Protocol B7461006

A PHASE 3, RANDOMIZED, OPEN-LABEL STUDY OF LORLATINIB (PF-06463922) MONOTHERAPY VERSUS CRIZOTINIB MONOTHERAPY IN THE FIRST-LINE TREATMENT OF PATIENTS WITH ADVANCED ALK-POSITIVE NON-SMALL CELL LUNG CANCER

Statistical Analysis Plan
(SAP)

Version: 2.0

Date: 26 May 2020

TABLE OF CONTENTS

LIST OF TABLES

1. VERSION HISTORY

This Statistical Analysis Plan (SAP) for study B7461006 is based on the protocol amendment 4 dated 04OCT2019.

Table 1. Summary of Major Changes in SAP Amendments

SAP Version	Date	Change	Rationale
1	21-Dec-2016	Original SAP	Not Applicable
2	26-May-2020	The group-sequential design of the study based on the primary PFS endpoint as assessed by BICR has been modified by replacing the planned futility analysis, that was to be conducted at 60% information fraction (IF), with an interim efficacy analysis to be conducted at approximately 75% IF. Added IC-DR and IC-TTR; Removed CBR. Added clarifications on a few algorithms Removed secondary analyses deemed unnecessary Added sensitivity analyses to be conducted to mitigate COVID impact	CCI [redacted] The list of the secondary endpoints has been revised to focus on the relevant ones for the compound.

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in study B7461006. This analysis plan is meant to supplement the study protocol. Any deviations from this analysis plan will be described in the Clinical Study Report.

The analyses of this study will include all data up to a data cutoff date which will be determined by the number of progression-free survival (PFS)/overall survival (OS) events occurred. All summaries and analyses will include all data pertaining to visits/assessments performed up to and including the data cutoff date.

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

2.1. Study Objectives

Primary Objective

- To demonstrate that lorlatinib as a single agent (Arm A) is superior to crizotinib alone (Arm B) in PFS in advanced ALK-positive non-small cell lung cancer (NSCLC) patients who are treatment naïve.

Secondary Objectives:

- To compare Arm A with Arm B in treatment-naïve advanced ALK-positive NSCLC patients with respect to OS;
- To evaluate the antitumor activity in each treatment arm;
- To evaluate the safety and tolerability in each treatment arm;
- To evaluate patient-reported outcomes (PROs) of health-related quality of life, disease/treatment-related symptoms of lung cancer, and general health status for each treatment arm;
- To evaluate candidate biomarkers of sensitivity or resistance to single-agent crizotinib or lorlatinib in pre-treatment tumor tissue;
- To evaluate candidate biomarkers of sensitivity or resistance to single-agent crizotinib or lorlatinib in peripheral blood.

Exploratory Objectives:

- [redacted]
- [redacted]
- [redacted]

2.2. Study Design

This is an open-label, multi-center, randomized, Phase 3 study of lorlatinib versus crizotinib in previously untreated patients with advanced ALK-positive NSCLC.

A total of 280 patients were to be randomized in a 1:1 fashion to receive lorlatinib or crizotinib. A total of 296 patients were actually randomized. Randomization was to be stratified by presence of brain metastases (Yes vs. No) and ethnic origin (Asian vs. non-Asian).

Patients will continue with the assigned study treatment until Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1 (Eisenhauer et al, European Journal of Cancer, 2009)[1] defined progression as determined by Blinded Independent Central Review (BICR), unacceptable toxicity, death, or withdrawal of consent. Patients may continue treatment beyond the time of RECIST-defined progression, at the discretion of the investigator if the patient is perceived to be experiencing clinical benefit. Crossover between treatment arms will not be permitted.

3. ENDPOINTS AND BASELINE VARIABLES: DEFINITIONS AND CONVENTIONS

3.1. Primary Endpoint(s)

PFS based on BICR assessment (RECIST v.1.1). PFS is defined as the time from date of randomization, to the date of the first documentation of disease progression (PD) per RECIST v1.1 based on BICR assessment, or death due to any cause, whichever occurs first. Refer to Section 6.1 for details on censoring rules.

3.2. Secondary Endpoint(s)

3.2.1. Efficacy Endpoints

OS is defined as the time from date of randomization to the date of death due to any cause. Refer to Section 6.2.1 for details on censoring rules.

PFS based on investigator's assessment. It is defined similarly to the primary endpoint but PD will be derived programmatically from the target lesion measurements, non-target lesion status, and new lesions recorded on the case report form (CRF).

Objective Response (OR) based on BICR assessment is defined as complete response (CR) or partial response (PR) according to RECIST v1.1 from date of randomization, until documented PD or start of new anti-cancer therapy. Both CR and PR must be confirmed by repeat assessments performed no less than 4 weeks after the criteria for response are first met.

OR based on investigator's assessment is defined similarly to OR based on BICR assessment but Best Overall Response (BOR) per investigator will be derived programmatically from the target lesion measurements, non-target lesion status, and new lesions recorded on the CRF.

Intracranial OR (IC-OR) based on BICR assessment is defined as above for OR but it is only based on intracranial disease in the subset of patients with at least 1 intracranial lesion.

IC-time to progression (IC-TTP) based on BICR assessment is defined as the time from date of randomization to the date of the first documentation of progression of intracranial disease, based on either new brain metastases or progression of existing brain metastases.

Duration of Response (DR) based on BICR assessment is defined, for patients with a confirmed OR per RECIST version 1.1, as the time from the first documentation of objective tumor response (CR or PR) to the first documentation of disease progression or death due to any cause, whichever occurs first.

IC-DR based on BICR assessment is defined as above for DR, only in patients with a confirmed IC-OR.

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

IC-TTR based on BICR assessment is defined as above for TTR, only in patients with a confirmed IC-OR.

PFS2 is defined as the time from randomization to the date of progression of disease on first subsequent systemic anti-cancer therapy (as provided by the investigator on the follow up systemic therapy page), or death from any cause, whichever occurs first

3.2.2. Safety Endpoints

Adverse events (AEs) (graded by NCI Common Terminology Criteria for Adverse Events (CTCAE) v.4.03 as provided by the investigator on the AE CRF page); laboratory abnormalities (graded by NCI CTCAE v.4.03 as programmatically derived based on laboratory values); vital signs (blood pressure, pulse rate) and body weight; electrocardiograms (ECGs); echocardiogram or multigated acquisition (MUGA) scan; ophthalmologic data.

3.2.3. PRO Endpoints

PROs as assessed by EORTC QLQ-C30, EORTC QLQ-LC13, EQ-5D-5L.

Time to Deterioration (TTD) in pain in chest, dyspnea, or cough individually from the EORTC QLQ-LC13 and as a composite endpoint will be defined as the time from randomization to the first time the patient's score shows a 10 point or greater increase after baseline in any of the 3 symptoms. Refer to Section 6.2.12 for details on censoring rules.

3.2.4. Other Endpoints

Tumor tissue biomarkers including, but not limited to, ALK gene rearrangement and/or mutation as measured by next generation sequencing (NGS) and/or immunohistochemistry (IHC).

Peripheral blood cfDNA (circulating free Deoxyribonucleic acid) biomarkers including, but not limited to, ALK gene rearrangement and/or ALK kinase domain mutations.

3.3. Exploratory Endpoints

CCI [redacted]

3.4. Baseline Variables

The date of first dose (start date) of study treatment is the earliest date of non-zero dosing of the study drug. The date of last dose of study treatment is the latest date of non-zero dosing of the study drug.

No windowing will be applied when defining baseline, except as noted in Section 5.2.9. Adequate Baseline Tumor Assessment. Any deviations from the protocol specified window will be documented as protocol deviations. A separate definition of adequate baseline will be provided for tumor assessment related efficacy endpoints.

For efficacy analyses and baseline characteristics associated with tumor assessments the last assessment prior to or on randomization will serve as the baseline assessment.

For safety (including Eastern Cooperative Oncology Group (ECOG) performance status) the last assessment performed on or prior to date of the first dose of study treatment (or prior to randomization for patients randomized but not dosed) will serve as the baseline assessment. If there are no observations meeting these criteria, then baseline is considered missing.

For PRO the last measurement prior to or on the first dose of study treatment will be used as the baseline measurement. If there are no observations meeting these criteria, then baseline is considered missing.

Triplicate ECGs are collected; therefore the baseline for each ECG measurement is the average of the pre-dose measurements on Cycle 1 Day 1 or, if not available, the average of the screening assessment. Unscheduled assessments will not be included in the calculation of the average. Most of the ECG parameters will not be derived as they are provided by sites on the CRF. QTcB (Bazett's correction) is the only parameter that will be derived based on RR and QT. The average of the replicate measurements will be determined after the derivation of the individual parameters at each timepoint.

3.4.1. Stratification

Randomization is stratified by the following factors:

- Presence of brain metastases (Yes vs No).

Yes means that the brain lesions can be either present and observed at patient entry or known from the patient medical history to have been present in the past even if no longer visible (eg, irradiated or surgically removed lesions, etc).

- Ethnic origin (Asian vs non-Asian).

This refers to the actual patient race, not the place where he/she lives or is treated.

All stratified analyses will utilize strata as defined in the randomization system. In addition, for PFS by BICR and OS, sensitivity analyses may be conducted including as strata the baseline ECOG PS in addition to the two stratification factors from the randomization system and/or replacing the stratification factors as recorded in the randomization system with the strata derived from data captured on the CRF.

3.5. Safety Endpoints

Safety endpoints will be summarized based on the on-treatment period unless otherwise specified.

On-treatment is defined as the time from the first dose of study treatment through end of study follow-up (infinite lag) or start of new anti-cancer therapy, whichever occurs first. New anti-cancer therapy comprises systemic therapy, non palliative radiotherapy or surgery. Adverse events occurring on the same day as the first dose of study treatment will be considered to have occurred during the on-treatment period. All other assessments which occur on the same day as the first dose of study treatment will be considered baseline assessments (see Section 3.4 for the definition of baseline).

Safety data collected outside the on-treatment period as described above will be listed but not summarized.

3.5.1. Adverse Events

An adverse event is considered treatment emergent (TEAE) relative to a given treatment if:

- The event occurs for the first time during the on-treatment period and was not seen prior to the start of treatment, or

- The event was seen prior to the start of treatment but increased in severity during the on-treatment period.

Adverse Events of Special Interest (AESI)

AESIs include: central nervous system (CNS) (mood, speech, cognition) events, Psychotic effects, Edema NEC, Weight increase, AV block, Peripheral neuropathy, Pancreatitis, Pneumonitis, Hypercholesterolemia, Hypertriglyceridemia, Vision disorders, Liver Test Increased and QTc prolongation. These events will be defined based on a list of Medical Dictionary for Regulatory Activities (MedDRA) Preferred Terms (PT) specified in the Safety Review Plan for lorlatinib. A final list will be provided to programming prior to database release.

3.5.2. Laboratory Data

Hematology, chemistry, lipid and coagulation tests results will be programmatically graded according to the NCI CTCAE version 4.03 for relevant parameters. Additional details are provided in Section 6.6.3.

4. ANALYSIS SETS

Data for all patients will be assessed to determine if patients meet the criteria for inclusion in each analysis population prior to unblinding and releasing the database and classifications will be documented per standard operating procedures.

4.1. Full Analysis Set

The full analysis (FA) set will include all patients who are randomized. Patients will be classified according to the treatment assigned at randomization. Randomized but not treated patients will be reported under their randomized treatment group for the full analysis set. The FA set will be the primary population for evaluating all efficacy endpoints and patient characteristics.

4.2. Per Protocol Analysis Set

Not applicable.

4.3. Safety Analysis Set

The safety analysis (SA) set will include all patients who receive at least 1 dose of study drug. A randomized but not treated patient will be excluded from this safety analysis set. Patients will be classified according to the treatment assigned at randomization unless the incorrect treatment(s) are received throughout the dosing period, in which case patients will be classified according to the first study treatment received. The SA set will be the primary population for evaluating treatment administration/compliance and safety.

