Official Title: A Phase 3, Multicenter, Randomized, Double-Blind, Placebo-Controlled Study of AG-120 in Previously Treated Subjects with Nonresectable or Metastatic Cholangiocarcinoma with an IDH1 Mutation

NCT Number: NCT02989857

Document Date: SAP Version 1: 01-April-2019

agios

STATISTICAL ANALYSIS PLAN

A Phase 3, Multicenter, Randomized, Double-Blind, Placebo-Controlled Study of AG-120 in Previously Treated Subjects with Nonresectable or Metastatic Cholangiocarcinoma with an IDH1 Mutation

STUDY DRUG: **AG-120**

PROTOCOL NUMBER: **AG120-C-005**

SAP Version: Version 1.0

Draft Date: 01 April, 2019

Prepared by: [redacted], Ph.D. Date

AG-120-005 Biostatistician

Approved by: [redacted], M.D. Date

AG-120-005 [redacted]

Approved by: [redacted], Ph.D. Date

[redacted], Biostatistics

TABLE OF CONTENTS

1. INTRODUCTION

This statistical analysis plan (SAP) describes the analysis for Protocol AG-120-C-005, "A Phase 3, Multicenter, Randomized, Double-Blind, Placebo-Controlled Study of AG-120 in Previously Treated Subjects with Nonresectable or Metastatic Cholangiocarcinoma with an IDH1 Mutation," Amendment 4, Version 5.0, which was issued on 04 April 2018. It contains definitions of analysis populations, derived variables, and statistical methods for the analysis of efficacy and safety.

This SAP provides a comprehensive and detailed description of the strategy, rationale, and statistical techniques to evaluate the necessary efficacy and safety endpoints. The purpose of this SAP is to ensure the credibility of the study findings by pre-specifying the statistical approaches to the analysis of study data prior to database lock. Pharmacokinetic (PK)/pharmacodynamic analysis and exploratory biomarker analyses are not in the scope of this SAP. This SAP will be finalized and approved prior to the database lock for the primary analysis.

2. STUDY OBJECTIVES

2.1. Primary Objective

The primary objective is:

- To demonstrate the efficacy of AG-120 based on progression-free survival (PFS) per Independent Radiology Center (IRC) assessment compared to placebo in subjects with nonresectable or metastatic cholangiocarcinoma with an isocitrate dehydrogenase 1 (IDH1) mutation.

2.2. Secondary Objectives

The secondary objectives are:

- To evaluate the safety and tolerability of AG-120 compared to placebo.
- To evaluate PFS per Investigator assessment.
- To compare the efficacy of AG-120 with placebo based on overall survival (OS), objective response rate (ORR), duration of response (DOR), and time to response (TTR), with response assessed per the Investigator and by the IRC.
- To evaluate health-related quality of life (HRQOL) with AG-120 compared to placebo as assessed by the European Organization for Research and Treatment of Cancer Quality of Life Questionnaires (EORTC-QLQ-C30 and EORTC-QLQBIL21), the Patient Global Impression of Change (PGI-C), and the Patient Global Impression of Severity (PGI-S).
- To evaluate health economic outcomes as assessed by the 5-level EuroQol five dimensions questionnaire (EQ-5D-5L).
- To evaluate the pharmacokinetics (PK) of AG-120.
- To evaluate the PK/pharmacodynamic relationship of AG-120 and 2-hydroxyglutarate (2-HG) in blood samples.

2.3. Exploratory Objectives

The exploratory objectives are:

- To evaluate, for the subgroup of placebo subjects who have crossed over to the AG-120 arm, the time from first dose of AG-120 to second documented progression on AG-120 or death, whichever occurs first (PFS2).
- To correlate baseline molecular and/or protein characteristics in tumor tissues with clinical response.
- To correlate baseline 2-HG levels in plasma samples with clinical response.
- To evaluate levels of mutant IDH1 and other genes in circulating tumor DNA obtained from plasma at baseline and over the course of the treatment.

- To correlate any PK variations with drug-metabolizing enzyme (DME) related genes, if the data are warranted.
- To explore additional biomarkers in blood for morphologic, functional, biologic, epigenetic, and metabolic changes over the course of treatment.

3. STUDY ENDPOINTS

3.1. Primary Endpoint

The primary endpoint is PFS, defined as the time from date of randomization to date of first documented disease progression (as assessed by the IRC per RECIST v1.1), or date of death due to any cause.

3.2. Secondary Endpoints

- Adverse events (AEs), serious adverse events (SAEs), AEs leading to discontinuation or death. The severity of AEs will be assessed by the National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE), version 4.03.
- Safety laboratory parameters, vital signs, 12-lead electrocardiograms (ECGs), evaluation of left ventricular ejection fraction (LVEF), Eastern Cooperative Oncology Group (ECOG) performance status (PS), and concomitant medications.
- Secondary efficacy endpoints include:
    - OS, defined as the time from date of randomization to date of death.
    - ORR, defined as the proportion of subjects with a best overall response (BOR) defined as CR or PR, as assessed by the Investigator and by the IRC per RECIST v1.1.
    - DOR, defined as the time from date of first documented complete response (CR) or partial response (PR) to date of first documented disease progression or death due to any cause, as assessed by the Investigator and by the IRC per RECIST v1.1.
    - TTR, defined as the time from date of randomization to date of first documented CR or PR for responders, as assessed by the Investigator and by the IRC per RECIST v1.1.
    - PFS as determined by the Investigator.
- HRQOL as assessed by validated instruments (EORTC-QLQ-C30, EORTC-QLQBIL21, PGI-C, and PGI-S).
- Health economic outcomes as assessed by the EQ-5D-5L instrument.
- Serial or sparse blood sampling at specified time points for determination of plasma concentration-time profiles and PK parameters of AG-120.
- Blood sampling at specified time points for determination of 2-HG levels to characterize the pharmacodynamic effects of AG-120.

4. OVERALL STUDY DESIGN

This is a Phase 3, multicenter, randomized, double-blind, placebo-controlled efficacy and safety study of orally administered AG-120 in subjects with advanced cholangiocarcinoma (nonresectable or metastatic). Subjects, all personnel involved in the evaluation of subjects' response to treatment (eg, Investigators, study coordinators, study pharmacists), and designated Sponsor team members will be blinded to study treatment until documented disease progression.

Subjects are required to have a histologically consistent diagnosis of IDH1 gene-mutated cholangiocarcinoma that is not eligible for curative resection, transplantation, or ablative therapies. Subjects must have documented progression of disease and have received treatment with at least 1 but not more than 2 prior treatment regimens for advanced disease (nonresectable or metastatic). At least 1 of the prior regimens must have included gemcitabine or 5-FU. Systemic adjuvant chemotherapy will be considered a line of treatment if there is documented disease progression during or within 6 months of completing the therapy, with Sponsor approval.

