BeiGene

STATISTICAL ANALYSIS PLAN

Study Protocol Number:	BGB-A317-304
Study Protocol Title:	A Phase 3, Open-Label, Multicenter, Randomized Study to Investigate the Efficacy and Safety of Tislelizumab (BGB-A317) (Anti-PD1 Antibody) Combined With Platinum-Pemetrexed Versus Platinum-Pemetrexed Alone as First-line Treatment for Patients With Stage IIIB or IV Non-Squamous Non-Small Cell Lung Cancer
Date:	Feb 18, 2020
Version:	1.0

SIGNATURE PAGE

Author: [redacted]

[redacted]

Approval

[redacted]

[redacted]

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

Abbreviation	Term
ADA	Anti-drug antibody
AE	adverse event
CI	confidence interval
CR	complete response
CRF	case report form
CT	computerized tomography
CTCAE	Common Terminology Criteria for Adverse Events
C_{trough}	minimum observed plasma concentration
DCR	disease control rate
DOR	duration of response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
IRC	independent review committee
KM	Kaplan-Meier
MedDRA®	Medical Dictionary for Regulatory Activities
NCI-CTCAE	National Cancer Institute Common Terminology Criteria for Adverse Events
ORR	objective response rate
OS	overall survival
PD	progressive disease
PD1	programmed cell death1
PFS	progression-free survival
PK	pharmacokinetics
PR	partial response
PT	preferred term
RECIST	Response Evaluation Criteria in Solid Tumors
SAE	serious adverse events
SAF	Safety Analysis Set
SAP	statistical analysis plan
SOC	system organ class
TC	tumor cell
TEAE	treatment-emergent adverse event
TMB	tumor mutation burden

1 INTRODUCTION

The purpose of this statistical analysis plan (SAP) is to describe the procedures and the statistical methods that will be used to analyze and report results for Protocol BGB-A317-304 "A Phase 3, Open-Label, Multicenter, Randomized Study to Investigate the Efficacy and Safety of Tislelizumab (BGB-A317) (Anti-PD1 Antibody) Combined With Platinum-Pemetrexed Versus Platinum-Pemetrexed Alone as First-line Treatment for Patients With Stage IIIB or IV Non-Squamous Non-Small Cell Lung Cancer". The focus of this SAP is for the planned primary, secondary and exploratory analysis specified in the study protocol.

The analysis details for exploratory biomarker analyses are not described within this SAP. Separate analysis plans will be completed for these specific analysis variables and attached to the clinical study report.

Reference materials for this statistical plan include the protocol amendment BGB-A317-304 (version 2.0, dated 24 January 2019). If the protocol or case report forms are amended or updated, then appropriate adjustments to the SAP may be made if changes are related to the planned analyses.

The SAP described hereafter is a priori plan. This is an open-label study with a planned interim analysis, the SAP will be finalized and approved before interim analysis. Statistical programming may occur as study data accumulate in order to have analysis programs ready at the time the study finishes.

2 STUDY OVERVIEW

Study Design

This is an open-label, randomized, multicenter Phase 3 study designed to compare the efficacy and safety of tislelizumab combined with cisplatin or carboplatin + pemetrexed (Arm A) and cisplatin or carboplatin + pemetrexed alone (Arm B) as first-line treatment in approximately 320 patients who have Stage IIIB or IV non-squamous NSCLC, whereby choice of platinum (cisplatin or carboplatin) will be at the investigator's discretion.

Patients who have not received chemotherapy with histologically confirmed non-squamous, locally advanced, or metastatic NSCLC (Stage IIIB or IV) are eligible. Histology of non-squamous NSCLC will be confirmed at the investigator's site. Patients with known *EGFR* mutation or *ALK* rearrangement are ineligible for the study; for patients without documented tissue-based documentation of EGFR status, fresh or archival tumor tissues are required for *EGFR* mutation assessment. Archival tumor specimens will be prospectively tested for PD-L1 expression by a central laboratory. If archived formalin-fixed paraffin-embedded (FFPE) tissue is not sufficient for PD-L1 analysis, a fresh biopsy sample will be needed. PD-L1 status will be characterized as PD-L1 membrane staining on tumor cells (TC) via the Ventana SP263 assay.

Patients will be stratified by disease stage (IIIB vs. IV), and PD-L1 expression (three levels: < 1% TC vs. 1%–49% TC vs. ≥ 50% TC). Patients whose tissues are unevaluable for PD-L1

expression will be included in the < 1% TC group. All patients will be randomized by a 2:1 ratio to receive one of the following treatment regimens:
Induction phase (4 to 6 cycles, Q3W):
Arm A: tislelizumab + carboplatin or cisplatin + pemetrexed
Arm B: carboplatin or cisplatin + pemetrexed
Maintenance phase (Q3W):
Arm A: tislelizumab + pemetrexed
Arm B: pemetrexed

Arm A patients who experience progressive disease per RECIST v1.1 during chemotherapy combination phase (induction or maintenance phase) or thereafter while receiving tislelizumab monotherapy will be permitted to continue tislelizumab monotherapy provided they meet additional criteria. (please see details in protocol section 3.3).

Arm B patients who develop radiographic disease progression per RECIST v1.1 (to be confirmed by the independent committee) will be given the option to cross over to receive tislelizumab monotherapy provided they meet additional criteria. (please see details in protocol section 3.3).

The study is composed of an initial screening phase (up to 28 days), a treatment phase (4 to 6 cycles of induction treatment followed by maintenance therapy until disease progression, unacceptable toxicity, or death), safety follow-up phase (30 days ± 7 Days after last dose of study drug including chemotherapy-only or before the initiation of a new anticancer treatment, whichever occurs first), and survival follow-up phase.