Efficacy endpoints may be assessed in this population as well.

4.4. Other Analysis Sets

CCI [redacted]

[redacted]

4.4.2. PRO Analysis Set

The PRO analysis set is defined as patients from the FA set who completed a baseline (last PRO assessment prior to the first dose of study treatment) and at least 1 post-baseline PRO assessment. The PRO analysis set will be the primary population for the analysis of change from baseline scores and time to deterioration (TTD) in patient-reported pain in chest, dyspnea, or cough. The analysis sets will be defined separately for EORTC QLQ-C30, EORTC QLQ-LC13, EQ-5D-5L and EQ-VAS.

4.4.3. Biomarker Analysis Set

The biomarker analysis set will include all patients in the SA set that have at least 1 screening biomarker assessment. Analysis sets will be defined separately for blood-based and tumor tissue-based biomarkers and will also be defined separately for non-paired and paired biomarkers.

The peripheral blood CNA analysis set is defined as patients from the SA set who have a blood-based biomarker assessment at baseline. The diagnostics concordance analysis set is defined as patients who have both a tumor tissue result captured in the CRF and a blood-based biomarker assessment at baseline.

5. GENERAL METHODOLOGY AND CONVENTIONS

5.1. Hypotheses and Decision Rules

5.1.1. Hypotheses and Sample Size Section

The primary objective of the study is to demonstrate that lorlatinib (Arm A) is superior to crizotinib (Arm B) in prolonging PFS by BICR assessment per RECIST v1.1. A key secondary objective of the study is to demonstrate that lorlatinib is superior to crizotinib in prolonging OS.

The study is designed to test H_0: $HR_{PFS} \geq 1$ vs. H_A: $HR_{PFS} < 1$, where HR_{PFS} is the hazard ratio (Arm A/Arm B) of PFS.

Approximately 280 patients (140 in each arm) were to be randomized using a 1:1 ratio stratified by presence of brain metastases and ethnic origin. As of February 28, 2019 enrollment was closed with 296 patients randomized. One hundred seventy seven (177) PFS events will be required to have at least 90% power to detect a HR of 0.611 using a one-sided stratified log-rank test at a significance level of 0.025 (one-sided), and a 2-look group-sequential design with a Lan-DeMets (O'Brien-Fleming) α-spending function to determine the efficacy boundaries.

The planned sample size was determined based on the assumption of a HR of 0.611 under the alternative hypothesis (under an exponential model, assuming median PFS of 11 months in the crizotinib arm and 18 months in the lorlatinib arm). The sample size further assumed a 15% drop-out rate within each treatment arm at 30 months, a non-uniform patient accrual over approximately 15 months and follow-up after the last patient is randomized of approximately 18 months.

This sample size will also allow comparison of OS between the 2 treatment arms, provided that superiority of lorlatinib over crizotinib with respect to PFS has been demonstrated. If the true HR is 0.70 under the alternative hypothesis (under an exponential model, assuming median OS of 48 months on the crizotinib arm and 68.6 months on the lorlatinib arm), a total of 198 deaths will be required to have 70% power using a one-sided stratified log-rank test at a significance level of 0.025 (one-sided), and a 3-look group-sequential design with Lan-DeMets (O'Brien-Fleming) α-spending function to determine the efficacy boundaries.

These calculations further assumed a 15% drop-out rate for OS on either treatment arm at 120 months, and a follow-up of approximately 110 months after the last patient is randomized.

5.1.2. Decision Rules

The original study design has been revised to have a PFS interim analysis for efficacy and a PFS final analysis. The interim analysis will be performed after approximately 133 (75%) PFS events have been documented by BICR assessment. The final analysis for PFS will occur when 177 PFS events have been documented by BICR, if the efficacy boundary has not been crossed at the interim analysis. Follow-up for OS will continue and a final OS analysis is planned to be performed when 198 deaths have occurred.

The nominal significance levels for the interim and final efficacy analyses of PFS will be determined by using the Lan-DeMets procedure with an O'Brien-Fleming stopping boundary. The overall significance level will be preserved at 0.025 (one-sided). The IA will be performed after approximately 133 PFS events (75% of the 177 events planned at the end of the study). If the value of the test statistic exceeds the efficacy boundary ($z < -2.337$, $p < 0.01$ if the interim analysis is performed after exactly 133 events), then the study is considered to have met its primary objective.

As the observed number of events at the interim analysis may not be exactly equal to the planned number of events, the efficacy boundaries will be updated based on the actual number of observed events using the pre-specified α-spending functions. Therefore, the observed Z-test statistic at the interim analysis will be compared with the updated efficacy boundaries. If the study continues to the final analysis, the p-value that will be used to declare statistical significance at the final analysis will be based on the actual number of PFS events documented at the data cutoff date for the final analysis and the α already spent at the interim analysis. If the interim analysis occurs exactly after 133 events and the study continues until the final analysis, the observed p-value for the comparison will have to be <0.022 ($Z < -2.012$) to declare statistical significance. If the number of events in the final

analysis deviates from the planned number of events, the final analysis null hypothesis rejection criterion will be determined based on the exact number of events so that the overall significance level across all analyses is maintained at one-sided 0.025.

The study will be considered positive if the primary objective is met as assessed by the stratified log-rank test for the primary PFS endpoint at the interim or final analysis.

The secondary OS endpoint will be analyzed using a hierarchical testing procedure, provided the primary PFS endpoint is statistically significant favoring the experimental arm. An α-spending function according to Lan-DeMets (O'Brien-Fleming) will be used to preserve the 0.025 overall level of significance and the repeated testing of the OS hypotheses at the interim and final analyses. The trial allows for the stopping of the study for a superior OS result, provided the primary PFS endpoint has already been shown to be statistically significant.

A maximum of three analyses are planned for OS:

1. A first interim analysis at the time of the interim/final PFS analysis (at the one that exceeded the efficacy boundary);

2. A second interim analysis when 139 deaths (70% of the total events planned for final OS analysis) are observed;

3. A final analysis when 198 deaths are observed.

The exact nominal p-values that will need to be observed to declare statistical significance at the time of these analyses for OS will depend on the number of OS events that have been observed at the time of these analyses and the α for OS already spent at the time of earlier analyses.

5.2. General Methods

Unless otherwise specified, baseline data will be summarized by treatment arm and for all treatment arms combined. Efficacy, exposure (including concomitant therapies), and safety data will be summarized by treatment arm only.

5.2.1. Data Handling After the Cut-off Date

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

5.2.2. Pooling of Data by Center

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

5.2.3. Definition of Study Day

Start day of study treatment is the day of the first dose of study treatment.

The study day for assessments occurring on or after the first dose of study treatment (eg, adverse event onset, laboratory date) will be calculated as:

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

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

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

For efficacy endpoints and for tumor assessment the study day will be calculated respect to the randomization date and not to the date of study treatment start.

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

5.2.4. Definition of Cycle and Cycle Day

Cycle start and end dates are derived per patient and not per study treatment. Both treatments will be dispensed at the beginning of every 28-day cycle.

- For Cycle X, the actual cycle start date for each subject is the earliest start date of dosing (dose>0 at that visit) in Cycle X visit CRF exposure page.
- For all but the last cycle,
  - Actual cycle stop date is calculated as the start date of the next cycle minus one day.
  - Actual cycle duration is calculated from Day 1 of a cycle to the day prior to Day 1 of the next cycle, as follows:

    Actual Cycle Duration (weeks) = (cycle stop date – cycle start date +1)/7
- For the last cycle, actual cycle duration is calculated as follows:
  - Actual Cycle Duration (weeks) = (last date of study treatment– cycle start date +1)/7

The cycle day will be calculated as:

Cycle day = Date of the assessment/event – cycle start date +1.

5.2.5. Date of Last Contact

The date of last contact will be derived for patients not known to have died at the data cutoff date using the latest complete date (ie, imputed dates will not be used in the derivation) among the following:

- All patient assessment dates (eg, blood draws (laboratory, Pharmacokinetics (PK)), vital signs, physical exam, performance status, ECG, Echocardiograms (ECHO)/MUGA/TRANS-THORACIC scans, tumor assessments, hemodynamic assessment);

- Start and stop dates of concomitant therapies including non-drug treatments or procedures;
- Completion dates for PRO Questionnaires;
- Start and end dates of anti-cancer therapies administered after study treatment discontinuation including systemic therapy, radiation, and surgeries;
- AE start and end dates;
- Last date of contact collected on the 'Survival Follow-up' CRF (do not use date of survival follow-up assessment unless status is 'alive');
- Study treatment start and end dates;
- Randomization date; and
- Date of discontinuation on disposition CRF pages (do not use if reason for discontinuation is lost to follow-up or death).

Only dates associated with actual examinations of the patient will be used in the derivation. Dates associated with a technical operation unrelated to patient status such as the date a blood sample was processed or dates data were entered into the CRF will not be used. Assessment dates after the data cutoff date will not be applied to derive the last contact date.

5.2.6. Measurable Disease

For BICR assessment, a patient with at least one target lesion identified by BICR will be considered to have measurable disease.

For investigator assessment, a patient will be considered to have measureable disease if there is at least one target lesion identified at baseline meeting the following criteria:

- Non lymph node lesions with longest diameter ≥10 mm by computed tomography (CT) scan;
- Non lymph node lesions with longest diameter ≥10 mm caliper measurement by clinical exam;
- CNS lesions with longest diameter ≥5 mm provided by gadolinium contrast enhanced magnetic resonance imaging (MRI); or
- Lymph nodes with short axis ≥15 mm when assessed by CT.

5.2.7. Tumor Assessment Date

Tumor assessment dates will be assigned differently for BICR assessment and for investigator's assessment of tumor data.

For analyses of BICR assessments, the date of tumor assessments will be provided by BICR and identified as the earliest scan/assessment date at each nominal timepoint. These dates, together with the assigned overall tumor response at each timepoint, will be used to programmatically identify the response/progression date and for censoring in time to event analyses; in fact no date of progression and response will be provided by BICR.

For analyses based on investigator's assessment when response/progression are derived programmatically from the target lesions measurements, non-target lesions status, and new lesions recorded on the CRF, the date of tumor assessment should also be derived as the earliest scan/assessment date. Since tumor assessments are captured on a log CRF page, a clustering algorithm for grouping scans will be used:

- Each cluster represents an actual visit. For each patient, the number of clusters is equal to the maximum number of assessments available among all lesions (target, non-target and new) and Investigator overall tumor assessment (IOTA) data. SAS procedure, Proc Fastclus, is applied to a variable that represents the days from the date of randomization to the unique date of the evaluation/assessment for lesion/IOTA that occurred close to each other in time will be assigned to the same cluster

5.2.8. Sum of Lesion Diameters

For lesions that are assessed as 'too small to measure', 5 mm will be imputed and used in the calculation of the sum of the lesion diameters.

5.2.9. Adequate Baseline Tumor Assessment

Adequate baseline is defined using the following criteria:

- All baseline assessments must be within 28 days prior to and including the date of randomization. However since on study tumor scans are to be performed within a ±1 week time window of the pre-specified calendar time, a similar logic is applied to the baseline scans, and all available data collected up to 5 weeks (35 days) prior to randomization will be considered for the baseline assessment.
- All documented lesions must have non-missing assessments (ie, non missing measurements for target lesions and non missing lesions status at baseline for non-target lesions).

5.2.10. Adequate Post Baseline Tumor Assessment

An adequate assessment is defined as an assessment where a response of CR, PR, SD, non-CR/non-PD, or PD has been provided by BICR or can be programmatically derived based on investigator's assessment of tumor data for the analyses by BICR or investigator

respectively. Timepoints where the response is not evaluable or no assessment was performed will not be used for determining the censoring date.

Sites of disease must be assessed using the same technique as baseline. If there is a change in technique BICR will continue to measure target lesions provided the imaging parameters do not render the measurements incomparable.

For the derivation of investigator's assessment of tumor data, change in technique is of concern if it affects target lesions.