A total of approximately 186 subjects will be randomized in a 2:1 ratio to the AG-120 and placebo arms, respectively, stratified by number of prior therapies (1 vs. 2). The primary analysis of PFS will occur once 131 PFS events have been determined by Investigator assessment. Overall survival will be analyzed twice, once at the time of PFS primary analysis and once at the occurrence of 150 OS events.

Subjects who meet all study eligibility criteria will be randomly assigned in a 2:1 ratio to receive AG-120 orally at a dose of 500 mg once daily (QD) or AG-120-matched oral placebo QD. Daily study treatment will begin on C1D1. Cycles are 28 days (+/-2 days) in duration and dosing is continuous. Study visits will be conducted every other week during Cycles 1-3 (Days 1 and 15), and Day 1 of each cycle thereafter. Radiographic assessment (computed tomography [CT] or magnetic resonance imaging [MRI]) for evaluation of disease response will be conducted every 6 weeks (±5 days) for the first 8 assessments (ie, through week 48) and every 8 weeks (±5 days) thereafter from C1D1, independent of dose delays and/or dose interruptions, and/or at any time when progression of disease is suspected.

Upon request by the Investigator, the subject and site staff will be unblinded to treatment assignment after documented disease progression (as assessed by the Investigator and after consultation with the Sponsor Medical Monitor), and subjects randomized to the placebo arm who continue to meet eligibility criteria determined in the End of Treatment (EOT) visit will be given the opportunity to cross over to the active treatment arm and receive AG-120. For those subjects who cross over, placebo will not be counted as a prior line of therapy for the purpose of eligibility. If subjects cross over, they will start again with study procedures as at C1D1. These subjects will continue to be evaluated for tumor response by the Investigator.

An independent data monitoring committee (IDMC) will review the safety data on a regular basis to ensure the safety of treatment and proper conduct of the study. The first interim safety review meeting will be conducted when approximately 20 subjects have completed 2 cycles of therapy or have discontinued earlier; thereafter, meetings will be conducted every 3-6 months or on an ad hoc basis. No formal interim analysis for efficacy will be conducted before the primary analysis.

5. SAMPLE SIZE ESTIMATION

A total of approximately 186 subjects will be needed for the study.

The primary objective of the study is to demonstrate the improvement in PFS by IRC assessment for subjects receiving treatment with AG-120 as compared to subjects receiving placebo.

Assuming a hazard ratio (HR) of 0.5 for PFS (equivalent to a median PFS of 3 months in the placebo arm vs. 6 months in the AG-120 arm, assuming an exponential distribution), a total of 131 PFS events are required to provide 96% power at a 1-sided alpha of 0.025 level of significance to reject the null hypothesis using a stratified log-rank test. Based on this, a total of approximately 186 subjects will need to be randomized in a 2:1 ratio to the AG-120 and placebo arms, respectively, assuming approximately a 22% dropout rate, an approximate 26-month randomization period, and an additional 6-month follow-up for PFS after the last subject is randomized. Therefore, the primary analysis of PFS will occur at approximately 6 months after the last subject is randomized.

Overall survival will be analyzed twice, once at the time of the final analysis for PFS and once at the occurrence of 150 OS events (final analysis for OS, which will occur at approximately 24 months after the last subject is randomized). Assuming an HR of 0.67 for OS (median OS of 8 months in the placebo arm vs. 12 months in the AG-120 arm, assuming an exponential distribution), a total of 150 OS events will provide 64% power at a 1-sided alpha of 0.025.

6. ANALYSIS POPULATIONS

The following subject populations (ie, analysis sets) will be evaluated and used for presentation of the data:

- Intent-to-Treat Set (ITT)
- Safety Analysis Set (SAS)
- Per-Protocol Set (PPS)
- Crossover Set (COS)

6.1. Intent-to-Treat Set (ITT)

All subjects who are randomized, with the treatment group designated according to the randomization. The ITT will be the primary analysis set for all analyses except for safety.

6.2. Safety Analysis Set (SAS)

All subjects who receive at least one dose of study drug (AG-120 or placebo). Subjects will be analyzed according to the actual treatment received. The SAS will be the primary analysis set for all safety analyses.

6.3. Per-Protocol Set (PPS)

All subjects in ITT who do not violate the terms of the protocol in a way that would significantly affect the study outcome, with treatment group designated according to the randomization.

In general, subjects who meet the following criteria will be excluded from this analysis set:

- Do not have histopathologically diagnosed nonresectable or metastatic cholangiocarcinoma
- Do not have documented IDH1 gene-mutated disease based on central laboratory testing
- Do not have any measurable lesion at baseline as defined by RECIST v1.1 as determined by IRC
- Three or more prior systemic therapy in an advanced setting (nonresectable or metastatic)
- Have received a prior IDH inhibitor.

6.4. Crossover Set (COS)

A subset of placebo subjects who cross over and receive AG-120 upon the radiographic documented progression (PD). The COS will be the analysis set for analyzing post-crossover data.

7. STATISTICAL ANALYSES

Unless specified otherwise, longitudinal data will be presented by "before crossover" and "after crossover" periods, defined as:

- **"After crossover"** contain data collected after placebo subjects crossed over to AG-120. COS will be the analysis set.
- **"Before crossover"** contain AG120 arm data and the Placebo arm data up till crossover.

7.1. Subject Disposition

The disposition table of subjects will include the following categories:

- Number and percentage of subjects who are randomized (ITT set)
  - Number and percentage of subjects who are randomized but not dosed
  - Number and percentage of subjects who are randomized and dosed
- An end-of-treatment disposition (still on treatment vs. discontinued from treatment) based on SAS set. The primary reason for treatment discontinuation will also be included.
- An end-of-study disposition (still on study vs. discontinued from study) based on ITT set. The primary reason for study discontinuation will also be included.
- Placebo subjects who crossed over to AG-120 based on COS set. An end-of-treatment disposition (still on AG-120 vs. discontinued from AG-120) and the primary reason for discontinuing AG-120 will be summarized.

A listing of subjects who failed screening, including the reason for screen failures will be provided.

A listing of subjects including the date of first dose, date of last dose, date of crossover (if any), reason of treatment discontinuation, date of death, cause of death, end of study date, and the reason of study discontinuation will be provided.

7.2. Protocol Deviations

A subject-level protocol deviation listing will be provided by site and subject based on ITT set.

Major subject-level protocol deviations will be summarized, including:

- Number and percentage of subjects with at least 1 major protocol deviation.
- Subjects by number of major protocol deviations (1, 2, ≥3).
- Number and percentage of subjects within each major protocol deviation category.

7.3. Demographics and Baseline Characteristics

Baseline/demographics-related tables will be presented per randomization (ITT).

7.3.1. Demographics

Demographic and other baseline variables will be summarized for each treatment arm based on all randomized subjects (ITT set) and for the two treatment arms combined (total).

Subject demographic data, including age, gender, race, regions, and ethnicity (if available) will be obtained during screening.