Patients will undergo tumor assessments at baseline and every 6 weeks (± 7 days) for the first 6 months, every 9 weeks (± 7 days) for the remaining 6 months of Year 1, and after completion of the Week 52 tumor assessment, tumor assessment will continue every 12 weeks (± 7 days) based on RECIST v1.1, regardless of dose delays to manage toxicities. Tumor response will be assessed by independent review committee (IRC) based on the Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1, and by investigators based on RECIST version 1.1. Patients will undergo tumor assessments until radiographic disease progression per RECIST v1.1, loss of clinical benefit (for tislelizumab-only patients who continue treatment after radiographic disease progression according to RECIST v1.1), withdrawal of consent, study completion by sponsor, start of a new anticancer therapy, or death, whichever occurs first.

Patients will be evaluated for adverse events (AEs) (all grades, according to National Cancer Institute Common Terminology Criteria for Adverse Event (NCI-CTCAE) version 5.0). All AEs will be recorded during the study (AEs from the time of the first dose and SAEs from the time of signing of informed consent) and for up to 30 days after the last dose of study drug, including platinum doublet chemotherapy, or until the initiation of another anticancer therapy, whichever occurs first. At the end of treatment, ongoing AEs considered related to study treatment will be followed until the event has resolved to baseline or ≤ Grade 1, the event is assessed by the investigator as stable, the patient is lost to follow-up, the patient withdraws consent, or it has been determined that study treatment or participation is not the cause of the AE.

Immune-related AEs (irAE) should be reported for 90 days after the last dose of tislelizumab, regardless of whether or not the patient starts a new anticancer therapy. The investigator should report any SAEs that are believed to be related to tislelizumab treatment at any time after treatment discontinuation until patient death, withdrawal of consent, or loss to follow up, whichever occurs first.

3 STUDY OBJECTIVES

3.1 PRIMARY OBJECTIVES

- To compare the progression-free survival (PFS) as assessed by the Independent Review Committee (IRC) per RECIST v1.1 in an Intent-To-Treat (ITT) analysis set between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone in chemotherapy-naive patients with Stage IIIB or Stage IV (as classified according to American Joint Committee Cancer 7th Edition of Cancer Staging Manual) non-squamous, non-small cell lung cancer (NSCLC).

3.2 SECONDARY OBJECTIVES

- To compare overall response rate (ORR) as assessed by the IRC and by the investigator per RECIST v1.1 between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone.
- To compare duration of response (DOR) as assessed by the IRC and by the investigator per RECIST v1.1 between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone.
- To compare overall survival (OS) between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone in the ITT analysis set
- To compare PFS as assessed by the investigator per RECIST v1.1 between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone in the ITT analysis set.
- To compare health-related quality of life (HRQoL) between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone.
- To evaluate the safety and tolerability of tislelizumab combined with platinum-pemetrexed compared with platinum-pemetrexed alone.
- To evaluate the correlation between programmed death-ligand1 (PD-L1) expression levels by immunohistochemistry (IHC) and antitumor activity of tislelizumab combined with platinum-pemetrexed.

3.3 EXPLORATORY OBJECTIVES

- To compare tumor assessment outcomes (eg, DCR, TTR) between tislelizumab combined with platinum-pemetrexed and platinum-pemetrexed alone as assessed by the investigator per RECIST v1.1.
- To assess tumor and blood biomarkers of tislelizumab response, resistance, and patient prognosis.
- To characterize the pharmacokinetics of tislelizumab when given in combination with platinum-pemetrexed.

- To assess host immunogenicity to tislelizumab.

4 STUDY ENDPOINTS

4.1 PRIMARY ENDPOINTS

- PFS as assessed by the IRC–the time from randomization to the first objectively documented disease progression, or death from any cause, whichever occurs first, as assessed by the IRC per RECIST v1.1 in the ITT analysis set.

4.2 SECONDARY ENDPOINTS

- ORR as assessed by the IRC–the proportion of patients who had complete response (CR) or partial response (PR) as assessed by the IRC per RECIST v1.1 in the ITT analysis set.
- DOR as assessed by the IRC–the time from the first occurrence of a documented objective response to the time of relapse, or death from any cause, whichever comes first, as assessed by the IRC per RECIST v1.1 in the ITT analysis set with documented objective responses.
- OS – the time from the date of randomization until date of death from any cause in the ITT analysis set.
- PFS as assessed by the investigator–the time from randomization to the first objectively documented disease progression, or death from any cause, whichever occurs first, as determined by the investigator per RECIST v1.1 in the ITT analysis set.
- ORR as assessed by the investigator–the proportion of patients who had CR or PR as determined by the investigator per RECIST v1.1 in the ITT analysis set.
- DOR as assessed by the investigator–the time from the first occurrence of a documented objective response to the time of relapse, or death from any cause, whichever comes first, as determined by the investigator per RECIST v1.1 in the ITT analysis set with documented objective responses.
- HRQoL–measured using the European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Lung Cancer (EORTC QLQ LC13) and Core 30 (EORTC QLQ-C30) as presented in patient-reported outcomes
- Incidence and severity of treatment-emergent AEs (TEAEs) graded according to National Cancer Institute-Common Terminology Criteria for Adverse Events (NCI-CTCAE), v5.0.
- PD-L1 expression by IHC as a predictive biomarker for response.

4.3 EXPLORATORY ENDPOINTS

- DCR–the proportion of patients who had a CR, partial response (PR), Non-CR/Non-PD, or stable disease as assessed by the investigator per RECIST v1.1.
- TTR–the time from randomization to the first occurrence of a documented objective response as assessed by the investigator per RECIST v1.1.
- Status of exploratory biomarkers, including but not limited to: PD-L1, tumor mutation burden (TMB), and immune-related gene expression profiling (GEP) in

archival and/or freshly obtained tumor tissues and blood (or blood derivatives) obtained before, during, or after treatment with tislelizumab or at progression and the association with disease status and/or response to tislelizumab in combination with chemotherapy.

- Summary of serum concentrations of tislelizumab.
- Assessments of immunogenicity of tislelizumab by determining the incidence of antidrug antibodies (ADAs).

5 SAMPLE SIZE CONSIDERATIONS

The sample size calculation was based on the number of events required to demonstrate the PFS superiority of Arm A to Arm B in the ITT analysis set.