- Any change in technique affecting NON-TARGET lesions are acceptable;
- For changes in techniques affecting TARGET lesions, impact will depend on technique change and on the lesion assessed, as shown below:

OTHER (ENHANCED CT) → Conventional CT scan – acceptable if organ=Chest, Not acceptable if ORGAN = Abdomen, Pelvis, Brain;

OTHER (ENHANCED CT) → GD MRI – acceptable if organ=Chest, Abdomen, and Pelvis, Not acceptable if ORGAN = Brain;

GD MRI →MRI – acceptable if organ=Chest, Abdomen, and Pelvis, Not acceptable if ORGAN = Brain;

GD-MRI→ CONVENTIONAL CT SCAN - – acceptable if organ=Chest, Not acceptable if ORGAN = Abdomen, Pelvis, Brain.

Changes in techniques considered not acceptable will make the timepoint response for target lesion not evaluable.

5.2.11. Unscheduled Assessments

Unless otherwise specified, unscheduled assessments will not be displayed in summary tables by nominal visit/timepoint. Unscheduled assessments will be used when deriving baseline and worst case on-treatment for safety and PRO analyses (except where noted for baseline ECGs). Additionally, unscheduled tumor assessments will be used for efficacy analyses (eg, defining date of progression/censoring, best overall response, date of last contact), CCI [redacted] will be used.

5.2.12. Standard Derivations and Reporting Conventions

The following conversion factors will be used to convert days into weeks, months or years: 1 week = 7 days, 1 month = 30.4375 days, 1 year = 365.25 days.

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

5.2.13. Analyses for Continuous Data

Continuous variables will be summarized using descriptive statistics ie, number of non-missing values and number of missing values [ie, n (missing)], mean, median, standard deviation (SD), minimum, maximum and first and third quartile (Q1 and Q3). Biomarker and PK summaries will also include coefficient of variation percent (%CV).

In case the analysis refers only to certain visits, percentages will be based on the number of patients with an assessment at that visit, unless otherwise specified.

5.2.14. Analyses for Categorical Data

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

In case the analysis refers only to certain visits, percentages will be based on the number of patients with an assessment at that visit, unless otherwise specified.

5.2.15. Analyses for Time to Event Data

The stratified log-rank test will be used for comparing treatments. Hazard ratios and the associated 95% two sided confidence intervals are estimated by Cox proportional hazards model. Repeated confidence interval will also be provided, to account for alpha spending, when appropriate.

Time to event endpoints will be summarized using the Kaplan-Meier method and estimated survival curves will be displayed graphically when appropriate. Graphs will describe the number of patients at risk over time. The median, first and third quartiles, and probabilities of an event at particular points in time will be estimated by the Kaplan-Meier method. Confidence intervals for medians and quartiles are based on the Brookmeyer-Crowley[2] method. Confidence intervals for the estimated probability of event at a particular timepoint will be generated using the log(-log) method with back transformation to a confidence interval on the untransformed scale. Summaries of the number and percentages of patients with an event and reason for censoring will also be provided on summary tables and/or figures.

5.3. Methods to Manage Missing Data

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

Any imputations will occur at the analysis dataset level. Additionally, in all patient data listings imputed values will be presented and flagged as imputed.

Missing statistics, eg, when they cannot be calculated, should be presented as ‘ND’ for not done, ‘NR’ for not reached or ‘NA’ for not applicable. For example, if N=1, the measure of variability cannot be computed and should be presented as ‘ND’ or ‘NA’.

5.3.1. Missing Dates

For purposes of data listings, dates will reflect only the information provided by the investigator on the CRF.

If start dates for adverse events or concomitant medications are completely missing a worst case approach will be taken whereby the events will be considered treatment emergent and the medications will be considered concomitant. If only partial information are available (eg, only a month and year or only a year) and the partial information provide sufficient information to indicate the dates are prior to the start of study treatment (eg, month/year less than month/year of first dose) then these will be considered to have started prior to treatment; otherwise a similar worst case approach will apply and these will be considered to have started after treatment.

Date of Last Dose of Study Treatment

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

- If the last date of study treatment is completely missing and there is no Disposition CRF page for the treatment phase and no death date, the patient should be considered to be ongoing and use the data cutoff date for the analysis as the last dosing date; or
- If the last date of study treatment is completely or partially missing and there is EITHER a Disposition CRF page for the treatment phase OR a death date available (on or prior to the data cutoff date), then impute this date as the last dose date:

  = 31DECYYYY, if only Year is available and Year < Year of min (Date of Completion/Discontinuation from the Disposition CRF page for the treatment phase, death date),

  = Last day of the month, if both Year and Month are available and Year = Year of min (Date of Completion/Discontinuation from the Disposition CRF page for the treatment phase, death date) and Month < the month of min (EOT date, death date), or

  = min (Date of Completion/Discontinuation from the Disposition CRF page for the treatment phase, death date), for all other cases.

Missing or Partial Death Dates

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

- If the entire date is missing it will be imputed as the day after the date of last contact (see derivation of date of last contact in Section 5.2.5); or

- If the day or month is missing, death will be imputed to the maximum of the full (non-imputed) day after the date of last contact and the following:

Missing day: 1st day of the month and year of death, or

Missing day and month: January 1st of the year of death.

Date of Start of New Anti-cancer Therapy

Incomplete dates for new anti-cancer therapy will be imputed as follows and will be used to determine censoring dates for efficacy analyses:

- The end date of new anti-cancer therapy will be included in the imputation for start date of new anti-cancer therapy. If the end data of new anti-cancer therapy is:

completely missing then it will be ignored in the imputations below,

partially missing with only year available then the imputations below will consider 31DECYYYY as the end date of the new anti-cancer therapy, or

partially missing with only month and year available then the imputations below will consider the last day of the month for MMMYYYY as the end date of the new anti-cancer therapy.

- For patients who have not discontinued study treatment at the time of the data cutoff date, last dose of study treatment is set to the data cutoff date in the imputations below.

- If the start date of new anti-cancer therapy is completely or partially missing then the imputed start date of new anti-cancer therapy is:

  = 31DECYYYY, if only Year is available and Year < Year of min [max (PD date + 1, last dose of study treatment +1), end date of new anti-cancer therapy]

  = Last day of the month, if both Year and Month are available and

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

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

  = min [max (PD date +1, last dose of study treatment +1), end date of new anti-cancer therapy], for all other cases.

AE Onset Date:

The following imputation rules apply if the event is unique for a patient or it is the first of a series of similar events; otherwise, the AE Onset Date will not be imputed:

- If the AE Collection Date is not missing, is less than the Date of First Exposure to Treatment, and is less than the AE Stop Date, then AE Onset Date is set to the Date of AE Collection.
- If the Previous Visit Date is greater than the Date of First Exposure to Treatment and less than the AE Stop Date, the AE Start Date is set to the previous visit date.
- If the Date of First Exposure to Treatment is greater than the previous visit date and less than the AE Stop Date, the AE Onset Date is set to the Date of First Exposure to Treatment.
- Otherwise AE Onset date is set to the AE Stop date.

AE Stop Date:

Ongoing events will have the AE Stop Date set to one of the following values:

- Date of Death, if the patient died and a date of death exists.
- Maximum of (Patient Withdraw date, AE Onset Date, AE Collection Date) if the patient withdrew from the study and a date of withdraw exists.
- Maximum of (AE Onset Date, Subject Summary Collection Date, AE Collection Date) if the Disposition CRF page for the long-term follow-up phase exists but a date of completion/discontinuation does not exists.
- Maximum of (Last Treatment Date, AE Onset Date) if no Disposition CRF page for the long-term follow-up phase exists.

Imputation will only occur if event is unique for the patient, or it is the last of a series of similar events; otherwise the Stop Date will not be imputed. Adverse Events are deemed similar if they have the same verbatim term.

Resolved events will have the AE Stop Date set to the maximum of the AE collection date and the AE Onset date.

Other Missing or Partial Dates

Imputation methods for other partial dates as follows:

- If the day of the month is missing for a start date used in a calculation, the first day of the month will be used to replace the missing date.
- If both the day and month are missing for a start date, the first day of the year is used.

- For stop dates, the last day of the month, or last day of the year is used if the day or day and month are missing, respectively.

- If the date is completely missing, no imputation will be performed.

5.3.2. Missing Toxicity Grade of Adverse Events

Prior to Study Treatment: If no toxicity grade is available or the grade is reported as unknown for an adverse event prior to the first study treatment, then Grade 1 will be assumed for purposes of defining a baseline grade for assessing if further occurrences are treatment emergent. However, if the patient experiences multiple episodes of the same AE prior to study treatment, then the maximum toxicity grade observed prior to study treatment will be utilized in assessing if further occurrences are treatment emergent.

During Study Treatment: If no toxicity grade is available or the grade is reported as unknown for an adverse event during the study treatment, then the event will be considered treatment emergent unless a baseline event was reported as Grade 4.

In summaries which present maximum toxicity grade, the maximum of non-missing grades will be displayed. Missing grade will only be displayed for cases where a patient reported only one event and the grade is missing.

5.3.3. Missing Pharmacokinetic (PK) data

Concentrations below the limit of quantification

For all calculations, figures, and estimation of individual pharmacokinetic parameters, all concentrations assayed as below the level of quantification (BLQ) will be set to zero. In log-linear plots these values will not be represented. The BLQ values will be excluded from calculations of geometric means and their confidence intervals. A statement similar to 'All values reported as BLQ have been replaced with zero' should be included as a footnote to the appropriate tables and figures. In listings BLQ values will be reported as below limit of quantification ("<LLOQ"), where LLOQ will be replaced with the corresponding value from the analytical assay used.

Deviations, Missing Concentrations and Anomalous Values

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

- A concentration has been reported as ND (ie, not done) or NS (ie, no sample);

- A deviation in sampling time is of sufficient concern or a concentration has been flagged as anomalous by the clinical pharmacologist.

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

5.3.4. Missing ECG Data

For ECG analyses, no values will be imputed for missing data (except RR values that can be derived from HR values, if present. If both RR and HR values are missing QTcB will not be dermined and only QTcF will be analyzed as provided from sites on the CRF). In case of missing RR value, RR (msec) will be derived as (60/HR (bpm))* 1000, if HR is collected.

If one or two of the triplicate measurements for an ECG parameter are missed, the average of the remaining two measurements or the single measurement can be used in the analyses. If all triplicate measurements are missing at a timepoint for an ECG parameter, no values will be imputed for this timepoint. If the triplicate needs to be repeated because of an artifact, then the repeated triplicate will be reported on an unscheduled CRF page. Based on a review of the data these unscheduled assessments may be used in place of the assessments at the nominal time. Data review and consultation with the study team is required to flag these cases.

5.3.5. Missing Biomarker Data

No missing data will be imputed.

6. ANALYSES AND SUMMARIES

6.1. Primary Endpoint(s) - Progression-free Survival as Assessed by BICR per RECIST v1.1

The primary efficacy analysis will compare PFS based on BICR assessment between the experimental arm and the control arm, and will be performed using a one-sided stratified log-rank test as described in Section 5.1. The primary population for analysis will be the Full Analysis set (FA). Strata will be based on those specified in the randomization system.

PFS is defined as the time from randomization to the date of the first documentation of PD per RECIST v1.1 as assessed by the independent oncologist or death due to any cause, whichever occurs first and will be summarized in months:

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

For purposes of censoring for time to event endpoints, anti-cancer therapies includes any systemic anti-cancer therapy (other than study treatment), radiation with curative intent or where treatment intent was not provided on the CRF, and/or surgical resection (surgery outcome: resected, partially resected, or missing) of target or non target lesions.

PFS data will be censored as follows:

- For patients who start a new anti-cancer therapy prior to an event, censoring will be at the last adequate tumor assessment (see Section 5.2.10) prior to the start of new anti-cancer therapy. Note: if date of progression occurs on the same date as the start of new anti-cancer therapy, the progression will be counted as an event. Similarly, if the start of new anti-cancer therapy corresponds to the date of last adequate tumor assessment, that tumor assessment will still be used for purposes of censoring for time to event endpoints.

