi-item sub-scale (mean score changes from baseline only).

Although no formal hypothesis was formulated, the p-value from the constrained longitudinal data Analysis (cLDA) model and Cox regression model will be provided for treatment comparisons of pembrolizumab in combination with chemotherapy vs. saline placebo with chemotherapy for key PRO endpoints 1 and 2, respectively. No multiplicity adjustment will be performed.

The mean score change for multi-item subscale(s) and single item scores from baseline to week 12 and 21 and the proportions of deterioration/stable/improvement QLQ-LC13 and QLQ-C30 will have a nominal p-values using the CMH test for general association to compare pembrolizumab in combination with chemotherapy vs. saline placebo with chemotherapy. For multi-item scale(s), the analysis will focus on the subscale score rather than each single item.

Scoring Algorithm:

QLQ-C30 Scoring: For each scale or item, a linear transformation will be applied to standardize the score as between 0 and 100, according to the corresponding scoring standard. For functioning and global health status/quality-of-life scales, a higher value indicates a better level of function; for symptom scales and items, a higher value indicates increased severity of symptoms.

According to the QLQ-C30 Manuals, if items $I_1, I_2,\ldots,I_n$ are included in a scale, the linear transformation procedure is as follows:

1. Compute the raw score: $RS = (I_1 + I_2 + \ldots + I_n)/n$

2. Linear transformation to obtain the score S:

$$\text{Function scales: } S = \left(1 - \frac{RS-1}{Range}\right) \times 100$$

$$\text{Symptom scales/items: } S = \frac{RS-1}{Range} \times 100$$

$$\text{Global health status/QoL: } S = \frac{RS-1}{Range} \times 100$$

Range is the difference between the maximum possible value of *RS* and the minimum possible value. If more than half of the items within one scale are missing, then the scale is considered missing, otherwise, the score will be calculated as the average score of those available items [5].

QLQ-LC13 scoring: The lung cancer questionnaire module comprises both multi-item and single-item measures of lung cancer-associated symptoms (i.e. coughing, haemoptysis, dyspnoea and pain) and treatment related symptom scores: sore mouth, dysphagia, peripheral neuropathy, and alopecia. A linear transformation will be applied to standardize the scores between 0 (worst) and 100 (best) as described above for the EORTC QLQ-C30 Scoring.

EQ-5D scoring: EQ-5D utility score will be calculated based on the European algorithm [6]. The five health state dimensions in this instrument include the following: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression (5).

The schedule for PRO data collection:

Table 3 provides the schedule for PRO data collection.

Table 3 PRO Data Collection Schedule

	Original Phase Treatment Week										DC[1]	Safety Follow-up[1]	Observation Phase
	0	3	6	9	12	21	30	39	48	60 and Q12W beyond			
EuroQol (EQ)-5D EORTC QLQ-C30 EORTC QLQ-LC13	C1	C2	C3	C4	C5	C8	C11	C14	C17	C21 and every 4 cycles beyond	X	X	Every 12 weeks until PD

C: Cycle.

DC: Discontinuation

[1] If the Discontinuation Visit occurs 30 days from the last dose of study treatment, at the time of the mandatory Safety Follow up Visit, PROs do not need to be repeated.

The general rule of mapping relative day to analysis visit is provided in Table 4 (due to space limit, only up to Week 48 is displayed).

Table 4 Mapping Relative Day to Analysis Visit

Treatment Week	Week								
	0	3	6	9	12	21	30	39	48
Day	1	21	42	63	84	147	210	273	336
Range	-59 to 1	2 to 31	32 to 52	53 to73	74 to 115	116 to 178	179 to 241	242 to 304	305 to 378

At each scheduled visit, three instruments, EQ5D, EORTC QLQ-C30, and EORTC QLQ-LC13, will be collected. If a patient does not complete the PRO instruments, the site staff will record the reason for the missing from the pre-defined choices. If there are multiple PRO collections within any of the stated time windows, we use the closest collection to the target day.

3.6.2.1 Analysis Populations

The primary analysis approach for the pre-specified exploratory PRO endpoints will be based on a quality of life related full analysis set (FAS) population following the intention-to-treat (ITT) principle and ICH E9 guidelines. This population consists of all randomized patients who have received at least one dose of study medication, and have completed at least one PRO assessment.