The estimates of the number of events required to demonstrate efficacy about PFS in the primary comparisons are based on the following assumptions:

1. Median PFS of 7 months in Arm B with exponential distribution assumption.
2. At a one-sided α of 0.025, 85% power to detect an HR of 0.65, corresponding to an improvement in median PFS from 7 months to 10.8 months, in the ITT analysis set.
3. Randomization ratio of 2:1.
4. One interim analysis of PFS planned in the ITT analysis set when approximately 71% of total PFS events occurred, with Lan-DeMets' approximation to O'Brien-Fleming boundary (O'Brien et al, 1979).

With these assumptions, a total of 215 PFS events is required for the ITT analysis set for the PFS final analysis. Assuming 320 patients are to be enrolled over an 8-month period at a constant enrollment rate, the PFS final analysis will occur approximately 19.2 months after the first patient is randomized.

6 INTERIM ANALYSIS

There will be one interim efficacy analysis of PFS performed in the ITT analysis set. The interim efficacy analysis of PFS will be performed when approximately 153 PFS events (71% of the targeted number of 215 PFS events) are observed in the ITT analysis set. It is estimated that it will take approximately 12.8 months to observe 153 PFS events.

The interim boundary for PFS is based on the Lan-DeMets approximation to O'Brien-Fleming boundary. The interim and final analysis timing and stopping boundaries based on approximately 153 interim PFS events are summarized in Table 1, and the exact time of each analysis will depend on actual number of events occurred. Meanwhile, the boundaries as presented in the Table 1 will be updated based on actual PFS events number at interim and final analyses according to the pre-specified spending function.

Table 1. Analysis Timing and Stopping Boundary for PFS in the ITT Analysis Set (overall one-sided hypothesis testing at α = 0.025)

Type of Analysis	Number of Events	Expected Time (Months)	Testing Boundary	
			P-value Boundary	Approx. HR Threshold
Interim analysis	153	12.8	0.0078	0.660
Final analysis	215	19.2	0.0226	0.748

7 STATISTICAL METHODS

7.1 ANALYSIS SETS

The ITT analysis set includes all randomized patients. Patients will be analyzed according to their randomized treatment arms. This will be the primary analysis set for efficacy analysis.

Per-Protocol analysis set (PP) includes patients in the ITT analysis set who had no important protocol deviations. Important protocol deviations are a subset of major protocol deviations impacting efficacy analysis. Criteria for exclusion from the PP will be determined and documented before the database lock for the primary analysis. This will be the secondary analysis set for efficacy analysis when there are over 10% ITT patients who had important protocol deviations.

The Safety analysis set (SAF) includes all randomized patients who received at least 1 dose of study drug; it will be the analysis set for the safety analyses. For first line treatment, patients who randomized to arm B (carboplatin or cisplatin + pemetrexed) but took any dose of Tislelizumab will be included in arm tislelizumab + carboplatin or cisplatin + pemetrexed in SAF. Patients who randomized to arm A (tislelizumab + carboplatin or cisplatin + pemetrexed) but did not take any dose of Tislelizumab will be included in carboplatin or cisplatin + pemetrexed in SAF.

The HRQoL analysis set includes all randomized patients who received at least 1 dose of study drug and completed at least one HRQoL assessment. This will be the analysis set for HRQoL analysis.

The PK analysis set includes all patients who receive at least 1 dose of tislelizumab per the protocol, for whom any post-baseline PK data are available.

The immunogenicity (ADA) analysis set includes all patients who received at least 1 complete dose of tislelizumab for whom both baseline ADA and at least 1 post-baseline ADA results are available.

7.2 DATA ANALYSIS GENERAL CONSIDERATIONS

Statistical programming and analyses will be performed using SAS® (SAS Institute, Inc., Cary, NC, USA), version 9.3 or higher, and/or other validated statistical software as required.

All descriptive statistics for continuous variables will be reported using mean, standard deviation (SD), median, 25 percentile (Q1), 75 percentile (Q3), minimum (Min), maximum (Max) and n. Categorical variables will be summarized as number (percentage) of patients.

The study Table Listing Graph shells will be provided in a separate document, which will show the content and format of all tables, listings, and graphs in detail.

7.2.1 Definitions and Computations

Baseline: Unless otherwise specified, a baseline value for ITT analysis set is defined as the last non-missing value collected before or at the time of randomization date, if not available, defined as last non-missing value collected before or at the time of first dose date. A baseline value for safety analysis set is defined as the last non-missing value collected before or at the time of first dose date.

Unscheduled Visits: Unscheduled measurements will not be included in by-visit table summaries and graphs, but will contribute to best/ worst case value where required (e.g. shift table). Listings will include scheduled and unscheduled data.

7.2.2 Conventions

Unless otherwise specified, the following conventions will be applied to all analyses:

- 1 year = 365.25 days. Number of years is calculated as (days/365.25) rounded up to 1 significant digit.
- 1 month = 30.4375 days. Number of months is calculated as (days/30.4375) rounded up to 1 significant digit.
- Age will be calculated as the integer part of (date of informed consent – date of birth + 1)/365.25
- P-values will be rounded to 4 decimal places. P-values that round to 0.0000 will be presented as '< 0.0001' and p-values that round to 1.000 will be presented as '> 0.9999'.

7.2.3 Handling of Missing Data

Missing data will not be imputed unless otherwise specified elsewhere in the SAP. Missing dates or partially missing dates will be imputed conservatively for adverse events and prior or concomitant medications/procedures.

By-visit endpoints will be analyzed using observed data, unless otherwise specified. For observed data analyses, missing data will not be imputed and only the observed records will be included.

7.2.4 Adjustment for Covariates

No adjustments for covariates are planned for primary, secondary and exploratory analyses in the study.

7.2.5 Multiplicity Adjustment

Multiplicity adjustment of interim analysis has been described in section 6.