- For patients with documented progression or death after two or more missing tumor assessments, censoring will occur at the last adequate tumor assessment prior to the missing assessments. In this study antitumor activity will be assessed through radiological tumor assessments conducted at screening and every 8 weeks until disease progression regardless of initiation of subsequent anti-cancer therapy. The allowable time window for disease assessments is ±1week while on treatment and whenever disease progression is suspected (eg, symptomatic deterioration). Therefore time without adequate assessment is defined as 126 days (16 weeks plus 2 weeks).

- For patients who do not have an adequate baseline tumor assessment or who do not have any post-baseline tumor assessments, censoring will occur on the date of randomization unless death occurred on or before the time of the second planned tumor assessment (ie, on or before day 126) in which case the death will be considered an event. Note for patients who died without any post baseline assessments and meet the definition of two or more missing assessment the reason for censoring will be documented as two or more missed assessments and censoring will occur on the date of randomization.

- All other patients alive without objective progression will be censored on the date of the last adequate tumor assessment.

The date of tumor response at each nominal timepoint based on the independent radiologist assessments will be used for determining the dates of last adequate assessment for censoring purposes.

The censoring and event date options to be considered for the PFS analysis are presented in Table 2.

Table 2. Outcome and Event Dates for PFS Analyses

Scenario	Date of event/censoring	Outcome
No adequate baseline assessment	Date of randomization	Censored [a]
Progression or death ≤126 days after last adequate tumor assessment or ≤126 days after date of randomization	Date of progression or death	Event
Progression or death >126 days after the last adequate tumor assessment [b]	Date of last adequate assessment [b] documenting no PD prior to anti-cancer therapy or missed assessments	Censored
No progression		
New anti-cancer therapy given prior to PD		

[a] if the patient dies ≤126 days after date of randomization the death is an event on the death date.

[b] if there are no adequate post-baseline assessments prior to PD or death, then the time without adequate assessment should be measured from the date of randomization; if the criteria were met the censoring will be on the randomization date.

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

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median PFS time with two-sided 95% confidence intervals (CI). The PFS rate at 12 months will be estimated with corresponding two-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley[2] (conftype=linear option in SAS Proc LIFETEST) and the CIs for the survival function estimates at the timepoints defined above will be derived using the log(-log) method according to Kalbfleisch and Prentice[3] (conftype=loglog default option in SAS Proc LIFETEST). The estimate of the standard error will be computed using Greenwood's formula.

Frequency (number and percentage) of patients with each event type (PD or death) and censoring reasons will be presented by treatment arm along with the overall event and censor rates.

Reasons for censoring should be summarized according to the categories in Table 3. If a patient meets multiple definitions for censoring the list will be used to define the hierarchy.

Table 3. Censoring Reasons and Hierarchy

Hierarchy	Condition	Censoring Reason
1	No adequate baseline assessment	No adequate baseline assessment
2	Start of new anti-cancer therapy before event.	Start of new anti-cancer therapy
3	Event more than 126 days from last adequate post-baseline tumor assessment/start date	Event after missing assessments[a]
4	No event and [withdrawal of consent date ≥ randomization date OR Disposition CRF page for the long-term follow-up phase = [WITHDRAWAL BY SUBJECT]	Withdrawal of consent
5	No event and lost to follow-up in any disposition page OR Disposition page for treatment phase present and Disposition page for the long-term follow-up phase absent and > 126 days between the last tumor assessment date and the data cutoff date	Lost to follow-up
6	No event and Disposition CRF page for the long-term follow-up phase and no adequate post-baseline tumor assessment	No adequate post-baseline tumor assessment
7	No event and none of the conditions in the prior hierarchy are met	Ongoing without an event

[a] more than 126 days after last adequate tumor assessment.

The PFS time or censoring time and the reasons for censoring will also be presented in a patient listing.

Time of Follow-Up for PFS

A plot will be generated to compare planned and actual relative day of tumor assessments by treatment arm. A Kaplan-Meier summary table for PFS follow-up duration will also be generated to assess the follow-up time in the treatment arms reversing the PFS censoring and event indicators.

6.1.1. Sensitivity Analyses for PFS as Assessed by BICR

The following sensitivity analyses will be performed to explore the robustness of the primary analysis results. These analyses are regarded as purely exploratory. The sensitivity analyses will repeat the primary analysis (p-value, HR and 95% CIs) on the FA described in Section 6.1 with the modifications below:

- PFS based on BICR assessment and counting all PDs and deaths as events regardless of missing assessments or timing of the event (ie, not censoring due to start of new anti-cancer therapy prior to event or due to missed assessments).

- PFS based on BICR assessment using an unstratified analysis.

- PFS based on BICR assessment analysed with stratified analysis using the two randomization stratification factors and baseline ECOG PS value from the CRF.

- Multivariable Cox proportional hazard models will also be used to explore the potential influences of baseline patient characteristics (such as age, gender, ethnic origin, presence of brain metastasis at baseline based on BICR intracranial assessment, smoking status, ECOG performance status, extent of disease, histology, etc.) on PFS. A stepwise selection procedure will serve to identify explanatory variables of potential prognostic values. Treatment will be forced into the model. Other variables are entered into and removed from the model in such a way that each forward selection step can be followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model or if the variable just entered into the model is the only variable removed in the subsequent backward elimination. The level of significance for an explanatory variable to enter the model is set to 0.15 (p-value of Score test) and the significance level for removing it is set to 0.40 (p-value of Wald test). The hazard ratios of all selected explanatory variables and of treatment effect will be reported including 2-sided 95% confidence intervals. No interactions will be considered.

Methods for evaluating the validity of model assumptions

Schoenfeld residuals for the stratified Cox proportional regression model will be plotted to investigate graphically violations from the proportional hazards (PH) assumption; a non-zero slope is evidence of departure from proportional hazards (PH). The PH assumption will be formally tested using Schoenfeld's residual test (Schoenfeld, 1982; Therneau & Grambsch, 2000).[4,5] Large departures from PH will be evidenced by a p-value <0.05.

In addition, the PH assumption will be checked visually within each stratum by plotting

log(-log(S(t))) versus log(t) ,

where S(t) is the estimated survival function for PFS at time t.

If there is evidence of large departures from proportional hazards, then PFS by BICR assessment will also be analyzed based on restricted mean survival time (RMST) differences (Zhang, 2013).[6]

The RMST up to time t* can then be interpreted as the expected survival time restricted to the common follow-up time t* among all patients. Analyses will be repeated using the following criteria to define t*:

- t^{*1} = min of (largest observed survival time for experimental arm, largest observed survival time for control arm) in months;
- t^{*2}= min of (largest event time for the experimental arm, largest event time for the control arm) in month;
- t^{*3} = the midpoint between the numbers t^{*1} and t^{*2}.

RMST can be estimated consistently by the area under the Kaplan-Meier curve over [0, t*].

The treatment effect between the experimental arm and the control arm will be assessed based on the difference in RMST. The associated 95% CI for the difference in means and one-sided p-value will be generated. RMST as a function of t* and the associated treatment effect between the experimental arm and the control arm will be plotted against time t*.

6.2. Secondary Endpoint(s)

6.2.1. Overall Survival

The primary population for analysis will be the Full Analysis set (FA). Strata will be those specified in the randomization system.

Overall survival (OS) is defined as the time from randomization to the date of death due to any cause. Patients last known to be alive will be censored at date of last contact (see Section 5.2.5). OS will be summarized in months:

$$\text{OS (months)} = [\text{date of death or censoring} - \text{start date} + 1]/30.4375$$

The primary analysis of OS will compare the OS time between the experimental arm and the control arm, and will be performed using a one-sided stratified log-rank test as described in Section 5.1.

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

OS time associated with each treatment arm will be summarized using the Kaplan- Meier method (product-limit estimates) and displayed graphically where appropriate. CIs for the 25th, 50th, and 75th percentiles will be reported. Repeated confidence interval will also be provided, to account for alpha spending. The Cox proportional hazards model will be fitted to compute the treatment HRs and the corresponding 95% CIs.

The OS rate at 12, 24 and 36 months will be estimated with corresponding two-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley[2] (conftype=linear option in SAS Proc LIFETEST) and the CIs for the survival function estimates at the timepoints defined above will be derived using the log(-log) method according to Kalbfleisch and Prentice[3] (conftype=loglog default option in SAS Proc LIFETEST). The estimate of the standard error will be computed using Greenwood's formula.

Frequency (number and percentage) of patients with an event (death) and censoring reasons will be presented by treatment arm. Censoring reasons are as follows:

- Alive;
- Withdrawal of consent;
- Lost to follow-up (Includes subjects deemed to be lost to follow-up by the Investigator and subjects with last follow-up >365 days prior to data cutoff date).

The OS time or censoring time and the reasons for censoring will also be presented in a patient listing.

Time of Follow-Up for OS

A Kaplan-Meier summary table for OS follow-up duration will also be generated to assess the follow-up time in the treatment arms reversing the OS censoring and event indicators.

6.2.1.1. Sensitivity Analyses for Overall Survival

The following sensitivity analyses will be performed to explore the robustness of the primary analysis results for OS. These analyses are regarded as purely exploratory. The sensitivity analyses will repeat the primary analysis (p-value, HR and 95% CIs) on the FA described in Section 6.2.1 with the modifications below:

- OS using an unstratified analysis.
- OS with stratified analysis using the two randomization stratification factors and baseline ECOG PS value from the CRF.

Multivariable Cox proportional hazard models will also be used to explore the potential influences of baseline patient characteristics (such as age, gender, race, ethnic origin, presence of brain metastasis at baseline based on BICR intracranial assessment,smoking status, ECOG performance status, disease stage, extent of disease, histology, etc) on OS. The same methodology described in Section 6.1.1 for PFS will be used for OS.

Methods for evaluating the validity of model assumptions

The same methodology described in Section 6.1.1 for PFS will be used for OS.

Finally, since majority of the patients enrolled in this study will likely receive additional anticancer systemic therapy upon PD by BICR, it is unclear if/how this additional therapy may impact the treatment effect estimate on OS. Based on type/frequency/distribution of follow-up therapy, analyses to correct for follow-up systemic therapy on treatment effect estimate for OS (eg 2-stage or other appropriate method) may be implemented.

6.2.2. Progression-free Survival as Assessed by Investigator per RECIST v1.1

The PFS primary analysis performed on BICR assessment will be repeated on PFS based on investigator assessment applying the same censoring rules.

BICR vs Investigator Assessment

A summary of the BICR assessment versus investigator assessment will be provided for the FA population including numbers of concordant and discordant assessments as well as the number of cases where PD was assessed at different timepoints by BICR and investigator.

The following categories will be summarized by treatment arm:

- Agreement on time and occurrence of PFS event (within 7 days) (a1);

- Agreement on PFS event but investigator event occurred later (by >7 days) (a2);
- Agreement on PFS event but investigator event occurred earlier (by >7 days) (a3);
- Investigator assessment of PFS event, BICR assessment of non-PFS event (b);
- Investigator assessment of non-PFS event, BICR assessment of PFS event (c); and
- Agreement on non-occurrence of PFS event (d).

A +/-7 day window will be used to assess the agreement on timing and occurrence of a PFS event.

Total Event Discrepancy Rate (b+c)/N, Early Discrepancy Rate (a3+b)/(a+b), Late Discrepancy Rate (a2+c)/(a2+a3+b+c), and Overall Discrepancy Rate (a2+a3+b+c)/N will be calculated where a=a1+a2+a3 (Amit et al. 2011).[7]

6.2.3. Objective Response

Objective Response Rate (ORR) is defined as the percentage of patients with a best overall confirmed response of CR or PR according to RECIST v1.1. Patients without documented CR or PR will be considered as non-responders. The evaluation of ORR will be provided both based on BICR and Investigator and will be relative to the FA population.

