STATISTICAL ANALYSIS PLAN

STUDY TITLE:

A Phase 3 Study of BBI-608 in combination with 5-Fluorouracil, Leucovorin, Irinotecan (FOLFIRI) in Adult Patients with Previously Treated Metastatic Colorectal Cancer (CRC)

VERSION: 3.0

DATE: December 15, 2020

STUDY DRUG: BBI-608 (aka BBI608, napabucasin)

PROTOCOL NUMBER: CanStem303C (aka BB608-303CRC)

SPONSOR:

This study is being conducted in compliance with good clinical practice, including the archiving of essential documents.

SIGNATURE PAGE

STATISTICAL ANALYSIS PLAN APPROVAL

Author:

_____Date: _____/_____/_____

[redacted]

_____Date: _____/_____/_____

[redacted]

Approved by:

_____Date: _____/_____/_____

[redacted]

_____Date: _____/_____/_____

[redacted]

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Term
5-FU	5-Fluorouracil
AE	Adverse Event
ALT	Alanine Aminotransferase
AST	Aspartate Transaminase
ATC	Anatomical Therapeutic Chemical
BBI	Boston Biomedical, Inc
BID	Twice a Day
BLQ	Below the Limit of Quantitation
BMI	Body Mass Index
BOR	Best Overall Response
BP	Blood Pressure
Bpm	beats per minutes
BSA	Body Surface Area
CfB	Change from Baseline
CI	Confidence Interval
CMH	Cochran-Mantel-Haenszel
CR	Complete Response
CRC	Colorectal Cancer
CRF	Case Report Form
CT	Computerized axial Tomography
CTCAE	Common Terminology Criteria for Adverse Events
CSR	Clinical Study Report
DBL	Database Lock
DCR	Disease Control Rate
DI	The actual dose intensity
DoR	Duration of Response
DSMB	Data Safety and Monitoring Board
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form
EGFR	Epidermal Growth Factor Receptor
EORTC-QLQ-C30	European Organization for Research and Treatment of Cancer Quality of Life Questionnaire Core 30 Item
FA	Final Analysis
FDA	Food and Drug Administration
FOLFIRI	5-Fluorouracil, Leucovorin, Irinotecan
GH	Global health status
HR	Hazard Ratio
IHC	immunohistochemical
IA	Interim Analysis
ITT	Intent-To-Treat
IV	intravenous
LLQ	Lower Limit of Quantitation
MD	Measurable Disease

Abbreviation	Term
MedDRA	Medical Dictionary for Regulatory Activities
mg/m^2	milligram / square meter
mmHg	Millimeter Mercury
MRI	Magnetic Resonance Imaging
NCI	National Cancer Institute
NE	Not Evaluable
OF	O'Brien-Fleming
ORR	Objective Response Rate
OS	Overall Survival
PD	Progressive Disease
PET	Positron Emission Tomography
PF	Physical Function
PFS	Progression Free Survival
PH	Proportional Hazard
PK	Pharmacokinetic
PP	Per-Protocol
PR	Partial Response
pSTAT3	Phosphorylated Signal Transducer and Activator of Transcription 3
PT	Preferred Term
q12h	Every 12 hour
QoL	Quality of Life
RDI	Relative dose intensity (%)
RECIST	Response Evaluation Criteria in Solid Tumors
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
SOC	System Organ Class
TEAE	Treatment Emergent Adverse Event
WHO-DD	World Health Organization Drug Dictionary

1. AMENDMENT HISTORY FROM PREVIOUS VERSIONS

This Statistical Analysis Plan (SAP) is amended (version 3.0) to incorporate clarifications prior to Database Lock (DBL). The high-level summary of the changes are as below:

- Clarify the baseline window for efficacy endpoints (Section 6.3);
- Add per-protocol analysis for overall survival and progression free survival (Section 5.6);
- Add summary of duration of response and time to response (Section 3.1.2.4);
- Add clustered AE definition and summary (Section 7.2.6.1);
- Update lab parameters to be summarized by CTCAE grading (Section 7.2.6.2);
- Update the definition for extent of exposure, dose intensity and dose changes (Section 7.2.1.6);
- Add sensitivity analysis for progression free survival, disease control rate and objective response rate for general population and pSTAT3(+) subpopulation (Section 7.2.4);
- A dedicated PRO analysis will be conducted in a separate standalone SAP. Part of QoL analysis has been removed from this SAP;
- Clarify combination functions approach for ORR/DCR endpoints (Section 4.5.2 and Appendix VI);
- Update company name to Sumitomo Dainippon Pharma Oncology, Inc. (SDPO).

The SAP has been amended (version 2.1) to incorporate FDA's feedback on the SAP version 2.0 per their communication on July 23, 2019 and also to update the statistical analysis for pSTAT3(+) subpopulation considering that the final pSTAT3 assay cut-section stability (CSS) window could be longer than 6 months (up to 15 months). Among others, the main changes of this SAP amendment (version 2.1) from SAP version 2.0 include:

- To update the analysis set definitions for pSTAT3(+)/pSTAT3(-) and pSTAT3(Unknown) subpopulation (See Section 5);

- To clarify the clinical cutoff date and analysis cutoff dates for both General Population and pSTAT3(+) subpopulation at the final analysis (See Section 6.1);

- To update the statistical analysis for primary endpoint and key secondary endpoints in the pSTAT3 (+) subpopulation at the final analysis (See Section 4.5.2, Section 7.2.2 , Section 7.2.3 , and Section 7.2.4);

- To introduce a small penalty for the interim analysis (IA) (1-sided alpha of 0.0001) and make appropriate multiplicity adjustment at the final analysis (FA) (See Section 4.5.2) per FDA's feedback;

- To pre-specify the clinically relevant baseline factors for the propensity score analysis and missing biomarker data imputation (See Section 7.2.9) per FDA's feedback;

- To include the sample SAS code for propensity scores and multiple imputation (See Appendix V) per FDA's feedback;

- To update the details of the tipping point analysis to evaluate the impact of missing biomarker data under worst case scenarios. (See Section 7.2.10) per FDA feedback.

The SAP version 2.0 was based on Clinical Protocol Study CanStem303C_Amendment 6.0_Submitted to Agency on 26 September 2018, Serial Number 0256. The main changes in the SAP version 2.0 from previous SAP version 1.0 were to align with protocol amendment 6.0 in terms of study design and analysis strategy. The changes included:

- Addition of a primary endpoint to assess the overall survival (OS) in the pSTAT3 (+) subpopulation as well as a change to an adaptive study design to identify the patient population that is most likely to benefit from BBI-608 treatment.
- Selection of key secondary endpoints.
- Multiplicity adjustment for the analysis of several endpoints (primary endpoints and key secondary endpoints), analysis of the 2 patient populations (General Population and pSTAT3(+) Subpopulation), and data-driven design changes.
- Blinding plans which are implemented by the Sponsor and by the CRO to minimize the operational bias and to maintain the integrity of the study data for the decision making at the interim analysis.
- An update to the potential outcomes from the interim analysis as well as final analysis timing.
- An update to the sensitivity analysis for overall survival in the pSTAT3(+) subpopulation.
- An update to the sensitivity analysis for missing biomarker data.

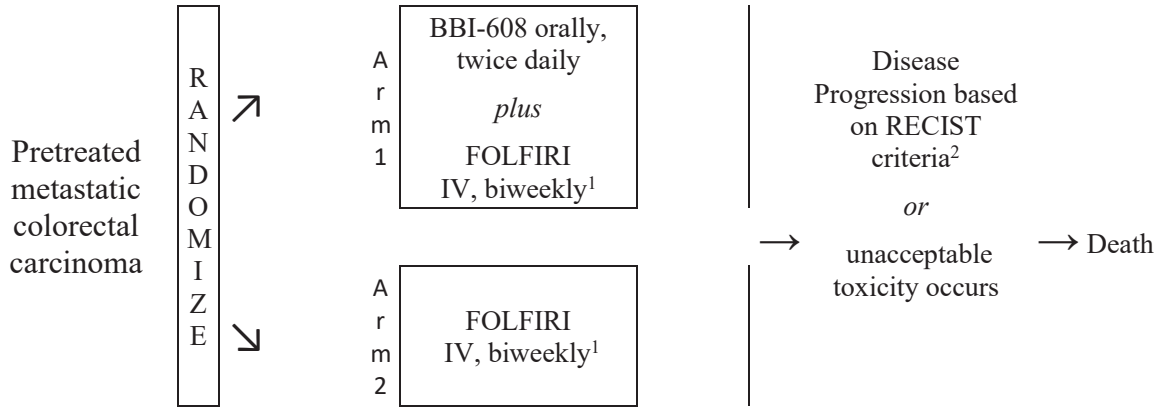
2. INTRODUCTION

This statistical analysis plan (SAP) describes the statistical methods that will be used during the analysis and reporting of data collected under Sumitomo Dainippon Pharma Oncology Protocol CanStem303C.This SAP should be read in conjunction with the study protocol and electronic case report forms (eCRFs). This version of the SAP has been developed using 226969 Mock CRF Report Form 20200731 v16.0.

2.1. Study Design

2.1.1. General Study Design and Plan

This is an international, multi-center, prospective, open-label, randomized, adaptive design Phase 3 trial of BBI-608 plus standard bi-weekly FOLFIRI (FOLinic acid-Fluorouracil-IRInotecan plus leucovorin) (Arm 1) *versus* standard bi-weekly FOLFIRI alone (Arm 2) in patients with previously treated metastatic colorectal cancer (mCRC). Addition of bevacizumab (bev) to the FOLFIRI regimen, per Investigator choice, will be permissible. The trial will enroll patients from North America, Europe, Australia, Asia, and Japan; it is anticipated that 25% to 30% of patients will be enrolled from the United States.



[1]Addition of bevacizumab to the FOLFIRI regimen, per Investigator choice, is permissible.
[2]If no other standard therapies are available at the time of disease progression, based on RECIST 1.1 criteria, and the patient has not experienced any adverse events requiring permanent discontinuation, BBI-608 may be continued as monotherapy.

2.1.2. Stratification

Patients will be stratified by:

1. Geographical region (North America/Western Europe/Australia, Japan/Korea *vs* Rest of the World)
2. Time to progression from start of first line therapy (<6 months vs. ≥6 months from start of first line therapy)
3. *Tumor RAS* status (mutant vs. wild type) [1]
4. Bevacizumab as part of study protocol treatment (yes vs. no)
5. Location of the primary tumor (left colon vs. right colon)[2]

[1]*RAS* mutant status refers to any known mutation in *KRAS* or *NRAS* mutations in exons 2, 3, and 4 resulting in decrease of response to anti-EGFR treatment.
[2]Lesions proximal to the splenic flexure will be considered right colon and lesions at splenic flexure and distal to it will be considered left colon.

2.1.3. Randomization

Patients will be randomized according to a 1:1 ratio using a permuted block randomization procedure to receive BBI-608 plus standard bi-weekly FOLFIRI or standard bi-weekly FOLFIRI alone. Addition of bevacizumab to the FOLFIRI regimen, per Investigator choice, is permissible.

Patients will be randomized to 1 of the following 2 arms:

Arm	Agent(s)	Dose and Route	Duration
1	BBI-608	240 mg orally 2 times daily[1,2]	Patients may continue to receive protocol therapy as long as they have not experienced any adverse events requiring permanent discontinuation of study medication and have not demonstrated disease progression based on RECIST 1.1 criteria. [4,5]
	FOLFIRI	Standard FOLFIRI IV, once every 2 weeks[3]	
2	FOLFIRI	Standard FOLFIRI IV, once every 2 weeks[3]	

[1] BBI-608 should be taken 1 hour before or 2 hours after a meal, 2 times daily, with approximately 12 hours between doses. BBI-608 administration will begin 2 days prior to the FOLFIRI infusion on Day 1 Cycle 1. These 2 days are referred to as Run-in Day 1 and Run-in Day 2. The run-in day period may be extended by up to 3 additional calendar days. Run-in Day 1 should occur within 2 calendar days of patient randomization.
[2] Patients should be encouraged to maintain sufficient fluid intake while on protocol treatment, such as taking BBI-608 with approximately 250 mL of fluid over the course of 30 minutes after the dose.
[3] Addition of bevacizumab to the FOLFIRI regimen, per Investigator choice, will be permissible. FOLFIRI chemotherapy infusion will start at least 2 hours following the first daily dose of BBI-608 and will be administered every 2 weeks. Irinotecan/leucovorin infusion will follow bevacizumab infusion in patients selected by the Investigator to receive standard dose of bevacizumab (5 mg/kg). Irinotecan 180 mg/m^2 together with leucovorin 400 mg/m^2 will be administered intravenously, over approximately 90 minutes and 2 hours, respectively, starting on Day 1 of Cycle 1, following bevacizumab infusion or at least 2 hours following the first dose of BBI-608 if bevacizumab is not administered. 5-FU 400 mg/m^2 bolus will be administered intravenously immediately following irinotecan/leucovorin infusion, followed by 5-FU 1200 mg/m^2/day (total 2400 mg/m^2) continuous infusion. This regimen will be repeated on Day 1 of every 14-day cycle.
[4] If any component of FOLFIRI is discontinued due to toxicity or any other reason deemed appropriate by the clinical investigators, BBI-608 should be continued until another criterion for stopping treatment is met. If BBI-608 is discontinued due to toxicity, FOLFIRI should be continued until another criterion for stopping treatment is met.
[5] If no other standard therapies are available at the time of disease progression, based on RECIST 1.1 criteria, and the patient has not experienced any adverse events requiring permanent discontinuation, BBI-608 may be continued as monotherapy.