7.3 PATIENT CHARACTERISTICS

7.3.1 Patient Disposition

The number and percentage of patients randomized, treated, permanently discontinued from study treatment, remained on treatment, discontinued from study, and remained on study will be summarized in the ITT analysis set. The primary reasons for study treatment discontinuation and study discontinuation will be summarized according to the categories in the CRF. Study follow-up time and primary reason for screen failure will be summarized.

7.3.2 Protocol Deviations

Important protocol deviations are a subset of major protocol deviations impacting efficacy analysis. Criteria for major protocol derivation and important protocol deviation impacting efficacy will be established and patients with major and important protocol deviations will be identified and documented before the database lock.

Both major protocol derivation and important protocol deviations will be summarized and listed by category in the ITT analysis set.

7.3.3 Demographic and Other Baseline Characteristics

Demographic and other baseline characteristics will be summarized in the ITT analysis set using descriptive statistics.
Continuous demographic and baseline variables include age, BMI (in kg/m^2) and body weight (in kg); categorical variables include sex, age group (<65 years, ≥65 years), ECOG performance status at baseline, smoking status, and stratification factors of disease stage and PD-L1 expression in tumor cell.

7.3.4 Disease History

Lung cancer disease history characteristics will be summarized in the ITT analysis set. Categorical disease characteristics variables include disease stage, histology, and baseline target lesion location. Continuous disease history variables include time from initial diagnosis to study entry and time from advanced/metastatic disease diagnosis to study entry.

Cancer associated symptoms at baseline will also be summarized by SOC, preferred term and CTCAE grade.

7.3.5 Prior Anti-Cancer Drug Therapies, Radiotherapy and Surgeries

The number and percentage of patients with prior anti-cancer drug therapies, time from end of last therapy to study entry, and type of prior anti-cancer drug therapy will be summarized in the ITT analysis set.

The number and percentage of patients with any prior anti-cancer radiotherapy, anatomical site, time from end of last radiotherapy to study entry will be summarized in the ITT analysis set.

The number and percentage of patients with any prior anti-cancer surgeries, curative intent (Yes, No), and time from last surgery to study entry will be summarized in the ITT analysis set.

7.3.6 Prior and Concomitant Medications

Prior and concomitant medications will be coded using the World Health Organization Drug Dictionary (WHO DD) of the version currently in effect at BeiGene at the time of database lock, and will be further classified to the appropriate Anatomical Therapeutic Chemical (ATC) code.

The number and percentage of patients who took prior and concomitant medications will be summarized respectively by ATC medication class and WHO DD preferred term (PT) in the safety analysis set. Prior medications will be defined as medications that received within 30 days before randomization and stopped before the day of first dose of study drug. Concomitant medications will be defined as medications that 1) started before the first dose of study drug and were continuing at the time of the first dose of study drug or 2) started on or after the date of the first dose of study drug up to 30 days after the patient's last dose (as of the Safety Follow-up visit). In addition, telephone contacts with patients should be conducted to assess irAEs and concomitant medications (if appropriate, ie, associated with an irAE or is a new anticancer therapy) at 60 and 90 days (± 14 days) after the last dose of study drugs regardless of whether or not the patient starts a new anticancer therapy.

7.3.7 Medical History

Medical history will be coded using MedDRA version currently in effect at BeiGene at the time of database lock. The number and percentage of patients reporting a history of any medical condition will be summarized by System Organ Class (SOC) and PT in the ITT analysis set.

7.4 EFFICACY ANALYSIS

If not specified otherwise, efficacy analyses described in this section will be based on the ITT analysis set.

7.4.1 Primary Efficacy Endpoints

The primary efficacy endpoint is PFS per the IRC.

PFS per the IRC is defined as the time from randomization to the first documented disease progression as assessed by the IRC with the use of RECIST v1.1, or death from any cause, whichever occurs first. The actual tumor assessment visit date will be used to calculate PFS.

The PFS censoring rules are specified in Appendix 10.1.

PFS per the IRC will be compared between tislelizumab with platinum-pemetrexed (Arm A) and platinum-pemetrexed alone (Arm B), the null and alternative hypotheses are set as follows:

$$H_0: \text{PFS in Arm A} \leq \text{PFS in Arm B}$$

$$H_a: \text{PFS in Arm A} > \text{PFS in Arm B}$$

The p-value will be calculated from a stratified log-rank test at one-sided significance level α=0.025 based on the stratification factors defined in Section 2 using the actual values from EDC (disease stage) and from central lab (PD-L1 expression).

Kaplan-Meier method will be used to estimate median and other quartiles of PFS along with its 95% confidence interval (constructed using Brookmeyer and Crowley method). Kaplan-Meier survival probabilities for each arm will be plotted over time. Event free rate at selected timepoints will be estimated with 95% confidence interval using Greenwood formula. Follow-up time will be estimated by the reverse Kaplan-Meier method.

A stratified Cox proportional hazard model with Efron's method of tie handling will be used to assess the magnitude of the treatment difference (i.e., hazard ratio) between the treatment arms, with the same stratification variables used for the stratified log-rank test. The hazard ratio and its 95% confidence interval from the stratified Cox model with Efron's method of tie handling and with a single treatment covariate will be reported.

In order to evaluate the robustness of the PFS per IRC, we will perform possible sensitivity analyses with different censoring rules if the number of patients with censor reason (i.e. the specific situation in Appendix 10.1) is greater than or equal to 5% of patients in the ITT analysis set. If the patient met multiple situations in Appendix 10.1, the censor reason will be the earliest situation.

The sensitivity analysis 1 is the same as the primary analysis except that it progressed at date of documented progression with protocol specified continued follow-up in all treatment arms or died at date of death whichever is earlier when new anticancer therapy was started.

The sensitivity analysis 2 is the same as the primary analysis except that it progressed at date of documented progression or died at date of death whichever is earlier after ≥2 missed tumor assessment.

When there are over 10% ITT patients who had important protocol deviations, sensitivity analysis 3 in PP analysis set will be implemented using primary PFS censoring rule.