The ORR on each treatment arm will be estimated by dividing the number of patients with OR (CR or PR) by the number of patients randomized to the respective treatment arm. The corresponding exact 2-sided 95% CIs will be provided. OR comparison between the 2 treatment arms will be assessed using Cochran-Mantel-Haenszel (CMH) test using the stratification factors and risk ratio and the corresponding 2-sided 95% CI will be provided as well.

Best overall response (BOR) will be determined based on reported overall responses at different evaluation timepoints, by the independent radiologist and by the investigator respectively, from the date of randomization until documented disease progression or start of new anti-cancer therapy, according to the following rules:

- CR = at least two determinations of CR at least 4 weeks apart and documented before progression and start of new anti-cancer therapy.
- PR = at least two determinations of PR or better (and not qualifying for a CR) at least 4 weeks apart and before progression and start of new anti-cancer.
- SD (for patients with at least one measurable lesion at baseline)= at least one SD assessment (or better and not qualifying for CR or PR) ≥6 weeks after date of randomization and before progression and the start of new anti-cancer therapy.

- Non-CR/Non-PD (for patients with only non-target disease at baseline) = at least one Non-CR/Non-PD assessment (or better and not qualifying for CR or PR) ≥6 weeks after date of randomization and before progression and the start of new anti-cancer therapy.

- PD = progression ≤18 weeks after date of randomizationand not qualifying for CR, PR or SD.

- Not Evaluable (NE) = all other cases.

Clinical deterioration will not be considered as documented disease progression.

The frequency (number and percentage) of patients with BOR of CR, PR, SD, PD, non-CR/non-PD (applicable only to patients with non-measurable disease at baseline), and NE (not-evaluable) will be tabulated.

6.2.4. Intracranial Objective Response

Assessment of intracranial disease will be made using a modified version of RECIST v.1.1.[8]

The IC-OR, based on BICR intracranial assessment, will be summarized similar to OR as described above in the subset of the FA population with at least 1 baseline intracranial lesion (based on BICR intracranial assessment). If data permits IC-OR will also be summarized in the subset of patients with at least 1 baseline measurable intracranial lesion (based on BICR intracranial assessment). Surgery or radiotherapy of extracranial lesions will not affect the determination of IC-OR.

6.2.5. Intracranial Time to Progression

IC-TTP based on BICR intracranial assessment will be calculated for the FA population and for subgroups of patients with and without brain metastases at baseline (based in BICR intracranial assessment). For the subgroup of patients without brain metastases at baseline only the new brain metastases will be considered events (based on BICR intracranial assessment).

For IC-TTP the same censoring rules descrived above for PFS will be implemented with the exception of patients that will die without PD since death in this analysis is not considered an event but a reason for censoring. In addition patients will not be censored in case of surgery or radiotherapy as described in Section 6.1, if surgery or radiotherapy involve an extracranial lesion.

Differences between treatment arms will be assessed by the 1-sided stratified log-rank test. IC-TTP associated with each treatment arm will be summarized using the Kaplan -Meier method and displayed graphically where appropriate. CIs for the 25th, 50th, and 75th percentiles will be reported. The Cox proportional hazards model will be fitted to compute the treatment HRs and the corresponding 95% CIs.

6.2.6. Probability of First Event Being a CNS Progression, Non CNS Progression, or Death

The probability of first event being a CNS progression (based on BICR intracranial assessment), a non CNS progression (based on BICR overall assessment), or Death will be evaluated with a Competing Risk approach by estimating Cumulative Incidence Functions, by treatment groups. All patients in the FA population will be included in this analysis, regardless of their baseline status of CNS metastases.

The time to first event being a Competing Event (either "CNS progression" or "non CNS progression" or "Death") is defined as time from randomization until the date of that specific event. Patients not known to have any of the Competing Events are censored on the date of the last adequate tumor assessment.

The event will be identified based on BICR data for overall assessment and intracranial assessment as presented in Table 4:

Table 4. Algorithm for CNS Progression, Non CNS Progression, or Death

IC-PD	Overall-PD	Additional condition	First event	Event date
Yes	Yes	IC-PD date <= overall PD date +7 days	CNS	IC-PD
Yes	Yes	IC-PD date > overall PD date +7 days	non-CNS	overall-PD
Yes	No	None	CNS	IC-PD
No	Yes	None	non-CNS	overall-PD
No	No	Patient died and death was a PFS event (death >126 days after the last adequate tumor assessment will not be considered)	Death	death
No	No	Patient alive at the data cutoff date	No event	NA

Patients who presented with one type of event are counted as a competing cause of failure for the analysis of other type of events. For example, in the estimation of time to first event being CNS progression, the patients having a non CNS progression (or those dying before the CNS progression occurs) are counted as competitive cause of failure.

To account for the competing risks inherent in the comparison of the probability of first event being a CNS progression between the lorlatinib and crizotinib groups, a stratified 1-sided logrank test will be computed on the basis of a cause-specific hazard function. HRs and 95% CI will be estimated by the stratified Cox proportional hazards model and plots for the cumulative incidences will be provided. The same analysis will be provided also on the probability of first event being a non-CNS progression and probability of death without prior CNS or non-CNS progression.

6.2.7. Duration of Response

Duration of Response (DR) based on BICR is defined, for patients from the FA population with a BOR of CR or PR, as the time from first documentation of objective response (CR or PR) per RECIST 1.1 based on BICR to the date of first documentation of objective progression of disease (PD) or death due to any cause. The censoring rules for DR are as described for PFS in Section 6.1.

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

Depending on the number of subjects who achieve a BOR of CR or PR and subsequently have an event in each arm, Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median DR time with two-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley.[2] Otherwise only listings will be provided. Inferential analyses (eg, a hazard ratio and p-value) will not be provided for duration of response.

6.2.8. Intracranial Duration of Response

The IC-DR, based on BICR intracranial assessment, will be summarized similar to DR as described above in the subset of the FA population with an IC BOR of confirmed CR or PR, as the time from first documentation of intracranial objective response (CR or PR) per modified RECIST 1.1 based on BICR to the date of first documentation of intracranial objective progression of disease (PD) or death due to any cause. If IC-OR will be summarized in the subset of patients with at least 1 measurable intracranial lesion, a corresponding IC-DR will be provided. The censoring rules for IC-DR are as described for PFS in Section 6.1, but surgery or radiotherapy that involve an extracranial lesion will not be considered.

6.2.9. Time to Tumor Response

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

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

6.2.10. Intracranial Time to Tumor Response.

Intracranial TTR based on BICR intracranial assessment is defined, for patients with a confirmed IC-OR, as the time from the date of randomization to the first documentation of intracranial objective response (CR or PR) which is subsequently confirmed.

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

6.2.11. Progression Free Survival 2

PFS2 is defined as the time from randomization to the date of progression of disease on first subsequent systemic anti-cancer therapy, or death from any cause, whichever occurs first. If no date of disease progression on first subsequent systemic anti-cancer therapy is available, patients will be censored at date of last contact.

Table 5. Outcome and Event Dates for PFS2 Analyses

	Date of event/censoring	**Outcome**
(No PD[a]) and (no death)	Date of last adequate tumor assessment documenting no PD	Censored
(No PD[a]) and death	Date of death	Event (Death)
(PD[a] date >SST start date) and (no death)	Date of last adequate tumor assessment documenting no PD prior to start date of SST	Censored
(PD[a] date > SST start date) and death	Date of death	Event (Death)
(PD[a] date ≤ SST start date) and (PD2 date > SST start date)	Date of PD2	Event (PD)
(PD[a] date ≤ SST start date) and (Death)	Date of death	Event (Death)
(PD[a] date ≤ SST start date) and (no PD2) and (no death)	Last contact date	Censored
(PD[a] and no SST) and (no death)	Last contact date	Censored
[a] PD is the first PD by investigator assessment per RECIST v1.1 as per PFS. SST is the subsequent systemic anti-cancer treatment. PD2 is the first PD after subsequent systemic anti-cancer treatment by investigator assessment.		

A sensitivity analysis will be conducted including the date of discontinuation from first subsequent systemic anti-cancer therapy as an event (if reason for treatment discontinuation is not COMPLETED THERAPY) when the date of disease progression is missing.

Differences between treatment arms will be assessed by the 1-sided stratified log-rank test. PFS2 time associated with each treatment arm will be summarized using the Kaplan- Meier method and displayed graphically where appropriate. CIs for the 25th, 50th, and 75th percentiles will be reported. The Cox proportional hazards model will be fitted to compute the treatment HRs and the corresponding 95% CIs.

6.2.12. Patient-Reported Outcomes

Patient reported outcomes of global QOL, functioning and cancer specific disease/treatment related symptoms, will be assessed with the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30, Version 3.0), and its corresponding module for lung cancer (QLQ-LC13, Version 3.0) and the EuroQol 5 dimension 5 level (EQ-5D-5L) questionnaire.

6.2.12.1. Scoring Procedure

EORTC QLQ-C30, EORTC QLQ-LC13 and EQ-5D-5L will be scored according to their respective validation papers and User's Guides. In summary, each scale of the EORTC QLQ-C30 and the QLQ-LC13 will be transformed so that scale scores will range from 0 to 100. After scores are transformed, higher scores on the EORTC QLQ-C30 will represent higher ("better") levels of functioning and/or a higher ("worse") level of symptoms. Higher scores on the EORTC QLQ-LC13 will represent higher ("worse") level of symptoms. In the EQ-5D-5L, patient's health state is measured on 5 dimensions and the respondent's self-rated health is assessed on a scale from 0 (worst imaginable health states) to 100 (best imaginable health state) by the EQ-VAS. Missing values in the instruments will also be handled following the guidance from their respective User Manuals.[9,10]

6.2.12.2. Instrument Completion Rate

For each treatment arm and at each time point, the number and percentage of patients who complete the EORTC QLQ-C30, EORTC QLQ-LC13 and EQ-5D-5L will be summarized, as will the reasons for non-completion of these measures.

6.2.12.3. Descriptive Summaries Over Time

For each treatment arm and at each assessment time point, summary statistics of the EORTC QLQ-C30, EORTC QLQ-LC13 and EQ-5D-5L at each time point will be reported.

For continuous endpoints, change from baseline will also be reported with N (number of non-missing values), mean, standard error, and 95% confidence intervals. Only patients with both a baseline score and a post-baseline score will be included in the summary of change from baseline scores.

6.2.12.4. Time-to-Event Endpoints

TTD in pain in chest, dyspnea, or cough individually and as a composite endpoint will be defined as the time from randomization to the first time the patient's score shows a 10 point or greater increase after baseline in any of the 3 symptoms.

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

Patients will be censored at the last time they completed a subscale assessment if they have not deteriorated. Death will not be considered an event for TTD.

TTD in each of the 3 symptoms as well as composite endpoint will be compared between treatment arms using an 1-sided log-rank test stratified by randomization stratification factors. The treatment effect will be estimated using a Cox Proportional Hazard model. Ties will be handled by replacing the proportional hazards model by the discrete logistic model (Ties=Exact option in SAS PROC PHREG).

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median TTD time with two-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley. Kaplan-Meier estimates of TTD in each of the 3 symptoms as well as composite endpoint will be graphically displayed.

6.2.12.5. Binary and Categorical Endpoint

Improvement, worsening, and stable categories will be determined over each cycle and summarized as an average by patient in the symptom scales, functioning scales, and in the global QOL scale. In the symptom scales, improvement is defined as a decrease of at least 10 points. I n the functioning and global QOL scales improvement is defined as an increase of at least 10 points. In the symptom scales worsening is defined as an increase of at least 10 points. In the functioning scale and global QOL scales, worsening is defined as a decrease of at least 10 points. Patients with global QOL, functioning scales, or symptom scales that have not improved nor worsened will be considered stable Osoba et al, established that a ≥10 point minimally important difference from baseline (ie, the first PRO measurement prior to the first dose of study treatment) on the scales of the EORTC QLQ-C30 would correlate with significant (moderate) change in disease symptoms and functioning.[11]

For the EQ-5D-5L health state profiles, the proportions of patients reporting "no", "slight", "moderate", "severe", or "extreme" problems will be reported at each time point, by treatment arm.