2.2. Study Objectives

2.2.1. Primary Objectives

- To compare OS in the General Population patients treated with BBI-608 plus biweekly FOLFIRI (Arm 1) *versus* biweekly FOLFIRI (Arm 2).
- To compare OS in the activated signal transducers and activators of transcription 3 (pSTAT3)-positive (pSTAT3(+)) Subpopulation patients treated with BBI-608 plus biweekly FOLFIRI (Arm 1) *versus* biweekly FOLFIRI (Arm 2).

2.2.2. Key Secondary Objectives

- To compare progression free survival (PFS) in the General Population patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.
- To compare PFS in the pSTAT3(+) Subpopulation patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.
- To compare disease control rate (DCR) in the General Population patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.
- To compare DCR in the pSTAT3(+) Subpopulation patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.
- To compare overall response rate (ORR) in the General Population patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.
- To compare ORR in the pSTAT3(+) Subpopulation patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.

2.2.3. Other Secondary Objectives

- To compare the Quality of Life (QoL), as measured using the European Organization for Research and Treatment of Cancer Quality of Life questionnaire (EORTC-QLQ-C30), in the General Population patients treated with BBI-608 plus bi-weekly FOLFIRI *versus* bi-weekly FOLFIRI.
- To compare the QoL, as measured using the EORTC-QLQ-C30, in the pSTAT3(+) Subpopulation patients treated with BBI-608 plus biweekly FOLFIRI *versus* biweekly FOLFIRI.
- To evaluate the safety profile of BBI-608 administered daily plus biweekly FOLFIRI with safety assessed according to the National Cancer Institute Common Toxicity Criteria for Adverse Events (NCI CTCAE) version 4.0 in the General Population and the pSTAT3(+) Subpopulation.

3. ENDPOINTS AND COVARIATES: DEFINITIONS AND COVENTIONS

3.1. Efficacy Endpoints

3.1.1. Primary Endpoints: Overall Survival in the General Population and the pSTAT3(+) Subpopulation

Overall Survival (OS), the primary endpoints of this study, in the General Population and the pSTAT3(+) Subpopulation is defined as the time from randomization to death from any cause. Patients who are still alive at the time of the interim or the final analysis data cutoff, or who have become lost to follow-up will be censored at their last date known to be alive on or before the date of data cutoff. All randomized patients will be included in the analysis of OS. Patients will be analyzed in the arm to which they are randomized regardless of the treatment they received (intent-to-treat principal).

OS in months is calculated as (date of death/last known to be alive – date of randomization +1)/30.4375. Patients lacking any dates which support patients still alive beyond randomization will have their OS censored at the date of randomization.

3.1.2. Key Secondary Endpoints

3.1.2.1. Progression Free Survival in the General Population and the pSTAT3(+) Subpopulation

Progression-Free Survival (PFS) is defined as the time from randomization to the first objective documentation of disease progression or death, whichever comes first. If a patient has not progressed or died at the time of interim or final analysis (up to the date of data cutoff), PFS will be censored on the date of the last tumor assessment (up to the date of data cutoff).

PFS in months is calculated as (first event date/censored date – date of randomization +1)/30.4375.

Tumor assessments will be performed every 8 weeks (every 56 days) after randomization (±5 days) until 6 months of treatment and every 12 weeks (every 84 days) thereafter until objective disease progression, lost to follow-up, withdrawal of consent or death.

Table 1 summarizes the censoring rules for the PFS analysis.

Table 1: PFS Definition: Events and Censoring Reasons and Hierarchy

Hierarchy Censoring	Situation	Date of Event or Censor	Event / Censor
1	**No baseline** radiological tumor assessment available	Date of randomization	Censor
5	**No post baseline** radiological tumor assessment available and no death reported within 2 scan intervals following randomization	Date of randomization	Censor
	No post baseline radiological tumor assessment available but death reported within 2 scan intervals following randomization	Date of death	Event
6	**No tumor progression** (per RECIST 1.1) and no death reported within 2 scan intervals following last adequate radiological tumor assessment	Date of last adequate radiological tumor assessment	Censor
	No tumor progression (per RECIST 1.1) but death reported within 2 scan intervals following last adequate radiological tumor assessment	Date of death	Event
	Tumor progression (per RECIST 1.1) documented within 2 scan intervals following previous adequate radiological tumor assessment	Earliest of the target, non-target and new tumor assessment dates	Event
3	**Tumor progression** (per RECIST 1.1) documented after 2 scan intervals following previous adequate radiological tumor assessment	Date of previous adequate radiological assessment	Censor
2	**New anticancer treatment** started and no tumor progression	Date of previous adequate radiological assessment immediately prior to start of new therapy	Censor
4	**No tumor progression** (per RECIST 1.1) and patient lost to follow-up or withdrawal of consent	Date of last adequate radiological Assessment	Censor

Notes: (1) Symptomatic deteriorations (i.e. symptomatic progressions, which are not radiographically confirmed) will not be considered as progressions.
(2) If target, non-target, and new lesion assessments have different dates within a visit, then the earliest of those dates will be considered as the date of the tumor assessment if the assessment for that visit is progressive disease (PD); otherwise the latest date will be used.
(3) Adequate radiographical tumor assessment refers to an assessment with overall response of CR, PR, SD, non-CR/non-PD, or PD.

Two scan intervals above indicate 16 weeks plus 10 days prior to the first 6 months of treatment, and 24 weeks plus 10 days thereafter. The censoring reason will be assigned according the hierarchy order in Table 1.

3.1.2.2. Disease Control Rate in the General Population and the pSTAT3(+) Subpopulation

Best Overall Response (BOR) is the best response recorded from randomization until disease progression or start of new anti-cancer therapy. Tumor scan assessment done after PD or after "new anti-cancer" treatment, but prior to PD will not be considered in the evaluation of BOR. BOR (based on unconfirmed response) is derived from the sequence of objective response determined by the following order:

- CR: One objective status of CR documented before progression or start of new anti-cancer therapy
- PR: One objective status of PR documented before progression or start of new anti-cancer therapy, but not qualifying as CR
- SD: At least 1 objective status of SD or better documented within at least 1 nominal scan interval (8 weeks – 5 days window = 51 days) after randomization date and before progression and the start of new anti-cancer therapy, but not qualifying as CR or PR
- PD: Progression documented within 2 nominal scan intervals (or 16 weeks + 5 days window = 117) after randomization date and not qualifying as unconfirmed CR, unconfirmed PR, or SD
- NE: All other cases will be categorized as NE. The reasons for NE will be summarized and the following reasons may be used:
  - Early death (Note: death prior to 8 weeks – 5 days window after randomization)
  - No post-baseline assessments
  - New anti-cancer therapy started before first post-baseline assessment
  - All post-baseline assessments have overall response NE
  - SD too early (<51 days after randomization date)
  - PD too late (>117 days after randomization date)

The assignment of reasons for NE will follow the above hierarchical order.

Disease Control Rate (DCR) is defined as the proportion of patients with BOR of a documented complete response, partial response, and stable disease (CR + PR + SD), based on RECIST 1.1, by investigator assessment. The estimate of DCR will be based on ODAS analysis set. In order to classify stable disease (SD) as the BOR, the assessment must be made a minimum of 8 weeks – 5 days window = 51 days from baseline where baseline is counted from randomization.

3.1.2.3. Objective Response Rate in the General Population and the pSTAT3(+) Subpopulation

Objective Response Rate (ORR) is defined as the proportion of patients with BOR of a documented complete response and partial response (CR + PR) based on RECIST 1.1 by investigator assessment. The estimate of ORR will be based on ODAS analysis set.

The patients will be considered as non-responders until proven otherwise. Thus, the patients who:

- Do not have CR or PR while on study; or
- Do not have a baseline or post-baseline tumor evaluation; or
- Do not have an adequate baseline tumor evaluation; or
- Receive new anti-cancer treatment other than the study medication prior to reaching a CR or PR; or
- Die, progress, or drop out for any reason prior to reaching a CR or PR

are considered non-responders.

3.1.2.4. Duration of Response in the General Population and the pSTAT3(+) Subpopulation

Duration of Response (DoR) is defined as the time from the first documentation of objective tumor response (CR or PR), as determined by the investigator evaluation, to the first documentation of objective tumor progression or death due to any cause, whichever occurs first. If tumor progression data include more than 1 date, the first date will be used. DoR in weeks will be calculated as (first date of PD or death − first date of CR or PR +1)/7. DoR will only be calculated for patients in the ODAS analysis set who have an objective tumor response. Censoring rules for DoR is identical to the censoring rules presented for PFS. Time to Response (TTR) is defined as the time from randomization to first documentation of objective tumor response (CR or PR) as determined by the investigator assessment. For patients proceeding from PR to CR, the onset of PR is taken as the onset of response. TTR will be calculated for the subgroup of patients with objective tumor response.

3.1.3. Other Secondary Endpoints

3.1.3.1. Quality of Life Endpoints in the General Population and the pSTAT3(+) Subpopulation

The Quality of Life (QoL) of patients will be assessed using EORTC QLQ-30 while the patient remains on study treatment (FOLFIRI with or without BBI-608). The EORTC QLQ-30 is a self-administered cancer specific questionnaire with multi-dimensional scales (Protocol Appendix 5). It consists of both multi-item scales and single item measures, including 5 functional domains, a global quality of life domain, 3 symptom domains, and 6 single items. Scoring of the EORTC QLQ-30 data will be completed following the procedures recommended by the EORTC Study Group on Quality of Life. For each domain or single item measure a linear transformation will be applied to standardize the raw score to range between 0 and 100.
The quality of life data will be analyzed to look for statistically and clinically significant differences between the BBI-608 plus FOLFIRI *versus* FOLFIRI groups. Questionnaire compliance rates will be ascertained for each group at each measurement time point.
Mean scores for each subscale will be summarized for each group at each analysis window. Change from Baseline will be summarized for each group at each post-baseline measurement time point. The endpoints in QoL analysis are the mean EORTC QLQ-C30 QoL change scores from baseline at Time 2 (Cycle 5 Day 1) and Time 4 (Cycle 9 Day 1) for the physical function and global health status/quality of life subscale scores.

The scoring method for EORTC QLQ-C30 is summarized below. In this summary Qi refers to the i-th question on the QLQ-C30.

Functional scale's scores:

· Physical functioning:(1 - ((Q1+Q2+Q3+Q4+Q5)/5 -1)/3) * 100
· Role functioning:(1 - ((Q6+Q7)/2-1)/3) * 100
· Emotional functioning:(1 - ((Q21+Q22+Q23+Q24)/4-1)/3) * 100
· Cognitive functioning:(1 - ((Q20+Q25)/2-1)/3) * 100
· Social functioning:(1 - ((Q26+Q27)/2-1)/3) * 100

Global health status score:

· Global QoL:((Q29+Q30)/2-1)/6 * 100

Symptom scale's scores:

· Fatigue:((Q10+Q12+Q18)/3-1)/3 * 100
· Nausea and vomiting:((Q14+Q15)/2-1)/3 * 100
· Pain:((Q9+Q19)/2-1)/3 * 100
· Dyspnea:((Q8-1)/3 * 100
· Insomnia:(Q11-1)/3 * 100
· Appetite loss:(Q13-1)/3 * 100
· Constipation:(Q16-1)/3 * 100
· Diarrhea:(Q17-1)/3 * 100
· Financial difficulties:(Q28-1)/3 * 100

Missing items in a scale will be handled by the following methods: values will be imputed for missing items by assuming that the missing items have values equal to the average of those items which are present for any scale in which at least half the items are completed. A scale in which less than half of the items are completed will be treated as missing.