3.6.2.2 Analysis Approaches

This section describes the planned analyses for the PRO endpoints. Table 5 gives an overview of the analyses planned for all PRO endpoints. Compliance will be summarized by treatment group and visit.

Table 5 Planned Statistical Analysis

Endpoint	Analysis	Primary Statistical Method	Report
Score change from baseline	Treatment effect estimation/comparison	Mixed effect model based on the missing at random (MAR) assumption	least-square mean score (95% C.I.) by treatment group and visit, least-square mean score change (95% C.I.) from baseline by treatment group and visit, pairwise between-group difference in score change from baseline (95% C.I., p-value).
	Disease progression effect estimation	Mixed effect model based on the MAR assumption	P-value of the interaction effect between progression and treatment discontinuation, by treatment group.
Time to deterioration	Treatment effect estimation/comparison	Kaplan-Meier plot, stratified log rank test and Cox proportional hazards model	Hazard Ratio (95% C.I.)
Proportion of deterioration/stable/improvement	Treatment effect estimation/comparison	Summary with multiple imputation based on the MAR assumption	Proportion by treatment group and visit

3.6.2.3 Treatment effect on PRO

To assess the treatment effects of treatment on the PRO, for each continuous endpoint defined, a constrained longitudinal data analysis (cLDA) model will be used as the primary analysis method, with the PRO score as the response variable, and treatment by study visit interaction, and stratification factors (Section 5.4 of the protocol) as covariates.

The cLDA model is specified as follows:
$E(Y_{ijt}) = \gamma_0 + \gamma_{jt} I(t > 0) + \beta_{it} X_i$, $j = 1, 2,$ $t = 0,1,2,3,...$, where Y_{ijt} is the PRO score for subject i, with treatment assignment j, at visit t, γ_0 is the baseline mean for all treatment groups, γ_{jt} is

the mean change from baseline for treatment group j at time t, X_i is the stratification factor vector for this patient, and β_{it} is the coefficient vector for stratification factor at time t.

Treatment effect on PRO score change from baseline will be evaluated at 12 weeks and potentially at 21 weeks, if the completion rate in the control arm is larger than 50%. Between-group comparison will be performed and the differences in the lsmean change from baseline will be reported, together with 95% C.I. and nominal p-value at the primary analysis time points. In addition, model-based lsmean score with 95% C.I. will be provided by treatment group and study visit.

Most of the patients without disease progression are expected to have complete data up to 12 and 21 weeks. Patients with disease progression confirmed or feeling worse due to drug-related AE may have missing PRO assessments. The missing data must be handled accordingly to obtain valid statistical inference. The cLDA model implicitly treats missing data as missing at random (MAR). Sensitivity analyses may be conducted in case the robustness of MAR assumption is questionable.

3.6.2.4 Effect of disease progression on PRO

A supportive analysis will be performed to assess the effects of disease progression and treatment on global health status/quality of life. A constrained longitudinal data analysis (cLDA) model will be applied, with the PRO score as the response variable, and treatment by study visit interaction, stratification, and progression status (time-varying) as covariates.

The cLDA model is specified as follows:
$E(Y_{ijt}) = \gamma_0 + \gamma_{jt} I(t > 0) + \beta X_i + \alpha_j Z_t, \quad j = 1, 2, \quad t = 0,1,2,3,..,$ where Y_{ijt} is the PRO score for subject i, with treatment assignment j, at visit t, γ_0 is the baseline mean for all treatment groups, γ_{jt} is the mean change from baseline for treatment group j at time t, X_i is the stratification factor vector for this patient, β is the coefficient vector for stratification factors, α_j is the disease progression effect for treatment j, and Z_t is the progression status at visit t.

The point estimate of the effect of disease progression on the lsmean PRO score change from baseline will be reported, together with 95% C.I., overall and by treatment groups.

3.6.2.5 Analysis of the Proportions of Deterioration/Stable/Improvement

Patients' post-baseline PRO score will be classified as "improved" "stable" or "deteriorated" according to a 10 points or greater change for each of the instrument/scale, as this magnitude of change is perceived by patients as being clinically significant [7]. The number and proportion of patients who "improved", "stable", or "deteriorated", from baseline will be summarized by treatment group and at week 12 and 21.