6.2.12.6. Continuous Endpoints

Longitudinal random intercept random slope mixed-effect model will be carried out for the total and subscale scores of the EORTC QLQ-C30/LC13 using PROC MIXED. Outcomes are PRO post-baseline scores and the predictors are the corresponding baseline PRO score, treatment, randomization stratification factor, time (treated as a continuous variable), and treatment-by-time interaction. Intercept and time are considered as random effects particular to each subject. All available data for each patient should be used in the analyses. All parameter estimates should be obtained using restricted maximum likelihood. The unstructured covariance structure should be used to define covariance between random effects (using option "Type=UN" as a part of the RANDOM statement in PROC MIXED). For the degrees-of-freedom calculations the Kenward and Roger algorithm should be used (using option "ddfm = kr" as a part of the MODEL statement in PROC MIXED).

6.3. Other Endpoint(s)

6.3.1. Biomarker Analysis for Secondary CCI [redacted] Endpoints

Biomarkers will be assessed separately for whole blood, serum, plasma, archival tumor tissue, and de novo tumor tissue biospecimens. In each case, summaries of baseline levels, changes from baseline (where appropriate) as measured by ratio to baseline, expression or genetic alterations will be reported. For continuous variables, summary statistics may include the mean, ratio to baseline, standard deviation, median, first and third quartiles, %CV

and minimum/maximum levels of biomarker measures; for categorical variables, summary may include number and percentage, odds ratio, frequency statistics, as appropriate.

No duplicate results for biomarker CCI [redacted] are expected. However, if more than one record is received for a particular time point, the duplicate records might be averaged.

Data from biomarker assays may be analyzed using graphical methods and descriptive statistics such as Wilcoxon Signed Rank, Wilcoxon Rank Sum Test, Kaplan-Meier estimates of efficacy parameters (eg, PFS, OS) with biomarkers as a covariate, or Box plots. The statistical approach will examine correlations of biomarker results with pharmacokinetic parameters and measures of anti-tumor efficacy.

CCI [redacted]

Analyses above will be conducted in the biomarker analysis sets that will be defined separately for blood-based and tumor tissue-based biomarkers and will also be defined separately for non-paired and paired biomarkers.

A concordance analysis between tumor tissue and peripheral blood cfDNA biomarker results for ALK rearrangements will also be conducted.

6.3.1.1. CNA Mutations

Analyses of CNA Mutations will be based on the peripheral blood CNA analysis set.

CNA mutations are measured at Screening, C2D1, C7D1 and EOT. Patients are expected to have none or one or more than one ALK kinase mutation, and patients may have a change in mutation during and/or after treatment.

Data will be summarized separately by treatment arm, lorlatinib or crizotinib. A frequency table will be provided for each timepoint for all mutations, and a frequency table of none vs ≥1 mutation will be provided. A summary table of the type of ALK kinase mutation(s) at Screening will be provided. A summary table of the type of EML4-ALK fusion variant at Screening will be provided. All data will be listed.

Five (5) EML4-ALK fusion subtypes will be analyzed:

- EML4-ALK v1;
- EML4-ALK v2;

- EML4-ALK v3;
- EML4-ALK other (will include v4, v5, v7, v8 and similar);
- ALK rearrangement other (all ALK fusion that does not involve EML4 as the fusion partner).

For treatment arm, frequency tables of the ALK kinase mutations and of the EML4-ALK fusion subtypes at Screening vs. OR (yes=objective response, and no=no objective response), and one vs ≥1 mutation will be provided.

PFS and OS analyses will be stratified none vs ≥1 mutation for ALK kinase mutation, and for EML4-ALK fusion variant, stratified by each of the 5 subtypes, and will follow the procedure described in Section 6, using Kaplan-Meier curves with appropriate summary statistics including median time and associated two-sided 95% CIs.

Diagnostic concordance assessment of ALK gene rearrangement status between the plasma-based Guardant Health G360 IUO test and the reference, central, IHC tumor tissue test result will be presented for the FA set. The reference, central test is the tumor tissue IHC testing performed at screening as part of the enrollment criteria, and all patients are therefore considered ALK gene rearrangement positive (ie, detected). Overall concordance will be summarized, by treatment arm and in total.

6.3.1.2. Tumor Tissue Analysis

Analyses of Tumor Tissues will be based on the Tumor Tissue Analysis Set.

DNA mutations in tumor tissue are measured at screening. Patients are expected to have none or one or more mutations.

The study will be summarized separately by treatment arm, loratinib or crizotinib. A frequency table of mutations at screening and none vs. ≥1 mutation will be provided. All data will be listed.

For each treatment arm, frequency tables of the mutations at Screening vs. OR (yes=objective response, and no=no objective response), and one vs >1 mutation will be provided.

PFS and OS analyses will be stratified none vs ≥1 mutation, and will follow the procedure described in Section 6.

6.3.2. Pharmacokinetic Analysis

CCI [redacted]

In addition, cerebrospinal fluid (CSF) samples (for both treatments), CCI [redacted] if available, will be listed by patients.

CCI [redacted]

6.4. Subset Analyses

All the subset analyses will be exploratory; no adjustment for multiplicity will be performed. Analyses will only be performed if there is sufficient sample size. The determination of whether or not there is sufficient sample size will be defined after enrollment is complete and prior to database lock. As a general rule, time to event analyses will not be performed on subgroups unless there are ≥10 events on each treatment arm within the defined subset and analyses of response rates and safety will only be performed if there are ≥20 patients on each treatment arm within the defined subset. Deviations from these analyses will be described in the clinical study report.

The following subset analyses will be performed for PFS and ORR, by BICR assessment, and OS based on the FA population:

> age (<65 years, ≥65 years), gender (male, female), ethnic origin (Asian vs non-Asian, from IVRS) , presence of baseline brain metastasis (Yes vs No, from BICR intracranial assessment), smoking status (never vs. current/former), ECOG performance status (0/1 vs. 2), extent of disease (locally advanced vs. metastatic), histology (adenocarcinoma vs. non adenocarcinoma).

Subset analyses for PFS, OS will use the primary censoring rules described in Sections 6.1 and 6.2.1. Treatment arms will be compared using a one-sided unstratified log rank test for each subset level and the unstratified HR and its corresponding 95% CI will be computed per subset level. The HR for PFS and OS and corresponding 95% CIs for all subgroups will also be presented in a forest plot.

Subset analyses according to age (<65 vs. ≥65), gender (male vs. female) and race (Asian vs non-Asian) will be performed for adverse events based on the SA population.

6.5. Baseline and Other Summaries and Analyses

6.5.1. Baseline Summaries

The following analyses will be based on the FA population overall and separately by treatment arm.

- **Demographic and Physical Characteristics**

  The following demographic and baseline characteristics will be summarized by number and percentage:

  - Gender (male, female);
  - Age (18-<45; 45- <65; ≥65);
  - Race (white, black, asian, other), Racial designation for Asian (Japanese, Korean, Chinese, other), ethnicity (Hispanic/Latino, non-Hispanic/Latino);
  - Eastern Cooperative Oncology Group (ECOG) Performance status;
  - Smoking classification (never smoker, former smoker, current smoker).

  Age (continuous) will be summarized with descriptive statistics (mean, median, standard deviation, minimum, and maximum).

- **Disease Characteristics**

  The following baseline disease characteristics will be summarized by number and percentage:

  - Histopathological Classification;
  - TMN stage at initial diagnosis and current stage;
  - Extent of disease (locally/advanced, metastatic);
  - Measurable disease at baseline (yes/no) by BICR (see Section 5.2.6);
  - Adequate baseline assessment (yes/no) by BICR (see Section 5.2.9);
  - Involved tumor sites at baseline by BICR;
  - Brain disease at baseline by BICR intracranial assessment.

  Involved tumor sites at baseline by BICR will be derived from target and non target lesions at baseline. Each patient will be counted once per organ "Other" will be counted as one organ site.

- **Stratification Factors**

  The number of patients per strata will be summarized based on strata reported in the randomization system. The following categories will be considered:

  - Presence of brain metastases Yes and ethnic origin Asian;
  - Presence of brain metastases No and ethnic origin Asian;
  - Presence of brain metastases Yes and ethnic origin non-Asian;
  - Presence of brain metastases No and ethnic origin non-Asian.

- **Medical History**

  - Medical history will be coded using the most current version of MedDRA and summarized by MedDRA's System Organ Class (SOC) and PT. Each patient will be counted only once within each PT or SOC. Summaries will be ordered by primary SOC and PT in descending order of frequency by the experimental treatment arm. Separate summaries will be provided for past and present conditions.

- **Prior Anti-Cancer Treatments**

  Prior anti-cancer treatments include systemic therapy, radiation, and surgery.

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

  - Patients with at least one prior anti-cancer systemictherapy;
  - Patients with at least one prior anti-cancer radiotherapy;
  - Patients with at least one prior anti-cancer surgery, excluding procedure=biopsy.

    Patients with prior anti-cancer systemic therapy will be further summarized based on therapy intent: Neo-Adjuvant/Adjuvant.

6.5.2. Study Conduct and Patient Disposition

The following analyses will be based on the FA population overall and separately by treatment arm.

6.5.2.1. Patient Disposition

A summary of the number of patients enrolled by country and site will be provided.

Discontinuation for each phase will be summarized separately and by reason for discontinuation. Discontinuations from study treatment will be summarized using the SA population and discontinuations from long-term follow-up will be summarized using the FA population.

6.5.2.2. Protocol Deviations

Protocol deviations will be compiled prior to database closure and will be summarized by category (n(%)) for the FA population by treatment arm. Categories will be assigned by the study Clinician.

6.5.3. Study Treatment Exposure

The following analyses will be based on the SA population.

6.5.3.1. Exposure to Lorlatinib

The summary of treatment exposure for lorlatinib will include the following information:

- Treatment duration (months);
- Cumulative dose (mg);
- Dose intensity (mg/week);
- Relative dose intensity (%).

The duration of lorlatinib (in months) is defined as:

$$\text{Treatment duration (months)} = (\text{last dose date} - \text{first dose date} + 1)/ 30.4375$$

The cumulative dose (mg) of lorlatinib is the sum of the actual dose levels that the patient received (ie, total dose administered (mg)). Lorlatinib is given once daily; therefore if all doses (eg, d mg/day) were received in a week the cumulative dose would be 7×100 mg.

The dose intensity (DI) and the relative dose intensity (RDI) of lorlatinib will be calculated for each patient during the study. The DI (mg/week) of lorlatinib during the study is defined as:

$$\text{DI (mg/week)} = [\text{cumulative dose (mg)}]/[\text{treatment duration (weeks)}]$$

The RDI of lorlatinib is defined as the ratio of the DI and planned dose intensity and expressed in %:

$$\text{RDI (\%)} = 100\times [\text{DI (mg/week]}/[7\times 100 \text{ (mg/week)}].$$

6.5.3.2. Exposure to Crizotinib

The summary of treatment exposure for crizotinib will include the following information:

- Treatment duration (months);
- Cumulative dose (mg);
- Dose intensity (mg/week);
- Relative dose intensity (%).

The duration of crizotinib (in months) is defined as:

Treatment duration (months) = (last dose date – first dose date +1)/30.4375

The cumulative dose (mg) of crizotinib is the sum of the actual dose levels that the patient received (ie, total dose administered (mg)). Crizotinib is given twice daily; therefore if all doses (eg, 2×250 mg/day) were received in a week the cumulative dose would be 7×2×250 mg.