Age (years), height (cm), weight (kg), body mass index (BMI) (weight [kg]/height [m^2]), and other continuous baseline characteristics will be summarized descriptively (number of subjects, mean, standard deviation [SD], median, minimum, and maximum). Age category (<45, ≥45 - <65, and ≥65 years), sex, race, and ethnicity and other categorical variables will be summarized by frequency tabulations (count, percent).

If age is not collected in the database, it will be calculated as follows: age = [ICF year – birth year].

Body mass index will be calculated as follows: BMI (kg/m^2) = weight in kg / (height in m)^2.

7.3.2. Baseline Characteristics

The following baseline characteristics will be summarized by treatment arms and for the two arms combined (total) based on ITT set. The variables will include:

- Number and percentage of subjects within each randomization strata (1 vs. 2 prior lines of therapy per randomization)
- IDH1 allele types per central IDH1 testing
- ECOG at baseline
- Cholangiocarcinoma type at diagnosis (Intrahepatic Cholangiocarcinoma, Extrahepatic Cholangiocarcinoma, Perihilar Cholangiocarcinoma, Unknown)
- T/N/M stage at initial diagnosis (Tx, Nx, Mx)
- Grade at initial diagnosis (Well differentiated, Moderately differentiated, Poorly differentiated, Undifferentiated, Unknown)
- Extent of disease at Screening
- Liver cirrhosis at Screening
- Presence of Biliary Stent at Screening
- Presence of Ascites at Screening
- Ascites related to Cholangiocarcinoma within the past 3 months
- Paracentesis with the past 3 months
- Pleural effusion related to Cholangiocarcinoma within the past 3 months
- Thoracentesis within the past 3 months

Baseline characteristic listing will be provided.

7.4. Prior Therapy and Medical History

7.4.1. Prior Systemic Anticancer Therapy

Prior therapies will be summarized for each treatment arm and for the two treatment arms combined (total) based on the ITT set. It will be summarized continuously as well as categorically. Types of therapies (adjuvant, palliative, etc.) will also be summarized.

This is the actual prior lines of systemic anticancer therapy subjects received in advanced setting in order to determine the study eligibility, not the stratification factor entered at the randomization. In majority of cases these two should match. However, it does not preclude the likelihood of mismatch due to mis-stratification, etc. A table will be provided to summarize the discrepancy between these two.

Duration of last line of therapy (in months) in advanced setting will be summarized, calculated as (end date of the last line of therapy – start date of the last line of therapy + 1)/(30.4375).

In case there are multiple drugs within 1 line of therapy, the outer window rule will be applied to calculate the duration: the earliest start date of all drugs will be the start date, and the latest end date will be the end date for that line.

The imputation of partial/missing dates is described in Section 10.6.

A data listing will be provided and the line that is contributed to the eligibility will be flagged.

7.4.2. Prior Local-Regional Therapy

A data listing will be provided for prior local-regional therapy, including the type of prior local-regional therapy, anatomical location, start date, end date, therapy setting, best response, and date of progression.

7.4.3. Prior Surgery for Cholangiocarcinoma

A data listing will be provided for prior surgery, including procedure, location, pathology finding, date of procedure, whether disease recur or progress after the procedure, and date of recurrence or progression.

7.4.4. Medical History

Medical history will be summarized by System Organ Class (SOC) and Preferred Term (PT), and sorted alphabetically in SOC and in PT. Medical history will be summarized by treatment arms and by both treatment arms combined (total) based on ITT set. A listing of medical history will also be provided.

7.5. Prior/Concomitant Medications and Procedures

A summary table of concomitant medications will be provided by treatment arms and total. Concomitant medications are those medications that were ongoing at the time of first dose or that were initiated after first dose but prior to last dose plus 28 days (inclusive). If an end date is missing or the medication is ongoing, the medication will be included as concomitant medications. Concomitant medications will be tabulated by anatomic therapeutic class (ATC;

level 4) and preferred term (PT). If the concomitant medication starts on or after the start of AG-120 upon crossover, it will be summarized in the "after crossover" table.

A summary table of concomitant procedures will be summarized by indication and preferred term. If the concomitant procedure starts on or after the start of AG-120 upon crossover, it will be summarized in the "after crossover" table.

In addition, the following data listings will be provided:

- A listing for prior and concomitant medications, with a flag to indicate prior vs. concomitant
- A listing of concomitant procedures

7.6. Exposure to Study Drug and Compliance

Extent of exposure will be summarized for each treatment arm based on SAS. The following variables will be summarized:

- Duration of treatment (in months): (Date of the last dose – Date of the first dose + 1)/30.4375
    - For subjects who are still on treatment at a data cut-off date, the data cut-off date will be the date of the last dose.
    - Duration of treatment will be summarized as a continuous as well as a discrete variable (≥1 month, ≥3 months, ≥6 months, ≥9 months, ≥12 months)
- Actual Dose Intensity (mg/day): Actual Total Dose/ Duration of Treatment (days)
- Relative Dose Intensity (%): Actual Dose Intensity/Planned Dose Intensity, where the planned total daily dose is the dose assigned at the study entry.
- Dose modifications (dose decreased, dose held) and the reasons for dose modifications.

In addition, a separate table for placebo subjects who crossed-over and received AG-120 will be provided by using the same analysis method.

8. EFFICACY ANALYSIS

For PFS, best overall response (BOR), ORR, DOR and other endpoints based on progression and/or response status, the primary analyses will be based on the independent central review by IRC. The endpoints based on Investigator assessment will be used as secondary analyses. Efficacy data after crossover will be analyzed separately.

After unblinding at local PD per Investigator assessment, subjects on AG-120 are allowed to stay on AG-120 if the treating Investigator deems they are clinically benefiting. The IRC will read all scans up to the discontinuation of AG-120. Subjects on placebo will no longer stay on placebo after radiographic disease progression and unblinding, hence no further scans will be read by IRC beyond the initial locally assessed disease progression imaging time point. Because of that, there is a potential systematic bias that subjects on the AG-120 arm could have PD called by IRC later than the locally assessed PD date; however, subjects on the placebo arm would systematically be censored by IRC analysis at the time of local PD if not concordant with local PD. To alleviate the potential imbalance between two arms, IRC assessments after the local PD from both arms will be excluded from the primary analysis of scan-related endpoints per IRC read. A sensitivity analysis by including all the scans read by IRC will be conducted.

All efficacy analyses will be based on the ITT set unless stated otherwise. If analyses are performed on more than one analysis set, the analyses on the ITT set will be considered primary. For each period, once the PD occurred, or the new anticancer therapy was initiated, all the subsequent tumor/response assessments will not be included in the tumor/response related summary analyses, but it will be included in the data listings.

Data after crossover (such as PFS) will be analyzed in a similar manner.

Listings for all efficacy endpoints will be provided.

8.1. Progression-Free Survival (PFS)

8.1.1. Primary Analyses of PFS

The primary analysis of PFS will be based on IRC assessment for ITT set.