The QoL assessment is performed prior to randomization and during protocol treatment after randomization. Since exact time of assessment may vary from patient to patient, it is necessary to provide a window for each QoL time point. Table 2 describes how to assign a questionnaire to a discrete time point, in which, target day is relative to randomization date:

Day from randomization = assessment date – randomization date,

week from randomization = (assessment date – randomization date) / 7

Table 2: Analysis Windows for QoL Endpoints

Time Point	Window	Target Day
Baseline	14 days prior to or on the randomization	Day 0
Week 4 (Cycle 3, Day 1)	0 -< 7 weeks	Day 29
Week 8 (Cycle 5, Day1)	7 weeks -< 10 weeks	Day 57
Week 12 (Cycle 7, Day 1)	10 weeks -< 14 weeks	Day 85
Week 16 (Cycle 9, Day 1)	14 weeks -< 20 weeks	Day 113
Week 24 (Cycle 13, Day 1)	20 weeks -< 28 weeks	Day 141

Each patient has 1 score selected from questionnaires collected from scheduled and unscheduled visits per scale per analysis window. If a patient had more than 1 questionnaire collected (scheduled and/or unscheduled) in an analysis window, the one collected closest to the target date is selected; if questionnaires had same days from the target date, the later one is selected; if questionnaires collected from the same date, the worst one is selected. For functional scales and global health status, the larger score the better; while for symptom scales, the smaller score the better.

The detailed analysis of QoL will be provided in a separate patient reported outcome analysis plan.

3.2. Safety Endpoints

Adverse Events (AEs) as characterized by type, frequency, severity (as graded by NCI CTCAE version 4.0), seriousness, and relationship to study therapy. Laboratory abnormalities as characterized by type, frequency, and severity (as graded by NCIC CTCAE version 4.0).

Other safety endpoints include physical examination, electrocardiogram (ECG), and vital sign data.

3.3. Other Endpoints

3.3.1. PK Endpoints

Plasma samples for sparse PK analysis will be obtained from all patients randomized to Arm 1 (BBI-608 with FOLFIRI) at the study visits occurring on Day 1 of Cycle 2 and Day 1 of Cycle 3 (corresponding to FOLFIRI infusion days). Patients randomized to Arm 2 (FOLFIRI) will not undergo plasma collection for PK analysis. The PK endpoint is:

- PK concentration of BBI-608

3.4. Covariates and Stratification Factors

3.4.1. Stratification

As described in the study design (Section 2.1.2), randomization in this study is stratified according to:

1. Geographical region (North America/Western Europe/Australia, Japan/Korea vs. Rest of the World)
2. Time to progression on first line therapy (<6 months vs. ≥6 months from start of first line therapy)
3. *RAS* mutation status (mutant vs. wild type)
4. Bevacizumab as part of study protocol treatment (yes vs. no)
5. Location of the primary tumor (left colon vs. right colon)

For patients who are incorrectly stratified at the time of randomization, either because of information available subsequent to randomization or due to clerical error, the actual baseline

value, as recorded in CRF will be used in the primary analysis. For patients enrolled prior to Protocol Amendment 3.0 which didn't stratify by "left colon vs. right colon", stratification will be derived from CRF data regarding location of the tumor and be used in primary analysis or sensitivity analysis. Any change in stratification that occurred after randomization will be noted in the clinical study report.

3.4.2. Covariates

In additional to the stratification factors, the potential influence of baseline patient characteristics on the primary endpoints (OS in the General Population and OS in the pSTAT3(+) Subpopulation will be evaluated such as (but not limited to) below:

Age (<65 years *versus* ≥ 65 years)
Age (<65 years, ≥65 to < 75 years, ≥75 to < 85 years *versus* ≥85 years)
Sex (Male *versus* Female)
Presence of liver metastases (Yes *versus* No)

Subgroup analysis for the primary endpoints (OS in the General Population and in the pSTAT3(+) Subpopulation) as sensitivity analyses will be conducted to address the benefit of BBI-608 between the treatment arms by stratification factors (Section 3.4.1), the above factors and the following (but not limited to):

ECOG performance status (0 *versus* 1)
Race (White, Black, Asian *versus* Other)
Country (Canada, USA, Western Europe, Australia, Japan, Korea, China *versus* Other)
Primary tumor site (Colon *versus* Rectum)
Prior bevacizumab (Yes *versus* No)

4. HYPOTHESIS AND DECISION RULES

4.1. Statistical Hypothesis

The primary study endpoints are OS for the General Population and the pSTAT3(+) Subpopulation. The hypotheses (the General Population and the pSTAT3(+) Subpopulation) for the study are as follows.

For the General Population, the null and alternative hypotheses are:

H_{10}: BBI-608 + FOLFIRI ≤ FOLFIRI in the General Population

H_{11}: BBI-608+FOLFIRI > FOLFIRI in the General Population

For the pSTAT3(+) Subpopulation, the null and alternative hypotheses are:

H_{20}: BBI-608 + FOLFIRI ≤ FOLFIRI in the pSTAT3(+) Subpopulation

H_{21}: BBI-608+FOLFIRI > FOLFIRI in the pSTAT3(+) Subpopulation

The study is designed to test for superiority of BBI-608 plus FOLFIRI over FOLFIRI in the General Population and the pSTAT3(+) Subpopulation. Superiority will be concluded if there is sufficient evidence to reject the null hypothesis.

4.2. Determination of Sample Size

Sample Size and Power in the General Population

For the General Population, this study is designed to have a power of 90% and a 1-sided $\alpha = 0.025$ to detect a 20% reduction in the risk of death (HR=0.80 which corresponds to an increase of median survival from 12.54 to 15.68 months). The above design assumptions account for the anticipated varying control hazard rates for the bevacizumab *versus* no-bevacizumab stratification levels. It is assumed that approximately 30% of subjects will receive bevacizumab with expected median OS (mOS) for the control arm FOLFIRI+bevacizumab subjects being 13.66 months while 70% of the subjects will not receive bevacizumab and have expected mOS of 12.06 months [Van Cutsem 2012]. Without adjusting for multiplicity, it is estimated that 850 events in the General Population will be required to detect this reduction, which would be observed by randomizing 1250 patients (General Population) over 26 months with patient follow up for an additional 14 months, for a total study duration of 40 months. It is anticipated that up to 5% dropout rate will occur over the entire study.

Sample Size and Power in the pSTAT3(+) Subpopulation

According to the preliminary data from Studies BBI608-101, CO.23, BBI608-201, BBI608-224, and BBI608-246, of the 599 patients with biomarker samples in the pooled analysis, 241 patients were pSTAT3(+), the pSTAT3(+) prevalence was estimated to be around 40% (95% CI: 36%, 44%). If considering a lower boundary of 36% as the pSTAT3(+) prevalence in the current study, and proportional to the total death events in the general population, there would be 310 (approximately 36% of 850) events in the pSTAT3(+) Subpopulation.

For the pSTAT3(+) Subpopulation, without adjusting for multiplicity, and assuming 310 events, there will be approximately 88% nominal power at a 1-sided $\alpha = 0.025$ to detect a 30% reduction

(HR [BBI-608+FOLFIRI *versus* FOLFIRI] = 0.70) in the risk of death in the pSTAT3(+) Subpopulation.

Overall Power of the Study

Assuming at the final analysis there are 850 events in the General Population and 310 events in the pSTAT3(+) Subpopulation, and considering the multiplicity adjustment aforementioned, the overall power for the entire study which succeeds either in terms of OS in the General Population or in the pSTAT3(+) Subpopulation or both, is approximately 90%, assuming that the true OS hazard ratio in the General Population is 0.80 and that in the pSTAT3(+) Subpopulation is 0.70.

4.3. Statistical Decision Rules for Interim and Final Analyses

An interim analysis will be conducted when 425 deaths occur (50% of total number of 850 deaths) in the General Population for the purpose of decision rules of futility, population and hypothesis selection, and event count adjustment. Interim results will be evaluated by the Data Safety and Monitoring Board (DSMB). Additionally, the DSMB will review safety during the conduct of the study. The role and responsibility of the DSMB will be defined in a separate charter. The recommendations from the DSMB on the study design and conduct will be based on the following decisions:

- **Futility stopping:** terminate the trial due to lack of efficacy (futility) or terminate pSTAT3(-) and pSTAT3 status unknown subpopulation
- **Patient population and hypothesis selection:** select the most appropriate patient population and hypothesis (hypothesis in General Population, and/or hypothesis in pSTAT3(+) Subpopulation) for evaluating the significance of the treatment effect at the final analysis
- **Event count adjustment:** maintain the current events size or increase the target number of events in 1 of or both pre-defined patient populations

The details decision rules are described as following:

Futility stopping rule: Perform an early assessment of the efficacy profile of BBI-608+FOLFIRI vs FOLFIRI in the 2 pre-defined patient populations and terminate the trial due to lack of efficacy (futility) if there is no evidence of treatment benefit in either population.

The decision rules will be applied simultaneously based on the observed hazard ratios at interim analysis:

- HR_0: Hazard Ratio (BBI-608+FOLFIRI vs FOLFIRI) in the General Population
- HR_+: Hazard Ratio (BBI-608+FOLFIRI vs FOLFIRI) in the pSTAT3(+) Subpopulation
- HR_-: Hazard Ratio (BBI-608+FOLFIRI vs FOLFIRI) in the pSTAT3(-) Subpopulation

If $HR_+ > c_1$ and $HR_0 > c_2$, then the trial may be stopped due to lack of efficacy.

If $HR_+ \leq c_1$ and $HR_- > c_3$, then continue the trial for pSTAT3(+) Subpopulation only and perform the FA in the pSTAT3(+) Subpopulation. Here $c_1 = c_2 = c_3 = 1$ are futility boundaries for the pSTAT3(+) Subpopulation, the General Population and the pSTAT3(-) Subpopulation respectively.

Patient population and hypothesis selection rule: Select the most appropriate patient population and hypothesis for evaluating the significance of the treatment effect at the FA.

The patient population selection rules are based on futility analysis and the interaction condition introduced [Millen 2012]. In this trial, if the General Population continues after interim analysis, the interaction condition will be evaluated by comparing the ratio of HR_+/HR_- with pre-specified threshold of c_4 where c_4 will be set to 0.9. If $HR_+/HR_- > c_4$ which indicates the treatment effect in the pSTAT3(+) Subpopulation is comparable to pSTAT3(-) Subpopulation, the hypothesis for the General Population only (the broad indication) would be appropriate.

On the other hand, if $HR_+/HR_- \leq c_4$, which indicate the pSTAT3(+) Subpopulation has substantial improvement compared to the pSTAT3(-) Subpopulation, both hypotheses of the pSTAT3(+) subpopulation and the General Population would be considered.

If only the pSTAT3(+) Subpopulation is continued, only the hypothesis for the pSTAT3(+) Subpopulation will be tested, and no interaction check will be performed.

Event count adjustment rule: Increase the target number of events in 1 of or both pre-defined patient populations to improve the probability of success in the trial.

Within each patient population (General Population or pSTAT3(+) Subpopulation), conditional power is defined as the probability of establishing a significant OS effect at the final analysis conditional upon the interim analysis data at a one-sided $\alpha = 0.025$. Conditional power is computed within each patient population using the population-specific hazard ratio observed at the interim analysis before an event count adjustment.

A closed-form expression for conditional power [Dmitrienko 2017] is given below:

$$\mathrm{CP} = \Phi\left(\sqrt{\frac{v_1}{v_2}}Z + \theta\sqrt{\frac{m_2 - m_1}{4}} - \frac{c}{\sqrt{v_2}}\right) \quad (1)$$

