STATISTICAL ANALYSIS PLAN

A Phase 3, Multicenter, Randomized, Double-blind, Placebo-Controlled Study of AG-881 in Subjects With Residual or Recurrent Grade 2 Glioma With an IDH1 or IDH2 Mutation

AG881-C-004

Version: 2.0

Date: 04-Feb-2021

Prepared by:

[redacted]

Name and Title (Printed)		Date (DD MMM YYYY)
[redacted]	[redacted]	[redacted]

Approved by:

Name and Title (Printed)		Date (DD MMM YYYY)
[redacted]	[redacted]	[redacted]
[redacted]	[redacted]	[redacted]
Name and Title (Printed)	Signature	Date (DD MMM YYYY)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

Abbreviation	Definition
AE	Adverse event
AESI	Adverse event of special interest
ALP	Alkaline phosphatase
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ATC	Anatomical Therapeutic Chemical
BIRC	Blinded Independent Review Committee
BMI	Body mass index
BrCS	Brain Cancer Subscale
CI	Confidence interval
CRF	Case report form
CRO	Clinical Research Organization
CS	Compound Symmetry
CTCAE	Common Terminology Criteria for Adverse Events
DI	Dose intensity
ECG	Electrocardiogram
eCRF	Electronic case report form
EOS	End of study
EOT	End of treatment
EQ-5D-5L	EuroQol 5 Dimensions 5 Levels
EQ-VAS	EQ-visual analog scale
EWB	Emotional Well-Being
FACT-Br	Functional Assessment of Cancer Therapy – Brain Tumor
FACT-G	Functional Assessment of Cancer Therapy – General
FWB	Functional Well-Being
FAS	Full Analysis Set
HRQoL	Health Related Quality of Life
IA	Interim analysis
IDMC	Independent Data Monitoring Committee
ITT	Intention-to-treat

IWRS	Interactive response system
LVEF	Left ventricular ejection fraction
LLN	Lower limit of normal
MedDRA	Medical Dictionary for Regulatory Activities
NE	Not evaluable
OS	Overall survival
PD	Pharmacodynamic
PGI-C	Patient Global Impression of Change
PGI-F	Patient Global Impression of Frequency
PGI-S	Patient Global Impression of Severity
PK	Pharmacokinetic
PPS	Per-Protocol Set
PT	Preferred Term
PWB	Physical Well-Being
QD	Once daily
QTc	Heart rate-corrected QT interval
QTcB	Heart rate-corrected QT interval using the Bazett's formula
QTcF	Heart rate-corrected QT interval using the Fridericia's formula
RCI	Repeated confidence interval
SAE	Serious adverse event
SAP	Statistical analysis plan
SD	Standard deviation
SOC	System Organ Class
SWB	Social/Family Well-Being
TEAE	Treatment-emergent adverse event
TTNI	Time to Next Intervention
ULN	Upper limit of normal
UN	Unstructured
VC	Variance Component
WBC	White blood cell
WHO	World Health Organization

1. VERSION HISTORY

This statistical analysis plan (SAP) describes the analysis associated with protocol AG881-C-004, Version 3.0 (dated 17-Dec-2020).

Table 1: Summary of Major Changes in Statistical Analysis Plan Amendments

Version	Version Date	Rationale and Summary of Changes
1.0	09-Sep-2020	Original version.
2.0	04-Feb-2021	This SAP was amended to reflect changes made from Version 2.0 to Version 3.0 of the protocol, as follows: • Section 2 "Introduction": updated the definition of end of study. • Section 3 "Trial Objectives and Endpoints": moved TTNI from secondary to key secondary and modified the definition of TTNI to include death as an event; added evaluation of CR+PR, time to CR+PR, duration of CR+PR, and vorasidenib's circulating metabolite AGI-69460 in plasma to secondary objectives and endpoints; broadened the exploratory HRQoL objectives and endpoints associated with PGI; added evaluation of TGR before and after treatment with vorasidenib and placebo to the exploratory objectives and endpoints. • Section 4 "Study Design": updated schedule of assessments for PFS. • Section 6.1 "Randomization, Blinding, Unblinding, and Crossover": updated the considerations associated with blinding and crossover. • Section 6.2 "Sample Size Determination and Decision Rules": updated the sample size calculations and operating characteristics to take into consideration a smaller hazard ratio for the primary endpoint, the change in key secondary endpoint, and the timing and type of interim analyses for each endpoint. • Section 6.3.4 "Baseline": expanded considerations to include additional HRQoL exploratory endpoints. • Section 7.3.1 "Demographics and Physical Measurements": added Tanner scale of sexual maturity. • Section 7.7.1 "Primary Endpoint": updated schedule of tumor assessments. • Section 7.7.2 "Key Secondary Endpoint": updated the definition of TTNI to include death as an event and added details regarding censoring. • Section 7.7.3 "Other Secondary Endpoints": updated considerations for TGR as removed from the testing strategy, added considerations for the analyses of the new endpoints (CR+PR, time to CR+PR, duration of CR+PR), provided more details for the analyses of HRQoL. • Section 7.7.4 "Exploratory Endpoints": added analyses for the new endpoint of pre- and post-treatment TGR, added analyses for the expanded endpoints for HRQoL as assessed by PGI. • Section 7.7.5 "Subgroup Analyses": removed TGR by BIRC as this is no longer a key secondary endpoint.

		• Section 7.10.2 "Interim Analyses and Summaries": was updated to add a subsection for the interim analysis for TTNI. In addition, • Section 7.7.4.2 "Pre- and Post-crossover TGR": updated details associated with statistical methodology. • Section 7.7.4.7 "Seizure Activity": added seizure AEs. • Section 7.9 "Biomarker Analyses": updated the details associated with the listings that will be provided. Minor edits and consistency corrections were implemented.

2. INTRODUCTION

This SAP provides the detailed methodology for summary and statistical analyses of the data collected in Study AG881-C-004, except for pharmacokinetic (PK) endpoints, which will be described in a separate SAP. This document may modify the plans outlined in the protocol; however, any major modifications of the primary endpoint definition or its analysis will also be reflected in a protocol amendment.

The clinical study report (CSR) will include all data up to a data cutoff date that is determined based on the number of events required for the final analysis of the primary endpoint [progression-free survival (PFS)] per blinded independent review committee (BIRC).

The data cutoff date will be defined prospectively once a data extract (before database lock) is available that indicates that all subjects have been randomized and that the required number of events for PFS is expected to occur by the cutoff date. The final number of events might deviate from the planned number; the data cutoff date will not be adjusted retrospectively in this case.

End of Study (EOS) is defined as the time at which all subjects have discontinued study treatment and completed the OS Follow-up period, died, withdrawn consent from overall study participation, are lost to follow-up, or the Sponsor ends the study, whichever occurs first.

3. TRIAL OBJECTIVES AND ENDPOINTS

3.1. Objectives

3.1.1. Primary Objective

The primary objective of the study is to demonstrate the efficacy of vorasidenib based on radiographic PFS per BIRC compared with placebo in subjects with residual or recurrent Grade 2 oligodendroglioma and astrocytoma with an IDH1 or IDH2 mutation who have undergone surgery as their only treatment.

3.1.2. Key Secondary Objective

The key secondary objective of the study is to demonstrate the efficacy of vorasidenib based on TTNI compared with placebo.

3.1.3. Other Secondary Objectives

The secondary objectives of the study are:

- To evaluate safety and tolerability of vorasidenib.
- To evaluate vorasidenib and placebo with respect to tumor growth rate (TGR) as assessed by volume per the BIRC.
- To evaluate the efficacy of vorasidenib and placebo based on objective response, CR+PR, time to response (TTR), time to CR+PR, duration of response (DoR), and duration of CR+PR, with response assessed per the BIRC and the Investigator.
- To evaluate vorasidenib and placebo with respect to OS.
- To evaluate vorasidenib and placebo with respect to health-related quality of life (HRQoL) as assessed by the Functional Assessment of Cancer Therapy – Brain (FACT-Br) questionnaire.
- To evaluate vorasidenib and placebo with respect to PFS per the Investigator assessment.
- To evaluate the PK of vorasidenib and its circulating metabolite AGI-69460 in plasma.

3.1.4. Exploratory Objectives

The following objectives are also to be explored:

- To evaluate, for subjects who cross over from placebo to the vorasidenib, the time from first dose of vorasidenib to documented progression on vorasidenib, as assessed by the Investigator, or death due to any cause, whichever occurs first.
- To evaluate TGR before and after treatment with vorasidenib among subjects who cross over from placebo to vorasidenib.
- To evaluate HRQoL with vorasidenib and placebo as assessed by the EuroQol 5 Dimensions, 5-level (EQ-5D-5L) questionnaire and the Patient Global Impression (PGI) questions.
- To evaluate neurocognitive function in subjects receiving vorasidenib and placebo as assessed by a validated battery of cognitive performance instruments.
- To evaluate seizure activity in subjects receiving vorasidenib and placebo.
- To evaluate the molecular and cellular markers that may be predictive of response and/or resistance, where feasible, in blood and archival tumor tissue.
- To evaluate TGR before and after treatment with vorasidenib and placebo.

- To evaluate time to malignant transformation and radiographic changes associated with histopathology-proven malignant transformation in subjects who have surgery or biopsy as an intervention.

3.2. Endpoints

3.2.1. Primary Endpoint

The primary endpoint is PFS, defined as the time from date of randomization to date of first documented radiographic PD (as assessed by the BIRC per modified Response Assessment for Neuro-oncology for Low-Grade Gliomas [RANO-LGG]), or date of death due to any cause, whichever occurs earlier.

3.2.2. Key Secondary Endpoint

The key secondary endpoint is TTNI, defined as the time from randomization to the initiation of the first subsequent anticancer therapy (including vorasidenib, for subjects randomized to placebo who subsequently cross over) or death due to any cause.

3.2.3. Other Secondary Endpoints

The secondary endpoints are:

- Adverse events, serious adverse events (SAEs), and AEs leading to discontinuation or death, and severity of AEs as assessed by the National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE), version 5.0.
- Safety laboratory parameters, vital signs, 12-lead electrocardiograms (ECGs), left ventricular ejection fraction (LVEF), Karnofsky Performance Scale (KPS)/Lansky Play-Performance Scale (LPPS), and concomitant medications.
- TGR as assessed by volume, defined as the percentage change in tumor volume every 6 months as assessed per BIRC.
- Objective response, defined as a best overall response CR, PR, or MR as assessed by the Investigator and by the BIRC per modified RANO-LGG.
- CR+PR, defined as a best overall response of CR or PR as assessed by the Investigator and by the BIRC per modified RANO-LGG.
- Time to response, defined as the time from the date of randomization to the date of first documented CR, PR, or MR for responders as assessed by the Investigator and by the BIRC per modified RANO-LGG.
- Time to CR+PR, defined as the time from the date of randomization to the date of first documented CR or PR for subjects with CR or PR as assessed by the Investigator and by the BIRC per modified RANO-LGG.
- Duration of response, defined as the time from the date of first documented CR, PR, or MR to the earlier of the date of death due to any cause or first documented

radiographic PD as assessed by the Investigator and by the BIRC per modified RANO-LGG.

- Duration of CR+PR, defined as the time from the date of first documented CR or PR to the earlier of the date of death due to any cause or first documented radiographic PD as assessed by the Investigator and by the BIRC per modified RANO-LGG.
- Overall survival, defined as the time from the date of randomization to the date of death due to any cause.
- HRQoL as assessed by the FACT-Br questionnaire.
- Progression-free survival as assessed by the Investigator per modified RANO-LGG.
- Serial or sparse blood sampling at specified time points for determination of plasma concentrations of vorasidenib and its circulating metabolite AGI-69460.

3.2.4. Exploratory Endpoints

The exploratory endpoints are:

- In the subset of subjects who crossover from placebo to vorasidenib after centrally confirmed radiographic PD by BIRC, the time from first dose to the date of documented progression on vorasidenib, as assessed by the Investigator per modified RANO-LGG, or death due to any cause, whichever occurs first.
- In the subset of subjects who crossover from placebo to vorasidenib after centrally confirmed radiographic PD by BIRC, TGR before and after treatment with vorasidenib.
- HRQoL as assessed by the EQ-5D-5L questionnaire and the PGI questions.
- Neurocognitive function as assessed by a validated battery of cognitive performance instruments measuring verbal learning, psychomotor function, working memory, attention, and executive function.
- Frequency, severity, and type of seizures, seizure AEs, number of antiepileptic drugs, and changes in antiseizure medications (dose, frequency).
- Baseline molecular and protein profiling in tumors, and morphologic, functional, epigenetic, biologic, and metabolic profiling in blood, plasma, and/or cerebrospinal fluid (CSF).
- TGR before and after treatment with vorasidenib and placebo.
- Time to malignant transformation and radiographic changes, defined as the time from the date of randomization to date of first histopathologic evidence of transformation and radiographic changes (eg, new enhancement, TGR changes, bidimensional changes) in subjects who have surgery or biopsy as an intervention.