Progression-free survival (PFS) is defined as the time in months from the randomization date to the date of the first documentation of disease progression as determined by the IRC per RECIST v1.1 or death due to any cause, whichever occurs first.

PFS = (Earliest Date of Disease Progression or Death – Randomization Date + 1) / 30.4375.

The approach regarding handling of missing response assessments and censoring is presented in Table 1: Handling of Missing Response Assessment and Censoring for the Primary Analysis of PFS.

A stratified log-rank test (one-sided) will be used to compare PFS of the AG-120 arm against the placebo arm at the time when 131 investigator-assessed events have occurred, with the one-sided significance level controlled at 0.025. The HR (AG-120/placebo) and the corresponding 2-sided 95% confidence interval (CI) will be estimated using a stratified Cox regression model. For both the stratified log-rank test and stratified Cox regression model, the strata will be those used for stratified randomization, that is, the number of prior lines of therapy at randomization.

Number of subjects with events, types of events (progression or death), and number of subjects censored, number of subjects for each reason of censoring, Kaplan-Meier estimates, and 95% CIs for the 25th percentile, median, and the 75th percentile for PFS will be presented by treatment group. Probabilities of event free at selected time points, such as 3-month, 6-month, 9-month, and 12-month, will be presented by treatment arms. Kaplan-Meier curves of PFS will be provided for each treatment arm with the number of subjects at risk over time included.

8.1.2. Additional Analyses of PFS

The following analyses of PFS will be performed by treatment arms:

For IRC assessments:

- Stratified log-rank test and Cox regression based on all scans read by IRC prior to crossover, including the scans after local PD from subjects who were originally assigned to AG-120 and continued to stay after unblinding.
- Stratified log-rank test and Cox regression based on PPS set.
- Unstratified log-rank test (one-sided) to compare the two treatment arms. An unstratified Cox regression will be used to estimate the HR (AG-120/placebo) with its 95% CI.
- Subgroup analysis with unstratified log-rank test and unstratified Cox regression model. The HR (AG-120/placebo) with its 95% CI will be displayed for all subgroups graphically in a forest plot. The subgroups will include:
    - The actual number of prior line of therapies in advanced setting (1 vs. ≥2)
    - Gender (female vs. male)
    - Extent of disease at screening (locally advanced vs. metastatic)
        - If subject had both local and metastatic, it would be considered as metastatic
    - Intrahepatic vs. extrahepatic
        - Perihilar will be included in extrahepatic category
    - ECOG at baseline (0 vs. >=1)
    - Regions (North America vs. Europe vs. Asia)

In addition, the following exploratory analyses may be performed:

- PFS analysis by excluding subjects who progressed early (≤47 days within randomization date, i.e., 6 weeks + 5 days of scan visit window).
- The Cox regression model by including treatment, sum of target lesions at baseline, and other possibly related parameters in covariates.

For Investigator assessments:

- Stratified log-rank test (one-sided) will be utilized to compare the two treatment arms. A stratified Cox regression model will be used to estimate the HR (AG-120/placebo)

with its 95% CI. The same censoring rule as described for the primary PFS based on IRC assessment will be used.

In addition, the concordance of PFS between investigator and IRC assessment will also be summarized by treatment arms and overall.

Table 1: Handling of Missing Response Assessment and Censoring for the Primary Analysis of PFS

Situation	Date of Censoring
No baseline assessment and no death.	Date of the randomization
Alternate anticancer systemic treatment started before documented progression (PD) or death.	Date of last adequate assessment prior to the start of anticancer treatment[1]
No adequate post-baseline assessment and no death.	Date of the randomization
No documented PD or death before data cutoff date.	Date of last adequate assessment[1]
Documented PD or death following a long gap from the previous adequate assessment (eg, 2 or more consecutive missed scheduled disease status assessments). The long gap is defined as ≥95 days (ie, 12 weeks + 10 days per the protocol defined visit window). If no adequate assessment prior to minimum of (PD, death), the long gap is calculated from the randomization date.	Date of last adequate assessment prior to the first occurrence of 2 or more consecutive missing scheduled assessments

[1] Adequate disease assessment is defined as a response assessment other than "not assessed" or "not evaluable." If there is no adequate response assessment prior to the start of anticancer treatment, it will be censored at the randomization date.

8.2. Objective Response Rate (ORR)

Objective response rate will be derived from BOR. Best overall response is defined as the best time point response that a subject achieves during the course of the study, with the response ranked according to the following order (from best to worst): CR>PR>SD>PD>UNK>Other (including Not Evaluable and Not Assessed). The number and proportion of ITT subjects with each category of BOR will be presented by treatment arms.

Per RECIST v1.1, SD that occurred within <38 days from the randomization date is assigned as UNK (6 weeks minus 5 days per protocol allowed visit window). In addition, the BOR of CR or PR requires a confirmation scan. The specific confirmation rule can be found in Appendix A: Best Overall Response Confirmation Rule.

The estimated ORR (percent of subjects with a BOR of CR or PR) and a 2-sided 95% CI (via exact binomial) will be provided by treatment arms. All subjects in the analysis set will be included in the denominator in the calculation of the percentage for each response category or ORR. If the subject has no response assessment, it will be treated as a non-responder in the analyses.

Objective response rate will be analyzed using one-sided Cochran-Mantel-Haenszel (CMH) test (see Appendix B: One-Sided Cochran–Mantel–Haenszel (CMH) Test) to compare the two treatment arms. The odds ratio and its 95% CI will be estimated. The strata will be those used for stratified randomization. If the number of responders in each arm is low (n<5), Fisher exact test will be used instead. The odds ratio (unstratified) and its 95% exact CI will be estimated.

In addition, the ORR (and BOR) without confirmation per IRC as well as per Investigator assessments will be analyzed similarly.

The waterfall plots of the maximum % reduction from baseline in target lesion measurement (sum of diameters) will be presented by treatment arms. The target lesions post-baseline vs. baseline have to be matched in order to calculate % change, except for the case where lesions disappeared (*Absent* or *Too Small to Measure*) or merged or split. If the lesion status is "Absent," 0 mm will be used as the measurement; if the lesion status is "too small to measure," 5 mm will be used per RECIST v1.1 guidance.

The swim lane plot of treatment duration per subject will be presented by treatment arms. Treatment ongoing status will be marked at the end of the lane. The BOR will be marked in colors. The swim lane plot will be based on ITT set.

8.3. Overall Survival (OS)

The OS analysis will be based on ITT set and will include all OS data, including data after crossover.

Overall survival is defined as the time in months from the randomization date to the date of death due to any cause. Subjects without documentation of death at the time of the data cutoff for analysis will be censored at the date the subject was last known to be alive, or the data cutoff date, whichever is earlier. The last known alive date is the last record in the study database. For example, this date may be the maximum of the last visit date or last contact date, including telephone follow-up where the subject is known to be alive.