Since missing data cannot be ignored at week 12 and week 21, additional summaries of deterioration/stable/improvement will be provided based on multiply-imputed data sets to handle the issue of missing data at week 12 and potentially at 21 weeks. Specifically, proportions will be calculated based on the multiply imputed data sets in the PRO score analysis, and then synthesized based on Rubin's rule. Proportions based on MAR imputation will be the primary result for proportions of deterioration/stable/improvement.

3.6.2.6 Analysis of the Time to Deterioration

The true time-to-deterioration is defined in as the time to first onset of 10 or more decrease from baseline with confirmation under right-censoring rule. The non-parametric Kaplan-Meier method will be used to estimate the deterioration curve in each group. The treatment difference in time-to-deterioration will be assessed by the stratified log-rank test. A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (hazard ratio) between treatment arms. The traditional time-to-deterioration, defined in as the time to first onset of 10 or more decrease from baseline without confirmation under right-censoring rule, may also be summarized.

3.6.2.7 Compliance Summary

Completion and compliance of QLQ-C30, QLQ-LC13 and EQ-5D by visit and by treatment will be described based on the ITT population and the PRO FAS population, respectively. Numbers and percentages of complete and missing data at each visit will be summarized for each of the treatment groups.

Completion rate in the ITT population is defined as the percentage of the number of subjects who complete at least one item over the number of randomized subjects at each time points. Completion rate in the FAS population is defined as the percentage of number of subjects who complete at least one item over the number of subjects in the FAS population at each time points.

The completion rate is expected to shrink in the later visit during study period due to the subjects who discontinued early. Therefore, another measurement, Compliance rate, defined as the percentage of observed visit over number of eligible subjects who are expected to complete the PRO assessment (not including the subjects missing by design (such as death, discontinuation, translation not available) will be employed as the support for completion rate).

The reasons of non-completion and non-compliance will also be provided. An instrument is considered complete if at least one valid score available according to the missing item rules outlined in the EORTC QLQ-C30 Manual for each functional and symptoms scale.

3.6.3 Statistical Methods for Safety Analyses

Safety and tolerability will be assessed by clinical review of all relevant parameters including adverse experiences (AEs), laboratory tests, vital signs, etc.

Adverse Events

Adverse events (AEs) will be coded using the standard MedDRA and grouped system organ class. Adverse events (AEs) will be graded by the investigator according to the Common Terminology Criteria for Adverse Events (CTCAE), version 4.0.

Tiered Approach

The analysis of safety results will follow a tiered approach (Table 6). The tiers differ with respect to the analyses that will be performed. "Tier 1" safety endpoints that will be subject to inferential testing for statistical significance with p -values and 95% confidence intervals provided for between-group comparisons. For this protocol, there are no Tier 1 events. Other safety parameters will be considered Tier 2 or Tier 3. Tier 2 parameters will be assessed via point estimates with 95% confidence intervals provided for between-group comparisons; only point estimates by treatment group are provided for Tier 3 safety parameters.

Adverse experiences (specific terms as well as system organ class terms) that are not pre-specified as Tier-1 endpoints will be classified as belonging to "Tier 2" or "Tier 3", based on the number of event s observed. Membership in Tier 2 requires that at least 4 subjects in any treatment group exhibit the event; all other adverse experiences and predefined limits of change will belong to Tier 3.

The threshold of at least 4 events was chosen because the 95% confidence interval for the between-group difference in percent incidence will always include zero when treatment groups of equal size each have less than 4 events and thus would add little to the interpretation of potentially meaningful differences. Because many 95% confidence intervals may be provided without adjustment for multiplicity, the confidence intervals should be regarded as a helpful descriptive measure to be used in review, not a formal method for assessing the statistical significance of the between-group differences in adverse experiences and predefined limits of change.

Continuous measures such as changes from baseline in laboratory, vital signs, that are not pre-specified as Tier-1 endpoints will be considered Tier 3 safety parameters. Summary statistics for baseline, on-treatment, and change from baseline values will be provided by treatment group in table format.