4. STUDY DESIGN

This is a Phase 3, global, multicenter, double-blind, randomized, placebo-controlled clinical study to evaluate the efficacy and safety of vorasidenib versus placebo in approximately 340 subjects with residual or recurrent Grade 2 glioma with an IDH1 or IDH2 mutation.

Eligible subjects will be randomized 1:1 to receive either vorasidenib film-coated tablet formulation, 40 mg QD or vorasidenib–matched oral placebo QD. A subset of subjects randomized in the beginning of the study will be provided with vorasidenib uncoated tablet formulation, 50 mg QD or vorasidenib–matched oral placebo QD; these subjects will be switched to vorasidenib film-coated tablet formulation 40 mg QD (if randomized to the vorasidenib arm) or vorasidenib–matched oral placebo QD (if randomized to the placebo arm), when vorasidenib film-coated tablet formulation is available.

Subjects will receive study treatment in continuous 28-day cycles.

Subjects may continue treatment with their assigned study treatment until centrally confirmed radiographic PD by the BIRC; development of unacceptable toxicity; need for initiation of chemotherapy, radiotherapy, or other anticancer therapy in the opinion of the Investigator in the absence of centrally confirmed radiographic PD by the BIRC; confirmed pregnancy; death; withdrawal of consent from treatment; lost to follow-up; or Sponsor ending the study, whichever occurs first.

Subjects randomized to placebo with centrally confirmed radiographic PD, and who are not in need of immediate chemotherapy or radiotherapy in the opinion of the Investigator, will have the option to crossover to receive vorasidenib. Subjects who crossover will restart the same schedule of assessments as in the blinded treatment phase.

Subjects who discontinue study treatment for reasons other than centrally confirmed radiographic PD by the BIRC or withdrawal of consent from treatment and overall study participation will enter PFS Follow-up with the same schedule of assessments as before study treatment discontinuation until radiographic PD is documented by the BIRC. Overall Survival Follow-up assessments will occur approximately 6 months (±4 weeks) after EOT (for subjects in PFS Follow-up, OS Follow-up will begin once PFS Follow-up has ended) and will continue for up to 5 years after the last subject is randomized, or all subjects have died, withdrawn consent from overall study participation, are lost to follow-up, or the Sponsor ends the study, whichever occurs first.

5. ANALYSIS DATA SETS

Only subjects who sign informed consent and are screened will be included in the analysis sets below.

The following analysis sets will be evaluated and used for presentation of the data:

- The Full Analysis Set (FAS) will include all subjects who are randomized. Subjects will be classified according to the randomized treatment arm according to the intent-to-treat (ITT) principle.

- The Safety Analysis Set will include all subjects who receive at least 1 dose of the study treatment. Subjects will be classified according to the treatment received; subjects randomized to placebo who receive at least one dose of vorasidenib prior to crossover, will be classified to the vorasidenib arm.
- The Per-Protocol Set (PPS) is a subset of the FAS. Subjects who meet any of the following criteria will be excluded from the PPS:
  - Do not receive at least 1 dose of the randomized treatment
  - Do not have any measurable lesions at baseline as assessed by the BIRC per modified RANO-LGG
  - Do not have histopathologically diagnosed Grade 2 oligodendroglioma or astrocytoma per WHO 2016 criteria (ie, do not meet Inclusion Criterion #3).
  - Have had any prior anticancer therapy other than surgery (biopsy, sub-total resection, gross-total resection) for treatment of glioma including systemic chemotherapy, radiotherapy, vaccines, small-molecules, IDH inhibitors, investigational agents, etc (ie, meet Exclusion Criterion #1).

Table 2 summarizes the use of the analysis sets.

Table 2: Analysis Sets for Each Endpoint

Endpoints	Full Analysis Set (FAS)	Per-Protocol Set (PPS)	Safety Analysis Set
Demographic and other baseline characteristics	✓		
Disposition	✓		
Major protocol deviations	✓		
Subsequent therapies	✓		
Exposure and concomitant therapies			✓
Efficacy	✓	✓ (PFS and TTNI)	
Safety			✓

6. GENERAL STATISTICAL CONSIDERATIONS

6.1. Randomization, Blinding, Unblinding, and Crossover

Subjects will be randomized in a 1:1 ratio to one of the following treatment arms:

- Vorasidenib = vorasidenib 40 mg QD
- Placebo = vorasidenib–matched oral placebo QD

A subset of subjects randomized prior to the introduction of the film-coated tablet received vorasidenib 50 mg QD or vorasidenib–matched oral placebo QD of the uncoated tablet formulation. Upon introduction of the film coated-tablet, these subjects were switched to vorasidenib film-coated tablet formulation 40 mg QD (if randomized to the vorasidenib arm) or vorasidenib–matched oral

placebo QD (if randomized to the placebo arm), while maintaining the study blind, with the film-coated tablets assigned and dispensed by the IWRS system.

Randomization assignment will be implemented by an Interactive Response System (IWRS) and stratified by:

- Chromosome 1p19q codeletion status (co-deleted or not co-deleted)
- Tumor size at baseline per local assessment (longest diameter of ≥2 cm or <2 cm)

This is a double-blind study where study subjects, Investigators, relevant clinical site staff, and the Sponsor will be blinded to study treatment assignment. The IWRS will assign each subject specific Medication ID–labeled study drug containers. Vorasidenib and placebo will be packaged and labeled identically so that the study pharmacist will remain blinded to treatment assignment.

Study subjects and relevant clinical site staff will remain blinded for the duration of study treatment until centrally confirmed radiographic PD by BIRC. The Sponsor, except for select identified individuals, will remain blinded to the treatment assignment and data until the final analysis for the primary endpoint.

For all subjects, after confirmation of radiographic PD by the BIRC, the subject's treatment assignment will be unblinded via the IWRS. At this time, the subject, Investigator, relevant clinical site staff, and clinical research organization (CRO) study members will be unblinded to the subject's treatment assignment.

Subjects randomized to placebo who have centrally confirmed radiographic PD by the BIRC, and who are not in need of immediate chemotherapy or radiotherapy in the opinion of the Investigator, will have the option to crossover to receive vorasidenib, provided the following eligibility criteria are met based on the EOT assessments: all initial screening eligibility criteria except Inclusion Criteria numbers 1, 3, 4, 5, 6, and 12 and Exclusion Criteria numbers 1, 6, and 8. Subjects randomized to vorasidenib who have centrally confirmed radiographic PD by the BIRC will not be permitted to continue vorasidenib.

6.2. Sample Size Determination and Decision Rules

In this section PFS refers to PFS as assessed by the BIRC per modified RANO-LGG.

6.2.1. Sample Size Determination

The following statistical hypotheses will be tested to address the primary objective:

H_{01}: $\Theta_1 \geq 0$ vs H_{11}: $\Theta_1 < 0$

where Θ_1 is the log hazard ratio of PFS in the vorasidenib arm versus the placebo arm.

In addition, the following statistical hypothesis will be tested to address the key secondary objective associated with TTNI :

H_{02}: $\Theta_2 \geq 0$ vs H_{12}: $\Theta_2 < 0$

where Θ_2 is the log hazard ratio of TTNI in the vorasidenib arm versus the placebo arm.

Approximately 340 subjects will be randomized to the treatment arms using a 1:1 randomization, stratified by chromosome 1p19q codeletion status (co-deleted or not co-deleted) and baseline tumor size per local assessment (longest diameter of ≥2 cm or <2 cm).

For the primary endpoint, a total of 164 PFS events will be required to have at least 90% power to detect a hazard ratio of 0.6 using a 1-sided log-rank test stratified by the randomization stratification factors at a significance level of 0.025, and a 3-look group sequential design with a Gamma family (-24) α-spending function to determine the efficacy boundaries and a Gamma family (-5) β-spending functions to determine the non-binding futility boundary.

For TTNI, a total of 152 TTNI events will be required to have approximately 80% power to detect a hazard ratio of 0.636 using a 1-sided log-rank test stratified by the randomization stratification factors at a significance level of 0.025, and a 2-look group sequential design with a Gamma family (-22) α-spending function to determine the efficacy boundaries. To preserve the overall type I error in the study, the fixed sequence testing (Westfall and Krishen, 2001) will be followed; TTNI will be tested only if PFS has reached statistical significance (at the time of interim analysis 2 for PFS or final analysis for PFS).

The sample size for this study is determined based on the following assumptions:

- The median PFS for subjects in the placebo arm is 18 months and the median PFS for subjects in the vorasidenib arm is 30 months; this corresponds to a hazard ratio of 0.6 under the exponential model assumption
- Assuming TTNI to be equal to PFS plus an additional 3 months to accommodate any required washout periods for subsequent anticancer therapy and to prepare for subsequent anticancer therapy, the median TTNI for subjects in the placebo arm is estimated to be 21 (18+3) months, and the median TTNI for subjects in the vorasidenib arm is estimated to be 33 (30+3) months; this corresponds to a hazard ratio of 0.636 under the exponential model assumption
- PFS and TTNI drop-out rate of approximately 10% at 12 months
- Non-uniform recruitment period of approximately 42 months

The data cutoff for the final PFS analysis will occur after all subjects have been randomized and the target number of PFS events has been reached.

The study will have met the primary objective if the stratified log-rank test for PFS is statistically significant at the time of IA2 or final analysis at the α level determined by the α-spending strategy.

6.2.2. Decision Rules

The interim and the final analyses for PFS will be performed based on the FAS after the target number of events has occurred as described below. A maximum of 3 distinct data cutoffs are planned in the study:

- Interim Analysis 1 (IA1, futility only): at the time when approximately 55 PFS events (33.5% of the expected 164 events) have occurred; this data cut will only

be used for a futility assessment of PFS although an α of $3x10^{-9}$ will be spent, per the α-spending function, to protect the integrity of the study

- Interim Analysis 2 (IA2, superiority and futility): at the time when approximately 123 PFS events (75% of the expected 164 events) have occurred and all subjects have been randomized in the study
- Final Analysis (FA): at the time when 164 PFS events have occurred and all subjects have been randomized in the study

Table 3 displays the maximum number of analyses expected, and the associated efficacy and futility boundaries for the primary endpoint, if the analyses are performed at the planned number of events as shown in the table.

- The futility boundaries are non-binding, but the study may be stopped for futility if at the time of IA1 or IA2, PFS crosses the futility boundary.
- If the efficacy boundary for PFS is crossed at IA2 or FA, then the primary objective of the study will have been demonstrated.

Table 3: Efficacy and Futility Boundaries for PFS

Analysis	IA1	IA2	FA
Number of events (Information fraction)	55 (33.5%)	123 (75%)	164(100%)
1-sided p-value (z-value) for efficacy	NA[a]	≤0.00006 (≤-3.838)	<0.025 (<-1.96)
1-sided p-value (z-value) for futility[b]	≥0.806 (≥0.864)	≥0.185 (≥-0.898)	NA

Abbreviations: IA1=interim analysis 1, IA2=interim analysis 2, FA=final analysis, NA=not applicable
[a] The study will not stop for efficacy at IA1. However, to preserve the integrity of the study, 1-sided α=$3x10^{-9}$ will be spent at the time of IA1.
[b] Non-binding.
The observed number of events at the interim analyses may not match the planned number of events. The efficacy and futility boundaries will be updated based on the actual number of observed events using the pre-specified α-and β-spending functions.

There are 2 planned analyses for TTNI to test for superiority, at the time of the PFS IA2 and FA, respectively, per the testing strategy outlined in Section 6.2.1.

The significance levels for the analyses of TTNI are determined by the hierarchical testing strategy and the α-spending function for TTNI (Gamma (-22)). Table 4 displays the analysis triggers for TTNI and the associated efficacy boundaries, if the analyses are performed at the planned number of events as shown in the table.

Table 4: Efficacy Boundaries for TTNI

Analysis	IA	FA
Analysis cutoff trigger	123 PFS events	164 PFS events
Number of TTNI events (Information fraction)[a]	110 (72.4%)	152 (100%)
1-sided p-value (z-value) for efficacy	<0.00006 (<-3.858)	<0.025 (<-1.96)

Abbreviations: FA = final analysis; IA = interim analysis.
[a] Number of events expected under H_{12} assuming a hazard ratio for TTNI of 0.636
The observed number of events at the interim analysis may not match the planned number of events. The efficacy boundary will be updated based on the actual number of observed events using the pre-specified α-spending function.

Because the observed number of events at IA1 for PFS, IA2 for PFS, or IA for TTNI may not be exactly equal to the planned number of events, the efficacy and, for the primary endpoint, futility boundaries will be updated based on the actual number of observed events using the pre-specified α-and β-spending functions. Therefore, the observed Z-test statistics at the interim analyses will be compared with the updated efficacy and, for the primary endpoint, futility boundaries. If the study continues to 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 events documented at the cutoff date for the final analysis, the α already spent at the interim analyses, and the hierarchical testing strategy. Further details are provided in Section 7.10.