As a primary analysis, a stratified one-sided log-rank test will be used to compare OS between the two treatment groups, with the one-sided significance level controlled at 0.025. The HR along with the 95% CI will be estimated using a stratified Cox model. For both the stratified log-rank test and the stratified Cox model, the strata will be those at randomization. A Kaplan-Meier plot for OS will be presented by treatment arm. Estimates and 95% CIs for the 25th percentile, median, and 75th percentile for OS will be presented by treatment arm (if estimable). Probabilities of survival at selected time points (3 months, 6 months, 9 months, and 12 months) may also be presented.

In addition, the following subgroup analyses will be performed:

- Subgroup analysis with unstratified log-rank test and unstratified Cox regression model. The HR (AG-120/placebo) with its 95% CI will be displayed for all subgroups graphically in a forest plot. The subgroups will include:
    - The actual number of prior line of therapies in advanced setting (1 vs. ≥2)
    - Gender (female vs. male)
    - Extent of disease at screening (locally advanced vs. metastatic)

        - If subject had both locally advanced and metastatic, it would be considered as metastatic
    - Intrahepatic vs. extrahepatic
        - Perihilar will be lumped into extrahepatic category
    - ECOG at baseline (0 vs. >=1)
    - Regions (North America vs. Europe vs. Asia)

Per the protocol, placebo subjects are allowed to cross over to AG-120 upon the radiographic disease progression provided the eligibility criteria continue being met. To adjust for the crossover effect from placebo to AG-120 on OS, advanced modeling methods, such as rank preserving structural failure time (RPSFT) method, will be explored. More details of this method will be in the Appendix C: Overall Survival Analyses using Rank Preserving Structural Failure Time Model (RPSFT).

8.4. Multiple Comparisons/Multiplicity Adjustment

Of the secondary endpoints, OS and ORR are designated as key secondary efficacy endpoints. The primary and key secondary endpoints will be tested at an overall one-sided Type I error rate at 2.5% level based on the fixed sequence testing procedure at the time of primary analysis. These endpoints will be tested in the following order:

- PFS based on IRC
- OS
- ORR based on IRC

In addition, a hierarchical testing procedure will be adopted for OS analyses only if the primary efficacy endpoint PFS is statistically significant. Two analyses are planned for OS: 1) an interim analysis at the projected time of the final analysis for PFS (provided PFS is significant); 2) a final analysis for OS when 150 deaths are observed. Overall survival at the interim will be tested with the alpha being determined using the gamma spending function (gamma=-8), and the overall Type I error rate will be controlled at the 1-sided 0.025 level. The log-rank test stratified by randomization stratification factor will be used to compare OS between the two treatment arms.

8.5. Time To Response (TTR)

Time to response is defined as the time (in months) from the randomization date to the date of first occurrence of confirmed/unconfirmed response per RECIST v1.1.

8.6. Duration of Response (DOR)

Among responders (subjects who have a best response of confirmed PR or CR), DOR in months will be calculated as the date of the first confirmed PR or CR to the date of the first PD or death, whichever is earlier. Duration of response = (Earliest Date of Death/Progression – Date of First Confirmed PR/CR + 1)/30.4375. The censoring rule will be the same as that for the PFS analysis

in Table 1: Handling of Missing Response Assessment and Censoring for the Primary Analysis of PFS.

For both TTR and DOR, a listing will be provided for responders only.

8.7. Health-Related Quality of Life (HRQOL)

In general, descriptive statistics will be used to summarize the individual items and sub-scale scores of HRQOL (ie, QLQ-C30 and QLQ-BIL21), anchor-based questions (ie, PGI-S, PGI-C) and the EQ-5D-5L data at each scheduled assessment time point. Data listings will also be provided by each endpoint and visit for each subject.

Upon receipt of FDA's feedback in May 2017, anchor questions (PGI-S and PGI-C) were added and HRQOL assessment frequency was increased from every 6 weeks prior to 1 year/every 8 weeks thereafter (to tie with scan schedule) to every cycle (every 4 weeks). The changes were implemented in Protocol Amendment 3.0, Version 4, published in September 2017.

8.7.1. Brief Descriptions of Questionnaires

QLQ-C30 Questionnaire

The QLQ-C30 contains 30 items in total and each item is a 4-point or 7-point Likert scale. These 30 items can be categorized into a global health status/HRQOL scale, 5 functional scales, 3 symptom scales, and 6 single-item scales. Each of the multi-item scales includes a different set of items. No item occurs in more than one scale:

- 1 global health status scale (2 items)
- 5 functional scales: physical function (5 items), role function (2 items), emotional function (4 items), cognitive function (2 items), social function (2 items)
- 3 symptom scales: fatigue (3 items), nausea and vomiting (2 items), pain (2 items)
- 6 single-item assessments relating to dyspnea, insomnia, appetite loss, constipation, diarrhea, and financial difficulties.

QLQ-BIL21 Questionnaire

The QLQ-BIL21 contains 21 items in total and each item is a 4-point Likert scale. These 21 items can be grouped into 5 scales and 3 single items:

- 5 scales (18 items): eating symptoms (4 items), jaundice symptoms (3 items), tiredness (3 items), pain symptoms (4 items), anxiety symptoms (4 items)
- 3 single-item assessments related to treatment side-effects, difficulties with drainage bag/tubes, concerns regarding weight loss

PGI-S

The anchor-based questionnaire PGI-S contains the following 3 items:

- The severity of the physical functioning decline over the past week
- The severity of the appetite decrease over the past week
- The severity of the pain over the past week

The possible outcomes include:

- None
- Mild
- Moderate
- Severe
- Very Severe

PGI-C

The anchor-based questionnaire PGI-C contains the following 3 items:

- The overall change in the physical functioning since the start of taking the study medication
- The overall change in the appetite since the start of taking the study medication
- The overall change in the pain since the start of taking the study medication

The possible outcome includes:

- Very much better
- Moderately better
- A little better
- No change
- A little worse
- Moderately worse
- Very much worse

EQ-5D-5L Questionnaires

The EQ-5D-5L contains 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression) as well as the EQ visual analogue scale (VAS). Five dimensions return categorical outcomes, and the VAS is a 0-100 number.

8.7.2. Statistical Analyses

8.7.2.1. QLQ-C30 and QLQ-BIL21 Analyses

Each scale will be transformed into a 0-100 score following EORTC guidelines:

- Estimate the average of the items that contribute to the scale; this is the *raw scale score*
- Use a linear transformation to standardize the *raw scale score,* so that it ranges from 0 to 100. *Scale score = (raw score – 1)/range *100*. The only exception is Physical Function in QLQ-C30, where the scale score per EORTC QLQ-C30 manual is *[1-(raw score – 1)/range]*100.* Here, range is the difference between the possible

maximum and minimum response to each individual item. For instance, if the item is scored from 1 to 4, then range=3.

Transformed scores for each scale and the absolute change from baseline will be summarized by treatment and scheduled visit. A mixed-effect model with repeated measurements on the change from baseline across visits for key *scale scores* (such as pain, appetite loss, and physical function) over the course of before crossover period may be explored with baseline score, treatment, visit, treatment-by-visit as fixed effects, and subject as random effects.