In addition, the broad clinical and laboratory AE categories consisting of the percentage of subjects with any AE, any drug related AE, any Grade 3-5 AE, any serious AE, any AE which is both drug-related and Grade 3-5, any AE which is both serious and drug-related, dose modification due to AE, and who discontinued due to an AE, and death will be considered Tier 2 endpoints. 95% confidence intervals (Tier 2) will be provided for between- treatment differences in the percentage of subjects with events; these analyses will be performed using the Miettinen and Nurminen method (1985), an unconditional, asymptotic method.

Table 6 Analysis Strategy for Safety Parameters

Safety Tier	Safety Endpoint	95% CI for Treatment Comparison	Descriptive Statistics
Tier 2	Any AE	X	X
	Any Serious AE	X	X
	Any Grade 3-5 AE	X	X
	Any Drug-Related AE	X	X
	Any Serious and Drug-Related AE	X	X
	Any Grade 3-5 and Drug-Related AE	X	X
	Dose Modification due to AE	X	X
	Discontinuation due to AE	X	X
	Death		
	Specific AEs, SOCs, or PDLCs (incidence ≥4 of subjects in one of the treatment groups)	X	X
Tier 3	Specific AEs, SOCs or PDLCs (incidence <4 of subjects in all of the treatment groups)		X
	Change from Baseline Results (Labs, ECGs, Vital Signs)		X

3.7 Summaries of Demographics, Baseline Characteristics and Other Analyses

The comparability of the treatment groups for each relevant characteristic will be assessed by the use of tables and/or graphs. No statistical hypothesis tests will be performed on these characteristics. The number and percentage of subjects randomized, and the primary reasons for discontinuation will be displayed. Demographic variables (e.g., age, gender) and baseline characteristics will be summarized by treatment either by descriptive statistics or categorical tables.

3.8 Interim Analyses

Two interim efficacy analyses are planned in addition to the final analysis.

- The first interim analysis (IA1) will evaluate PFS and OS. It will be performed after enrollment is complete and approximately 370 PFS events have been observed, approximately 19 months after first patient enrolled. It is estimated that approximately 242 deaths will be observed. ORR will be tested at this interim analysis if both the PFS and OS test results are significant.
- The second interim analysis (IA2) is the final analysis for PFS and will be performed after approximately 468 PFS events have been observed, approximately 26 months after first patient enrolled. An interim analysis of OS will be performed; it is estimated that approximately 332 deaths will be observed.
- The final analysis (FA) will evaluate OS only and will be performed after approximately 416 deaths have been observed. It is estimated that this will occur approximately 35 months after enrollment begins.

The analyses planned, endpoints evaluated and drivers of the timing are summarized in Table 7.

Table 7 Analyses Planned, Endpoints Evaluated, and Drivers of Timing

Analysis	Endpoint(s)	Timing
IA1	PFS; OS; ORR if both PFS and OS are positive	~370 PFS events (~242 OS events expected at this time)
IA2	PFS; OS	~468 PFS events (~332 OS events expected at this time)
FA	OS	~416 OS events

Decisions to stop the trial early will be based on DMC recommendations with review by the Executive Oversight Committee.

Type I error control for the efficacy analyses as well as efficacy bounds are described in the next section.

3.9 Multiplicity

The trial uses the graphical method of Maurer and Bretz to provide strong multiplicity control for multiple hypotheses as well as interim analyses.

Figure 1 shows the initial one-sided α-allocation for each hypothesis in the ellipse representing the hypothesis. The weights for reallocation from each hypothesis to the others are represented in the boxes on the lines connecting hypotheses. This is further explained below.

See Figure 1 for the type I error reallocation strategy for endpoints PFS, OS, and ORR.

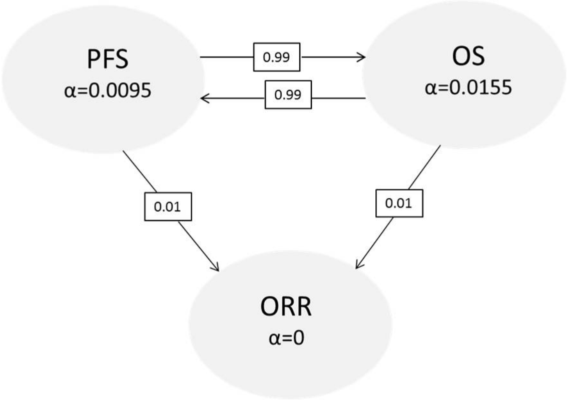

Figure 1 Type I Error Reallocation Strategy

PFS

PFS hypothesis will be tested at α=0.0095. When OS test is significant, the PFS hypothesis may be tested at α=0.025 (re-allocated α). The Lan-DeMets O'Brien-Fleming spending function was used to control the type I error in the interim analysis. Table 8 below demonstrates the bounds and boundary properties for PFS hypothesis testing. The table will be updated using the actual number of PFS events at the interim and final PFS analyses.