The dose intensity (DI) and the relative dose intensity (RDI) of crizotinib will be calculated for each patient during the study. The DI (mg/week) of crizotinib during the study is defined as:

DI (mg/week) = [cumulative dose (mg)]/[treatment duration (weeks)]

The RDI of crizotinib is defined as the ratio of the DI and planned dose intensity and expressed in %:

RDI (%) = 100× [DI (mg/week]/[7×2×250 (mg/week)].

6.5.3.3. Dose Reductions, Interruptions, and Delays

A dose reduction is defined as a non-zero dose that is less than the prior dose.

The number and percentage of patients with at least one dose reduction as well as a breakdown of dose reductions (1/2/≥3) will be summarized by treatment arm.

Reasons for dose reductions will also be summarized. Patients can contribute to more than one reason if multiple dose reductions occurred for different reasons, but will only be counted once per reason. Percentages will be calculated based on the total number of patients in the safety analysis set.

An interruption is defined a 0 mg dose administered for more than 1 day. The number and percentage of patients with dose interruptions and the corresponding reasons will be summarized by treatment arm. Patients can contribute to more than one reason if multiple dose interruptions occurred for different reasons, but will only be counted once per reason. Percentages will be calculated based on the total number of patients in the safety analysis set.

6.5.4. Concomitant Medications

Concomitant medications received by patients during the study will be summarized for the safety analysis population by treatment arm.

Concomitant medications refer to all medications which started prior to first dose of study treatment and continued during the on-treatment period (see Section 5.2) as well as those started during the on-treatment period. Concomitant medications will be coded in the WHO Drug coding dictionary and will be tabulated by preferred term in descending order of frequency in the lorlatinib arm. In case of equal frequency regarding drug class (respectively drug name), alphabetical order will be used. A patient will be counted only once within a given drug name, even if he/she received the same medication at different times.

6.5.5. Subsequent Anti-Cancer Therapies/Procedures

Subsequent Anti-Cancer Therapies and Procedures are defined as therapies entered on the 'Follow-up Cancer Therapy', 'Follow-up Radiation Therapy', and 'Follow-up Surgery' CRF pages. The number and percentage of patients within each category (medication therapy, radiation therapy, and surgeries) will be provided by treatment arm.

Medications will be coded using the WHO Drug coding dictionary and will be tabulated by drug category and preferred term in descending order of frequency on the lorlatinib arm.

Analyses will be based on the FA population by treatment arm.

6.6. Safety Summaries and Analyses

Unless otherwise specified, summaries of AEs and other safety parameters will be based on the safety analysis population by treatment arm.

6.6.1. Adverse Events

All analyses will be based on treatment emergent events unless otherwise specified. Treatment emergent is defined in Section 3.5.1. AEs not considered treatment emergent will be flagged in data listings.

A high level summary of adverse events will include the number and percent of patients with:

- Any Adverse Event;
- Serious AE;

- Adverse Events with CTCAE Grade 3-4;
- Grade 5 events;
- AEs leading to dose interruptions;
- AEs leading to dose reductions;
- AEs leading to withdraw.

Each unique adverse event at the PT level in each treatment arm of the study for a patient is included in the count.

Seriousness, toxicity grade, action taken (interruption, reduction, and withdraw) are as reported by the investigator on the adverse event CRF.

Summaries by treatment arm, SOC and/or PT in decreasing frequency based on the frequencies observed in the lorlatinib arm will be provided for:

- Treatment Emergent Events by Maximum Toxicity Grade (All Causality);
- Treatment Emergent Events by Maximum Toxicity Grade (Treatment Related);
- Serious Treatment Emergent Events by Maximum Toxicity Grade (All Causality);
- Serious Treatment Emergent Events by Maximum Toxicity Grade (Treatment Related);

An event will be considered treatment related if the investigator considered the event related to the study drug or this information is unknown.

Because the frequency of certain medical concepts or conditions may have been underestimated by reliance on a single MedDRA preferred term (PT), selected PTs were aggregated in cluster terms.

The following summaries will be provided by treatment arm and PT/clusters of AEs (ie, summaries will not include SOC) in decreasing frequency based on the frequencies observed in the lorlatinib arm for:

- Treatment Emergent Events (All Causality) Experienced by ≥10% of patients in either treatment arm;
- Treatment Emergent Events (All Causality) by Preferred Term and Maximum Toxicity Grade;

- Treatment Emergent Grade 3-5 Events (All Causality) by Preferred Term and Maximum Toxicity Grade;
- Treatment Emergent Adverse Events Leading to Dose Interruptions by Maximum Toxicity Grade (All Causality);
- Treatment Emergent Adverse Events Leading to Dose Reductions by Maximum Toxicity Grade (All Causality);
- Treatment Emergent Adverse Events Leading to Permanent Withdraw by Maximum Toxicity Grade (All Causality);
- Serious Treatment Emergent Events (All Causality).

Each patient will be counted only once within each PT.

As described in Section 5.3.2, in case a patient has events with missing and non missing toxicity grades, the maximum of the non-missing grade will be displayed. Missing grade will only be displayed in the event that only one event has been reported for a patient and the grade is missing.

6.6.1.1. Basic Results

For basic results disclosures in the US and EU the following additional summaries will be provided:

- Treatment Emergent Non Serious Adverse Events by SOC and PT in >5% of patients in either treatment arm;
- Treatment Emergent Serious Adverse Events by SOC and PT; and
- Fatal Adverse Events by SOC and PT.

Each of the above summaries will include a count of the number of patients with all causality events and the number of patients with treatment related events.

6.6.1.2. Adverse Events of Special Interest

These analyses will be performed for treatment emergent AEs of special interest as specified in 3.5.1.

- **Time to AE Onset** (in days) is defined as the time from the date of the first dose to the onset date of the AE, regardless of grade. If a patient has multiple episodes of an AE, the date of the first occurrence is used. Time to AE onset (in days) will be calculated as (AE start date – first dose date +1). Time to onset is calculated for the subgroup of patients who had the specific AE.

- **Time to Grade 2 or 3 or 4 AE Onset** (in days) is defined similarly as time to AE onset for Grade 2 or 3 or 4 AEs.

- **Duration of AE** (in days) is defined as the cumulative duration across episodes of the AE, regardless of grade, where duration for each episode is the time from the AE start date to the AE end date. For one episode, duration (in days) = AE end date – AE start date +1. If a patient has multiple episodes of an AE, cumulative duration across all episodes will be used adjusting for any overlap. If a patient has an AE that was ongoing at the time of analysis, the time is censored at the last available on treatment visit date. Duration is calculated for the subgroup of patients who had the specified AE.

Descriptive statistics will be presented for time to AE onset (days), time to Grade 2 or 3 or 4 AE onset, and duration of AEs for the subgroup of patients with the AE.

Person-year exposure analyses may also be considered in case of relevant difference in treatment duration between treatment arms.

6.6.2. Deaths

The frequency (number and percentage) of patients in the safety analysis set who died and who died within 28 days after last dose of study treatment as well as the primary reason for death, will be tabulated based on information from the 'Notice of Death' and 'Survival Follow-Up' CRFs, by treatment arm.

The frequency (number and percentage) of patients in the safety analysis set who died within 30 days of first dose of study treatment will also be provided.

Date and cause of death will be provided in individual patient data listing together with selected dosing information (study treatment received, date of first/last administration, dose).

6.6.3. Laboratory Data

Laboratory results will be converted to International System of Units (Système International d'unités, SI) units which will be used for applying toxicity grades and for all summaries.

As described in Section 3.4, baseline will be defined as the last assessment performed on or prior to date of the first dose of study treatment (or prior to randomization for patients randomized but not dosed). If there are multiple assessments that meet the baseline definition on the same day without the ability to determine which was truly last, then the worst grade will be assigned as the baseline grade. Since a few CTCAE terms (including Hypo/Hypercalcemia and Activated Partial Thromboplastin) can be derived using several laboratory tests (analytes) refer to the section 2.3.7 of the "Pfizer Oncology CTCAE Grading Implementation Guidance for Laboratory Data" for determination of baseline CTCAE grade in this situation.

Laboratory results will be programmatically classified according to NCI-CTCAE version 4.03 grade. Non-numerical qualifiers will not be taken into consideration in the derivation of grade (eg, hypokalemia Grade 1 and Grade 2 are only distinguished by a non-numerical qualifier and therefore Grade 2 will not be derived). In summary statistics the number and percentage of patients corresponding to grades that only include non-quantitative criteria will be displayed as a blank or NA (not assessed) rather than 0. If there is any overlap between grade criteria (eg, CTCAE grading criteria for Creatinine Increased – a value can fall into one range based on comparison to ULN and another range based on comparison to baseline), the highest (worst) grade would be assigned to that record. Grade 5 is defined in the CTCAE criteria guidance as an event with an outcome of death. Since laboratory data does not collect an outcome, Grade 5 is not used when programmatically grading laboratory data.

Grade 0 or Outside Toxicity Reference (OTR) is not defined specifically by in the CTCAE guidance. However, programmatically this is used as a category to represent those patients who did not meet any of the Grades 1 to 4 criteria. If the laboratory value is evaluable for CTCAE criteria grading (numeric value is present, valid units and ranges are present as required to allow conversion to standard units and grading), and does not qualify for any of the Grade 1-4 criteria for a given lab test, then the value is assigned as Grade 0 or OTR.

Abnormalities will be described using the worst grade post-baseline. When determining the maximum post-baseline grade for a given patient and CTCAE test, the maximum across all analytes and assessments contributing to that CTCAE test will be used. Several laboratory tests have bi-directional grading criteria defined so that both low (hypo) and high (hyper) values can be graded separately. Each criterion will be summarized separately. In the cases where a value is graded as a Grade 1, 2, 3, or 4 for one of the directions, that value will also be assigned as a Grade 0 for the opposite direction for that test. For example, a value meeting the criteria for Grade 3 Hypercalcemia will be classified as a Grade 0 Hypocalcemia. For CTCAE terms that can be derived using one of several laboratory tests, the maximum post-baseline grade for a given patient and CTCAE term will be the maximum across all possible laboratory tests.

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

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

If the investigator reports both the absolute and % value for Neutrophils or Lymphocytes from the same laboratory sample date and patient, ONLY the absolute value will be graded. The % value will not be graded in this scenario.

If the % value is converted to the differential absolute count for grading and the LLN for the differential absolute count is not available (only LLN for % is available) then Grade 1 will be assigned if the following conditions are met:

- Lymphocyte count decreased:

derived absolute count does not meet Grade 2-4 criteria, and

% value < % LLN value, and

derived absolute count ≥800/mm^3.

- Neutrophil count decreased

derived absolute count does not meet Grade 2-4 criteria, and

% value < % LLN value, and

derived absolute count ≥1500/mm^3.

For **calcium**, CTCAE grading is based on Corrected Calcium and Ionized Calcium. Corrected Calcium is calculated from Albumin and Calcium as follows:

- Corrected Calcium (mg/dL) = Calcium (mg/dL) – 0.8 [Albumin (g/dL)-4].

Laboratory toxicities will be tabulated using descriptive statistics (number of patients and percentages):

- Summary of laboratory parameters by worst CTCAE grade post-baseline table will include number and percentage of patients with Grade 0, 1, 2, 3, 4, and Grade 1-4 laboratory abnormalities.

- Shift table will summarize baseline CTCAE grade versus the worst post-baseline CTCAE grade.

- Patients who developed toxicities of grade ≥3 will also be listed

Drug Induced Liver Toxicity

Alanine aminotransferase (ALT), aspartate aminotransferase (AST), Alkaline Phosphatase (ALP), and total bilirubin (TBILI) are used to assess possible drug induced liver toxicity. The ratios of test result over the ULN will be calculated and classified for these three parameters during the on-treatment period.