7.4.2 Secondary Efficacy Endpoints

Overall Survival

OS is defined as the time from randomization to death from any cause. Data for patients who are not reported as having died at the time of analysis will be censored at the date last known to be

alive. Data for patients who do not have post-baseline information will be censored at the date of randomization.

Similar methodologies used to evaluate PFS per the IRC will be applied to OS analysis.

Progression-free survival per the investigator
PFS per the investigator is defined as the time from randomization to the first objectively documented disease progression, or death from any cause, whichever occurs first, as determined by the investigator per RECIST v1.1 in the ITT analysis set. Similar methodology used to evaluate PFS per the IRC will be applied to analysis of PFS per the investigator. Concordance of PFS per the investigator and PFS per IRC in the ITT analysis set will also be assessed.

Overall response rate per the IRC
ORR (confirmation not required according to RECIST v1.1) is the proportion of patients who had a CR or PR as assessed by the IRC per RECIST v1.1 in ITT analysis set. Patients without any post-baseline assessment will be considered non-responders. Patients without measurable disease at baseline will also be considered as non-responders. The difference in ORR between arms will be evaluated using the Cochran-Mantel-Haenszel (CMH) chi-square test with the actual stratification factors as strata. The two-sided 95% CIs for the odds ratio and the difference in ORR will be calculated, as well as Clopper-Pearson 95% CIs for the ORR within each arm.

A waterfall plot of best percent change in sum of target lesion diameters from baseline per IRC will be provided by treatment arm. The patients in each arm will be ordered by the percentage, patients with the largest percentage will be presented on the right.

Overall response rate per the investigator
ORR (confirmation not required according to RECIST v1.1) is the proportion of patients who had a CR or PR as determined by the investigator per RECIST v1.1 in ITT analysis set. Patients without any post-baseline assessment will be considered non-responders. Similar methodology used to evaluate ORR per the IRC will be applied to analysis of ORR per the investigator.

Duration of response per the IRC
DOR per the IRC is defined for patients with an objective response as the time from the first documented objective response to documented disease progression as assessed by the IRC using the RECIST v1.1, or death from any cause, whichever occurs first. Only the subset of patients who show a complete response or partial response will be included in the DOR analysis. Data for patients who are alive and who have not experienced disease progression at the time of analysis will be censored at the date of the last tumor assessment. If no tumor assessments were performed after the date of the first occurrence of the objective response (CR or PR), DOR will be censored at the date of the first occurrence of the objective response. Median DOR and corresponding 95% CIs will be estimated using Kaplan-Meier methodology for each treatment arm. Comparisons of DOR per the IRC between treatment arms will be made using the log-rank test for descriptive purposes only.

Duration of response per the investigator
DOR per the investigator is defined for patients with an objective response as the time from the first documented objective response to documented disease progression as determined by the investigator using the RECIST v1.1, or death from any cause, whichever occurs first. Data for patients who are alive and who have not experienced disease progression at the time of analysis will be censored at the date of the last tumor assessment. If no tumor assessments were performed after the date of the first occurrence of the objective response (CR or PR), DOR will be censored at the date of the first occurrence of the objective response. Similar methodology used to evaluate DOR per the IRC will be applied to analysis of DOR per the investigator.

Health-Related Quality of Life

Analysis Method

There is no multiplicity adjustment for the HRQoL analysis. Descriptive statistics will be used for all HRQoL analyses. All HRQoL analysis will be in HRQoL analysis set. Both EORTC QLQ-C30 and EORTC QLQ-LC13 instruments have been extensively tested for reliability and validity (Bergman et al, 1994; Osoba et al, 1994; Groenvold et al, 1997). Two items measuring overall health status and quality of life are graded on a 7-point Likert scale, while all remaining items are graded on a 4-point scale ranging from 1 (Not at all) to 4 (Very much).

Compliance

Compliance for the EORTC QLQ-C30 and EORTC QLQ-LC13 modules, defined as the proportion of questionnaires actual received out of the expected number (i.e, number of patients on treatment), in the HRQoL analysis set will be summarized for each assessment time point and treatment arm.

Change from Baseline by Visit

For each scale or item of EORTC QLQ-C30 and EORTC QLQ-LC13, summary statistics at each assessment time point and change from baseline will be presented by treatment arm in tables. Boxplot depicting the mean scores over time of global health status/quality of life will be provided for each treatment arm.

Details of HRQoL scoring are specified in Appendix 11.2 according to the algorithm described in the EORTC QLQ-C30 and EORTC QLQ-LC13 scoring manual (Fayers 2001).

Time to Deterioration (TTD) of HRQoL

Time to deterioration (TTD is defined as the time from randomization to first onset time at which deterioration was clinically meaningful that was confirmed at a subsequent clinically meaningful deterioration. The minimum clinically meaningful important difference on the QLQ-C30 and LC13 is at least 10 points (Osoba et al, 1998; King, 1996; Maringwa et al 2011). The clinically

meaningful deterioration in symptoms of QLQ-C30 and QLQ-LC13 is defined as ≥10 points increase from baseline. And clinically meaningful deterioration in function and global health status/quality of life is defined as ≥10 points decrease from baseline. The median TTD of QLQ-C30 global health status/quality of life and composite of cough, chest pain, and dyspnea in the QLQ-LC13 will be calculated using Kaplan-Meier estimates, and presented with 2-sided 95% CIs.

PD-L1 expression as a predictive biomarker for response

Exploratory biomarker analyses will be performed to examine the relationship between tumor tissue PD-L1 expression and measures of efficacy. Additionally, predictive and prognostic exploratory biomarkers in tumor tissue and/or blood will be examined for their association with disease status and clinical outcomes. These exploratory analyses will not be included in the CSR for this study.