Table 8 Boundary properties for planned analyses of PFS based on potential alpha-levels to be used for testing

Analysis	Value	α=0.0095	α=0.025
IA 1: 79%	Z	2.6946	2.2676
N: 616	p (1-sided)	0.0035	0.0117
Events: 370	HR at bound	0.7427	0.7790
Month: 19.2	P(Cross) if HR=1	0.0035	0.0117
	P(Cross) if HR=0.7	0.7160	0.8439
IA2*	Z	2.3895	2.0222
N: 616	p (1-sided)	0.0084	0.0216
Events: 468	HR at bound	0.7910	0.8204
Month: 26	P(Cross) if HR=1	0.0095	0.0250
	P(Cross) if HR=0.7	0.9040	0.9550

*The final analysis of PFS will be performed at IA2.

If the OS superiority null hypotheses are rejected at an interim or final analysis, each PFS interim analysis test may be compared to its updated rejection boundary for formal testing.

OS

The OS hypothesis will be tested at α=0.0155. When PFS test is significant, the OS hypothesis may be tested at α=0.025 (re-allocated α). Table 9 demonstrates the bounds and boundary properties for OS hypothesis testing. The HR of OS between the experimental group and control group is assumed to be 0.7. The table will be updated using the actual number of OS events at the interim and final OS analyses. The Lan-DeMets O'Brien-Fleming spending function was used to derive the design in α columns.

Table 9 Boundary properties for planned analyses of OS

Analysis	Value	α=0.0155	α=0.025
IA 1: 58%	Z	2.9715	2.7215
N: 616	p (1-sided)	0.0015	0.0032
Events: 242	HR at bound	0.6662	0.6893
Month: 19.3	P(Cross) if HR=1	0.0015	0.0032
	P(Cross) if HR=0.7	0.3711	0.4695
IA 2: 80%	Z	2.4926	2.2831
N: 616	p (1-sided)	0.0063	0.0112
Events: 332	HR at bound	0.7482	0.7666
Month: 26.2	P(Cross) if HR=1	0.0068	0.0122
	P(Cross) if HR=0.7	0.7332	0.7991
Final	Z	2.2155	2.0300
N: 616	p (1-sided)	0.0134	0.0212
Events: 416	HR at bound	0.7940	0.8095
Month: 35	P(Cross) if HR=1	0.0155	0.0250
	P(Cross) if HR=0.7	0.9000	0.9300

Unplanned OS analysis might be requested by DMC for assessment of risk and benefit during the trial. These analyses are not to declare a positive efficacy finding. If requested by regulators, a sensitivity analysis of OS will be performed to count these DMC required OS analyses as interim analyses and adjust Type I error by the pre-specified alpha spending function in the protocol.

If the PFS superiority null hypotheses are rejected at any interim analyses, each OS interim and final analysis test may be compared to its updated rejection boundary for formal testing.

ORR

When both PFS and OS tests are significant, α=0.025 will be allocated to the ORR test. If the OS test does not achieve statistical significance at IA1, the p value of the ORR test from IA1 will be compared to 2.5% if the null hypotheses for PFS and OS are rejected at a later time.

3.10 Sample Size and Power Calculation

This trial is well-powered for the primary and key secondary endpoints. Enrollment of 570 subjects is assumed to occur over 12 months at 2:1 ratio between the experimental and control groups. The actual enrollment is 616 subjects within 13 months. With 370 PFS events at IA1, the study has ~72% power for detecting a HR of 0.7 at 0.0095 (one-sided) and ~84% power for detecting a HR of 0.7 at 0.025 (one-sided). With 468 PFS events at IA2, the study has ~90% power for detecting a HR of 0.7 at 0.0095 (one-sided) and ~96% power for detecting a HR of 0.7 at 0.025 (one-sided). The duration of PFS in the control group is assumed to follow an exponential distribution with a median of 6.5 months based on historical data. The assumed follow-up time after last patient enrolled is 13 months for IA2. An exponential dropout rate of 0.35% per month is assumed.