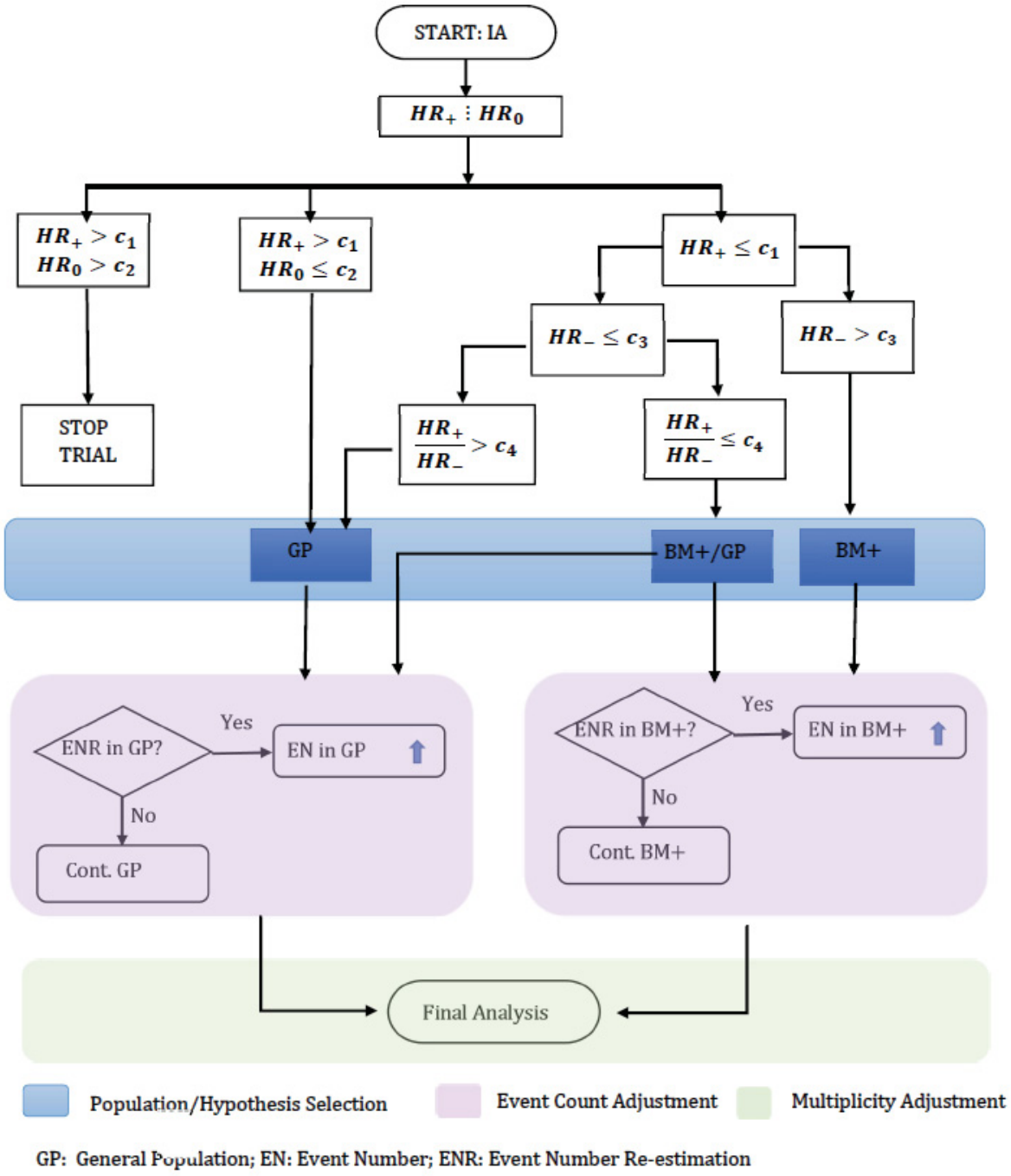
where

$$v_1 = \frac{m_1}{m_2},\ v_2 = \frac{m_2 - m_1}{m_2},$$

Z is the 1-sided log-rank test statistic computed at the interim analysis within the selected patient population and c = 1.96 is the significance level to be applied at the final analysis. m_1 and m_2 denote the number of OS events at the interim analysis and at the final analysis. θ is the observed effect size (log-hazard ratio) at the interim analysis.

The event count adjustment rule will be defined by identifying an optimal balance between the competing goals of increasing conditional power and reducing the number of events. The event count adjustment will also consider other factors including potential clinical benefit and operational feasibility. The details of the event count adjustment rule will not be described here in the SAP, but will be provided to the unblinded reporting team and DSMB in a separate document by the study statistician, and will not be shared with the study team, site and other blinded partners. This procedure is introduced to minimize the risk of the treatment effect being revealed to blinded partners by the event number re-estimation.

Figure 1: Adaptive Study Design Schema



GP: General Population; EN: Event Number; ENR: Event Number Re-estimation

4.4. Outcome of Interim Analysis, Final Analyses Timing and Hypothesis Testing Strategy

Based on interim analysis possible study actions may be as follows:

- Terminate the study for futility (lack of efficacy)
- Continue the study as is
- Continue the study in the pSTAT3(+) Subpopulation only; stop treatment of pSTAT3(-) patients; no increase in target number of events in pSTAT3(+) patients
- Continue the study in the pSTAT3(+) Subpopulation only; stop treatment of pSTAT3(-) patients; increase target number of events in pSTAT3(+) patients
- Continue the study in both General population and pSTAT3(+) population; increase target number of events in pSTAT3(+) patients

For the pSTAT3(+) Subpopulation, the total number of events is to be increased to no more than 434 events which leads to 85% power at 1-sided $\alpha = 0.025$ assuming HR = 0.75 (40% increase from original planned 310 events). This may result in enrolling further pSTAT3(+) patients and/or a longer follow-up period.

Please note that an increased event number in the General Population might also be considered, although not mentioned in the above likely IA outcomes. However, since this study is statistically powered under the assumption of the true hazard ratio of 0.8 for the General Population, which is considered to be close to the boundary of clinically meaningful benefit in OS, the increase of the number of events in the General Population will need to be justified by a careful assessment of potential clinical benefit.

The IA timing will be triggered at ~425 events (50% information) in the General Population. Table 3 summarizes the final analysis timing based on different IA outcomes.

Table 3: Final Analysis Timing by Different Outcomes from IA

Outcomes from IA	Final Analysis Timing	Patient Enrollment
Only General Population hypothesis is selected	850 events in General Population if no event count adjustment*$	1250 patients enrolled
Only pSTAT3(+) hypothesis is selected	310 to 434 events* in pSTAT3(+) depending upon IA decision	May enroll more pSTAT3(+) patients after IA or expand the follow-up period
Both General Population and pSTAT3(+)	850 events in General Population*$ or 310 to 434	May enroll more General Population or pSTAT3(+)

hypotheses are selected	events* in pSTAT3(+) whichever comes later.	patients after IA or expand the follow-up period

*The event count adjustment may also consider other factors including potential clinical benefits and operational feasibility.
$For the General population, the number of events may not be increased unless other factors such as potential clinical benefits will drive the increase.

4.5. Multiplicity

To address multiplicity introduced by the analysis of several endpoints (primary endpoints and key secondary endpoints), analysis of the 2 patient populations (General Population and pSTAT3(+) Subpopulation), and data-driven design changes, a Hochberg-based gatekeeping procedure will be implemented [Dmitrienko and Tamhane (2011, 2013), Kordzakhia et al. (2018b)]. The hypothesis testing for the key secondary endpoints will follow the order of PFS, DCR and ORR in both the General Population and the pSTAT3(+) Subpopulation.

The gatekeeping procedure will be applied in conjunction with the combination function approach at the final analyses to ensure strong Type I error rate control.

4.5.1. Hochberg-based Gatekeeping Procedure

The first component of the multiplicity adjustment that accounts for the first 2 sources of multiplicity (analysis of several endpoints in several patient populations) relies on a Hochberg-based gatekeeping procedure. This procedure is defined using the mixture-based approach developed in Dmitrienko and Tamhane [*2011, 2013*] and later enhanced in Kordzakhia et al. [*2018b*]. This gatekeeping procedure will be applied to the 8 null hypotheses of no effect based on the 4 endpoints that are evaluated in the 2 patient populations:

- Family 1: Hypothesis H_{10} (null hypothesis of no OS effect in the General Population) and Hypothesis H_{20} (null hypothesis of no OS effect in the pSTAT3(+) Subpopulation).
- Family 2: Hypothesis H_{30} (null hypothesis of no PFS effect in the General Population) and Hypothesis H_{40} (null hypothesis of no PFS effect in the pSTAT3(+) Subpopulation).
- Family 3: Hypothesis H_{50} (null hypothesis of no DCR effect in the General Population) and Hypothesis H_{60} (null hypothesis of no DCR effect in the pSTAT3(+) Subpopulation).
- Family 4: Hypothesis H_{70} (null hypothesis of no ORR effect in the General Population) and Hypothesis H_{80} (null hypothesis of no ORR effect in the pSTAT3(+) Subpopulation).

The hypotheses will be organized into 2 branches (i.e. the null hypotheses in the General Population [H_{10}, H_{30}, H_{50} and H_{70}] and the null hypotheses in the pSTAT3(+) Subpopulation [H_{20}, H_{40}, H_{60} and H_{80}]), and tested sequentially within each branch.

Key features of the Hochberg-based gatekeeping procedure include:

- The gatekeeping procedure accounts for the clinically relevant logical restrictions (i.e. for the fact that the null hypotheses are organized into branches and tested sequentially within each branch).

- The gatekeeping procedure utilizes powerful Hochberg-type tests (regular and truncated Hochberg tests) for testing the hypotheses within each family. These tests account for positive correlations between the test statistics within each family, which is induced by the fact that the pSTAT3(+) Subpopulation is nested within the General Population. The regular and truncated Hochberg tests control the Type I error rate within each family since if the test statistics within each family follow a bivariate normal distribution with a non-negative pairwise correlation [Sarkar (*2008*)].
- The gatekeeping procedure uses truncated Hochberg tests in the first 3 families because they serve as gatekeepers for other families. The regular Hochberg test is applied in Family 4 since it is the last family in the sequence. The truncated Hochberg tests used in Families 1, 2, and 3 will be defined using a pre-specified truncation parameter γ=0.8.

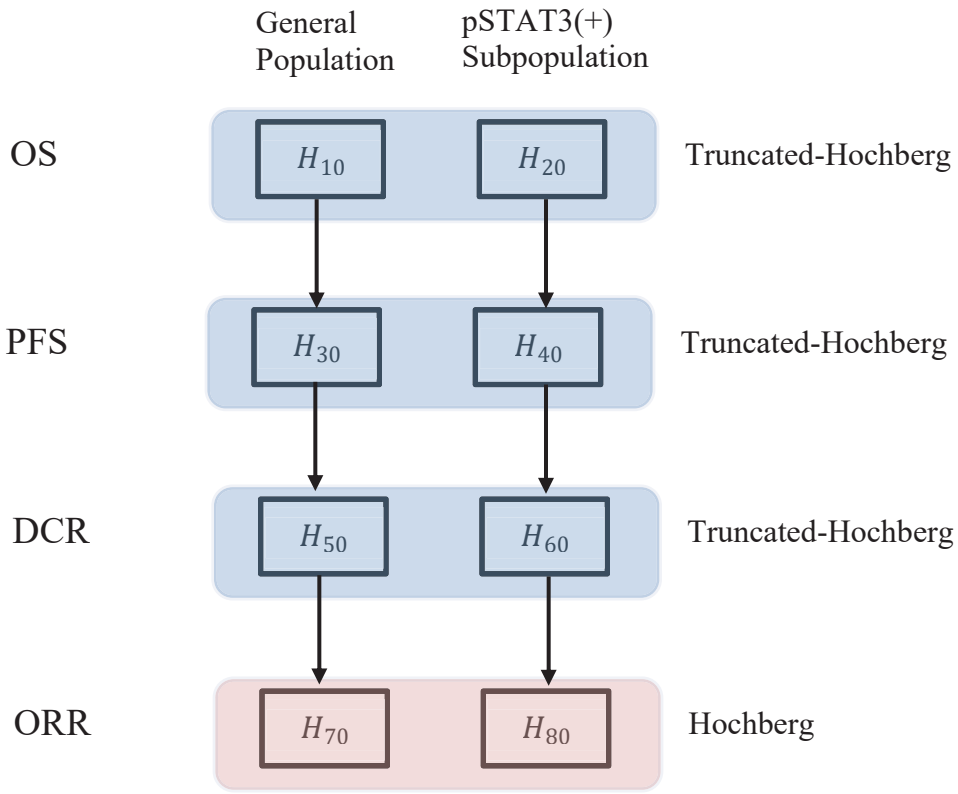
To define the regular and truncated Hochberg tests, consider, for example, Family 1 which includes the hypotheses H_{10} and H_{20}. Let p_1 and p_2 denote the corresponding 1-sided p-values and let $p_{(1)} < p_{(2)}$ denote the ordered p-values. Finally, let α_1 denote the 1-sided significance level applied within this family (α_1 may not be equal to $\alpha = 0.025$). The regular Hochberg test relies on the following 2-step testing algorithm:

- Step 1: The test rejects both H_{10} and H_{20} if $p_{(2)} \leq \alpha_1$.
- Step 2: If $p_{(2)} > \alpha_1$, the test rejects the hypothesis corresponding to the smaller p-value if $p_{(1)} \leq \alpha_1/2$.

The truncated Hochberg test utilizes the following 2-step testing algorithm:

- Step 1: The test rejects both H_{10} and H_{20} if $p_{(2)} \leq (\gamma+(1-\gamma)/2)\alpha_1$.
- Step 2: If $p_{(2)} > (\gamma+(1-\gamma)/2)\alpha_1$, the test rejects the hypothesis corresponding to the smaller p-value if $p_{(1)} \leq \alpha_1/2$.

Figure 2: Hochberg-based Gatekeeping Procedure



4.5.2. Combination P-Value Approach

The combination function approach will be applied using the weighted inverse-normal combination function to address the third source of multiplicity (evaluation of the treatment effect at several decision points). The joint application of the combination function approach and gatekeeping procedure relies on ideas presented in Sugitani, Bretz and Maurer [*2016*], Kordzakhia et al [*2018a*], and Sugitani et al. [*2018*].