Summary of liver function tests will include the following categories. The number and percentage of patients with each of the following during the on-treatment period will be summarized by treatment arm:

- ALT ≥3×ULN, ALT ≥5xULN, ALT ≥10×ULN, ALT ≥20×ULN;
- AST ≥3×ULN, AST ≥5xULN, AST ≥10×ULN, AST ≥20×ULN;
- (ALT or AST) ≥3×ULN, (ALT or AST) ≥5×ULN, (ALT or AST) ≥10×ULN, (ALT or AST) ≥20×ULN;
- TBILI ≥2×ULN;
- Concurrent ALT ≥3×ULN and TBILI ≥2×ULN;
- Concurrent AST ≥3×ULN and TBILI ≥2×ULN;
- Concurrent (ALT or AST) ≥3×ULN and TBILI ≥2×ULN;
- Concurrent (ALT or AST) ≥3×ULN and TBILI ≥2×ULN and ALP >2×ULN;
- Concurrent (ALT or AST) ≥3×ULN and TBILI ≥2×ULN and ALP ≤2×ULN or missing;
- Concurrent measurements are those occurring on the same date.

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

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

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

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

6.6.4. Vital Signs

Vital signs data includes weight, pulse, systolic blood pressure, and diastolic blood pressure. Measurements were only to be provided once per timepoint. If multiple assessments are provided per timepoint, the maximum value will be used for reporting.

The number and percent of patients in each of the following minimum and maximum blood pressure, body weight, and pulse categories will be presented:

- Increase in Systolic Blood Pressure ≥40 mmHg;
- Decrease in Systolic Blood Pressure ≥40 mmHg;
- Decrease in Systolic Blood Pressure ≥60 mmHg;
- Increase in Diastolic Blood Pressure ≥20 mmHg;
- Decrease in Diastolic Blood Pressure ≥20 mmHg;
- Decrease in Diastolic Blood Pressure ≥40 mmHg;
- Decrease in Body Weight ≥10%;
- Increase in Body Weight ≥10%;
- Increase in Body Weight ≥20%;
- Maximum Pulse Rate >120 bpm;
- Minimum Pulse Rate <50 bpm;
- Maximum increase in pulse rate ≥30 bpm;
- Maximum decrease in pulse rate ≥30 bpm.

All assessments, including unscheduled assessments will be considered. A patient can be included in multiple categories if different criteria are met at different timepoints.

6.6.5. Electrocardiogram

Triplicate ECGs are required, with different frequency between arms, at each assessment, with the exception of pre-dose on Day 1 of Cycle 3 and every other cycle thereafter. A mean score is calculated for any replicate measurements having the same nominal visit. The mean measurement is reported.

QT, RR, HR, PR, QRS, QTcF (Fridericia correction data will be provided by sites on the ECG page of the CRF), and QTcB (Bazett's correction will be programmatically derived using the following formula:

$$QTcB(m\sec) = \frac{QT(m\sec)}{\sqrt{RR}(\sec)},$$

where RR represents the RR interval of the ECG, in seconds

QTcF and QTcB will be summarized by maximum post-baseline values using the following categories

- <450 msec;
- ≥450 msec but <480 msec;
- ≥480 msec but <500 msec;
- ≥500 msec.

Based on these categories shift tables will be provided for baseline vs worst on study QTcF and QTcB.

Unscheduled assessments will be utilized in addition to planned assessments.

And maximum increases from baseline in QTcF and QTcB (including scheduled and unscheduled assessments) will be also summarized based on the following categories:

- Change ≥60 msec;
- Change ≥30 msec but <60 msec;
- Change <30 msec.

In the event that neither Fridericia's nor Bazett's correction adequately adjusts for heart rate, an additional correction such as a population or patient-specific baseline correction could be used and should be fully justified.

For PR and QRS maximum increases from baseline, the following categories will be applied:

- PR change ≥50% if absolute baseline value was <200 msec;
- PR change ≥25% if absolute baseline value was ≥200 msec;
- QRS change ≥50% if absolute baseline value was <100 msec;
- QRS change ≥25% if absolute baseline value was ≥100 msec.

In addition, the number and percentage of patients with maximum postdose PR interval (<160, 160 to <180, 180 to <200, 200 to <220, 220 to <240, 240 to <260, and ≥260 msec) will be provided and based on these categories shift tables for baseline vs worst on study will be generated.

If more than one ECG is measured at a nominal time post-dose (eg, triplicate ECGs within 2-4 minutes), the mean will be used to represent a single observation per patient and time post-dose. If any of the three individual ECGs results in a QTc ≥500 msec and the mean is not ≥500 msec, then that patient's data will be described in the safety section in the study report in order to place the ≥500 msec value in appropriate clinical context. On the other hand, such individual ≥500 msec value within a triplicate will not be included in the categorical analysis unless the average from that triplicate is also ≥500 msec.

6.6.6. Left Ventricular Ejection Fraction (LVEF)

LVEF% will be summarized as frequency (number and percentage) of patients with:

- a shift from baseline normal to at least one result below the institutional lower limit of normal during the on-treatment period;
- >20-point decrease from baseline in LVEF%.

A patient will be included in the categories above if any post-baseline assessment (including unscheduled assessments) meet the criteria.

6.6.7. Mood and Suicidal Ideation and Behavior Analyses

- **The Beck Depression Inventory-II (BDI-II) - Assessment of Mood**

An assessment of mood will be administered to patients via the Beck Depression Inventory-II (BDI-II) scale at the timepoints described in the Schedule of Activities of the Study Protocol. This is a 21 item self-reported scale, with each item rated by patients on a 4 point scale (ranging from 0-3). The scale includes items capturing mood, (loss of pleasure, sadness, and irritability), suicidal ideation, and cognitive signs (punitive thoughts, self- criticism, self-dislike pessimism, poor concentration) as well as somatic signs (appetite, sleep, fatigue, libido).

Scores are obtained by adding up the total points from the series of answers. Higher total scores indicate more severe depressive symptoms. The standardized cutoffs are as follows:

- 0–13: minimal depression;
- 14–19: mild depression;
- 20–28: moderate depression;
- 29–63: severe depression.

The following descriptive statistics will be provided: n, arithmetic mean and SD. Frequencies and percentages will be displayed on items capturing mood, (loss of pleasure, sadness, irritability), suicidal ideation, and cognitive signs (punitive thoughts, self- criticism,

self-dislike, pessimism, poor concentration) as well as somatic signs (appetite, sleep, fatigue, libido) on the BDI-II scale will be summarized.

- **Columbia Suicide Severity Rating Scale (C- SSRS)**

Frequencies and percentages will be displayed for patients with suicidal ideation, suicidal behavior, and self-injurious behavior without suicidal intent based on the C-SSRS during treatment.

7. CONSIDERATION ON COVID-19 IMPACT

Enrolment in this study was concluded almost one year before COVID-19 emergency started, therefore there was no impact on randomization and assessment of baseline characteristics.

Based on data presented at the 5th Safety eDMC meeting that occurred on December 12th, 2019, approximately 50% of patients were on treatment and 80% of patients were on study as of the data cut off date of October 21st, 2019.

As of the date of finalization of the current SAP, no COVID-related deaths have been reported as well as no patient has been reported to have symptoms of COVID-19 and/or tested positively for SARS-CoV-2. Most of the scheduled visits occurred as per plan, with only a few delayed and a couple of visits missed; therefore it is not anticipated any major impact on the analysis due to missing data.

Sensitivity analyses censoring COVID related deaths from PFS analysis and/or additional safety analyses excluding safety data after date of diagnosis for patients with COVID will be however carried out, if the number of patients that will report events related to COVID-19 will be more than 10% of the population enrolled in this study.

8. INTERIM ANALYSES

8.1. Introduction

The goals of the interim analyses are to allow early stopping of the study for efficacy. The interim analysis of PFS and OS will be performed as described in Sections 5.1.1 and 5.1.2 using the methodology described in Section 6.1 for PFS and 6.2.1 for OS.

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

The final PFS analysis will be performed by the Sponsor's clinical team but interim analyses of OS will continue to be performed by an independent statistician. Unblinded results from each of the interim analyses for OS will not be communicated to the Sponsor's clinical team or to any party involved in the study conduct (apart from the independent statistician and E-DMC members) until the E-DMC has determined that either (i) OS analysis has crossed the prespecified boundary for efficacy or (ii) the study needs to be terminated due to any cause.

All patients will continue to be followed for OS until the final OS analysis (or earlier if OS reaches statistical significance at an interim analysis).

8.2. Interim Analyses and Summaries

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

8.2.1. Interim Analysis for PFS

Let $u(t_1)$ and $u(t_F)$ denote the upper critical boundaries based on the test statistics Z_1 and Z_F for efficacy at the interim and the final analysis, respectively.

The critical values $u(t_1)$ for the interim analysis of PFS is determined such that:

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1)$$

where P_0 denotes the probability under the null hypothesis and $\alpha(t_1)$ denotes the α spent, based on the predefined spending functions at information fraction t_1 (t_1 is calculated as the ratio of the number of PFS events observed at the time of the data cutoff for the interim analysis and the total number of PFS events targeted for the final analysis).

The boundary for the final efficacy analysis is derived such that:

$$\alpha(t_1) + P_0(Z_1 < u(t_1), Z_F \geq u_F) = 0.025.$$

As described in Section 5.1.2, if the number of PFS events in the final analysis deviates from the target number of PFS events, the final analysis criteria will be determined as above taking into account the actual alpha spent at the interim analysis and the actual correlation between the two test statistics Z_1 and Z_F, so that the overall one-sided significance level across all analyses and comparisons is preserved at 0.025.

8.2.2. Interim Analysis for OS

Let $u(t_i)$ and $u(t_F)$ denote the upper critical boundaries based on the test statistics Z_i and Z_F for efficacy at the i^{th} interim and the final analysis, respectively, where i=1, 2.

In what follows P_0 denotes the probability under the null hypothesis, and $\alpha(t_i)$ denotes the α spent based on the predefined α-spending function at information fraction t_i (t_i is calculated as the ratio of the number of OS events observed at the time of the data cutoff for the i^{th} interim analysis and the total number of OS events targeted for the final analysis).

The critical value $u(t_1)$ for the interim analysis of OS is determined such as:

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1).$$

The critical value for the additional interim analysis is calculated recursively as follows:

$u(t_2)$ is derived such that $\alpha(t_1) + P_0(Z_1 < u(t_1), Z_2 \geq u(t_2)) = \alpha(t_2)$.

The boundary for the final efficacy analysis is derived such that:

$$\alpha(t_2) + P_0(Z_1 < u(t_1), Z_2 < u(t_2), Z_F \geq u_F) = 0.025.$$

As described in Section 5.1.2, if the number of OS events in the final analysis deviates from the target number of OS events, the final analysis criteria will be determined as above taking into account the actual alpha spent at the interim analyses and the actual correlation between the three test statistics Z_i (i=1, 2) and Z_F, so that the overall one-sided significance level across all analyses and comparisons is preserved at 0.025.

9. REFERENCES

1. Eisenhauer EA, et al. New response evaluation criteria in solid tumors: Revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan;45(2):228-47.

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

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

4. Schoenfeld, D. A. (1982), "Partial Residuals for the Proportional Hazards Regression Model," Biometrika, 69, 239–241.

5. Therneau, T. M. and Grambsch, P. M. (2000). Modeling Survival Data: Extending the Cox Model, Springer-Verlag, New York.

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

7. Amit O, et al. Blinded independent central review of progression in cancer clinical trials: Results from a meta-analysis and recommendation from a PhRMA working group. European Journal of Cancer 47:1772-1778, 2011.

8. Long G et al. Dabrafenib in patients with Val600GLu or Val600Lys BRAF-mutant melanoma metastatic to the brain (BREAK-MB): a multicentre, open-label, Phase 2 trial. Lancet Oncology 2012; 13:1087-1095.

9. Fayers P et al. EORTC QLQ-C30 Scoring Manual. 2001. EORTC, Brussels.

10. The EuroQol Group. EQ-5D-5L: User Guide. V 2.1, April 2015. EuroQol Research Foundation, Rotterdam.

11. Osoba D et al, Interpreting the significance of changes in health-related quality of life scores. 3. Cin Oncology. 1998. 16:139-144.