6.3. Definitions

6.3.1. Study Drug and Study Treatment

In this study both study drug and study treatment are defined as vorasidenib or placebo.

6.3.2. Start and End Dates of Study Drug and Study Treatment

The **start of study treatment** (vorasidenib or placebo) is the earliest date of administration of a non-zero dose of the study drug.

The **end of study treatment** (vorasidenib or placebo) is the latest date of administration of a non-zero dose of the study drug on or before the data cutoff date. For subjects randomized to placebo who crossover to receive vorasidenib, vorasidenib is not considered in the derivation of end of study treatment.

6.3.3. Study Day

The study day for assessments or events occurring on or after the start of study treatment (eg, AE onset, disease/response assessment) will be calculated as:

Study day=Date of the assessment or event−start of study treatment+1.

The study day for assessments or events occurring before the start of study treatment (eg, baseline characteristics, medical history) will be negative and calculated as:

Study day=Date of the assessment or event−start of study treatment.

There is no study day 0. The study day will be displayed in data listings.

6.3.4. Baseline

For efficacy evaluations, the last adequate assessment on or before the date of randomization will be used as the baseline. Per protocol, the first assessment for HRQoL-FACT-Br, HRQoL-EQ-5D-5L, PGI-F, PGI-S and neurocognitive function is planned to occur on Cycle 1 Day 1 before the start of study treatment. Therefore, for assessments of HRQoL-FACT-Br. HRQoL-EQ-5D-5L, PGI-F, PGI-S and neurocognitive function only, if there is no value available on or before the date of randomization, then the last measurement on or before the start of study treatment will be used as the baseline. For seizure activity, baseline will be based on data collected on C1D1 and include assessment of the subject's seizure history over the previous 30 days.

For summaries of baseline characteristics based on the FAS, baseline will be defined as follows:

- For subjects randomized and not dosed: the last assessment on or before the date of randomization
- For subjects randomized and dosed: the last assessment on or before the start of study treatment

For safety evaluations, the last assessment on or before the start of study treatment will be used as the baseline.

If, per protocol, an assessment is to be performed on study day 1, before the first dose of study treatment, and the assessment time, time of first dose of study treatment, or both, is missing (or not collected), it will be assumed that the assessment is performed before study treatment administration. Unscheduled assessments will be used in the determination of baseline; however, an unscheduled assessment on study day 1 will be considered to have been obtained after study treatment administration.

For biomarker evaluations, the last assessment on or before the date of the start of study treatment will be used as the baseline.

If no assessment meets the definition of baseline for an evaluation, the baseline will be set to missing.

6.3.5. On-Treatment Period

The on-treatment period starts on the date of the start of study treatment and ends at min(date of end of study treatment+28, date of start of subsequent anticancer therapy−1), where the start of subsequent anticancer therapy is defined in Section 6.3.6.

Data listings will include all assessments and events, with those that occur outside of the on-treatment period flagged.

6.3.6. Start of Subsequent Anticancer Therapy

The start of subsequent anticancer therapy is used in censoring for efficacy analyses. The earliest start date, on or after randomization, captured in the Prior and On Study Medications and On Study Procedures eCRF pages (with Subsequent Anti-cancer Therapy box checked) will be used in the analyses as the start of subsequent anticancer therapy.

6.3.7. Last Contact Date

The last contact date will be derived using the last complete date in the eCRF on or before the data cutoff date, from among the following:

- All assessment dates (eg, vital signs assessment, ECG assessment)
- Dates of administration of study drug, concomitant medications, and subsequent anticancer therapies
- Start and end dates of AEs
- Last contact date collected on the Overall Survival Follow-up eCRF when the subject status is alive
- Randomization date
- Withdrawal of consent date
- Date of discontinuation on disposition eCRF pages. If the option "Lost to Follow-up" is selected, the corresponding date will not be used in the derivation of last contact date.

Notes:

- Only dates associated with actual examinations of the subject will be used in the derivation. Dates associated with a technical operation unrelated to subject status, such as the date a blood sample is processed, will not be used.
- Assessment dates after the data cutoff date will not be used to derive the last contact date.

6.4. General Methods

6.4.1. Data Handling After Cutoff Date

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

6.4.2. 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, and 1 year=365.25 days.

The following derivations will be implemented.

- Age (years): (year of informed consent – year of birth)
- BMI (kg/m^2)=weight (kg)/height $(m)^2$
- Duration (in days) from a reference date (eg, randomization date, start date of study treatment)=
  - date of event−reference date + 1, if the date of the event is on or after the reference date
  - date of event−reference date, if the date of the event is before the reference date

Reporting conventions will be as follows:

- Mean and median will be displayed to one more decimal place than the raw data.
- Standard deviation (SD) will be displayed to two more decimal places than the raw data.
- Percentages will be displayed to 1 decimal place (however, percentages corresponding to 0 counts will be reported as 0 rather than 0.0 and 100 percent will be reported as 100 rather than 100.0).
- p-values will be reported with 4 decimal places; all p-values should be specified to be 1-sided or 2-sided.
- Unless otherwise specified, rounding will be performed to the closest integer/first decimal using the common mid-point between the two consecutive values, eg, 5.11 to 5.14 will be rounded to 5.1, and 5.15 to 5.19 will be rounded to 5.2.
  - Non-zero percentages that are <0.1 before rounding will be displayed as "<0.1", eg, 0.09 will be reported as <0.1 rather than as 0.1.
  - p-values<0.0001 before rounding will be displayed as "<0.0001", eg, a p-value of 0.00009 will be displayed as <0.0001 rather than as 0.0001.

6.4.3. Pooling of Data Across Sites

In order to provide overall estimates of treatment effects, data will be pooled across sites. The "site" factor will not be considered in statistical models or subgroup analyses given the high number of participating sites in contrast to the anticipated small number of subjects randomized at each site.

6.4.4. Continuous and Categorical Variables

Continuous variables will be summarized using descriptive statistics, ie, number of non-missing values, mean, SD, median, quartiles, minimum, and maximum. Time-to-event endpoints in the presence of censoring will be estimated using Kaplan-Meier methodology.

Categorical variables will be summarized by frequency distributions (number and percentage of subjects within a given category in the analysis data set). Unless otherwise specified, the calculation of percentages will include the "missing" category. Therefore,

counts of missing observations will be included in the denominator and presented as a separate category. For summaries by visit, percentages will be based on the number of subjects with data available for that visit, unless otherwise specified.

6.4.5. Unscheduled Visits

Generally, data collected at unscheduled visits will be included and summarized for both safety and efficacy analyses in the same manner as the data collected at scheduled visits. Descriptive statistics (mean, SD, median, quartiles, minimum, and maximum) by nominal visit or time point for safety endpoints such as laboratory measurements, ECG parameters and vital signs will include only data from scheduled visits. Summaries of outliers [eg, worst value, worst change from baseline, worst Common Terminology Criteria for Adverse Events (CTCAE) grade] during the on-treatment period for safety endpoints such as AEs, laboratory measurements and ECG parameters will include data from both scheduled and unscheduled visits.

6.5. Methods for Handling Missing Data

6.5.1. Adverse Event and Concomitant Medication Start Dates

If the end date is non-missing and the imputed start date is after the end date, the end date will be used as the start date.

(1) Missing day only

- If the month and year are the same as the month and year of the date of the start of study treatment, the date of the start of study treatment will be used.
- If the month and year are before the month and year of the date of the start of study treatment, the last day of the month will be used.
- If the month and year are after the month and year of the date of the start of study treatment, the first day of the month will be used.

(2) Missing day and month

- If the year is the same as the year of the date of the start of study treatment, the date of the start of study treatment will be used.
- If the year is before the year of the date of the start of study treatment, 31 December will be used.
- If the year is after the year of the date of the start of study treatment, 01 January will be used.

(3) Missing day, month, and year

- The date of the start of study treatment will be used.

6.5.2. Adverse Event and Concomitant Medication End Dates

If the start date is non-missing and the imputed end date is before the start date, the start date will be used as the end date. If the death date is available and the imputed end date is after

the death date, the death date will be used as the end date. If an imputation for an AE end date results in an AE end date that is after the data cutoff date, the AE will be considered as ongoing at the data cutoff date.

(1) Missing day only

- The last day of the month will be used.

(2) Missing day and month

- 31 December will be used.

(3) Missing day, month, and year

- The event will be regarded as ongoing.

6.5.3. Exposure

No imputation will be done for the date of the first dose of study drug.

If the date of the last dose of study drug is missing or partially missing, it will be imputed as follows (separately for each study drug):

- If the last date of study drug is completely missing and there is no End of Treatment Disposition eCRF page for the study drug AND there is no death date, the subject should be considered to be ongoing and the data cutoff date for the analysis will be used as the last dosing date.
- If the last date of study drug is completely or partially missing and there is EITHER an End of Treatment Disposition eCRF page for the study drug OR a death date (on or before the data cutoff date), then the imputed last dose date is:

  =Last day of the year, if only the year is available and Year<Year of min(EOT date, death date)

  =Last day of the month, if both the year and month are available and Year=Year of min(EOT date, death date) and Month <Month of min(EOT date, death date)

  =min(EOT date, death date), for all other cases

6.5.4. Death Date

Missing or partial death dates will be imputed based on the last contact date (as derived in Section 6.3.7), as follows:

- If the death date is missing it will be imputed as the day after the date of last contact.
- If the day is missing or both the day and month are missing, the death date will be imputed as follows:
  - Missing day only: max(1st day of the month and year of death, last contact date+1)
  - Missing day and month: max(01 January of the year of death, last contact date+1)

If the imputed death date is after the data cutoff date, the subject will be considered to be alive at the time of the data cutoff date.

7. STATISTICAL ANALYSES

7.1. Subject Disposition

For all subjects screened in the study, the following will be summarized:

- Number of subjects screened in the study
- Frequency (number and percentage) of subjects who discontinued the study before randomization, overall and by reason for discontinuation. Percentages will be calculated based on the number of subjects screened in the study.

In addition, the frequency of subjects in each of the analysis sets described in Section 5 will be summarized by treatment arm. Percentages will be calculated only for analysis sets that are a subset of the FAS or a subset of the safety analysis set.

The following summaries will be presented by treatment arm based on the FAS:

- Frequency of subjects in each randomization strata and combination of randomization strata (per IWRS)
  - Chromosome 1p19q codeletion status (co-deleted or not co-deleted)
  - Tumor size at baseline (longest diameter of ≥2 cm or <2 cm)
- Frequency of subjects in each randomization strata and combination of randomization strata (as derived from data in the eCRF)
  - Chromosome 1p19q codeletion status (co-deleted or not co-deleted) as reported in the Molecular Classification and Gene Mutation Analysis eCRF
  - Tumor size at baseline (longest diameter of ≥2 cm or <2 cm) as reported in the Screening Target Lesion eCRF
- Frequency of subjects randomized/treated in each geographic region, country, and site
- Frequency of subjects randomized and not treated, overall and by reason for discontinuation
- Frequency of subjects with study drug ongoing in the treatment (blinded) epoch
- Frequency of subjects who discontinued study drug in the treatment (blinded) epoch, overall and by the reason for discontinuation
- For each of the subsequent epochs [PFS follow-up, treatment (crossover), PFS follow-up (crossover), OS follow-up]:
  - Frequency of subjects who entered the epoch
  - Frequency of subjects ongoing

- Frequency of subjects who discontinued the epoch or the study, overall and by reason for discontinuation

The frequency of subjects with disposition reason, in each epoch, due to reasons associated with COVID-19 pandemic will further be summarized under the main reason for discontinuation.

In addition, the following cross-tabulations will be performed

- Cross-tabulation of randomization strata by IWRS vs randomization strata as derived from data in the eCRF
- Cross-tabulation of subjects randomized (vorasidenib, placebo) vs subjects who have received at least 1 dose of study drug prior to crossover (vorasidenib, placebo)

Disposition for all screened subjects and randomization data will be provided in by-subject listings.

7.2. Protocol Deviations

All major protocol deviations that impact the safety of the subjects, the conduct of the study, or the evaluation of the study results will be reported by treatment arm and overall, based on the FAS. These will include:

- Subjects randomized/treated despite not satisfying the eligibility criteria
- Subjects who develop withdrawal criteria while on the study but are not withdrawn
- Subjects who receive a study drug different from that assigned at randomization
- Subjects who are randomized under the wrong stratification factor(s)
- Subjects who receive an excluded concomitant medication

In addition, for each category of major protocol deviations, those related to COVID-19 will be summarized.

Major protocol deviations will be provided in a by-subject listing.

7.3. Demographic and Other Baseline Characteristics

The following summaries will be presented by treatment arm and overall based on the FAS, unless otherwise specified.