7.4.3 Subgroup Analyses

Subgroup analysis of primary endpoint of PFS per the IRC will be conducted to determine if the treatment effect is consistent across various subgroups, the HR estimates of PFS from an unstratified Cox model and its 95% CI will be estimated and plotted within each category of the following classification variables (a subgroup will not be analyzed if it includes <10% of the ITT analysis set):

- Age (< 65 year versus ≥ 65years)
- Sex (female versus male)
- ECOG Performance Status (0 versus 1)
- Smoking status (Former or Current versus Never)
- Disease stage (IIIB versus IV)
- Brain metastases at baseline (Yes versus no)
- Liver metastases at baseline (Yes versus no)
- PD-L1 expression in TC (≥ 50% TC versus 1%-49% TC versus < 1% TC)
- ALK rearrangement (Negative versus unknown)

Forest plot of subgroup analysis in PFS per the IRC will be provided. Additional subgroup analysis may also be conducted per additional prognosis factors as suggested.
Subgroup analysis of secondary endpoints may also be conducted.

7.4.4 Exploratory Efficacy Endpoints

Disease control rate per the IRC and per the investigator
DCR is defined as the proportion of patients with objective response (CR or PR), Non-CR/Non-PD, or stable disease maintained for ≥ 6 weeks using the RECIST v1.1. Both DCR per the IRC and per the investigator will be analyzed. Similar methodologies for analysis of ORR will be applied.

Time to response per the IRC and per the investigator
TTR per the investigator is defined for patients with an objective response as determined by the investigator as the time from randomization to the first occurrence of a CR or PR as determined

by the investigator using the RECIST v1.1. TTR per the IRC is defined for patients with an objective response as determined by the IRC as the time from randomization to the first occurrence of a CR or PR as determined by the IRC using the RECIST v1.1. Only the subset of patients who show a complete response or partial response will be included in the TTR analysis. TTR will be summarized for descriptive purposes.
The mean, SD, median, and range of TTR will be provided.

Time to first subsequent anti-cancer systemic therapy (TFST)

TFST is defined for patients with the use of subsequent anti-cancer systemic therapy as the time from end of study treatment to first dose of subsequent anti-cancer systemic therapy. The mean, SD, median, and range of TFST will be provided.

Second Progression-free survival per the investigator
Analysis of progression-free survival after next line of treatment (PFS2) is defined as the time from randomization to second/subsequent disease progression, or death from any cause, whichever occurs first. Patients alive and for whom a second objective disease progression has not been observed will be censored at the last time known to be alive and without second objective disease progression. PFS2 analysis will be provided when data is mature. Similar methodology used to evaluate PFS per the IRC will be applied to analyze PFS2 per the investigator.

Subsequent anti-cancer systemic therapy
Subsequent anti-cancer systemic therapy will be summarized by category and PT in the ITT analysis set.

7.5 SAFETY ANALYSES

All safety analyses will be based on safety analysis set. The incidence of treatment-emergent adverse events (TEAEs, Section 7.5.2) and SAEs will be summarized. Laboratory test results, vital signs, ECG, ECOG and their changes from baseline will be summarized using descriptive statistics. Abnormal values will be flagged.

Safety analyses of patients who crossover to receive tislelizumab monotherapy may be separately summarized if data allows. When crossover rate (the rate among the patients who crossed over to receive at least one dose of tislelizumab monotherapy) is equal to or larger than 10% among the patients who actually randomized into Arm B, additional analysis for patients who crossed over will be provided, including but not limited to tislelizumab extent of exposure, overall summary of TEAE and concomitant medication.

7.5.1 Extent of Exposure

Extent of exposure to each study drug will be summarized descriptively by the following variables:

- Number of treatment cycles

- Duration of exposure (weeks) is defined as:

  (date of last dose of study drug + 21 days – date of first dose of study drug)/7, with censored by death date and cutoff date, without censoring when calculating actual dose intensity.

- Cumulative dose (mg): the sum of all actual dose of study drug, given from first to last administration

- Actual dose intensity (ADI) in mg/cycle is defined as

  Cumulative dose (mg) received by a patient / Duration of exposure (cycles)

- Relative dose intensity (RDI) in % is defined as:

$$100 \times \frac{\text{ADI (mg/cycle)}}{\text{Planned Dose Intensity (mg/cycle)}}$$

  Where planned dose intensity (PDI) of tislelizumab equals to 200mg/cycle and for other chemotherapy:

  PDI = $\frac{\sum_1^K Planned\ Dose\ in\ Cycle\ K}{K}$, where K = total number of cycles patient has received

7.5.2 Adverse Events

AEs will be graded by the investigators using CTCAE version 5.0 and coded using the Medical Dictionary for Regulatory Activities (MedDRA) currently in effect at BeiGene at the time of database lock.

A treatment-emergent AE (TEAE) is defined as an AE that had an onset date or a worsening in severity from baseline (pre-treatment) on or after the first dose of study drug up to 30 days after the last dose of study drug or initiation of new anticancer therapy, whichever occurs first. For tislelizumab arm, the TEAE classification also applies to irAEs that are recorded up to 90 days after the last dose of tislelizumab, regardless of whether the patient starts a new anticancer therapy. Only those AEs that were treatment emergent will be included in summary tables. All AEs, treatment emergent or otherwise, will be presented in patient data listings. AEs that were treatment emergent for patients who crossover to receive tislelizumab monotherapy may be separately summarized if data allows.

An overview table of patients with at least one TEAE will be presented with the incidence of:

- patients with any TEAE
- patients with any ≥ grade 3 TEAE
- patients with any serious TEAEs

- patients with any TEAE leading to death
- patient with any TEAE leading to permanent discontinuation of any component of study treatment
- patients with any TEAE related to any component of study treatment
- patients with any TEAE related to tislelizumab
- patients with any TEAE related to tislelizumab and ≥ grade 3
- patients with any TEAE ≥ grade 3 and related to any component of study treatment
- patients with any serious TEAEs and related to any component of study treatment
- patients with any potential irTEAE

Treatment-related TEAEs include those events considered by the investigator to be related or with missing assessment of the causal relationship. For patients with multiple occurrences of the same event will be counted only once, and the maximum grade per CTCAE v5.0 will be used.