Combination Function Approach for General Population

For the General Population, the combination function approach remains the same as previous version of SAP 2.0. The test statistics for evaluating the significance of the treatment effect or the corresponding p-values will be computed from the data collected up to the interim analysis (Stage 1) and after the interim analysis (Stage 2). The stage-wise p-values for the primary endpoint (OS) will be obtained from the increments of the log-rank test statistics using the general approach developed in Shen and Cai [*2003*] and Wassmer [*2006*]. The increments of the log-rank test statistics that correspond to the 2 trial stages are defined as follows:

$$Z_{11}^{*} = Z_{11}, Z_{21}^{*} = \frac{\sqrt{k_{21}}\, Z_{21} - \sqrt{k_{11}}\, Z_{11}}{\sqrt{k_{21} - k_{11}}} \quad (2)$$

- Z_{11} and Z_{21}: the 1-sided log-rank test statistic [*Jennison and Turnbull, 2000*] at the IA and FA for the General Population.
- k_{11} and k_{21}: the number of events at the IA and FA for the General Population.

The 1-sided stage-wise p-values referred as the Stage 1 and Stage 2 p-values for testing the null hypothesis H_{10}, will be computed from these log-rank test statistics as

$$p_1 = 1 - \Phi(Z_{11}^{*}), q_1 = 1 - \Phi(Z_{21}^{*}) \quad (3)$$

where $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution.

The two stage-wise test statistics are independent of each other under the null hypothesis of no treatment effect of H_{10} and thus the final treatment effect p-value can be found by combining the stage-wise p-values, i.e.,

$$s_1 = c_2(p_1, q_1) = 1 - \Phi\left(\sqrt{w_1}\Phi^{-1}(1 - p_1) + \sqrt{1 - w_1}\Phi^{-1}(1 - q_1)\right) \quad (4)$$

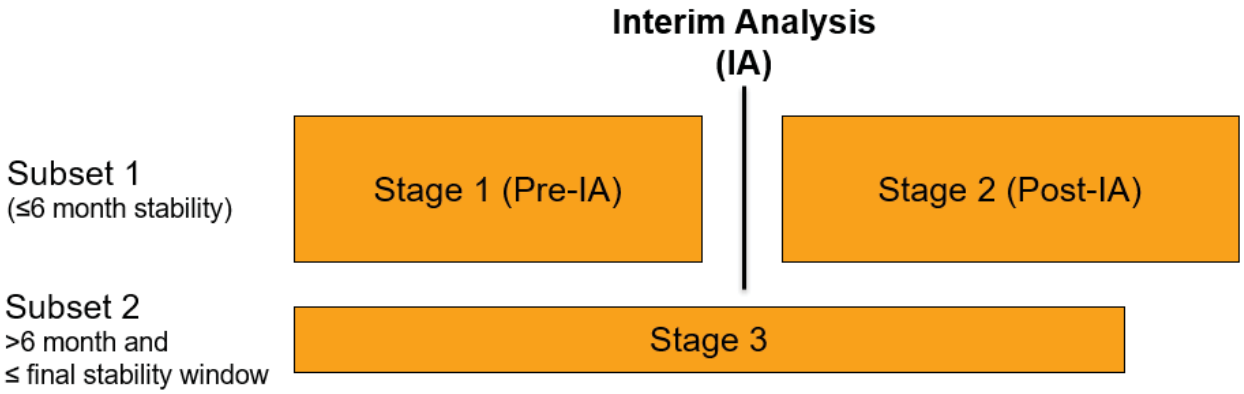
where w_1 and $1 - w_1$ are the pre-defined stage weights assigned to Stages 1 and 2 for the null hypothesis H_{10}.

Combination Function Approach for pSTAT3(+) Subpopulation

At the IA, based on preliminary assessments of the ongoing CSS study, patients with specimens that were within the stability window of 6 months are included for pSTAT3 Subpopulations designations. Due to the availability of additional patients with pSTAT3 results within the final CSS window [See Section 5.1 for detailed explanations], the test statistics at the FA for the primary and the key secondary endpoints in the pSTAT3(+) Subpopulation will be computed

across two subsets of patients data, i.e., Subset 1 of the patients with pSTAT3 positive status within specimen age up to 6-months (including both Stage 1 pre-IA data and Stage 2 post-IA data), and Subset 2 (Stage 3) of patients with pSTAT3 positive status with specimen age for longer than 6 months and within the final CSS window. Figure 3 presents the data that will be used for the primary and the key secondary endpoints in the pSTAT3(+) Subpopulation at the FA.

Figure 3: Data in FA of the pSTAT3(+) Subpopulation



Subset 2 data may also be treated as Stage 3 data for the pSTAT3(+) Subpopulation.

For the pSTAT3(+) Subpopulation, the statistics test for evaluating the significance of the treatment effect or the corresponding p-values in terms of OS at the FA will be computed from the following three stages of data from the two subsets of patients defined above. The three stages of data are:

- Stage 1 (Subset 1 Stage 1) Data: OS data collected up to the IA from Subset 1.
- Stage 2 (Subset 1 Stage 2) Data: OS data collected after the IA from Subset 1.
- Stage 3 (Subset 2) Data: OS event data from Subset 2.

The final treatment effect p-value for the null hypothesis of OS in the pSTAT3(+) Subpopulation can be computed by combining the stage-wise p-values from the three parts of the data.

Since Subset 1 and Subset 2 are two mutually exclusive sets of patients, Subset 2 data may also be treated as Stage 3 data for the pSTAT3(+) Subpopulation to describe multi-stage combination function approach.

Subset 1 Stage 1 and Stage 2 p-values for the primary endpoint (OS) in the pSTAT3(+) Subpopulation will be obtained from the increments of the log-rank test statistics in a similar fashion as conducted for the General Population:

$$Z_{12}^{*} = Z_{12}\,, Z_{22}^{*} = \frac{\sqrt{k_{22}}\,Z_{22} - \sqrt{k_{12}}\,Z_{12}}{\sqrt{k_{22} - k_{12}}} \qquad (5)$$

- Z_{12} and Z_{22}: the 1-sided log-rank test statistic at the IA and FA for the pSTAT3(+) Subpopulation with specimen age up to the 6-month stability window (Subset 1).

- k_{12} and k_{22}: the number of events at the IA and FA for the pSTAT3(+) Subpopulation with specimen age up to the 6-month stability window (Subset 1).

Furthermore, let Z_{32}^* denotes the 1-sided log-rank test statistic for the pSTAT3(+) Subpopulation in Subset 2. To ensure the test statistic Z_{32}^* be independent of the stage wise test statistics Z_{12}^* and Z_{22}^* under the null hypothesis of no treatment effect [Magirr et al., 2016], the test statistic Z_{32}^* needs to be computed based on the data available by the analysis cutoff date (a pre-determined calendar date) for Subset 2 (See Section 6.1 for details of the analysis cut dates).

The stage-wise p-values for the pSTAT3(+) Subpopulation are defined as follows:

$$p_2 = 1 - \Phi(Z_{12}^*), \quad q_2 = 1 - \Phi(Z_{22}^*), r_2 = 1 - \Phi(Z_{32}^*). \tag{6}$$

As stated above, since the length of the patient follow-up period is pre-defined in Subset 2, the test statistic Z_{32}^* is independent of the stage wise test statistics Z_{12}^* and Z_{22}^* under the null hypothesis of no treatment effect. As a result, the final treatment effect p-value for the null hypothesis H_{20} can be computed by combining the stage-wise p-values, i.e.,

$$s_2 = c_3(p_2, q_2, r_2) = 1 - \Phi\left(\sqrt{v_2}\sqrt{w_2}\Phi^{-1}(1-p_2) + \sqrt{v_2}\sqrt{1-w_2}\Phi^{-1}(1-q_2) + \sqrt{1-v_2}\Phi^{-1}(1-r_2)\right), \tag{7}$$

where w_2 and $1 - w_2$ are the pre-defined weights assigned to Stages 1 and 2 and, in addition, v_2 and $1 - v_2$ are the pre-defined weights assigned to Stages 1+2 and 3 for the null hypothesis H_{20}.

Combination Function Approach for Key Secondary Endpoints

The stage-wise test statistics for PFS are defined in the same manner. To define the stage-wise test statistics for DCR and ORR, the log-rank test needs to be replaced by the Z test for proportions and the number of events need to be replaced by the numbers of patients. Considering ORR/DCR are binary endpoints, in order to perform statistical inferences for ORR and DCR using combination function approach, two distinct subsets of patients for stage 1 and stage 2 will be defined in Appendix VI to obtain two independent test statistics from stage 1 and stage 2 based on minimum 36 weeks duration. The DCR 36 weeks and ORR 36 weeks are defined based on 36 weeks from randomization. The p-values used in combination functions will be calculated based on DCR 36 weeks and ORR 36 weeks.

The 1-sided stage-wise p-values for PFS, DCR, and ORR are defined as follows:

- Hypothesis H_{30}: Let p_3 and q_3 denote the 1-sided p-values computed from the null distributions of the stage-wise test statistics (Stage 1 and 2) for PFS in the General Population.
- Hypothesis H_{40}: Let p_4, q_4 and r_4 denote the 1-sided p-values computed from the null distributions of the stage-wise test statistics (Subset 1 Stage 1, 2 and Subset 2) for PFS in the pSTAT3(+) Subpopulation.

- Hypothesis H_{50}: Let p_5 and q_5 denote the 1-sided p-values computed from the null distributions of the stage-wise test statistics (Stage 1 and 2) for DCR in the General Population.
- Hypothesis H_{60}: Let p_6, q_6 and r_6 denote the 1-sided p-values computed from the null distributions of the stage-wise test statistics (Subset 1 Stage 1, 2 and Subset 2) for DCR in the pSTAT3(+) Subpopulation.
- Hypothesis H_{70}: Let p_7 and q_7 denote the 1-sided p-values computed from the null distributions of the stage-wise test statistics (Stage 1 and 2) for ORR in the General Population.
- Hypothesis H_{80}: Let p_8, q_8 and r_8 denote the 1-sided p-values computed from the null distributions of the stage-wise test statistics (Subset Stage 1, 2 and Subset 2) for ORR in the pSTAT3(+) Subpopulation.

To apply the combination function principle, the stage-wise p-values for OS, PFS, DCR, and ORR will be combined using an appropriately defined weighted inverse-normal combination function, i.e., $c_2(p_i, q_i)$ for the null hypothesis H_{i0}, $i = 3,5,7$, and $c_3(p_i, q_i, r_i)$ for the null hypothesis H_{i0}, $i = 4,6,8$. The hypothesis-specific stage weights used in these combination functions are defined in Table 4.

Table 4: Prespecified Stage Weights for Different Endpoints

	w_i	v_i
H_{10}: OS in General Population	0.5	
H_{20}: OS in pSTAT3(+) Subpopulation	0.5	0.95
H_{30}: PFS in General Population	0.65	
H_{40}: PFS in pSTAT3(+) Subpopulation	0.65	0.95
H_{50}: DCR in General Population	0.9	
H_{60}: DCR in pSTAT3(+) Subpopulation	0.9	0.95
H_{70}: ORR in General Population	0.8	
H_{80}: ORR in pSTAT3(+) Subpopulation	0.8	0.95

The inferences for the 8 hypotheses corresponding to OS, PFS, DCR, and ORR will be performed at the FA using the composite *p*-values that are defined as follows:

- Hypothesis H_{10}: The composite *p*-value is defined as $s_1 = c_2(p_1, q_1)$
- Hypothesis H_{20}: The composite *p*-value is defined as $s_2 = c_3(p_2, q_2, r_2)$
- Hypothesis H_{30}: The composite *p*-value is defined as $s_3 = c_2(p_3, q_3)$.

- Hypothesis H_{40}: The composite *p*-value is defined as $s_4 = c_3(p_4, q_4, r_4)$
- Hypothesis H_{50}: The composite *p*-value is defined as $s_5 = c_2(p_5, q_5)$.
- Hypothesis H_{60}: The composite *p*-value is defined as $s_6 = c_3(p_6, q_6, r_6)$
- Hypothesis H_{70}: The composite *p*-value is defined as $s_7 = c_2(p_7, q_7)$
- Hypothesis H_{80}: The composite *p*-value is defined as $s_8 = c_3(p_8, q_8, r_8)$

If the IA outcome will be the treatment effect is evaluated only in the General Population at the FA and thus the hypotheses H_{20}, H_{40}, H_{60} and H_{80} are dropped after the IA, the corresponding Stage 2 and Stage 3 (Subset 2) p-values (i.e. q_2, q_4, q_6, q_8, r_2, r_4, r_6 and r_8) will be set to 1. Similarly, if the treatment effect is evaluated only in the pSTAT3(+) Subpopulation at the FA and thus the hypotheses H_{10}, H_{30}, H_{50} and H_{70} are dropped after the IA, the corresponding Stage 2 p-values (i.e., q_1, q_3, q_5 and q_7) will be set to 1.