7.3.1. Demographics and Physical Measurements

Demographic characteristics and physical measurements at baseline will be summarized as follows:

- Demographic characteristics
  - Sex: male, female

- Race: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, White, other, unknown
- Ethnic origin: Hispanic or Latino, not Hispanic or Latino, not reported
- Age (years): summary statistics
- Age categories:
  - <16 years
  - 16-<18 years
  - 18-<40 years
  - 40-<65 years
  - ≥65 years

- Physical measurements
  - Height (cm)
  - Weight (kg)
  - BMI (kg/m^2)
  - Tanner Stage of Sexual Maturity (Stage 1 to 5)

Demographic data for all screened subjects will be provided in a by-subject listing.

7.3.2. Disease Characteristics

The following baseline characteristics of the underlying disease will be summarized based on the data entered in the eCRF:

- Histological subtype
- Karnofsky Performance Scale score at baseline (100, 90-80, 70-60, 50-40, 30-10)
- Lansky Play-Performance Scale (LPPS) score at baseline (100, 90-80, 70-60, 50-40, 30-10)
- Time since initial diagnosis of tumor to randomization
- Laterality at initial diagnosis
- Location of tumor at initial diagnosis
- MGMT promoter status (methylated, unmethylated, unknown)
- TERT promoter status (yes, no, unknown)
- ATRX mutation status (yes, no, unknown)
- P53 mutation status (yes, no, unknown)
- Pre-treatment tumor growth (descriptive statistics, and categorically for <4, 4-<8, ≥8 mm/year)

Data on disease characteristics will be provided in by-subject listings.

7.3.3. Medical History

Medical history will be summarized in frequency tabulations according to the latest version of the Medical Dictionary for Regulatory Activities (MedDRA) by System Organ Class (SOC) and Preferred Term (PT).

Medical history will be provided in by-subject listings.

7.3.4. Prior Therapies

Prior medications are defined as medications (from the Prior and Concomitant Medications eCRF) that are started before start of study treatment.

All non-study medications will be coded according to the Anatomical Therapeutic Chemical (ATC) code and PT using the latest version of the World Health Organization Drug Dictionary (WHODD). All prior medications will be summarized in frequency tabulations according to the WHO ATC third level and PT by treatment arm and overall, based on the safety analysis set.

Prior surgeries for glioma are defined as surgeries (from the Prior Surgery for Glioma eCRF) that are started before the randomization.

The prior surgeries will be summarized by treatment arm and overall, based on the FAS, as follows:

- Frequency of subjects with prior surgery for glioma overall and by
  - type of prior surgery
  - laterality
  - brain site
- Time from last surgery for glioma to randomization (descriptive statistics, and categorically for >1 - 2, >2 - 4, >4 years)

7.4. Exposure to Study Drug and Compliance

The following summaries will be presented by treatment arm based on the safety analysis set. The derivations are provided for vorasidenib or matched placebo. In Section 7.4.1 and Section 7.4.2, vorasidenib refers to vorasidenib or matched placebo, depending on the study treatment classification in the safety analysis set.

Exposure and compliance will be listed and summarized by treatment arm. In addition, a listing of exposure for subjects who receive vorasidenib uncoated tablet formulation 50 mg QD or matched placebo will be provided, indicating the start and end date of dosing with 50 mg QD or matched placebo.

7.4.1. Treatment Duration and Exposure

Exposure will be summarized as dose received [cumulative dose, actual dose intensity (DI)] and as dose received relative to planned dose [relative dose intensity (RDI)]. Duration of exposure to each study drug will be summarized as a continuous variable as well as in

categories (> 0 - 6, >6 - 12, >12 - 18, >18 - 24, >24 - 30, >30 - 36, >36 - 40, and >40 months).

Exposure to vorasidenib

- Duration of exposure (month)=(last non-zero dose date−first non-zero dose date+1)/30.4375
- Cumulative dose (mg) = sum of the actual doses
- Planned DI (mg/month) = planned cumulative dose (mg)/duration of exposure (month). Switching from 50 mg QD vorasidenib uncoated tablet formulation to 40 mg QD vorasidenib film-coated tablet formulation will be considered as planned for subset of subjects who will be provided the uncoated formulation at the beginning of the study. Therefore, the planned cumulative dose will be calculated as the sum of planned dose of the uncoated tablet formulation before the switch plus the sum of planned dose of the film-coated tablet formulation after the switch.
- Actual DI (mg/month)=cumulative dose (mg)/duration of exposure (month)
- RDI (%)=100×Actual DI (mg/month)/Planned DI (mg/month)

7.4.2. Dose Modifications

The summary of dose modifications will include:

- The frequency of subjects with at least 1 dose reduction
- Summary of the number of days with dose reductions
- The frequency of subjects with at least 1 interruption of study drug
- Summary of the number of days with interruptions of study drug

Dose reduction is defined as an administered non-zero dose that is lower than the planned dose.

An interruption of study drug is defined as a 0 mg dose of vorasidenib or matched placebo on one or more days. What follows defines how dose interruptions will be counted in the case of multiple dose interruptions.

- If an interruption occurs consecutively for at least two days then it will be counted only once.
- If an interruption occurs for more than one day, but the days are not consecutive, ie, there is at least one dosing day in between, then each dose interruption will be counted as a different occurrence (eg, if the actual dose on days 1, 3 and 5, is 10 mg and the actual dose on days 2 and 4 is 0 mg, the total number of dose interruptions is 2).

7.5. Concomitant Therapies

The following summaries will be presented by treatment arm based on the safety analysis set.

Concomitant medications are defined as non-study medications (from the Prior and Concomitant Medications eCRF) that are started during the on-treatment period or are started before the start of the study treatment and end or remain ongoing during the on-treatment period.

All non-study medications will be coded according to ATC code and PT using the latest version of the WHO Drug Dictionary. All concomitant medications will be summarized in frequency tabulations according to WHO ATC third level and PT.

Concomitant procedures are defined as procedures (from the On-Study Procedures eCRF) that are started during the on-treatment period or are started before the start of the study treatment and end or remain ongoing during the on-treatment period.

The concomitant procedures will be coded by the latest version of MedDRA by SOC and PT and will be summarized in frequency tabulations by SOC and PT.

7.6. Subsequent Therapies

The following summaries will be presented by treatment arm based on the FAS. Subsequent therapies are defined as therapies that are started after the last dose of study treatment (for subjects randomized and dosed) or after randomization (for subjects randomized and not dosed).

- Frequency of subjects with any subsequent anti-cancer therapy (eg, drug, surgery, radiation) and by type of anti-cancer therapy
- Frequency of subjects with any subsequent anti-cancer drug therapies overall and by type of drug therapy. Subsequent anti-cancer drug therapies will be coded according to ATC code and PT using the latest version of the WHO Drug Dictionary
- Frequency of subjects randomized to placebo who crossover to receive vorasidenib will be summarized separately.

7.7. Efficacy Analyses

The following analyses will be based on the FAS using the IWRS randomization stratification factors, unless otherwise specified.

7.7.1. Primary Endpoint

7.7.1.1. Primary Analyses

PD in this section refers to documented radiographic progression of disease as assessed by the BIRC per modified RANO-LGG.

Progression-Free Survival (PFS) is defined as the time from randomization to the first documentation of PD or death due to any cause, whichever occurs first.

Schedule of tumor assessments:

- Screening
- Every 12 weeks (±7 days) beginning at C4D1 until PD by BIRC assessment

- Beginning at Cycle 37, tumor assessments will be conducted less frequently: every 6 months for the next 2 years, and annually after that until PD by BICR assessment
- Subjects who discontinue study treatment for reasons other PD by BIRC assessment or withdrawal of consent for further study participation will have tumor assessments performed at the EOT visit and every 12 weeks (±7 days) thereafter until PD by BIRC is documented, or the subject withdraws consent for further study participation
- Subjects randomized to placebo who crossover to receive vorasidenib will follow the same schedule of response assessments outlined above starting at Cycle 1 Day 1 after crossover.

Adequate assessments:

- An adequate baseline assessment is defined as
  - an assessment within 35 days before or on the date of randomization
  - all documented lesions must have non-missing assessments (ie, non-missing measurements for target non-enhancing lesions)
- An adequate postbaseline assessment is defined as an assessment where complete response (CR), partial response (PR), minor response (MR), stable disease (SD), or PD can be determined. Time points where the response is not evaluable (NE) or no assessment was performed will not be used for determining censoring dates.

Censoring:

- Subjects without an event or with an event after 2 or more inadequate or missing postbaseline tumor assessments will be censored on the date of the last adequate tumor assessment that documented no PD; regardless, deaths within 24 weeks after randomization for subjects who did not initiate subsequent anticancer therapy will be considered an event.
- If a subsequent anticancer therapy is started prior to an event, the subject will be censored on the date of the last adequate tumor assessment that documented no PD prior to the start of the subsequent anticancer therapy.
- Subjects with no adequate baseline tumor assessment or with no adequate postbaseline tumor assessments within 24 weeks after randomization will be censored on the date of randomization, unless the subject dies within 24 weeks after randomization, in which case, death will be an event on date of death.

The censoring and event date options to be considered for the PFS analysis are presented in Table 5.

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

Table 5: Outcome and Event Dates for Primary Analysis of PFS

Scenario	Date of Event or Censoring	Outcome
No adequate baseline assessment	Date of randomization [a]	Censored [a]
PD or death • after at most 1 missing or inadequate postbaseline tumor assessment, or • ≤ 24 weeks after randomization	Date of PD or death	Event
PD or death • after 2 or more missing or inadequate tumor assessments	Date of last adequate tumor assessment [b] documenting no PD prior to subsequent anticancer therapy or missed tumor assessments	Censored
No PD		
Subsequent anticancer given prior to PD or death		

[a] If the subject dies ≤24 weeks after randomization and did not initiate subsequent anticancer therapy, the death is an event with date on death date.
[b] If there are no adequate postbaseline tumor assessments prior to the PD or death, then the time without adequate assessment should be measured from randomization; if the criteria are met the censoring will be on the date of randomization.

The primary efficacy analysis for PFS will compare the PFS time based on BIRC assessment between the experimental arm (vorasidenib) and the control arm (placebo) and will be performed using a 1-sided stratified log-rank test as described in Section 6.2.1.

The hazard ratio will be estimated using a Cox's Proportional Hazard (PH) 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 by replacing the proportional hazards model by the discrete logistic model (Ties=Discrete option in SAS PROC PHREG).

In order to account for the group sequential design in this study, the repeated confidence interval (RCI) method (Jennison and Turnbull, 2000) will be used to construct the 2-sided RCIs for the hazard ratio at the interim and the final analyses of PFS. In addition, the unadjusted 95% CIs for the hazard ratio will also be reported at the interim and the final analyses for PFS.

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 2-sided 95% confidence intervals (CIs). In particular, the PFS rate at 3, 6, 12, 18, 24, 30, 36, 42, 48 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice, 2002 (conftype=loglog default option

in SAS PROC LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

The frequency (number and percentage) of subjects with each event type (PD or death) and censoring reasons will be presented by treatment arm. Reasons for censoring will be summarized according to the categories in Table 6 following the hierarchy shown.

Table 6: Censoring Reasons and Hierarchy for Primary Analysis of PFS

Hierarchy	Condition	Censoring Reason
1	No adequate baseline assessment	No adequate baseline assessment
2	Start of subsequent anticancer therapy before event	Start of subsequent anticancer therapy
3	Event after 2 or more missing or inadequate postbaseline tumor assessments/date of randomization	Event after 2 or more missing or inadequate postbaseline assessments
4	No event and (EOS date ≥date of randomization when reason for EOS=Withdrawal by Subject)	Withdrawal of consent
5	No event and lost to follow-up in any disposition page or survival follow-up page	Lost to follow-up
6	No event and EOS date not missing and no adequate postbaseline tumor assessment	No adequate postbaseline tumor assessment
7	No event and none of the conditions in the prior hierarchy are met	Ongoing without an event

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

Time of Follow-Up for PFS

A Kaplan-Meier plot 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. Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median time of follow-up for PFS with 2-sided 95% CIs. In particular, the rate at 3, 6, 12, 18, 24, 30, 36, 42, 48 months will be estimated with corresponding 2-sided 95% CIs.

7.7.1.2. Sensitivity Analyses

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, hazard ratio and 95% CI) described in Section 7.7.1.1 with the modifications below:

- PFS based on BIRC assessment and counting all PD and deaths as events regardless of missing assessments or timing of the event
- PFS based on BIRC assessment on the PPS

- PFS based on BIRC assessment using an unstratified analysis
- PFS based on BIRC assessment using strata derived according to eCRF data instead of those entered in IWRS
- PFS based on BIRC assessment modifying the censoring rules in Table 5 to consider all deaths as events
- PFS based on BIRC assessment modifying the censoring rules in Table 5 with initiation of subsequent anticancer therapy not used as a censoring reason
- PFS based on BIRC assessment on the FAS excluding subjects who received vorasidenib uncoated tablet formulation 50 mg QD or matched placebo

Methods to Evaluate the Validity of the Model Assumption

The PH assumption will be checked visually by plotting log(-log(PFS)) versus log(time) within each randomization stratum.