If the grade is missing for one of the treatment-emergent occurrences of an adverse event, the maximal grade on the remaining occurrences with the same preferred term of the same patient will be used. If the patient has no other TEAE with the same preferred term, then impute as the maximum grade on all TEAE with the same preferred term; If the grade is missing for all the occurrences, do not impute, a "missing" category will be added in the summary table.

The incidence of following TEAEs will be reported by SOC and PT, sorted by decreasing frequency of the SOC and PT:

- TEAE by maximum grade
- TEAE leading to permanent discontinuation of any component of study treatment
- TEAE leading to death
- ≥ grade 3 TEAE
- TEAE related to any component of study treatment
- Serious TEAE

All deaths and cause of death will be summarized, including those occurred during the study treatment period and those reported during the survival follow-up period after treatment discontinuation.

7.5.3 Laboratory Values

Descriptive summary statistics (eg, n, mean, standard deviation, median, minimum, maximum for continuous variables; n [%] for categorical variables) for selected laboratory parameters described in Table 2 and their changes from baseline will be summarized by visit.

Selected laboratory parameters that are graded in NCI-CTCAE v5.0 will be summarized by shift from baseline CTCAE grades to maximum post-baseline grades, parameters with CTCAE grading in both high and low directions will be summarized separately.

Patient data listings of selected laboratory parameters will be provided.

Table 2 Clinical Laboratory Tests

Serum Chemistry	Hematology	Thyroid Function
Alanine aminotransferase (ALT)	Hemoglobin	Free Triiodothyronine (FT3)
Aspartate aminotransferase (AST)	White blood cell (WBC) count	Free Thyroxine (FT4)
Creatinine	Neutrophil (Absolute)	Thyroid Stimulating
Potassium	Platelet count	Hormone (TSH)
Sodium		
Creatine kinase (CK)		
Creatine kinase-cardiac muscle isoenzyme (CK-MB)		

7.5.4 Vital Signs

Descriptive statistics for vital sign parameters (systolic and diastolic blood pressure, pulse rate, body temperature) and weight and changes from baseline will be summarized by visit.

7.5.5 Electrocardiograms (ECG)

ECG will be performed during baseline and multiple time post-baseline points (refer the time points to the protocol study assessments and procedures schedule). Postbaseline abnormal QTc observations will be summarized.

7.5.6 ECOG

A shift table from baseline to worst post-baseline in ECOG performance score will be summarized.

7.6 PHARMACOKINETIC ANALYSES

PK samples will be collected only in patients randomized to receive tislelizumab. Tislelizumab post-dose and C_{trough} (pre-dose) will be tabulated and summarized for each cycle at which these concentrations are collected. Descriptive statistics will include means, medians, ranges, standard deviations, coefficient of variation (CV%), geometric mean and geometric CV%, as appropriate.

Additional PK analyses, including population PK analyses and exposure-response (efficacy, safety endpoints) analyses may be conducted as appropriate and the results from these analyses will be reported separately from the CSR.

7.7 IMMUNOGENICITY ANALYSIS

Human anti-drug antibodies (ADA) to tislelizumab will be assessed during the study as defined in the protocol.

ADA attributes:

- Treatment boosted ADA is defined as ADA positive at baseline that was boosted to a 4-fold or higher level following drug administration.
- Treatment-induced ADA is defined as ADA negative at baseline and ADA positive post-baseline.
- Persistent ADA response is defined as Treatment-induced ADA detected at 2 or more time points during treatment or follow-up, where the first and last ADA positive samples are separated by 16 weeks or longer; or detected in the last time point.
- Transient ADA response is defined as Treatment-induced ADA detected only at 1 time point during treatment or follow-up, excluding last time point; or detected at 2 or more timepoints during treatment or follow-up, where the first and last positive samples (irrespective of any negative samples in between) are separated by less than 16 weeks and the last time point is negative. Transient ADA is a treatment-induced response that is not considered persistent.
- Neutralizing ADA is defined as ADA that inhibits or reduces the pharmacological activity.

ADA response endpoints:

- ADA incidence is defined as sum of treatment-emergent ADA, which include both treatment-induced and treatment-boosted ADA-positive patients, as a proportion of the ADA evaluable population.
- ADA prevalence is defined as proportion of all patients that are ADA positive, including pre-existing ADA, at any time point.

The immunogenicity results will be summarized using descriptive statistics by the number and percentage of patients who develop detectable ADA. The incidence of positive ADA and neutralizing ADA will be reported for evaluable patients. The effect of immunogenicity on PK, efficacy, and safety may be evaluated if data allow.

8 CHANGES IN THE PLANNED ANALYSIS

Not applicable.

9 REFERENCES

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

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

Collett D. Modelling survival data in medical research. New York: Chapman & Hall; 1994.

Common Terminology Criteria for Adverse Events (CTCAE). Version 4.0. United States Department of Health and Human Services, National Institutes of Health, National Cancer Institute, Washington, DC, USA, May 28, 2009.

Eisenhauer EA, Therasse P, Bogaerts J Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). Eur J Cancer. 2009; 45:228-47.

Food and Drug Administration Center for Drug Evaluation Research CDER and Center for Biologics Evaluation and Research. FDA Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics, 2007.

Greenwood M. "The natural duration of cancer". Reports on Public Health and Medical Patients (London: Her Majesty's Stationery Office). 1926; 33:1-26.

Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. Washington, DC: United States Food and Drug Administration; 2007.

Kaplan E, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958; 53:457-81.

Fayers PM, Aaronson NK, Bjordal K,Groenvold M, Curran D, Bottomley A, on behalf of the EORTC Quality of Life Group. Brussels: EORTC, EORTC QLQ-C30 Scoring Manual (3rd edition). 2001. ISBN: 2-9300 64-22-6.

Bergman B, Aaronson NK, Ahmedzai S, Kaasa S, Sullivan M (1994) The EORTC QLQ-LC13: a modular supplement to the EORTC Core Quality of Life Questionnaire (QLQ-C30) for use in lung cancer clinical trials. EORTC Study Group on Quality of Life. Eur J Cancer 30A: 635–642.