8.7.2.2. Anchor Based Approach to Find Clinically Meaningful Change

Estimation of meaningful change is achieved by computing the change scores for the EORTC-QLQ-C30 and EORTC-QLQ-BIL21 scores between baseline and follow-up associated with minimal improvement groups defined on PGI-C and change from baseline to post-baseline PGI-S (ΔPGI-S) ratings as anchors. This approach will be taken for only before-crossover data. Minimally sufficient improvement is directly defined from the PGI-C by analyzing those subjects who report a minimal level of improvement on the anchors. Minimally sufficient improvement for the PGI-S is operationalized by taking the difference in PGI-S between baseline and follow-up and selecting participants with a one-point improvement. Average EORTC-QLQ-C30 and EORTC-QLQ-BIL21 scores will be computed for each of the anchor-based minimally sufficient improvement group definitions.

Relevant subscales of the EORTC-QLQ-C30 and EORTC-QLQ-BIL21 will be paired with their most-appropriate PGI-C/PGI-S rating (eg, the physical functioning PGI-C/S will be employed for the EORTC-QLQ-C30 physical function subscale). Similarly, the EORTC-QLQ-C30 appetite item and EORTC-QLQ-BIL21 eating symptoms subscale will be paired with the appetite PGI-C/S. Finally, the EORTC-QLQ-C30 and EORTC-QLQ-BIL21 pain subscales will each be paired with the PGI-C/S as anchors.

EORTC-QLQ-C30 and EORTC-QLQ-BIL21 change scores will be plotted as cumulative distribution functions (CDFs) for change. Cumulative distribution functions will be stratified on their respective anchor variables (PGI-C and ΔPGI-S). In addition, after unblinding, the CDFs will be stratified by treatment-arm.

8.7.2.3. EQ-5D-5L Analysis

The EQ-5D contains a descriptive system with one response for each of 5 dimensions (Mobility, Self-Care, Usual Activities, Pain/Discomfort, and Anxiety/Depression). The first response is coded as 1 (indicating no problems), the second response is coded as a 2 (indicating slight problems), the third response is coded as a 3 (indicating moderate problems), the forth response is coded as a 4 (indicating severe problems), and the fifth response is coded as a 5 (indicating unable to function). Ambiguous responses (eg, more than one response in a dimension) are treated as missing values. For each dimension, the number and percentage of subjects in each category will be summarized by treatment and visits.

The EQ-5D-5L also contains a utility scale of health state (VAS) ranging between 0 (worst health) and 100 (best health). Summary statistics for VAS scores and the absolute change from baseline will be reported by treatment and visits.

8.8. Post Crossover

Upon the radiographic progression of disease, subjects will be unblinded. If it's determined that they received placebo, subjects will be allowed to cross over to AG-120, provided they continue to be eligible for the study. Data after crossover will be analyzed separately, but in a similar fashion as outlined above.

In particular, tumor assessments after crossover will be based on Investigator assessment. Tumor measurements will be re-baselined prior to the start of crossover. The corresponding endpoints include:

- PFS per Investigator assessment (PFS2)
- Objective response rate /best overall response per Investigator assessment. Note that SD<38 days of the first AG-120 dose date will be assigned as UNK.
- Time To Response and DOR per Investigator assessment

9. SAFETY ANALYSIS

Unless specified otherwise, the safety data will be summarized by treatment based on SAS. Specifically, it will include placebo before crossover, AG-120 as dosed, AG-120 after crossover, and total AG-120. The total AG-120 includes all the subjects who have ever been dosed with AG-120 (that is, the sum of AG-120 as dosed and AG-120 after crossover). If a subject receives at least one dose of AG-120, the actual arm will be AG-120 as dosed. The same rule applies to the after crossover.

9.1. Adverse Events

Treatment emergent adverse events (TEAEs) are defined as any AEs that begin or worsen on or after the start of study drug through 28 days after the last dose of study drug. All AEs will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) dictionary, unless otherwise specified. The severity will be graded based on the NCI CTCAE. All AEs will be listed. Only TEAEs will be summarized and will be referred to as AEs hereafter.

If a subject experiences multiple AEs under the same PT within an SOC, then the subject will be counted only once for that PT within that SOC. If a subject experiences the same AE more than once with different intensity or grade, then the event with the highest grade will be tabulated in "by grade" tables.

Tables summarizing the incidence of AEs will be generated by treatment arms for each of the following:

- Overall summary of AEs
- All AEs by SOC and PT
- Most common AEs by PT (ie, those events reported by ≥5% of subjects in any treatment group)
- Most common Treatment-Related AEs by PT (ie, those events reported by ≥5% of subjects in any treatment group)
- Grade 3 or higher AEs by SOC and PT
- Most common Grade 3 or higher AEs by PT (ie, those Grade 3 or higher AEs reported by ≥5% of subjects in any treatment group)
- Related AEs by SOC and PT
- Related Grade 3 or higher AEs by SOC and PT.
- SAEs by SOC and PT
- Related SAEs by SOC and PT
- AEs leading to study drug discontinuation by SOC and PT
- AEs leading to study drug reduced by SOC and PT
- AEs leading to study drug held by SOC and PT
- Related AEs leading to study drug discontinuation by SOC and PT

- Related AEs leading to study drug reduced by SOC and PT
- Related AEs leading to study drug held by SOC and PT
- On-treatment death due to AE by SOC and PT, where on-treatment is defined as within 28 days of last dose.
- On-treatment death due to related AE by SOC and PT, where on-treatment is defined as within 28 days of last dose.

All AEs will be listed by subject. By-subject listings will be provided for AEs leading to on-treatment death, SAEs, and AEs leading to discontinuation of treatment.

9.2. Adverse Events of Special Interest

QT prolongation AEs will be summarized for the following:

a) QT prolongation all grades by PT

b) QT prolongation Grade 3 or higher by PT

c) Treatment-related QT prolongation all grades by PT

Time to first QT prolongation AE will be summarized continuously and by ≤15 days, 16-30, 31-45, 46-60, and >60 days for all grade and Grade 2 or higher AEs.

9.3. On-Treatment Death

All on-treatment deaths (≤28 days of last dose) and the reason of deaths will be summarized by treatment based on SAS set.

In addition, all-cause mortality will be summarized including the deaths ≤30 days of the first and last dose, and ≤60 days of the first and last dose.

9.4. Laboratory Data

Unless specified otherwise, lab data will be summarized by treatment and scheduled visits. Data after crossover will be summarized by scheduled visits in a separate table or listing. The baseline used in the "after crossover" outputs will be the closest assessment prior to subjects receive open-label AG-120.

9.4.1. Hematology

For hematologic parameters, the actual values and the change from baseline will be summarized by treatment and scheduled visits.