Schoenfeld residuals for the stratified Cox PH model will be plotted to investigate graphically violations from the PH assumption; a non-zero slope is evidence of departure from PH. The PH assumption will be formally tested using Schoenfeld's residual test (Schoenfeld, 1980; Therneau and Grambsch, 2000). Large departures from PH will be evidenced by a p-value <0.05.

If these show large departures from PH, then PFS by BIRC assessment will also be analyzed based on restricted mean survival time (RMST) differences.

The hazard ratio estimate from the Cox PH model is routinely used to empirically quantify the between-arm difference under the assumption that the ratio of the two hazard functions is constant over time. When this assumption is plausible, such a ratio estimate captures the relative difference between two survival curves. However, the clinical meaning of such a ratio estimate is difficult, if not impossible, to interpret when the underlying PH assumption is violated (ie, the hazard ratio is not constant over time).

The RMST is a robust and clinically interpretable summary measure of the survival time distribution. Unlike median survival time, it is estimable even under heavy censoring. There is a considerable body of methodological research (Royston and Parmar, 2011; Uno et al, 2014; Zhang, 2013) about the use of RMST to estimate treatment effects as an alternative to the hazard ratio approach.

The RMST methodology is applicable independently of the PH assumption and can be used, at a minimum, as a sensitivity analysis to explore the robustness of the primary analysis results. However, when large departures from the PH assumption are observed, the log-rank test is underpowered to detect differences between the survival distributions for the treatment arms, and a test of the difference between the RMST for the experimental arm and the control arm may be more appropriate to determine superiority of the experimental arm compared to the control arm with respect to the time-to-event endpoint.

As it pertains to the **cutoff point (τ)** to evaluate the RMST, the cutoff point should not exceed the minimum of the largest observed time for both treatment arms so that the RMST of all treatment arms being evaluated can be adequately estimated and comparison between

treatments is feasible; τ should be clinically meaningful and closer to the end of the study follow-up so that the majority of survival outcomes will be covered by the time interval.

> τ=min(largest observed survival time for the experimental arm, largest observed survival time for the control arm).

The RMST up to time τ can then be interpreted as the expected survival time restricted to the common follow-up time τ among all subjects.

In this section, "survival" is meant to denote PFS.

The treatment effect between the experimental arm and the control arm will also be assessed based on the difference in RMST. The associated 95% CI for the difference in RMST and 1-sided p-value will be generated.

Investigator vs BIRC assessment

PFS based on investigator assessment will be analyzed as a secondary endpoint using the same methodology that is described in Section 7.7.1.1 referring to PD by investigator assessment instead of PD by BIRC assessment. In addition, a summary based on investigator assessment vs BIRC assessment will be provided including numbers of concordant and discordant assessments as well as the number of cases where PFS event was assessed at different timepoints based on investigator and BIRC assessments.

Table 7 outlines the possible outcomes by investigator and BIRC (Amit et al, 2011).

Table 7: Possible Outcomes for Investigator vs BIRC

		BIRC	
		Event	**No Event**
Investigator	**Event**	a = a1 + a2 + a3	b
	No Event	c	d

a1: number of agreements on timing and occurrence of event;
a2: number of times agreement on event but INV declares event later than BIRC;
a3: number of times agreement on event but INV declares event earlier than BIRC;
N= a+b+c+d.
The timing agreement of event is defined as a window of +7 days

The following measure of discordance will be calculated for each treatment arm:

- Total Event Discrepancy Rate: (b+c)/N
- Early Discrepancy Rate (EDR): (a3+b)/(a+b)
- Late Discrepancy Rate (LDR): (a2+c)/(a2+a3+b+c)
- Overall Discrepancy Rate: (a2+a3+b+c)/N

The EDR represents the positive predictive value of investigator assessment and quantifies the frequency with which the investigator declares PFS event earlier than BIRC within each treatment arm as a proportion of the total number of investigator-assessed events.

The LDR quantifies the frequency with which the investigator declares PFS event later than BIRC as a proportion of the total number of discrepancies within the treatment arm.

Discordance metrics are calculated for each treatment arm and, for each metric, the difference in discordance between the experimental and control arms is used to evaluate potential bias. If the discordance is similar across the treatment arms, then this suggests the absence of evaluation bias favoring a particular treatment arm. A negative differential discordance for EDR and/or a positive differential discordance for LDR may be indicative of investigator evaluation bias in favor of the experimental arm (Amit et al, 2011).

7.7.2. Key Secondary Endpoint

Time to next intervention is the time from randomization to the initiation of first subsequent anticancer therapy (including vorasidenib for subjects randomized to placebo who subsequently crossover to vorasidenib) or death due to any cause. If a subject does not initiate subsequent anticancer therapy or does not die by the data cutoff date, TTNI will be censored at the last known alive date.

TTNI (months) = (date of event or censoring – date of randomization +1)/30.4375,

where event is the first subsequent anticancer therapy or death, whichever occurs first.

The hazard ratio for TTNI will be estimated using a Cox's PH model stratified by the randomization strata. 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 by replacing the proportional hazards model by the discrete logistic model (Ties=Discrete option in SAS PROC PHREG).

In order to account for the group sequential design in this study, the RCI method (Jennison and Turnbull, 2000) will be used to construct the 2-sided RCIs for the hazard ratio at the interim and the final analyses of TTNI. In addition, the unadjusted 95% CIs for the hazard ratio will also be reported at the interim and the final analyses for TTNI.

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median TTNI with 2-sided 95% CIs. In particular, the TTNI rate at 3, 6, 12, 18, 24, 30, 36, 42, 48 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice, 2002 (conftype=loglog default option in SAS PROC LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

The frequency (number and percentage) of subjects with each event type (first subsequent anticancer therapy or death) and censoring reasons will be presented by treatment arm. Reasons for censoring will be summarized according to the categories in Table 8 following the hierarchy shown.

Table 8: Censoring Reasons and Hierarchy for TTNI

Hierarchy	Condition	Censoring Reason
1	No event and (EOS date ≥date of randomization when reason for EOS=Withdrawal by Subject)	Withdrawal of consent
2	No event and lost to follow-up in any disposition page or survival follow-up page	Lost to follow-up
3	No event and none of the conditions in the prior hierarchy are met	Ongoing without an event

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

7.7.3. Other Secondary Endpoints

The following analyses will be based on the FAS by treatment arm unless otherwise specified. Analyses for tumor-related endpoints will be performed separately based on BIRC assessment and based on Investigator assessment per modified RANO-LGG.

7.7.3.1. Tumor Growth Rate (TGR)

TGR is defined as the on-treatment percentage change in tumor volume every 6 months. The difference in TGR between the vorasidenib and placebo arms will be assessed by slope of tumor growth over time using a linear mixed effects model as implemented in the SAS procedure PROC GLIMMIX.

Tumor volume will be measured by the BIRC at baseline and after randomization following the schedule of tumor assessments outlined in Section 7.7.1.1. Under the assumption of exponential growth, log transformation of tumor volume will be modeled as follows.

Let y_{ij} be the tumor volume in the natural logarithmic scale for the i^{th} subject at the j^{th} measurement.

$$y_{ij} = \beta_0 + \beta_1 Group_i + \beta_2 Time_{ij} + \beta_3 Group_i \times Time_{ij} + \beta_4 LBTV_i + \beta_5 CODEL_i + u_{0i} + u_{1i} Time_{ij} + \varepsilon_{ij},$$

where

- $LBTV_i$: log of tumor volume at baseline for subject i, a fixed effect
- $CODEL_i$: codeletion randomization stratification stratum for subject i, a fixed effect
- $Group_i$: assigned treatment arm for subject i (0=control arm, 1= experimental arm) a fixed factor
- $Time_{ij}$: time in months from randomization to the date of the j^{th} measurement on the i^{th} subject
- u_{0i}: intercept coefficient associated with subject i, a random effect
- u_{1i}: slope coefficient associated with subject i, a random effect

- β_0: overall intercept
- β_j: regression coefficients for fixed effect factors, j=1, …, 5.
- ε_{ij}: random error associated with the subject i (residuals)

$\begin{bmatrix} u_{0i_i} \\ u_{1i} \end{bmatrix}$ and ε_{ij} are assumed to be independent and

$\varepsilon_{ij} \sim N(0, \sigma_e^2)$

$\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix} \right)$.

While the error term will have a diagonal covariance matrix, there will be a correlation between u_{0i} and u_{1i}. The model will take this into account using an unstructured covariance matrix (UN) between u_{0i} and u_{1i}. Should the estimation algorithm not converge, then a Compound Symmetry matrix (CS) will be considered. In this case, the assumption is that the intercept and slope have the same variance.

$$\begin{bmatrix} u_{0i_i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \sigma_{12} \\ \sigma_{21} & \sigma^2 \end{bmatrix} \right)$$

If the model fails to converge using CS, the variance component (VC) structure will be used by assuming $\sigma_{12} = \sigma_{21} = 0$. The log-likelihood ratio test will be used to test for the homogeneity between the residuals across treatment groups. If the homogeneity of the test is rejected at the 2-sided 0.05 significance level, different residual variances, σ_e^2, will be estimated for each treatment arm.

The least-square (LS) mean of TGR every 6 months with its 95% CI will be summarized by treatment arm based on the FAS. TGR every 6 months will be estimated for each treatment arm as follows:

- Vorasidenib: $\text{TGR} = 100 \times (e^{6(\beta_2+\beta_3)} - 1)$
- Placebo: $\text{TGR} = 100 \times (e^{6\beta_2} - 1)$

Log-transformed TGR is assumed to be normally distributed and this will be verified visually with a QQ-plot of residuals, using the stored SAS output obtained from fitting the linear mixed model.

7.7.3.2. Best Overall Response and Objective Response

Best overall response (BOR) will be assessed based on responses at different evaluation time points from randomization until the first documentation of PD, according to the following rules. Only tumor assessments performed on or before the start date of any subsequent anticancer therapies will be considered in the determination of BOR.

- CR = at least one determination of CR.
- PR = at least two determinations of PR at least 4 weeks apart and before first documentation of PD or PR sustained for at least 4 weeks.
- MR= at least one determination of minor response.

- SD = at least one SD assessment or better before first documentation of PD (and not qualifying for CR, PR, MR and PD)
- PD = PD after randomization (and not qualifying for CR, PR, MR and SD).
- NE: all other cases

Objective Response (OR) is defined as a BOR of CR, PR, or MR.

Subjects who do not have a postbaseline tumor assessment due to early PD, who receive subsequent anticancer therapies prior to achieving a CR, PR, MR, or who die, have a PD or drop out for any reason prior to achieving a CR, PR or MR will be counted as non-responders in the assessment of OR. Each subject will have an objective response status (0: no OR; 1: OR). OR rate (ORR) is the proportion of subjects with OR in the analysis set.

ORR by treatment arm will be calculated along with the 2-sided 95% CI using the Clopper-Pearson method (exact CI for a binomial proportion as computed by default by the FREQ procedure using the EXACT option).

In addition, the frequency (number and percentage) of subjects with BOR of CR, PR, MR, SD, PD, and NE will be tabulated. Subjects with BOR of NE will be summarized by reason for having NE status. The following reasons will be used:

- No baseline assessment
- No evidence of disease at baseline
- No postbaseline assessments due to death
- No postbaseline assessments due to other reasons
- All postbaseline assessments have overall response NE
- Subsequent anticancer therapy started before first postbaseline assessment

The association of study treatment and OR will be tested by the General Association Statistic of the Cochran-Mantel-Haenszel (CMH) test with the randomization strata taken into account. The null hypothesis of no association in any of the randomization strata is tested against the alternative, which specifies that there is an association between study treatment and tumor response at least in one randomization stratum. The CMH test will be performed at 2-sided alpha level of 0.05.

The stratified odds ratio in terms of OR will also be estimated along with its 95% CI to compare study treatments. The odds ratio is defined as the odds of OR with experimental treatment divided by the odds of OR with control treatment. The Breslow-Day test will be used to check the homogeneity of the odds ratio across the randomization strata. It tests the null hypothesis that odds ratios in all strata are equal against the alternative hypothesis that at least in one stratum the odds ratio is different.

In case the null hypothesis of homogeneity of odds ratios across strata is not rejected at the 2-sided alpha level of 0.05, the common odds ratio will be determined using the Mantel-Haenszel estimate (by the FREQ procedure using CMH option in SAS); if the null hypothesis of homogeneity of odds ratio across all strata is rejected, the odds ratio per stratum will be calculated with the corresponding exact CI.