Finally, to compute the multiplicity adjusted *p*-values for the hypotheses of interest at the FA, a closed family associated with these null hypotheses needs to be introduced. The closed family contains 2^8 - 1= 255 intersections. Let $H(I)$ denote an arbitrary intersection hypothesis from the closed family, which is associated with the index set I. A local *p*-value, also known as the intersection p-value, will be computed for the intersection hypothesis $H(I)$ as shown below using the composite p-values defined above (i.e. s_1 through s_8).

To ensure that the final gatekeeping procedure is consistent with the logical relationships among the null hypotheses, the corresponding restrictions need to be imposed on the hypotheses within each intersection. These restrictions ensure, for example, that the hypothesis H_{30} cannot be rejected if the hypothesis H_{10} is not rejected. Let I^* denote the restricted index set which accounts for the logical relationships among the null hypotheses for the intersection hypothesis $H(I)$.

It is easy to show that, for any intersection hypothesis $H(I)$ in the closed family, there can be at most two indices left in the restricted index set I^* and therefore it is sufficient to consider the following 2 cases:

Case 1. Suppose that there are 2 indices in the restricted index set I^* that are denoted by i and j, i.e., $I^* = \{i, j\}$. If the corresponding hypotheses are included in the same family, the local *p* value- for this intersection hypothesis is given by

$$p(I) = \min(2s_{(i)}, s_{(j)}) \quad (8)$$

$s_{(i)}$ and $s_{(j)}$ denote the ordered p-values, i.e., $s_{(i)} < s_{(j)}$.

If the corresponding hypotheses are included in 2 different families and the index i corresponds to the more important family, the local *p*-value for this intersection hypothesis is given by

$$p(I) = 2\min(s_i/(1+\gamma), s_j/(1-\gamma)). \quad (9)$$

Case 2. Suppose that there is only 1 index in the restricted index set I^*, denoted by i, i.e., $I^* = \{i\}$. In this case, the local *p*-value for the intersection hypothesis is simply given by

$$p(I) = s_i \quad (10)$$

The resulting intersection *p*-values will be utilized for computing the multiplicity-adjusted *p* values for the 8 null hypotheses at the final analysis. Per communication from FDA on

23 July 2019, it was suggested to sponsor to introduce a small penalty for the IA because of the unblinded analysis of the efficacy data even though the IA were not to stop for efficacy. Therefore, a small penalty for the IA (1-sided alpha of 0.0001) is introduced here and a null hypothesis will be rejected if the intersection p-values for all intersection hypotheses containing this particular null- hypothesis are less than or equal to a 1-sided $\alpha = 0.0249$.

The proposed decision rules in the adaptive design ensure overall Type I error rate control with respect to the primary and key secondary endpoints. The Type I error rate control is maintained even if the target number of events is increased or a decision to restrict the patient enrollment to the subpopulation of pSTAT3(+) patients is made at the IA.

4.6. Blinding/Unblinding

This is an open-label study; however, to minimize the risk of potential bias, procedures for blinding have been implemented to ensure the integrity of the study data (Sponsor blinding plan and CRO blinding plan documents, respectively). Every effort will be made per the blinding plans and the DSMB charter to maintain the firewall between blinded and unblinded teams at SDPO and partners and to minimize the potential operational bias. An overview of the blinding plan is as follows:

- SDPO Blinding Plan (Section 9): SDPO has developed an internal blinding plan and trained the entire SDPO staff on the procedures outlined in the plan. SDPO has ensured that mechanisms are in place to ensure that the staff follow these established procedures. In addition, a key measure has been put in place involving QA oversight of activities that are governed by the SDPO blinding plan.
- CRO Blinding Plan (Section 9): An additional blinding plan is in place for communications between the CRO and SDPO to ensure blinded team members at SDPO do not receive unblinding communications. The CRO is tasked with blinding data prior to distributing it to BBI, and follows their own SOPs with regard to maintaining a blind.
- The majority of the study team members will remain completely blinded to all clinical data and will not participate in any review or analysis of these data until the final database lock or termination of the study. Some members may be unblinded on a patient level as required by the scope of their work. Specific firewalls exist between the blinded and unblinded team members to ensure no transfer of clinical data from an unblinded team member to a blinded team member.
- To support regular safety reviews and the interim analysis, the analyses will be conducted by an independent third party (i.e., an unblinded team at the CRO, Parexel). The unblinded statistician and programmer from the third/independent party will send the unblinded data and analyses directly to the DSMB (the roles and responsibilities of the independent DSMB are detailed in a separate charter). Only blinded aggregate data for the study population as a whole (i.e., not by Arm) will be shared with the Sponsor. Pharmacovigilance safety review will be conducted by internal, unblinded personnel.
- All data collected for this study are contained in the electronic data capture system (EDC; Oracle InForm) with audit trail capability consistent with 21 CFR Part 11; access to these data is password protected and limited to individuals responsible for reviewing and monitoring individual patient safety data or overseeing protocol adherence.

5. ANALYSIS SETS

5.1. Intent-to-Treat Analysis Set

5.1.1. Intent-to-Treat Analysis Set in the General Population (ITT-GP)

The intent-to-treat (ITT) analysis set in the General Population (ITT-GP) will consist of all randomized patients with study drug assignment designated according to initial randomization. ITT-GP will be used for the analysis of the General Population for evaluating efficacy endpoints, patients' characteristics and patient disposition.

5.1.2. Intent-to-Treat Analysis Set in the pSTAT3(+) Subpopulation, the pSTAT3(-) Subpopulation and the pSTAT3(Unknown) Subpopulation

The pSTAT3 Subpopulations will be defined by the results of a Clinical Trial Assay (CTA) for specimen age within a defined cut-section stability (CSS) window (6 months at IA or final CSS window at FA). Specimen age is the time interval between sectioning and staining.

A patient with positive pSTAT3 status within a defined CSS window is the one with pSTAT3 positive designated by CTA assay with specimen age up to 6-month at IA or up to final CSS window at FA.

A patient with negative pSTAT3 status within a defined CSS window is the one with pSTAT3 negative designated by CTA assay with specimen age up to 6-month at IA or up to final CSS window at FA.

A patient with unknown pSTAT3 status within a defined CSS window is the one with no specimen submitted for testing, specimen identified as non-evaluable upon pathology evaluation, tissues test out of a defined CSS window or missing due to any other reasons.

5.1.2.1. Intent-to-Treat Analysis Set in the pSTAT3(+) Subpopulation (ITT-pSTAT3(+))

The intent-to-treat analysis set in the pSTAT3(+) Subpopulation (ITT-pSTAT3(+)) will consist of all randomized patients with study drug assignment designated according to initial randomization and having positive pSTAT3 status tested within a defined CSS window (6 months at IA or Final CSS window at FA). ITT-pSTAT3(+) will be used for the analysis for the pSTAT3(+) Subpopulation for evaluating efficacy endpoints, patients' characteristics and patient disposition.

5.1.2.2. Intent-to-Treat Analysis Set in the pSTAT3(-) Subpopulation (ITT-pSTAT3(-))

The intent-to-treat analysis set in the pSTAT3(-) Subpopulation (ITT-pSTAT3(-)) will consist of all randomized patients with study drug assignment designated according to initial randomization and having negative pSTAT3 status tested within a defined CSS window (6 months at IA or Final CSS window at FA). ITT-pSTAT3(-) will be used for the analysis for the pSTAT3(-) Subpopulation for evaluating efficacy endpoints, patients' characteristics and patient disposition at the IA and/or FA depending on decision from the IA.

5.1.2.3. Intent-to-Treat Analysis Set in the pSTAT3(Unknown) Subpopulation (ITT-pSTAT3(Unknown))

The intent-to-treat analysis set in the pSTAT3(Unknown) Subpopulation (ITT-pSTAT3(Unknown)) will consist of all randomized patients with study drug assignment designated according to initial randomization and having unknown pSTAT3 status. ITT-pSTAT3(Unknown) will be used for the analysis for the sensitivity analysis for evaluating impact of missing biomarker status of the primary and key secondary endpoints.

5.1.2.4. Evolving Knowledge of Cut-Section Stability (CSS) for pSTAT3 Biomarker Assay

The pSTAT3 IHC D3A7 CTA was used for this study to identify the pSTAT3(+) patient population. This biomarker assay is an immunohistochemistry-based method intended for use in detection of pSTAT3 protein in cut tissue sections. A CSS study is ongoing to establish the final pSTAT3 CSS window which is expected to be completed by Feb 2020.

At the time of the IA in the CanStem303C study, CSS preliminary assessments from the ongoing CSS study were indicative of a stability of at least 6 months for slides stored at -20 degrees Celsius, the same manner in which CanStem303C specimens have been stored. The IA, therefore, included pSTAT3 biomarker status results for cut slide specimens age up to 6 months. That is, the IA assumed a CSS window of 6 months.

However, after the IA in the CanStem303C study, the biomarker data monitoring from the ongoing CSS study indicates that samples will likely remain stable at -20 degrees Celsius for greater than 6 months. That is, the final CSS window could be longer than 6 months and up to 15 months. Consequently, more patients with pSTAT3 biomarker results could be included in the FA of the CanStem303C study. Table 5 presents preliminary estimated number (% of ITT-GP) of enrolled patients for all possible CSS windows in the CanStem303C study at the time this document was being prepared. Actual number may vary at the final analysis.

Table 5: Preliminary Estimated Number of Specimens tested within Various Stability Windows in the CanStem303C study

Number of patient specimens collected within specified stability window	n (%) for CTA	Increment % from a CSS of 6-month
Total number of patients enrolled	1253	
Number of patient specimens collected	1127 (90.2%)	
Stability window of 6-month	859 (68.6 %)	
Stability window of 7-month	876 (70.1 %)	1.4%
Stability window of 8-month	891 (71.1 %)	2.6%
Stability window of 9-month	904 (72.1 %)	3.6%
Stability window of 10-month	918 (73.3 %)	4.7%

Stability window of 11-month	944 (75.3 %)	6.8%
Stability window of 12-month	962 (76.8 %)	8.2%
Stability window of 13-month	987 (78.8 %)	10.2%
Stability window of 14-month	1004 (80.1 %)	11.6%
Stability window of 15-month	1023(81.6 %)	13.1%
Note: Number of patient specimens with pSTAT3 status determined by testing specimens stored at the respective time-frames; 1 specimen reported per patient. The projected increase in the specimens represents pSTAT3(+), pSTAT3(-) and Non-Evaluable (NE) upon pathology evaluation.		

5.1.2.5. Subsets of ITT-pSTAT3(+) and ITT-pSTAT3(-) at the FA

As a result of evolving knowledge of CSS window, ITT-pSTAT3(+) at the FA consists of patients from two subsets:

- ITT-pSTAT3(+)-Subset-1 will consist of ITT-pSTAT3(+) patients with specimen age up to a 6-month CSS window.
- ITT-pSTAT3(+)-Subset-2 will consist of ITT-pSTAT3(+) patients with specimens age > 6 months and <= final CSS window.

Similarly, ITT-pSTAT3(-) at the FA consists of patients from two subsets:

- ITT-pSTAT3(-)-Subset-1 will consist of ITT-pSTAT3(-) patients with specimen age up to a 6-month CSS window.
- ITT-pSTAT3(-)-Subset-2 will consist of ITT-pSTAT3(-) patients with specimens age > 6 months and <= final CSS window.

A patient with unknown pSTAT3 status due to out of 6-month CSS window at IA may have either pSTAT3 positive or pSTAT3 negative status at the FA when considering the final CSS window established wider than IA. Therefore, it is expected that there will be fewer patients with unknown pSTAT3 status within the final CSS window at the FA, compared to the IA with 6-month CSS window.

Figure 4 below shows the pSTAT3 Subpopulations components at the IA and at the FA.

Figure 4: pSTAT3 Subpopulations Components at Interim Analysis and Final Analysis

	Interim Analysis			Final Analysis		
Specimen Age (≤6m)*	+	-	~~NE~~	+	-	~~NE~~
Specimen Age (>6m and ≤ final CSS window)				+	-	~~NE~~
Specimen Age (> final CSS window)						
No Specimen or Other Reasons for Missing	**Missing**	**Missing**	**Missing**	**Missing**	**Missing**	**Missing**

+	Positive by CTA Readout
-	Negative by CTA Readout
NE	Not Evaluable by CTA Readout
Missing	Missing or Not Applicable
*	Approximately 20 additional specimens collected post IA
	Considered as Not Valid Readout

Definitions of the subpopulations

pSTAT3(+) subpopulation at IA (**Subset 1:** ≤6m CSS window)
FA (**Subset 1:** ≤6m CSS window + **Subset 2:** >6m to ≤ final CSS window).

pSTAT3(-) subpopulation at IA (≤6m CSS window) or at FA (≤ final CSS window).

pSTAT3(Unknown) subpopulation include patients with no specimen submitted for testing, specimens identified as non-evaluable upon pathology evaluation, tissues tested out of the defined CSS window or missing due to any other reasons.