Shift tables will be presented based on CTCAE grade by treatment, including

- WBC (low)
- Absolute Neutrophil Count (low)
- Hgb (low)
- Platelets (low)

9.4.2. Clinical Chemistry

For chemistry parameters, the actual values and the change from baseline will be summarized by treatment and scheduled visits.

Shift tables from baseline to worst value based on CTCAE grade will be presented by treatment, including

- Bilirubin (high)
- Alkaline Phosphatase (high)
- ALT (high)
- AST (high)
- Sodium (low)
- Potassium (low/high)
- Phosphate (low)

In addition, subjects with liver dysfunction will be presented in a listing. Liver dysfunction is defined as elevated ALT or AST of ≥ 3x ULN and associated with increase in total bilirubin ≥2x ULN (±10 days).

9.4.3. CA19-9 Level

The actual values and the change from baseline for CA19-9 level will be summarized by treatment arm and scheduled visits. A corresponding data listing will be provided.

9.4.4. Coagulation Analysis

The following coagulation tests will be presented in a listing: prothrombin time (PT), activated partial thromboplastin time (aPTT), and international normalized ratio (INR).

9.4.5. Urinalysis

Urinalysis results will be listed.

9.4.6. Pregnancy Tests

The pregnancy test results will be listed.

9.5. Physical Examination

A by-subject listing will be presented for any clinically significant findings from physical examination.

9.6. Vital Signs

Vital signs (except height) will be presented as both actual values and changes from baseline by treatment and scheduled visits. Vital sign measurements will also be listed.

9.7. Electrocardiograms (ECGs)

Descriptive statistics for the actual values and changes from baseline over time will be summarized for each ECG parameter: QT (msec), RR interval (msec) and QTcF (Fridericia's correction). QTcF is calculated as QTcF = QT/cube root of RR interval.

The proportion of subjects with maximum post-baseline absolute QTcF intervals that fall into following categories will be presented:

- ≤450 msec
- >450 and ≤480 msec
- >480 to ≤500 msec
- >500 msec

The proportion of subjects who have a maximum post-baseline change from baseline in QTcF intervals of the following categories will be presented:

- ≤0 msec
- >0 to ≤30 msec
- >30 to ≤60 msec
- >60 msec

A listing of subject-level ECG measurement will be provided.

9.8. Left Ventricular Ejection Fraction (LVEF)

The LVEF data will be listed.

9.9. ECOG Performance Status (PS)

The ECOG PS will be summarized by treatment arms and visits, and by shift tables from baseline to worst values across all visits by treatment arms.

A by-subject listing will be provided.

9.10. Interim Analysis

There will be no formal interim analyses of the data. Interim safety reviews will be conducted by an IDMC on a regular basis. Details are specified in the protocol and IDMC charter.

10. GENERAL METHODS

Summary statistics will be presented by treatment and scheduled visit, unless stated otherwise.

Unless otherwise specified, descriptive statistics for continuous data will include the number of subjects with data to be summarized (n), mean, standard deviation, median, minimum, and maximum.

Descriptive statistics for categorical/qualitative data will include frequency counts and percentages. The total number of subjects in the treatment arm will be used as the denominator for percent calculations, unless stated otherwise.

Descriptive statistics associated with time-to-event analyses will include the number of events, the number of subjects censored, 25% quartile, median, 75% quartile, and 95% CI for median. These statistics will be presented for all time-to-event analyses, unless stated otherwise.

Listings will be provided for selected endpoints. Listings will be sorted by site, treatment, and subject IDs.

10.1. General Rules

All data listings that contain an evaluation date will contain a relative study day (Rel Day).

Study days are numbered relative to the day of the first dose of study medication (designated as Day 1) for safety related outputs, and relative to the randomization date for the efficacy related outputs. The preceding day is Day −1, the day before that is Day −2, etc.

Other conventions include the following:

- Data from all study centers will be combined in the analysis, unless specified otherwise.
- Confidence intervals, if presented, will be 2-sided 95% CIs unless otherwise specified.
- All analyses and summary tables will have the analysis population sample size for each treatment group (treatment arm, overall, etc.) in the column heading (ie, number of subjects).

10.2. Unscheduled Visits and Visit Windows

Unscheduled visits will not be mapped and will not be included in the summary statistics per visit tables. However, they will be included in the data listing as well as the shift tables.

10.3. Baseline Values

Unless otherwise specified, the baseline value is defined as the value collected at the time closest to, and prior to, the start of study drug administration. In case subjects are not dosed, the latest assessment will be considered as baseline. Values collected at unscheduled visits prior to the start of the study drug administration will be included in the calculation of baseline values. In majority of cases, C1D1 assessment should be treated as the baseline as the study protocol requires assessment to be done pre-dose at C1D1. For visits after crossover, the closest

assessment prior to the start of AG-120 will be considered as the baseline for crossover data analysis.

10.4. Computing and Coding Standards

Activities will be performed using the following tools:

Table, listing, and figure production	SAS Version 9.2 or higher
Coding	
Adverse Events	MedDRA Version 21.1 or higher
Medical Histories	MedDRA Version 21.1 or higher
Prior and Concomitant Medications	WHODrug Version 2018 Q1 or higher
Grading	
AEs	CTCAE Version 4.03
Labs	CTCAE Version 4.03

10.5. Missing/Partial Dates in Adverse Events and Concomitant Medications/Therapies

Missing/Partial Start Dates:

If the stop date is non-missing and the imputed start date is greater than the stop date, the stop 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 first dose date, the first dose date will be used
- If the month and year are before the month and year of the first dose date, the last day of the month will be assigned to the missing day
- If the month and year are after the month and year of the first dose date, the first day of the month will be assigned to the missing day

2) Missing day and month

- If the year is the same as the year of the first dose date, the first dose date will be used
- If the year is prior to the year of the first dose date, December 31 will be assigned to the missing fields
- If the year is after the year of the first dose date, January 1st will be assigned to the missing fields

3) Missing day, month, and year

- The first dose date will be used

If the first dose date is missing, no imputation will be performed.

Missing/Partial Stop Dates:

1) Missing day only
   - The last day of the month will be assigned as the missing day
2) Missing day and month
   - December 31 will be assigned to the missing fields
3) Missing day, month, and year
   - The event will be regarded as ongoing

All the imputed dates will be compared against the End of Study date, death date, and data cutoff date. The earliest date will be chosen.

10.6. Missing/Partial Dates in On-study Anticancer Therapies

On-study anticancer therapy with start dates that are completely or partially missing will be imputed as follows:

1. If the start date has month and year but day is missing, the 15 of the month will be imputed.
2. If the start date has year, but day and month are missing, July 1 will be imputed.

On-study anticancer therapy with stop dates that are completely or partially missing will be imputed as follows:

1. If the stop date has month and year but day is missing, the last day of the month will be assigned.
2. If the stop date has year, but day and month are missing, December 31 will be assigned.
3. If the stop date misses day, month and year, the event will be considered as ongoing.