BIRC vs Investigator Assessment:

Table 9 outlines the possible BOR outcomes by investigator and BIRC.

Table 9: Possible BOR Outcomes for Investigator vs BIRC

BOR		BIRC Assessment					
		CR	PR	MR	SD	PD	NE
Investigator Assessment	CR	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}
	PR	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}
	MR	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{36}
	SD	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}
	PD	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}	n_{56}
	NE	n_{61}	n_{62}	n_{63}	n_{64}	n_{65}	n_{66}

$\sum_{i=1}^{6}(n_{ii})$ is the number of agreements on BOR between BIRC and Investigator

$\sum_{i,j=1}^{6}(n_{ij})$ for i ≠ j is the number of disagreements on BOR between BIRC and Investigator

N= $\sum_{i,j=1}^{6}(n_{ij})$

The following measures of concordance will be calculated for each treatment arm:

- Concordance rate for BOR = $\sum_{i=1}^{6}(n_{ii})\ /N$
- Concordance rate for response = $[\sum_{i,j=1}^{3}(n_{ij}) + \sum_{i,j=4}^{6}(n_{ij})]/N$

Concordance rates are calculated for each treatment arm and, for each metric, the difference in concordance between the experimental and control arms are used to evaluate potential bias. A similar concordance across the treatment arms suggests the absence of evaluation bias favoring a particular treatment arm.

7.7.3.3. CR+PR

The endpoint CR+PR is defined as best overall response of CR or PR. The analysis of CR+PR will be conducted using the same methodology described in Section 7.7.3.2 for objective response by counting subjects who achieve MR as nonresponders in the assessment of CR+PR.

7.7.3.4. Time to and Duration of Response

Time to response (TTR) is defined, for subjects with OR, as the time from randomization to the first documentation of objective response (CR, PR, or MR) which is subsequently confirmed.

TTR (in months) = (first date of OR – date of randomization +1)/30.4375

TTR will be summarized, by treatment arm, using simple descriptive statistics.

Duration of Response (DoR) is defined, for subjects with OR, as the time from the first documentation of objective response (CR, PR, or MR) to the first documentation of PD or death due to any cause.

The outcome, event dates and reasons for censoring for DoR will match those for the analysis of PFS (Table 5) with the exception that subjects will not be censored for no adequate baseline assessment or for no adequate postbaseline assessment, as only subjects with an OR are included in the analysis of DoR.

DoR (months) = (date of event or censoring– first date of OR +1)/30.4375

Kaplan-Meier estimates (product-limit estimates) will be presented by treatment arm together with a summary of associated statistics including the median DoR with 2-sided 95% CIs. In particular, the DoR rate at 3, 6, 12, 18, 24, 30, 36, 42, 48 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice, 2002 (conftype=loglog default option in SAS PROC LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

DoR will be displayed graphically and analyzed using Kaplan-Meier methodology. If the number of subjects with OR is small, the Kaplan-Meier method may not provide reliable estimates. In this case, only descriptive statistics or listings will be provided.

7.7.3.5. Time to and Duration of CR+PR

Time to CR+PR is defined, for subjects with CR or PR, as the time from randomization to the first documentation of CR or PR.

Time to CR+PR (in months) = (first date of CR or PR – date of randomization +1)/30.4375

Duration of CR+PR is defined, for subjects with CR or PR, as the time from the first documentation of CR or PR to the first documentation of PD or death due to any cause.

The outcome, event dates and reasons for censoring for duration of CR+PR will match those for the analysis of PFS (Table 5) with the exception that subjects will not be censored for no adequate baseline assessment or for no adequate postbaseline assessment, as only subjects with a CR or PR are included in the analysis of duration of CR+PR.

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

The analysis of both endpoints time to and duration of CR+PR will be conducted using the same methodology described in Section 7.7.3.4 for time and duration of response by counting subjects who achieve MR as nonresponders in the assessment of CR+PR.

7.7.3.6. Overall Survival

Overall survival is defined as the time from date of randomization to the date of death due to any cause. If a subject is not known to have died by the data cutoff date, then OS will be censored at the date of last contact (see Section 6.3.7).

OS (months) = (date of death or censoring– date of randomization +1)/30.4375

The hazard ratio for OS will be estimated using a Cox's PH 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 by replacing the proportional hazards model by the discrete logistic model (Ties=Discrete 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 OS time with 2-sided 95% CIs. In particular, the OS rate at 3, 6, 12, 18, 24, 30, 36, 42, 48 months will be estimated with corresponding 2-sided 95% CIs. The CIs for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice, 2002 (conftype=loglog default option in SAS PROC LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

The frequency (number and percentage) of subjects with an event (death) and censoring reasons will be presented by treatment arm. Reasons for censoring will be summarized according to the categories in Table 10 following the hierarchy shown.

Table 10: OS Censoring Reasons and Hierarchy

Hierarchy	Condition	Censoring Reason
1	No event and [withdrawal of consent date ≥ date of randomization OR End of study (EOS) = Subject refused further follow-up]	Withdrawal of consent
2	No event and [lost to follow-up in any disposition page OR data cutoff date – last contact date > 30 weeks]	Lost to follow-up
3	No event and none of the conditions in the prior hierarchy are met	Alive

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

7.7.3.7. HRQoL as Measured by the FACT-Br

The FACT-Br is a patient-reported outcome (PRO) measure designed to assess HRQoL for patients with brain malignancies. The measure is comprised of the 27 items from the Functional Assessment of Cancer Therapy-General (FACT-G) and a 23-item disease-specific subscale. The FACT-Br contains the following subscales and scores: Brain cancer subscale (BrCS; 23 items), Physical Well-Being (PWB; 7 items), Social/Family Well-Being (SWB; 7 items), Emotional Well-Being (EWB; 6 items), Functional Well-being (FWB; 7 items), FACT-G total score (ie, PWB+SWB+EWB+FWB; 27 items), the FACT-Br total score (ie, PWB+SWB+EWB+FWB+BrCS; 50 items), and a Trial Outcome Index (TOI; PWB+FWB+BrCS; 37 items). Participants rate each item on a 5-point rating scale (0 = "Not at all", 1 = "A little bit", 2 = "Somewhat", 3 = "Quite a bit", and 4 = "Very much").

The scoring manual outlines several items that are reverse coded. Scores are generated for the BrCS, PWB, SWB, EWB, and FWB subscales by: 1) summing responses, 2) multiplying by the number of items in the subscale, and 3) dividing by the number of items answered. This prorated scoring procedure is acceptable provided that more than 50% of the items in a subscale are answered by each subject (eg, a minimum of 4 of 7 items, 4 of 6 items). The FACT-G total score is calculated as the sum of the four core subscale scores, provided the overall item response is ≥80% (ie, at least 22 of the 27 items were answered), and has a possible range of 0-108 points. The FACT-Br total score is the sum of all components of the FACT-G plus the BrCS score with possible range of 0-184 points.

Descriptive statistics will be used to summarize the individual items, subscale scores, total scores and change from baseline in the total and subscale scores at each scheduled assessment time point by treatment arm. The analyses described below will focus on five key FACT-Br scores: FACT-Br total score, BrCS, PWB, FWB, and TOI subscale scores, which are expected to be most sensitive to change in the target population. The frequency of subjects with missing assessments at each timepoint will be summarized by treatment arm.

Longitudinal mixed models for key FACT-Br scores:

Change in FACT-BR total and subscale scores from baseline at subsequent assessments will be analyzed separately using a linear mixed model for repeated measures. Each model will include treatment arm, time and the interaction time by treatment arm as fixed effect factors; and baseline score, and randomization stratification factors as fixed covariates. Both intercept and slope of time will be included in the model as random effect factors. An unstructured (UN) variance-covariance structure will be used to model the variance-covariance matrix of the random factors. A heterogeneous model with different residual variances across treatment arms will be considered if the test for homogeneity is rejected at 2-sided 0.05 level. If the model fails to converge using UN structure, a reduced variance-covariance structure will be considered (eg, CS or VC).

Comparison between the treatment arms at each assessment time point (eg, 1-, 2-, 3-, 4-months and every 3 months after until EOT) will be conducted using a t-test on LS-means of changes from baseline scores. The LS-mean difference between the experimental and control arm will be estimated with 95% CIs at each timepoint. Graphical display of LS-mean estimates over time will be generated by treatment arm for the total and subscale scores.

Change from baseline of total and subscale scores are assumed to be normally distributed and this will be verified visually with a QQ-plot of residuals, using the stored SAS output obtained from fitting the linear mixed model.

The assumption of linearity between changes from baseline scores, and the independent variables time and baseline tumor size will be evaluated using Pearson's correlation coefficients and visually using scatter plots. Locally weighted scatterplot smoothing methods may be considered to explore appropriate functions that best fit the data and/or for linearization. Polynomial regression models (eg, adding quadratic term of time in the linear model) or spline models will be considered.

Meaningful change thresholds:

The analysis of meaningful change will focus on establishing the range of scores that reflect meaningful change with respect to deterioration from baseline in the treatment arm. Primary analyses of meaningful change will be established using anchor-based methods and the PGI questions. Distribution-based approaches will supplement these primary analyses to contextualize measure variability in the key FACT-Br scores. Details of both the anchor-and distribution-based approaches follow.

Anchor-based approach

Anchor-based estimates of clinically meaningful change thresholds will be obtained for the key FACT-Br scores by using a linear model to estimate average change scores within strata defined by the PGI anchors. Estimates, effect sizes, and 95% CIs will be provided. Relevant subscales or items of the FACT-Br will be paired with their most appropriate PGI rating (e.g., the glioma symptoms PGI-S will be employed for the Brain cancer subscale). See description of PGI anchor questions in Section 7.7.4.5.

Meaningful changes for the PGI-S and PGI-F questions will be operationalized by taking the difference in the PGI-S or PGI-F rating between baseline and follow-up for each anchor and identifying subjects with a one-level worsening (eg, “mild” symptoms at baseline to “moderate” symptoms at follow-up). The follow-up timepoint will correspond to the timepoint used for modeling change in FACT-Br score. For the PGI-C, contrast variables will be created using the “No change” group as the reference group. Given the prognosis of this disease, even under treatment, subjects are not expected to report they are doing better on the PGI-C. As such, the intended meaningful change estimate will be defined using the point estimate for the group of subjects who indicated they were “A little worse” on the PGI-C at a timepoint corresponding to the timepoint used for modeling change in FACT-Br score. In the event that the cell size for this group is limited, the PGI-C rating will be recoded by collapsing the “A little worse” rating with the next worse rating. Linear models will be re-estimated to establish the estimate for meaningful change.

To supplement the anchor-based estimates of meaningful change and to graphically reflect the magnitude of the meaningful change detectable by the FACT-Br scores, empirical cumulative distribution function (eCDF) curves and probability density function (PDF) curves, described below, will be provided. The eCDF curves will be stratified on anchor categories (ie, subject’s ratings on the PGI-C, and changes in ratings on the PGI-S and PGI-F from baseline).

The eCDF graphs will be generated stratified by treatment arm. A vertical line will be plotted that best represents the meaningful change threshold. Graphs will be inspected for complete separation of the eCDF curves between arms, expected ordering of curves, and a treatment effect in the range of the clinically meaningful change threshold.

Distribution-based approach

Distribution-based estimates will be generated for the key FACT-Br subscales to support interpretation of score changes using proportion of SDs and estimates of the standard error of measurement (SEM). Thresholds for the SD of the scores will be established using 0.5 and 1 SDs, respectively. The SEM will be used to establish a threshold for change scores

that exceed measurement error of the FACT-Br. The SEM will be calculated as $SEM = S\sqrt{1 - r_{xx}}$, where S is the baseline SD of the FACT-Br scores, and r_{xx} is the reliability of these scores, which in this are the documented reliability indices for each score reported by the developers. A criterion of 2 SEMs will be used to demarcate a threshold for detectable change scores. Descriptive statistics at each of these distribution-based thresholds will be calculated.

Similar to the eCDF plots, probability distribution functions (PDFs) curves for the FACT-Br scores will be plotted by anchor categories to better understand the implications of different distributional statistics. The estimated SEM will be represented by a vertical line in each plot so that the density of change scores for a particular anchor category can be examined with respect to the measurement variability. Using unblinded data, PDF plots will also be created stratified for treatment groups.

Descriptive summaries for subjects' data after crossover:

Data from subjects in the placebo arm who crossover to vorasidenib will be summarized to assess changes in HRQoL after receiving vorasidenib.

To understand whether changes in HRQoL are evident, descriptive summaries for individual items, subscale scores, and total scores of the FACT-Br, as well as potential data visualizations will be provided. Scores will be summarized beginning with most recent FACT-Br scores prior to crossover, and at each available assessment post crossover. Given that these subjects will restart the schedule of assessments at C1D1, summaries of their scores post crossover will be compared to the corresponding Cycle and Day summary from the treatment group.