Osoba D, Zee B, Pater J, Warr D, Kaizer L, Latreille J (1994) Psychometric properties and responsiveness of the EORTC quality of Life Questionnaire (QLQ-C30) in patients with breast, ovarian and lung cancer. Qual Life Res 3: 353–364.

Groenvold M, Klee MC, Sprangers MA, Aaronson NK (1997) Validation of the EORTC QLQ-C30 quality of life questionnaire through combined qualitative and quantitative assessment of patient-observer agreement. J Clin Epidemiol 50: 441–450.

Osoba D, Rodrigues G, Myles J, Zee B, Pater J (1998) Interpreting the significance of changes in health-related quality-of-life scores. J Clin Oncol 16: 139–144.

King MT. The interpretation of scores from the EORTC quality of life questionnaire QLQ-C30. Qual Life Res 1996; 5: 555–567.

Maringwa J, Quinten C, King M, et al. Minimal clinically meaningful differences for the EORTC QLQ-C30 and EORTC QLQ-BN20 scales in brain cancer patients. Annals of Oncology 2011; 22: 2107–2112.

10 APPENDIX

10.1 PFS CENSORING RULES

Definition of Progression Date: Progression date is assigned to the first time when tumor progression was documented.

The PFS derivation rules in this SAP follow the Table C1 and C2 described in Appendix C of Food and Drug Administration (FDA) "Guidance for Industry Clinical Trial Endpoints for the Approval of Non-Small Cell Lung Cancer Drugs and Biologics (2015)", which includes documented progression only.

Censoring rules for primary and sensitivity analyses are summarized in Table 3.

Table 3 Censoring Rules for Primary and Sensitivity Analysis of PFS Per RECIST version 1.1

No.	Situation	Primary Analysis	Sensitivity Analysis
1	Incomplete or no baseline tumor assessments	Censored at randomization date	
2	No postbaseline tumor assessment and no death	Censored at randomization date	
3	No postbaseline tumor assessment and death	Died at date of death	
4	Progression documented between scheduled visits	Progressed at date of documented progression	
5	No progression	Censored at date of last adequate tumor assessment with no documented progression	
6	New anticancer treatment started	Censored at date of last adequate tumor assessment before date of new anticancer treatment	Progressed at Date of documented progression with protocol specified continued follow-up in all treatment arms or died at date of death whichever is earlier
7	Death between adequate assessment visits	Died at date of death	
8	Death or progression after ≥2 missed tumor assessment visit	Censored at date of last adequate tumor assessment prior to the ≥2 missed tumor assessments	progressed at date of documented progression or Died at date of death whichever is earlier

10.2 HEALTH RELATED QUALITY OF LIFE

The QLQ-C30 consists of thirty questions that are associated with five functional scales (physical, role, cognitive, emotional, and social), three symptom scales (fatigue, pain, and nausea and vomiting), a global health status and six single items (dyspnea, insomnia, appetite loss, constipation, diarrhea, and financial difficulties). QLQ-C30 and QLQ-LC13 scale scores will be calculated as described below.

Scoring Process
The principle for scoring applies to all scales/scores: Raw scores are calculated as the average of the items that contribute to the scale.

A linear transformation to standardize the raw scores is utilized, so that the scores are ranged from 0 to 100. Increases in scores for functional domains (e.g., physical, role, social, emotional, etc.) are improvements while increases in scores for symptoms (e.g., fatigue, vomiting and nausea, diarrhea, pain, etc.) are deteriorations.

Missing Items
If at least half of the items for a scale are answered, then all the completed items are used to calculate the score. Otherwise, the scale score is set to missing.

In practical terms, if items I_1, I_2, ... I_n are included in a scale, the procedure is as follows:

Raw Score
For all scores, the raw score (RS), is the mean of the component items
RS = $(I_1+I_2+\ldots+I_n)/n$

Derived Scale
The derived scales are obtained from the raw scores as defined in the EORTC manual. The derived scales have a more intuitive interpretation: larger function scale or global health status / QoL are improvements while larger symptom scales (e.g., pain, nausea, etc.) are deteriorations.

The derivation formulas are as follows.

Linear transformation
Apply the linear transformation to 0-100 to obtain the score S,

Functional scales: $S = \left\{1 - \frac{(RS-1)}{range}\right\} \times 100$

Symptom scales / items: $S = \{(RS-1)/range\} \times 100$

Global health status / QoL: $S = \{(RS-1)/range\} \times 100$

Table 4 Scoring of QLQ-C30

	Scale	Number of items	Item range	Item Numbers
Global health status/ QoL Global health status/QOL	QL2	2	6	29,30
Functional Scales				
Physical functioning	PF2	5	3	1, 2, 3, 4, 5
Role functioning	RF2	2	3	6, 7
Emotional functioning	EF	4	3	21, 22, 23, 24
Cognitive functioning	CF	2	3	20, 25
Social functioning	SF	2	3	26, 27
Symptom Scales/ items				
Fatigue	FA	3	3	10, 12, 18
Nausea and vomiting	NV	2	3	14, 15
Pain	PA	2	3	9, 19
Dyspnoea	DY	1	3	8
Insomnia	SL	1	3	11
Appetite loss	AP	1	3	13
Constipation	CO	1	3	16
Diarrhoea	DI	1	3	17
Financial Difficulties	FI	1	3	28

Table 5 Scoring of QLQ-LC13

	Scale	Number of items	Item range	Item Numbers
Symptom scales/items				
Dyspnoea	LCDY	3	3	3,4,5
Coughing	LCCO	1	3	1
Haemoptysis	LCHA	1	3	2
Sore mouth	LCSM	1	3	6
Dysphagia	LCDS	1	3	7
Peripheral neuropathy	LCPN	1	3	8
Alopecia	LCHR	1	3	9
Pain in chest	LCPC	1	3	10
Pain in arm or shoulder	LCPA	1	3	11
Pain in other parts	LCPO	1	3	12