If the imputed date is before the last dosing date, the last dosing date will be used as the on-study anticancer therapy start date. If the imputed date is after the death date then the death date will be used. If the imputed stop date is before start date, the stop date will be used as the start date.

This imputation rule will only be used in efficacy endpoints derivation (such as PFS, OS). The listing of post-study anticancer therapy will not apply imputation rules.

10.7. Missing/Partial Dates at Screening Visits

The following rules apply to dates recorded during the screening visits (eg, prior anticancer therapy).

1) Missing day only
   - 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 earliest inform consent date at screening, the 1st of January will be used.
   - If the year is before the year of the earliest inform consent date at screening, the 15th of June will be used.
3) Missing day, month and year
   - No imputation will be applied.

If the imputed end date is before the start date, the start date will be used as the end date. If the imputed dates are after the inform consent date at screening, then the inform consent date will be chosen.

11. CHANGES TO ANALYSES SPECIFIED IN THE PROTOCOL

Below is a summary of changes to the analyses specified in the protocol:

- Pharmacokinetic Analysis Set (PAS) as well as PK-related analyses are not described here and will be in a stand-alone PK/pharmacodynamic report.

12. APPENDICES

12.1. Appendix A: Best Overall Response Confirmation Rule

CR CONFIRMATION:

To confirm a CR (the first CR), another CR (the confirming CR) at least 4 weeks after the first CR will be looked for. It is required that there is no other overall response, except CR or NE or NA, between the first CR and the confirming CR.

- If such confirming CR exists, then the subject BOR is confirmed CR.
- If such confirming CR is guaranteed non-existing (eg, the first CR followed by PD or off-study), then the confirmed BOR is SD if the first CR was assessed ≥38 days after the randomization date, otherwise the first CR will be replaced by UNK and the BOR will be derived subsequently.

PR CONFIRMATION:

To confirm a PR (the first PR), another CR or PR at least 4 weeks after the first PR will be looked for. It is required that there is no other overall response, except CR or PR or NE or NA, between the first PR and the confirming CR or PR.

- If such confirming CR or PR exists, then the subject's BOR is PR (confirmed).
- If such confirming CR or PR is guaranteed non-existing (eg, the first PR followed by PD or off-study), then the confirmed BOR is SD if the first PR was assessed ≥ 38 days after the randomization date, otherwise the first PR will be replaced by UNK and the BOR will be derived subsequently.

12.2. Appendix B: One-Sided Cochran–Mantel–Haenszel (CMH) Test

The Cochran–Mantel–Haenszel (CMH) test compares binary responses of two treatment groups, adjusting for stratification factors. In the CMH test, the data are arranged in a series of associated

2 × 2 contingency tables, the null hypothesis is that the observed response is independent of the treatment used in any 2 × 2 contingency table.

Let O_{hij} be the observed frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h and E_{hij} be the expected frequency of treatment i (i=1 or 2) with outcome j (j= 1 or 2) in stratum h, where

$$E_{hij} = \frac{(O_{hi1} + O_{hi2}) \cdot (O_{h1j} + O_{h2j})}{O_{h11} + O_{h12} + O_{h21} + O_{h22}}$$

Also, let P_{hij} be the probability of outcome being j given treatment being i (i=1 or 2, j= 1 or 2) in stratum h. To test the following hypothesis,

$$H_0 : P_{h11} = P_{h21} \text{ for all } h \in \{1, \ldots, H\}$$

versus

$$H_1 : P_{h11} > P_{h21} \text{ for at least one } h \in \{1, \ldots, H\}$$

The one-sided Cochran-Mantel-Haenszel Test Statistics is constructed as:

$$Z_{CMH} = \frac{\sum_{h=1}^{H} (O_{h11} - E_{h11})}{\sqrt{V_{11}}} \sim Z$$

Where

$$V_{11} = Var(\sum_{h=1}^{H} (O_{h11} - E_{h11})) = \sum_{h=1}^{H} \frac{E_{h11} \bullet E_{h22}}{O_{h11} + O_{h12} + O_{h21} + O_{h22} - 1}$$

The test statistic Z_{CMH} will be compared with standard normal distribution to obtain p-value of the CMH test.

12.3. Appendix C: Overall Survival Analyses using Rank Preserving Structural Failure Time Model (RPSFT)

Subjects who received placebo per randomization is allowed to crossover and receive AG-120 upon radiographic PD. To account for the potential bias introduced by crossover, the RPSFT model (Robins and Tsiatis, 1991; White et al, 1997 & 1999) is utilized as a sensitivity analysis.

The RPSFT model and assumptions

RPSFT assumes that the AG-120 after the switch is acting by multiplying survival time by a given factor (acceleration factor) relative to placebo, and assumes the treatment effect is the same for all subjects regardless of when treatment is received (common treatment effect).

Specifically, let U_i denote the latent survival time if subject i were assigned to the placebo arm, adhere to it and discontinue only after the event (also called counter-factual event time),

$$U_i = T_i^{off} + T_i^{on} \exp(\psi_0)$$

where T_i^{off} is the time that subject i is off treatment, and T_i^{on} is the time that subject i is on treatment; $\exp(\psi_0)$ is the acceleration factor which denotes the amount by which a subject′s survival time is 'increased' by the active treatment. A positive (negative) ψ_0 value corresponds to a harmful (beneficial) treatment effect. Specifically, for

- AG-120 subjects at randomization: $U_i(\psi_0) = T_i^{ag120} \exp(\psi_0)$;
- placebo subjects who crossed over to AG-120: $U_i(\psi_0) = T_i^{pbo} + T_i^{ag120} \exp(\psi_0)$;
- placebo subjects without crossover: $U_i(\psi_0) = T_i^{pbo}$.

In order to estimate ψ_0, we assume that U_i is independent of randomized treatment assignment and can be viewed as baseline characteristics. Thus, if we conduct a hypothesis test (such as log-rank test) for the treatment difference on $U_i(\psi_0)$, we shall obtain a p-value close to 1 with a sufficiently large sample size. RPSFT works by reconstructing the survival time of subjects, as if they have never received active treatment. A grid search within a reasonable range will then be performed in order to find the estimated ψ_0 with the largest p-value. The corresponding point estimate of HR between the two arms will be reported, with the 95% CI generated from bootstrapping method.

Re-censoring

Administrative censoring refers to the censoring where the event is not observed by the time of data cutoff. Unfortunately its time scale cannot be adjusted in the same way as event, as potential bias could be introduced because censoring would be dependent on time spent on treatment and thus treatment arm (informative censoring). To overcome this problem, the counter-factual event times are re-censored by the minimum U_i that could have been observed for individuals (with and without events) across their possible treatment changes.

Let C_i be the potential censoring time for a subject i. The subject is then re-censored at the minimum possible censoring time:

$$D_i^*(\psi_0) = \min(C_i, C_i \exp(\psi_o)).$$

If $D_i^* < U_i$, then U_i is replaced by D_i^* and the subject is censored. For treatment arm where switching didn't occur, re-censoring is not applied.