7.7.4. Exploratory Endpoints

The following analyses will be based on the FAS unless otherwise specified.

7.7.4.1. Progression-Free Survival After Crossover

Progression-free survival after crossover is defined as the time from first dose of vorasidenib after crossover to second documented PD based on investigator assessment or death due to any cause, whichever occurs earlier. This endpoint will be summarized for the subset of subjects from FAS who crossover from placebo to vorasidenib.

Kaplan-Meier estimates (product-limit estimates) will be presented together with a summary of associated statistics including the median time to malignant transformation with 2-sided 95% CI. In particular, the rate at 3, 6, 12, 18, 24, 30, 36, 42, 48 months will be estimated with corresponding 2-sided 95% CIs. The CI for the median will be calculated according to Brookmeyer and Crowley (1982) and the CIs for the survival function estimates at the time points defined above will be derived using the log-log transformation according to Kalbfleisch and Prentice, 2002 (conftype=loglog default option in SAS PROC LIFETEST) with back transformation to a CI on the untransformed scale. The estimate of the standard error will be computed using Greenwood's formula.

7.7.4.2. Pre- and Post-crossover TGR

For subjects randomized to placebo who crossover to vorasidenib, the TGR before and after the crossover may be different. To allow two different growth profiles before and after the crossover, the following piece-wise linear mixed model will be used.

Let y_{ij} be the tumor volume in the natural logarithmic scale for the i^{th} subject at the j^{th} measurement.

$$y_{ij} \sim N(\mu_{ij}, \sigma_y^2),$$

$$\mu_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i})t_{ij} + (\beta_2 + u_{2i})t'_{ij} + \beta_3 CODEL_i + \beta_4 LBTV_i,$$

where

- t_{ij}: time in months (since the crossover to vorasidenib) of the j^{th} measurement on the i^{th} subject before and after the crossover with $t_{ij} = 0$ at the crossover, a fixed effect
- $t'_{ij} = \begin{cases} 0 \ if \ t_{ij} \leq 0 \\ t_{ij} \ if \ t_{ij} > 0 \end{cases}$
- $LBTV_i$: log of tumor volume at baseline for subject i, a fixed effect
- $CODEL_i$: codeletion status for subject i, a fixed effect
- β_3, β_4: regression coefficients of $LBTV$ and $CODEL$
- $\beta_0, \beta_1, \beta_2$: overall intercept and slopes
- u_{0i}, u_{1i}, u_{2i}: random intercept and slopes of subject i

The random vector (u_{0i}, u_{1i}, u_{2i}) is assumed to follow a multivariate normal distribution with mean $(0, 0, 0)$ and an unstructured variance-covariance matrix (UN). If the model fails to converge using UN, a reduced variance-covariance structure (eg, CS or VC) will be considered. The pre and post-crossover TGR every 6 months will be estimated as follows:

- Before crossover: $TGR = 100 \times (e^{6\beta_1} - 1)$
- After crossover: $TGR = 100 \times (e^{6(\beta_1 + \beta_2)} - 1)$

The point estimates will be provided along with their associated 95% CIs. The analysis will be carried out using the SAS procedure PROC GLIMMIX.

7.7.4.3. Pre- and Post-Treatment TGR

The pre- and posttreatment TGR as assessed by volume will be estimated by treatment using the following piece-wise linear mixed model.

Let y_{ij} be the tumor volume in the natural logarithmic scale for the i^{th} subject at the j^{th} measurement.

$$y_{ij} \sim N(\mu_{ij}, \sigma_y^2),$$

$$\mu_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i})t_{ij} + (\beta_2 + u_{2i})t'_{ij} + \beta_3 Group_i + \beta_4 Group_i \times t_{ij} + \beta_5 Group_i \times t'_{ij} + \beta_6 CODEL_i + \beta_7 LBTV_i,$$

where

- t_{ij}: time in months (since the first dose) of the j^{th} measurement on the i^{th} subject before and after the first dose with $t_{ij} = 0$ at the first dose, a fixed effect.
- $t_{ij}^{'} = \begin{cases} 0 \ if \ t_{ij} \leq 0 \\ t_{ij} \ if \ t_{ij} > 0 \end{cases}$
- $Group_i$: assigned treatment arm for subject i (0=control arm, 1= experimental arm), a fixed factor
- $LBTV_i$: log of tumor volume at baseline for subject i, a fixed effect
- $CODEL_i$: codeletion status for subject i, a fixed effect
- $\beta_3, \beta_4, \ldots, \beta_7$: regression coefficients of fixed effects
- $\beta_0, \beta_1, \beta_2$: overall intercept and slopes
- u_{0i}, u_{1i}, u_{2i} : random intercept and slopes of subject i

The random vector (u_{0i}, u_{1i}, u_{2i}) is assumed to follow a multivariate normal distribution with mean $(0, 0, 0)$ and an unstructured variance-covariance matrix (UN). If the model fails to converge using UN, a reduced variance-covariance structure (eg, CS or VC) will be considered. The difference in TGR every 6 months between posttreatment and pretreatment (post – pre) will be estimated as follow:

- Vorasidenib: $TGR = 100 \times (e^{6(\beta_2+\beta_5)} - 1)$
- Placebo: $TGR = 100 \times (e^{6\beta_2} - 1)$

The point estimates will be provided along with their associated 95% CIs. The analysis will be carried out using the SAS procedure PROC GLIMMIX.

7.7.4.4. Time to Malignant Transformation

Time to malignant transformation is defined as the time from randomization to the date of histopathologic evidence of malignant transformation as assessed by the investigator in subjects who have surgery or biopsy as an intervention.

Time to malignant transformation, for subjects with malignant transformation, will be summarized, by treatment arm, using simple descriptive statistics.

7.7.4.5. HRQoL as Assessed by EQ-5D-5L and PGI Questions

HRQoL will further be characterized using the EQ-5D-5L questionnaire and the PGI-C, PGI-F, and PGI-S questions. The EQ-5D-5L, which includes the EQ-visual analog scale (EQ-VAS), will be scored according to published scoring guidelines. The EQ-5D-5L contains a descriptive system with one categorical response for each of 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). Ambiguous responses (eg, more than one response in a dimension) are treated as missing values. The EQ-VAS ranges between 0 (worst health) and 100 (best health).

Five PGI questions will be administered to aid in the interpretation of HRQoL and other patient-reported and performance outcome endpoints. The PGI-S is a self-rated evaluative instrument that will be administered to assess static, current-state severity of symptoms as perceived by the subject on a 4-point scale ranging from "none" to "severe." Three concepts will be assessed in separate PGI-S questions: glioma symptoms, neurocognitive functioning, and seizures. The PGI-F is a self-rated evaluative instrument that will be administered to assess static, current-state frequency of seizures as perceived by the subject on a 4-point scale ranging from "none" to "very often." The PGI-C is a self-rated evaluative instrument that will be administered to assess change in overall health/status as perceived by the subject on a 7-point scale ranging from "very much worse" to "very much improved."

Descriptive statistics and/or frequency tabulations will be used to summarize the individual items and VAS score of the EQ-5D-5L and the PGI questions. Change in scores from baseline will also be tabulated. Summaries will be reported by treatment arm and visit as appropriate.

7.7.4.6. Neurocognitive Function

Neurocognitive function will be assessed using a validated battery of 5 neurocognitive performance tests measuring verbal learning, psychomotor function, working memory, attention, and executive function. The battery of tests consists of the Detection test, which measures psychomotor functioning; the Groton Maze Learning test, which assesses executive function using a maze learning paradigm; the Identification test, which measures reaction time; the International Shopping List test, which measures verbal learning using a word list learning paradigm; and the One Back test, which measures working memory.

Descriptive statistics and/or frequency tabulations will be used to summarize individual neurocognitive performance test scores. Changes in scores from baseline will also be tabulated. Summaries will be reported by treatment arm and visit as appropriate.

7.7.4.7. Seizure Activity

Seizure activity will be assessed including the patient-reported monthly frequency and severity of seizures, type of seizures as assessed by the investigator, seizure AEs, and changes in anti-seizure medications (including dose and frequency).

Descriptive statistics and/or frequency tabulations will be used to summarize individual aspects of seizure activity. Changes from baseline will also be tabulated. Summaries will be reported by treatment arm and visit as appropriate.

7.7.5. Subgroup Analyses

Subgroup analyses to be performed for PFS by BIRC assessment, TTNI, and OR by BIRC are presented in Table 11.

Table 11: Subgroup Analyses

Subgroup	Categories
Chromosome 1p19q codeletion status (IWRS)	Co-deleted, Not co-deleted
Tumor size at baseline (IWRS)	Longest diameter of ≥2 cm, <2 cm
Gender	Male, Female
Race	Asian, Black or African American, White, Other
Ethnicity	Hispanic or Latino, Not Hispanic or Latino
Geographic Region	North America, Western Europe, rest of the world
Age	<18, 18-<40, 40-<65, ≥65 years
Pre-treatment tumor growth	<4, 4-<8, ≥8 mm/year
Number of prior surgeries	≤1, ≥2
Type of most recent surgery	Gross total, Subtotal or biopsy
Time from last surgery to randomization	<2, 2-<4, ≥4 years
Location of tumor at initial diagnosis	Frontal, Non-Frontal
MGMT hypermethylation	Yes, No, Unknown
TERT promoter mutation	Yes, No, Unknown
ATRX mutation	Yes, No, Unknown

If there is a low number of subjects within a category (<5% of the subjects in the FAS, the categories will be pooled (if 3 or more categories are pre-specified for the subgroup) or the subgroup will not be analyzed (if only 2 pre-specified categories in the subgroup). Efficacy analyses in subgroups will be purely exploratory and are intended to evaluate the consistency of treatment effect.

Subset analyses for each of the endpoints will use the methodology outlined in Sections 7.7.1.1, 7.7.2, and 7.7.3.2 without taking into consideration randomization stratification

The hazard ratios for PFS and TTNI, and the corresponding 95% CIs for all subgroups will be presented in a forest plot.

The ORR odds ratio for each subgroup and corresponding 95% exact CIs will also be presented in a forest plot for each treatment arm.

If there is a low number of subjects within a category (<5% of the subjects in the FAS), the categories will be pooled (if 3 or more categories are pre-specified for the subgroup) or the subgroup will not be analyzed (if only 2 pre-specified categories in the subgroup). Efficacy analyses in subgroups will be purely exploratory and are intended to evaluate the consistency of treatment effect.

7.8. Safety Analyses

Summaries of safety data will be presented by treatment arm based on the safety analysis set.

7.8.1. Adverse Events

Treatment-emergent adverse events (TEAEs) are AEs with a first onset date during the on-treatment period or worsening from baseline. All summaries described below will be based on TEAEs, if not otherwise specified.

All AEs will be listed by subject and AEs with onset outside of the on-treatment period will be flagged in the listings. Unless otherwise specified, TEAEs will be summarized according to the latest version of MedDRA by SOC and/or PT, severity (based on CTCAE v5.0 grading), seriousness, and relation to study treatment in decreasing frequency based on the frequencies observed for vorasidenib.

Each subject will be counted only once within each SOC or PT. If a subject experiences multiple TEAEs under the same PT within a SOC for the same summary period, only the TEAE assessed as related or with the worst severity, as applicable, will be included in the summaries of relationship and severity. If a subject has TEAEs with missing and non-missing grades, the maximum of the non-missing grades will be displayed. No imputation of missing grades will be performed.

The following will be summarized:

- TEAEs by SOC and PT
- TEAEs by SOC, PT, and worst grade
- Most common TEAEs and Grade ≥3 TEAEs by PT; these will include TEAEs (any grade) reported in ≥10% of subjects in either treatment arm or Grade ≥3 TEAEs reported in ≥5% of subjects in either treatment arm. These thresholds may be changed based on the observed data without an amendment to this SAP.
- Treatment-related TEAEs, by SOC and PT
- Treatment-related TEAEs, by SOC, PT, and worst grade
- Grade ≥3 TEAEs, by SOC and PT
- Treatment-related Grade ≥3 TEAEs, by SOC and PT
- Serious TEAEs, by SOC and PT
- Treatment-related Serious TEAEs, by SOC and PT
- TEAEs leading to discontinuation of study drug, by SOC and PT
- TEAEs leading to interruption of study drug, by SOC and PT
- TEAEs leading to dose reduction, by SOC and PT
- TEAEs leading to death, by SOC and PT
- Treatment-related TEAEs leading to death, by SOC and PT

7.8.1.1. Adverse Events of Special Interest

The following are considered AESIs:

- Elevated Liver Transaminases (see "vorasidenib-specified Safety Search Criteria" for the criteria used to identify the relevant AEs)

The following will be summarized for each AESI category:

- AESIs by PT
- AESIs by PT and worst grade
- Grade ≥3 AESIs by PT
- AESIs leading to discontinuation of study drug by PT
- Serious AESIs by PT
- AESIs leading to death by PT

7.8.1.2. Adverse Events Associated with COVID-19

The selection of AEs associated with COVID-19 will be based on the MedDRA MSSO list of PTs. The following will be summarized:

- TEAEs associated with COVID-19, by SOC and PT
- Grade ≥3 TEAEs associated with COVID-19, by SOC and PT
- Serious TEAEs associated with COVID-19, by SOC and PT
- TEAEs associated with COVID-19 leading to discontinuation of study drug, by SOC and PT
- TEAEs associated with COVID-19 leading to interruption of study drug, by SOC and PT
- TEAEs associated with COVID-19 leading to dose reduction, by SOC and PT
- TEAEs associated with COVID-19 leading to death, by SOC and PT

7.8.2. Death

The frequency of subjects in the safety analysis set who died, along with the cause of death, will be tabulated based on information from the Death Report eCRF. Cause of death will be summarized for the following categories:

- On-treatment death: Deaths within 28 days after the last dose of study treatment (ie, deaths during the on-treatment period)
- Post-treatment death: Deaths more than 28 days after the last dose of study treatment (ie, deaths after the end of the on-treatment period)
- Overall: All deaths

In addition, for each cause of death reported in the eCRF, those related to COVID-19 will be summarized.