With 242 deaths at IA1, the study has ~37% power for detecting a HR of 0.7 at 0.0155 (one-sided) and ~47% power for detecting a HR of 0.7 at 0.025 (one-sided) when the PFS test is significant. With 332 deaths at IA2, the study has ~73% power for detecting a HR of 0.7 at 0.0155 (one-sided) and ~80% power for detecting a HR of 0.7 at 0.025 (one-sided) when the PFS test is significant. With 416 deaths at FA, the study has ~90% power for detecting a HR of 0.7 at 0.0155 (one-sided) and ~93% power for detecting a HR of 0.7 at 0.025 (one-sided) when the PFS test is significant. The duration of OS in the control group is assumed to follow an exponential distribution with a median of 13 months based on historical data. The exponential dropout rate assumed for OS is 0.1% per month.

The sample size of 616 yields 89% (α=0.025) power to detect an ORR difference from an underlying 45% response rate in the control group to 60% in the experimental group, when both PFS and OS tests are significant. The approximate treatment difference required to reach the bound (Δ ORR) is 0.0575.

Power and interim analysis calculations were performed using EAST 5 and the gsDesign R package.

3.11 Subgroup Analyses and Effect of Baseline Factors

To determine whether the treatment effect is consistent across various subgroups, the between-group treatment effect for PFS, OS and ORR (with a nominal 95% CI) will be estimated and plotted within each category of the following classification variables:

- Age category (≤65, >65 years)
- ECOG Performance Scale (0, 1)
- Sex (female, male)
- Race (white, non-white)
- Geographic region (US, Ex US)
- Geographic region (EU, Ex EU)Smoking status (never, former/current)
- Brain metastasis status at baseline (yes, no)
- PD-L1 expression (unknown, TPS <1%, or TPS ≥1%)
- PD-L1 expression (unknown, TPS <50%, or TPS ≥50%)
- PD-L1 expression (unknown, TPS <1%, 1%≤TPS≤49%, or TPS ≥50%)
- Platinum chemotherapy (cisplatin, carboplatin)

For PFS and OS, the stratified Cox model will be used. For ORR, the unstratified Miettinen and Nurminen method will be used. The consistency of the treatment effect will be assessed descriptively via summary statistics by category for the classification variables listed above. If any level of a subgroup variable has fewer than 10% of the ITT population, above analysis will not be performed for this level of the subgroup variable. If a subgroup variable has two levels and one level of the subgroup variable has fewer than 10% of the ITT population, then this subgroup will not be displayed in the forest plot.

The EU region includes countries from both EU member states (2016) and EFTA members.

3.12 Extent of Exposure

The extent of exposure will be summarized as duration of treatment in cycles.

4.0 REFERENCES

1. Robins JM, Tsiatis AA. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Commun Stat-Theor M 1991;20(8):2609-31.

2. Bezjak A, Tu D, Seymour L et al., Symptom Improvement in Lung Cancer Patients Treated With Erlotinib: Quality of Life Analysis of the National Cancer Institute of Canada Clinical Trials Group Study BR.21 J Clin Oncol 24:3831-3837.

3. Blackhall F, Kim DW, Besse B, et al. Patient-reported outcomes and quality of life in PROFILE 1007: A randomized trial of crizotinib compared with chemotherapy in previously treated patients with ALK positive NSCLC. J Thorac Oncol. 2014;9: 1625–1633.

4. Yang JC, Hrish V, Schuler M, et al. Symptom Control and Quality of Life in LUX-Lung 3: A Phase III Study of Afatinib or Cisplatin/Pemetrexed in Patients With Advanced Lung Adenocarcinoma With EGFR Mutations. J Clin Oncol 31:3342-3350.

5. The EORTC QLQ-C30 Manuals, Reference Values and Bibliography.

6. EQ-5D-3L User Guide, Oct 2013.

7. 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; 16:139-44.

8. Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. Stat Biopharm Res 2013;5(4):311-20.

Revision History

Date	Summary of Change
7MAY2016	Original Document
16NOV2017	Original document is updated to align with the protocol amendment v07 for the global study with regard to statistical analysis plan.