Note: All specimens will be tested only one time regardless of final stability window.

5.1.2.6. Biomarker Analysis Set (BAS)

The biomarker analysis set will consist of all randomized patients who have 1 evaluable result of pSTAT3 status (positive or negative) within a defined CSS window. It will consist of both ITT-pSTAT3(+) and ITT-pSTAT3(-) patients.

5.2. Safety Analysis Set

5.2.1. Safety Analysis Set in the General Population (SAS-GP)

The safety analysis set in the General Population (SAS-GP) will include all randomized patients who receive at least 1 dose of study drug (BBI-608 and/or FOLFIRI) with treatment assignment designated according to the actual study treatment received. SAS-GP will be used for all safety and drug exposure evaluation in the General Population.

5.2.2. Safety Analysis Set in the pSTAT3(+) Subpopulation (SAS-pSTAT3(+))

The safety analysis set in the pSTAT3(+) Subpopulation (SAS-pSTAT3(+)) will include all randomized patients who receive at least 1 dose of study drug (BBI-608 and/or FOLFIRI) with treatment assignment designated according to the actual study treatment received and having positive pSTAT3 status within a defined CSS window (6 months at IA or Final CSS window at FA). SAS-pSTAT3(+) will be used for all safety and drug exposure evaluation in the pSTAT3(+) Subpopulation.

5.2.3. Safety Analysis Set in the pSTAT3(-) Subpopulation (SAS-pSTAT3(-))

The safety analysis set in the pSTAT3(-) Subpopulation (SAS-pSTAT3(-)) will include all randomized patients who receive at least 1 dose of study drug (BBI-608 and/or FOLFIRI) with treatment assignment designated according to the actual study treatment received and having negative pSTAT3 status within a defined CSS window (6 months at IA or Final CSS window at FA). SAS-pSTAT3(-) will be used for all safety and drug exposure evaluation in the pSTAT3(-) Subpopulation at IA and/or FA depending on decision from the IA.

5.2.4. Safety Analysis Set in the pSTAT3(Unknown) Subpopulation (SAS-pSTAT3(Unknown))

The safety analysis set in the pSTAT3(Unknown) Subpopulation (SAS-pSTAT3(Unknown)) will include all randomized patients who receive at least 1 dose of study drug (BBI-608 and/or FOLFIRI) with treatment assignment designated according to the actual study treatment received and having unknown pSTAT3 status.

5.3. ORR/DCR Analysis Set

5.3.1. ORR/DCR Analysis Set in the General Population (ODAS-GP)

The ORR/DCR analysis set in the General Population (ODAS-GP) will include all randomized patients with measurable disease by RECIST 1.1 at randomization with study drug assignment designated according to initial randomization. ODAS-GP will be used for evaluating BOR, ORR, DCR and DoR in the General Population.

5.3.2. ORR/DCR Analysis Set in the pSTAT3(+) Subpopulation (ODAS-pSTAT3(+))

The ORR/DCR analysis set in the pSTAT3(+) Subpopulation will include all randomized patients with measurable disease by RECIST 1.1 at randomization with study drug assignment designated according to initial randomization and having positive pSTAT3 status within a defined CSS window (6 months at IA or Final CSS window at FA). ODAS-pSTAT3(+) will be used for evaluating BOR, ORR, DCR and DoR in the pSTAT3(+) Subpopulation.

- ODAS-pSTAT3(+)-Subset-1 will consist of ODAS-pSTAT3(+) patients with specimen age up to a 6-month CSS window.
- ODAS-pSTAT3(+) -Subset-2 will consist of ODAS-pSTAT3(+) patients with specimens age > 6 months and <= final CSS window.

5.3.3. ORR/DCR Analysis Set in the pSTAT3(-) Subpopulation (ODAS-pSTAT3(-))

The ORR/DCR analysis set in the pSTAT3(-) Subpopulation will include all randomized patients with measurable disease by RECIST 1.1 at randomization with study drug assignment designated according to initial randomization and having negative pSTAT3 status within a defined CSS window (6 months at IA or Final CSS window at FA). ODAS-pSTAT3(-) will be used for evaluating BOR, ORR, and DCR in the pSTAT3(-) Subpopulation at IA and/or FA depending on decision from the IA.

- ODAS-pSTAT3(-)-Subset-1 will consist of ODAS-pSTAT3(-) patients with specimen age up to a 6-month CSS window.
- ODAS-pSTAT3(-) -Subset-2 will consist of ODAS-pSTAT3(-) patients with specimens age > 6 months and <= final CSS window.

Note that ODAS-pSTAT3(+) and ODAS-pSTAT3(-) at the IA are subsets of those at the FA, as explained in Section 5.1.4.

5.3.4. ORR/DCR Analysis Set in the pSTAT3(Unknown) Subpopulation (ODAS-pSTAT3(Unknown))

The ORR/DCR analysis set in the pSTAT3(Unknown) Subpopulation will include all randomized patients with measurable disease by RECIST 1.1 at randomization with study drug assignment designated according to initial randomization and having unknown pSTAT3 status.

5.4. QoL Analysis Set

5.4.1. QoL Analysis Set in the General Population (QoL-GP)

The QoL Analysis Set in the General Population (QoL-GP) will include all randomized patients who have at least 1 QoL assessment in the General Population. QoL-GP will be used for QoL endpoint analysis in the General Population. QoL assessments from 14 days prior to randomization and afterward are considered.

5.4.2. QoL Analysis Set in the pSTAT3(+) Subpopulation (QoL-pSTAT3(+))

The QoL Analysis Set in the pSTAT3(+) Subpopulation (QoL-pSTAT3(+)) will include all randomized patients who have at least 1 QoL assessment and having positive pSTAT3 status within a defined CSS window (6 months at IA or Final CSS window at FA). QoL-pSTAT3(+) will be used for QoL endpoint analysis in the pSTAT3(+) Subpopulation.

5.4.3. QoL Analysis Set in the pSTAT3(-) Subpopulation (QoL-pSTAT3(-))

The QoL Analysis Set in the pSTAT3(-) Subpopulation (QoL-pSTAT3(-)) will include all randomized patients who have at least 1 QoL assessment and having negative pSTAT3 status within a defined CSS window (6 months at IA or Final CSS window at FA). QoL-pSTAT3(-) will be used for QoL endpoint analysis in the pSTAT3(-) Subpopulation at IA and/or FA depending on decision from the IA.

5.5. PK Analysis Set

5.5.1. PK Analysis Set in the General Population (PK-GP)

The PK analysis set in the General Population (PK-GP) will include patients in the safety analysis set who have at least 1 quantifiable concentration in the General Population. PK-GP will be used for PK concentration presentations in the General Population.

5.5.2. PK Analysis Set in the pSTAT3(+) Subpopulation (PK-pSTAT3(+))

The PK analysis set in the pSTAT3(+) Subpopulation (PK-pSTAT3(+)) will include patients in the safety analysis set who have at least 1 quantifiable concentration and having positive pSTAT3 status tested within a defined CSS window (6 months at IA or Final CSS window at FA). PK-pSTAT3(+) will be used for PK concentration presentations in the pSTAT3(+) Subpopulation.

5.5.3. PK Analysis Set in the pSTAT3(-) Subpopulation (PK-pSTAT3(-))

The PK analysis set in the pSTAT3(-) Subpopulation (PK-pSTAT3(-)) will include patients in the safety analysis set who have at least 1 quantifiable concentration and having negative pSTAT3 status tested within a defined CSS window (6 months at IA or Final CSS window at FA). PK-pSTAT3(-) will be used for PK concentration presentations in the pSTAT3(-) Subpopulation.

5.6. Per-Protocol Analysis set

The Per-Protocol analysis set will consist of all ITT patients who do not have major protocol deviations that would impact the study outcome significantly as determined by the medical monitor and biostatistician prior to database unblinding.

Depending on the medical review, the major protocol deviations may include the following:

- Patients that are dosed in the study despite not satisfying the inclusion criteria;
- Patients that develop withdrawal criteria but are not withdrawn;
- Patients that receive the wrong treatment or an incorrect dose;
- Patients that receive an excluded concomitant medication;
- Deviation from good clinical practice (GCP).

Protocol deviations will be identified and maintained in a separate document, and the list of patients with protocol deviations including those with major protocol deviations that may be excluded from efficacy analysis will be finalized before database lock and unblinding.

Per-protocol analysis sets are defined for Overall Survival and for Progression Free Survival endpoints separately. The detailed exclusion was defined in Protocol Deviation Specs Version (226969_Protocol_Deviation_Specification_Form_PDSF_20200521_V4.0) dated 21 May 2020 by blinded review. The decisions for each patient to be included or excluded from the per protocol analysis sets will be determined by the medical monitor and biostatistician prior to database unblinding.

5.6.1. Per-Protocol Analysis Set in the General Population for Overall Survival (PPAS-GP-OS)

The Per-Protocol analysis set in the General Population for Overall Survival (PPAS-GP-OS) will consist of all ITT patients who do not have major protocol deviations that would impact the overall survival endpoint. Patients who randomized but not treated will be excluded from PPAS-GP-OS.

5.6.2. Per-Protocol Analysis Set in the pSTAT3(+) Subpopulation for Overall Survival (PPAS-pSTAT3(+)-OS)

The Per-Protocol analysis set in the pSTAT3(+) Subpopulation for Overall Survival (PPAS-pSTAT3(+)-OS) will consist of PPAS-GP-OS patients who have positive pSTAT3 status tested within the final CSS window.

5.6.3. Per-Protocol Analysis Set in the General Population for Progression Free Survival (PPAS-GP-PFS)

The Per-Protocol analysis set in the General Population for Progression Free Survival (PPAS-GP-PFS) will consist of all ITT patients who do not have major protocol deviations that would impact the PFS endpoint. Patients who randomized but not treated will be excluded from PPAS-GP-PFS.

5.6.4. Per-Protocol Analysis Set in the pSTAT3(+) Subpopulation for Progression Free Survival (PPAS-pSTAT3(+)-PFS)

The Per-Protocol analysis set in the pSTAT3(+) Subpopulation for Progression Free Survival (PPAS-pSTAT3(+)-PFS) will consist of PPAS-GP-PFS patients who have positive pSTAT3 status tested within final CSS window.

5.7. Treatment Allocation

Patients who were randomized, but not treated with study drug will be reported under their randomized treatment arm for efficacy analyses. However, by definition they are excluded from the safety analyses.

Patients who are randomized but received incorrect treatment will be reported under their randomized treatment for efficacy analyses and will be reported under the treatment they actually received for all safety analyses and treatment evaluations.

Table 6: High Level Summary of Analysis and Analysis Sets at Final Analysis

		General Population	pSTAT3(+)*	pSTAT3(-)*	pSTAT3(Unknown)!
Disposition/Baseline /MH		ITT-GP	ITT-pSTAT3(+)$	ITT-pSTAT3(-)	ITT-pSTAT3(Unknown)
Efficacy	OS/PFS	ITT-GP	ITT-pSTAT3(+)$	ITT-pSTAT3(-)	ITT-pSTAT3(Unknown)
	DCR/ORR/BOR	ODAS-GP	ODAS-pSTAT3(+)$	ODAS-pSTAT3(-)	ODAS-pSTAT3(Unknown)
	QoL	QoL-GP	QoL-pSTAT3(+)	QoL-pSTAT3(-)	
Safety	Safety/PE/VS/LB	SAS-GP	SAS-pSTAT3(+)	SAS-pSTAT3(-)	

*: pSTAT3(+) and pSTAT3(-) will be based on pSTAT3 status from all patients with specimen age <= final CSS window.

!: includes patients with unknown pSTAT3 status based on the final CSS window; patients with no specimen submitted for testing, specimen identified as non-evaluable upon pathology evaluation, tissues test out of a defined CSS window or missing due to any other reasons.

$: combined Subset 1 and Subset 2 patients will be used for primary analysis, sensitivity analysis based on Subset 1 will be performed.

6. DATA HANDLING

6.1. Clinical Cutoff and Analysis Cutoff

Clinical Cutoff date for the IA will be identified for this study based on when approximately 425 deaths (50% of total target number of deaths in ITT-GP) occur. Clinical cutoff date for the FA may vary depending on the outcome of the IA. When outcome from IA are to continue both hypotheses, clinical cutoff date will be based on when the total target number of deaths from ITT-pSTAT3(+) or 850 deaths in ITT-GP occur, whichever comes later. This means that the clinical cutoff date for the study at the FA will be a single calendar date, not differentiated by analysis set.