Deaths for all screened subjects will be provided in a by-subject listing.

7.8.3. Clinical Laboratory Data

Clinical laboratory test results will be expressed in SI units.

For all laboratory tests (chemistry, hematology, coagulation), the actual values and the changes from baseline will be summarized by study visit and time point.

For each laboratory test performed in the study, a by-subject listing of laboratory test results will be presented with the corresponding CTCAE grades (if applicable), laboratory normal ranges, and flags for values below lower limit of normal (LLN) or above upper limit of normal (ULN).

Parameters with CTCAE grades available:

Clinical laboratory test results will be graded according to CTCAE v5.0 as applicable. Grading will be derived based on the numerical thresholds defined by the CTCAE criteria. Non-numerical qualifiers will not be taken into consideration in the derivation of CTCAE grading.

Laboratory test results classified according to CTCAE will be described using the worst grade. For parameters graded with 2 separate toxicity criteria, such as potassium (hypokalemia/hyperkalemia), the toxicities will be summarized separately. Low direction toxicity (eg, hypokalemia) grades at baseline and postbaseline will be set to 0 when the variables are derived for summarizing high direction toxicity (eg, hyperkalemia), and vice versa.

The frequency of subjects with laboratory toxicities during the on-treatment period will be tabulated as follows. The denominator used to calculate percentages for each laboratory test is the number of subjects evaluable for CTCAE grading for that parameter (ie, those subjects for whom a Grade of 0, 1, 2, 3 or 4 can be derived).

- The summary of laboratory parameters by CTCAE grade will include the number and percentage of subjects with Grade 1, 2, 3, 4; Grade 3-4; and Any Grade (Grades 1-4) during the on-treatment period. The highest CTCAE grade during the on-treatment period is considered the worst grade.
- The shift table will summarize baseline CTCAE grade versus worst CTCAE grade during the on-treatment period. The highest CTCAE grade during the on-treatment period is considered the worst grade.
- Newly occurring or worsening laboratory abnormalities (Any Grade, Grade 3-4) during the on-treatment period will also be summarized.

Parameters with CTCAE grades not available:

Results of laboratory tests that are not part of CTCAE will be presented according to the following categories: below the LLN, within normal limits, and above the ULN according to the laboratory normal ranges.

Shift tables will display the frequency of subjects with shifts from baseline missing, <LLN, normal, or >ULN to each of <LLN, normal or >ULN during the on-treatment period.

7.8.3.1. Hematology

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

Derived differential absolute count=(WBC count)×(Differential % value/100)

If the range for the differential absolute count is not available (ie, the range is only available for the percentage) then Grade 1 will be attributed as follows:

- 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 ≥1,500/mm^3

7.8.3.2. Chemistry

Liver function tests: Alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), and total bilirubin are used to assess possible drug-induced liver toxicity. The ratios of test result to ULN will be calculated and categorized for these parameters during the on-treatment period.

The summary of liver function tests will include the following categories. The frequency of subjects with each of the following during the on-treatment period will be summarized by treatment arm:

- ALT >3×ULN, ALT >5×ULN, ALT >10×ULN, ALT >20×ULN
- AST >3×ULN, AST >5×ULN, 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
- total bilirubin >2×ULN
- Concurrent ALT >3×ULN and total bilirubin >2×ULN
- Concurrent AST >3×ULN and total bilirubin >2×ULN
- Concurrent (ALT or AST) >3×ULN and total bilirubin >2×ULN
- Concurrent (ALT or AST) >3×ULN and total bilirubin >2×ULN and ALP ≥2×ULN

- Concurrent (ALT or AST) >3×ULN and total bilirubin >2×ULN and (ALP <2×ULN or missing)

Concurrent measurements are those occurring on the same date.

Categories will be cumulative, ie, a subject with an AST >10×ULN will also appear in the categories >5×ULN and >3×ULN. Liver function test 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 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 total bilirubin, ALT, AST, and ALP values for subjects with a postbaseline total bilirubin >2×ULN, ALT >3×ULN, or AST >3×ULN will be provided.

For **calcium**, CTCAE grading is based on corrected calcium and ionized calcium. Corrected Calcium is calculated from albumin and calcium as follows:

Corrected calcium (mmol/L)=measured total calcium (mmol/L)+0.02×[40−serum albumin (g/L)]

7.8.3.3. Pregnancy Tests

Pregnancy test results will be presented in a by-subject listing.

7.8.4. Vital Signs and Physical Measurements

For all physical measurements and vital sign assessments (height, weight, BMI, systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, temperature) the actual values and the changes from baseline will be summarized by study visit.

Vital signs and physical measurements will be presented in a by-subject listing.

7.8.5. Left Ventricular Ejection Fraction

LVEF% will be summarized using simple descriptive statistics of actual values (screening and EOT) and change from baseline.

LVEF% will be presented in a by-subject listing.

7.8.6. Electrocardiograms

ECG summaries will include all ECG assessments from the on-treatment period. QTcB interval will be derived based on RR and QT interval (see below).

Selecting Primary QT Interval Correction for Heart Rate

The analysis of QT interval data is complicated by the fact that the QT interval is highly correlated with heart rate. Because of this correlation, formulas are routinely used to obtain a

corrected QT interval, denoted QTc, which is independent of heart rate. This QTc is intended to represent the QT interval at a standardized heart rate. Several correction formulas have been proposed in the literature. For this analysis several of those methods of correction will be used, as described below. The QT interval corrected for heart rate by Bazett's formula, QTcB, is defined as

$$QTcB = \frac{QT}{\sqrt{RR}}$$

and the QT interval corrected for heart rate by the Fridericia's formula, QTcF, is defined as

$$QTcF = \frac{QT}{\sqrt[3]{RR}},$$

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

Although Bazett's correction is the historical standard, it does not perform well when heart rate fluctuates. Fridericia's formula may perform better under these conditions. If QTcB and QTcF do not adequately correct for heart rate and there are a sufficient number of subjects (>30) with baseline ECGs, an alternative correction (QTcP) to achieve the goal of getting uncorrelated QTc and RR is based on a linear regression method which yields, theoretically, uncorrelated QTc and RR.

Linear regression method:

- Fit a model QT (ms)=a+b×RR (sec) to baseline data
- Use the estimated slope, $\hat{b}$, to correct QT
- Corrected QT for heart rate will be derived as follows:

  QTcP (ms)=QT (ms)+ $\hat{b}$×[1−RR(sec)]

Data will be summarized using QTcF and QTcB. However, if these are not appropriate for the data set because of an observed large correlation between corrected QT and heart rate using the baseline assessments, the results will also be summarized using QTcP.

ECG Summaries

The following analyses will be performed for each applicable ECG parameter (RR, QT, and QTc) during the on-treatment period. The denominator used to calculate percentages for each category is the number of subjects evaluable for the category.

- Pearson correlation between QT and RR interval, QTc (QTcB, QTcF, and QTcP) and RR interval using individual (non-averaged) baseline assessments
- Frequency of subjects with notable ECG values, defined as those in the following categories:
  - QT/QTc interval increase from baseline >30 ms, >60 ms
  - QT/QTc interval >450 ms, >480 ms, >500 ms

All ECG assessments and qualitative ECG abnormalities will be presented in by-subject listings.

7.8.7. Performance Scores

The Lansky/Karnofsky score shift from baseline to lowest score during the on-treatment period will be summarized.

Lansky/Karnofsky scores will be presented in a by-subject listing.

7.9. Biomarker Analyses

Exploratory analyses will be performed for biomarker endpoints, based on the available data. There may be circumstances when a decision is made to stop collection, not perform, or discontinue the analysis of biomarker samples due to either practical or strategic reasons (eg, inadequate sample numbers, issues related to the quality of samples, or issues related to the assay that precludes the analysis of samples). Under such circumstances, the sample size may be too small to perform any data analysis and the available data will only be listed.

Listings will be presented for molecular, protein, and morphological (ie, appearance of microvascular proliferation, histological features of anaplasia, mitotic activity) profiling in tumors, and functional, epigenetic, biologic, and metabolic profiling in blood, plasma, and/or CSF. Summary tabulations or graphical presentations may be presented depending on the available data. If feasible, exploratory correlation analyses with clinical response or outcomes might be performed.

7.10. Interim Analyses

7.10.1. Introduction

The interim analyses will be performed by an independent statistician. The goals of the interim analyses for PFS are to allow early stopping of the study for futility or efficacy. The interim analyses will be performed as described in Sections 6.2.1 and 6.2.2.

Unblinded results from interim analyses will not be communicated to the Agios study team, the investigators or any other personnel involved in the study conduct, other than the independent statistician and IDMC members, until the IDMC has determined that:

- PFS has crossed the pre-specified boundary for efficacy, or
- the study needs to be terminated due to any cause, including futility or safety reasons.

Further details will be described in the IDMC charter.

7.10.2. Interim Analyses and Summaries

At each analysis time point, the critical boundaries for the group sequential test will be derived from the predefined spending function(s) as described in Section 6.2. The calculations of boundaries will be performed using EAST v6.5 or later.

7.10.2.1. Interim Analysis for PFS

Throughout this section, PFS refers to PFS as assessed by the BIRC per modified RANO-LGG.

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 for PFS, respectively, and let $l(t_i)$ and $l(t_F)$ denote the lower critical boundary for futility at the i^{th} interim and final analyses for PFS, respectively, where i=1, 2. For the final analysis, $l(t_F)=u(t_F)$.

In what follows P_0 and P_a denote the probabilities under H_{01} and H_{11} respectively, and $\alpha(t_i)$ and $\beta(t_i)$ denotes respectively the α and β spent based on the predefined spending functions at information fraction t_i; t_i is calculated as the ratio of the number of PFS events observed at the time of the cutoff for the i^{th} interim analysis and the total number of PFS events targeted for the final analysis.

The critical values $u(t_1)$ and $l(t_1)$ for the 1^{st} interim analysis of PFS will be derived so that the following criteria are met:

$$P_0(Z_1 \geq u(t_1)) = \alpha(t_1) \text{ and } P_a(Z_1 \leq l(t_1)) = \beta(t_1),$$

Critical boundaries for the IA2 and FA for PFS are calculated recursively as follows

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

$l(t_2)$ is derived such that $\beta(t_1) + P_a(Z_1 > l(t_1), Z_2 \leq l(t_2)) = \beta(t_2)$,

The boundary for the final efficacy analysis will be derived so that the following criteria are met:

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

7.10.2.2. Interim Analysis for TTNI

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 for TTNI, respectively.

In what follows P_0 denotes the probabilities under H_{02}, and $\alpha(t_1)$ denotes the α spent based on the predefined spending function at information fraction t_1; t_1 is calculated as the ratio of the number of TTNI events observed at the time of the cutoff for the interim analysis and the total number of TTNI events expected at the time of the final analysis.

The critical values $u(t_1)$ for the interim analysis of TTNI will be derived so that the following criterion is met:

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

The boundary for the final efficacy analysis of TTNI will be derived so that the following criterion is met:

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

8. REFERENCES

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.

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

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

Jennison C, Turnbull BW. Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC. 2000

Kalbfleisch JD, Prentice, RL. The Statistical Analysis of Failure Time Data. Second Edition. John Wiley & Sons, Inc. 2002

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

Royston P, Parmar MK. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Statistics in medicine, 2011 30 (19),pp. 2409-2421

Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. Biometrika 1980; 67:145-53.

Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Statistics for Biology and Health 2000.

Uno H, et al.. Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis. J Clin Oncol 2014; 32.

Westfall PH, Krishen A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. Journal of Statistical Planning and Inference. 2001;99(1):25-40

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.