Based on the current event prediction for both General Population and pSTAT3(+) Subpopulation in a blinded fashion by an independent outside vendor who has access to biomarker status of individual patients but no treatment assignment information, the target number of deaths in pSTAT3(+) Subpopulation based on 6-month CSS window will happen prior to 850 death in General population. In this scenario, the analysis cutoff date of both ITT-pSTAT3(+) and General Population will be the same as clinical cutoff date, however, analysis cutoff date of ITT-pSTAT3(+)-Subset-2 needs to be a pre-specified calendar date before the final analysis cutoff date. This is to ensure the independence of the test statistics from the three components in the combination function approach as described in Section 4.5.2 .

Table 7 below presents the clinical cutoff dates in the study and the analysis cutoff dates for the two analysis populations in the scenario mentioned above.

Table 7: Clinical and Analysis Cutoffs in the scenario that 310 deaths in ITT-pSTAT3(+) happen earlier than 850 in ITT-GP

	Clinical Cutoff	Analysis Cutoff
General Population	Date when 850 deaths occur in GP*	Date when 850 deaths occur in GP*
pSTAT3(+) Subpopulation	Date when 850 deaths occur in GP*	Subset 1: Date when 850 deaths occur in GP* Subset 2: A pre-determined calendar date before Date when 850 deaths occur in GP

* Date when 850 deaths occur in GP will be pre-determined by prediction prior to database lock.

To operationally facilitate database lock process, the calendar dates for the clinical cutoff and analysis cutoff dates will be pre-determined based on continuous blinded data monitoring before the database lock.

All summaries and analyses will include all data pertaining to visits/assessments performed up to these analysis cutoff dates. Sensitivity analysis to include all data update to the clinical cutoff date will be performed as well.

6.2. Handling of Missing Values

6.2.1. Missing or Partial Death Dates

It is recommended that the database be designed to mandate a complete death date. If there is a record for death, but the date is missing or is partial, it will be imputed based on the last contact date.

- If the entire date is missing, the death date will be imputed as the day after the date of last contact.
- If the day or both day and month are missing, the death date will be imputed to the maximum of the full (non-imputed) day after the date of last contact, 1st day of the month and year of death, if day of death is missing OR January 1st of the year of death, if both day and month of death are missing.

6.2.2. Partial or Missing Start Date or End Date for Adverse Event or Medications

For the patient data listings, no imputation of incomplete dates will be applied. The listings will present the incomplete dates without any change.

The missing day of onset of an AE or start date of a therapy will be imputed as:

- first dose date if the month and year of the event is the same as first dose date;
- otherwise, the 1st of the month that the event occurred.

The missing day of resolution of an AE or end date of a therapy will be imputed as:

- the last day of the month of the occurrence. If the patient died in the same month, then set the imputed date as the death date.

Missing both day and month of an AE or start date of a therapy will be imputed as:

- the date of the first dose, if the onset year is the same as the year of the first dose date;
- otherwise, January 1 of the year of onset.

If the resolution date of an AE or end date of a therapy is missing both the day and month, the date will be imputed as:

- December 31 of the year of occurrence. If the patient died in the same year, then set the imputed date as the death date.

If the date is completely missing, then no imputation will be done and the event will be considered as treatment emergent (for AEs) or concomitant (for medications) unless the end date rules out the possibility.

6.2.3. Partial or Missing Date for New Anticancer Treatment

For complete missing of start date for new anti-cancer treatment, the date of last dose of study drug + 1 will be used as start date.

For partial dates of start date for new anti-cancer treatment,

- If day is missing, the 1st of the month will be assigned. Compared with date of last dose of study drug +1 and take the one which come later.
- If day and month is missing, the 1st of Jan will be assigned. Compared with the date of last dose of study drug +1, and take the one which come later.

6.2.4. Partial Date for Pathological Diagnosis

Dates missing the day, or both the day and month of the year will adhere to the following conventions in order to calculate the months from the first pathological diagnosis:

- The missing day of the first pathological diagnosis will be set to the first day of the month that the diagnosis occurred.
- If the date of the first pathological diagnosis is missing both the day and month, the diagnosis date will be set to January 1 of the year of diagnosis.

6.2.5. Missing Efficacy Endpoints

For primary and secondary efficacy analyses no values will be imputed for missing data. For time to event endpoints, non-event observations will be censored and for ORR/DCR, patients with no post-baseline tumor evaluations or missing baseline tumor evaluation will be counted as non-responders.

6.2.6. Missing PK/PD Values

The handling of BLQ values for nonlinear mixed-effects modeling (aka Population-PK, or Pop-PK) will be described in a separate analysis plan. For the purpose of data tabulation, summary statistics, and by-time point graphics, BLQ values will be treated as zero.

6.2.7. Missing QoL Data

For the EORTC QLQ-C30, if less than half of the constituent items on the QLQ-C30 have been answered for a multi-item subscale, that subscale will be considered as missing. Single-item subscales will be considered missing if the constituent item is incomplete.

6.2.8. Missing Biomarker Data

When analyzing the pSTAT3(+) Subpopulation, patients without biomarker data will be left as is. Imputation methods such as multiple imputation methods will be applied as a sensitivity analysis if needed.

6.3. General Data Handling

- Age will be computed from July 1 of the year of birth to the date of Informed Consent, as: (Date of Informed Consent - July 1 of the year of Birth +1)/365.25 round down to the nearest integer.

- Baseline Measurements:
  - Efficacy: The last non-missing measurement on or prior to the date of randomization will serve as the baseline measurement. In the event such a value is missing, the last assessment completed prior to or on the date of the first study drug administration (either napabucasin or FOLFIRI, whichever was administered first) will be used as the baseline assessment so long as this assessment was taken within 35 days of randomization (for example, baseline ECOG, baseline LDH). The tumor scans outside the time window (within 35 days of randomization) will be considered a protocol deviation and will be evaluated on a case by case basis.
  - Safety: The last non-missing assessment prior to or on the date of the first study drug administration (either napabucasin or FOLFIRI, whichever was administered first) will be used as the baseline assessment. In the case that no any study drug administrated, the last non-missing assessment prior to or on the randomization date will be used as baseline assessment.
  - Other baseline characteristics: The last non-missing measurement on or prior to the date of randomization will serve as the baseline measurement. In the event such a value is missing, the last non-missing assessment completed prior to or on the date of the first dose of study drug administration (either napabucasin or FOLFIRI, whichever was administered first) will be used as the baseline assessment.

- For the safety parameters: unless otherwise specified, post baseline reporting period includes non-missing records collected from after first dose of study treatment until 30 days following last dose of study treatment (napabucasin and/or FOLFIRI).

- Schedule and unscheduled visits post baseline: unless otherwise specified, for the summary by visit post baseline, only scheduled visits will be included; for summary of the maximum or maximum changes post baseline, all scheduled visits and unscheduled visits will be included. All records (scheduled/unscheduled) regardless post baseline reporting period or not will be listed in the data listings.

- Study Day:
  - For safety analysis: Study day is calculated as:
    - Assessment date – first dose date + 1; if the assessment was performed on or after the first dose day.
    - Assessment date – first dose date; if the assessment was performed prior to the first dose date.
  - For efficacy (time to event) analysis: Study day is calculated as:
    - Assessment date – randomization date + 1; if the assessment was performed on or after the randomization date.
    - Assessment date – randomization date; if the assessment was performed prior to the randomization date.

- Time-to-event: The event or censoring time (days) is calculated as:

Date of event/censoring – Date of randomization + 1

- Duration: Duration (except for duration of study treatment) is calculated as:

  - Duration in days: (End Date – Start Date + 1)
  - Duration in weeks: (End Date – Start Date + 1) / 7
  - Duration in months: (End Date – Start Date + 1) / 30.4375; Average days in months = average number of days in a year /12
  - Duration in years: (End Date – Start Date + 1) / 365.25; Average days in a year = 365.25, reflecting the Julian Year of 3 years with 365 days each and 1 leap year of 366 days.

7. STATISTICIAL METHODOLOGY AND STATISTCAL ANALYSES

7.1. Statistical Methods

Whilst every effort has been made to pre-specify all analyses in this statistical analysis plan, if any additional exploratory analyses are found to be necessary, the analyses and the reasons for them will be detailed in the clinical study report (CSR).

7.1.1. Analyses of Time to Event endpoints

Time to event endpoints will be summarized using the Kaplan-Meier method and displayed graphically when appropriate. Median event times and 2-sided 95% confidence interval for each median will be provided [Brookmeyer R and Crowley JJ, 1982] with log-log transformation.

Time to event endpoints will be compared between the treatments using log-rank tests. Cox proportion-hazards (PH) model will be fitted and the estimated hazard ratio and 2-sided 95% confidence interval will be provided.

Please see Section 7.2.2 for details on how log-rank tests and Cox-PH models will be applied for the General Population and the pSTAT3(+) Subpopulation differently.

In addition, Cox-PH models will be used to explore the potential influences of baseline stratification factors (listed in Section 3.4.1) and other baseline characteristics on time to event endpoints.

7.1.2. Analyses of Binary Endpoints

The rates of binary efficacy endpoints for the 2 arms will be compared using a Cochran-Mantel-Haenszel (CMH) test stratified for baseline stratification factors (listed in Section 3.4.1). In addition, point estimates of the rates for each treatment arm will be provided along with the corresponding exact 2-sided 95% CI based on the Clopper Pearson Method. The difference of point estimates between arms will be provided along with the 2-sided 95% CI based on Harmonic Means method adjusting stratification factor or normal approximation (unstratified).

7.1.3. Analyses of Continuous Data

Descriptive statistics, including the mean, standard deviation, median, minimum and maximum values will be provided for continuous endpoints.

7.1.4. Analyses of Categorical Data

The number and percentage of patients in each category will be provided for categorical variables. Change from baseline for Categorical data will be evaluated by shift table if appropriate. Chi-square test may be used for comparison if needed.

P-values greater than or equal to 0.0001 will be presented to 4 decimal places. P-values less than 0.0001 will be presented as “<0.0001”.

7.2. Statistical Analysis

7.2.1. Standard Analysis

All the standard analysis will be summarized for the General Population and the pSTAT3(+) Subpopulation at IA and FA. Similar information will be summarized for the pSTAT3(-) Subpopulation at the IA and/or FA depending on the IA outcomes.

7.2.1.1. Disposition of Patients

All patients enrolled (signed informed consent) will be summarized for the General Population and the pSTAT3(+) Subpopulation. The number and percentage of study patients will be tabulated for different analysis sets as specified in Section 5 .

The number and percentage of patients enrolled by site will be summarized by treatment arm in ITT-GP and/or ITT-pSTAT3(+) and/or ITT-pSTAT3(-) and/or ITT-pSTAT3(Unknown).

The number and percentage of patients reaching end of treatment (BBI-608, FOLFIRI, and bevacizumab if provided)) and end of study, along with discontinuation reasons will be summarized in ITT-GP and/or ITT-pSTAT3(+) and/or ITT-pSTAT3(-) and/or ITT-pSTAT3(Unknown).

7.2.1.2. Demographic and Other Baseline Characteristics

The following demographic and baseline characteristics will be summarized by treatment arm for patients in ITT-GP and/or ITT-pSTAT3(+) and/or ITT-pSTAT3(-) and/or ITT-pSTAT3(Unknown).

- Age, age group (<65 years vs. ≥65 years), Age (<65 years, ≥ 65 – 75 years, ≥ 75 – 85 years and ≥85 years), sex, race, ethnicity, ECOG status, and smoking history
- Weight (kg), height (cm), child-bearing potential, and body mass index (BMI)

Demographic and baseline characteristics will also be listed for each patient.

7.2.1.3. Disease Characteristics, Medical History, and Primary Therapy

Disease characteristics including, but not limited to, medical history, and prior therapy will be summarized by treatment arm for the ITT-GP and/or ITT-pSTAT3(+) and/or ITT-pSTAT3(-) and/or ITT-pSTAT3(Unknown).

- Primary Type of Cancer (Colon vs Rectal); Primary Tumor Present or Resected; Primary Tumor Location, Histology Sub-classification
- Disease Measurability (Present vs Not Present); all target and non-target lesions: presence of target lesions, site of disease, numbers of disease sites; target lesions: number of target lesions, largest measure, site of disease, method of assessment
- KRAS Mutation Status (Mutant vs Wild Type), NRAS Mutation status (Mutant vs Wild Type, Unknown)
- Prior radiotherapy (Yes vs No); Anatomical Target of Procedure; Indication

- Prior cancer system medication (Yes vs No); Prior Systemic Treatment: 1st line, neoadjuvant/adjuvant and 1st line, neoadjuvant/adjuvant, and other; prior exposure to specific types of therapy: Oxaliplatin, Fluoropyrimidine, and Bevacizumab; number of unique therapeutics per patients: therapeutic received (based on ingredients coded from Who Drug Coding Dictionary)
- Prior Cancer Surgery by SOC and Preferred Terms of Medical Dictionary for Regulatory Activities (MedDRA)
- The number and percentages of patients within different specimen age duration (≤ 6 month, > 6 month to Final CSS window and Patients out of final CSS window) will be summarized by Biomarker status
- Reasons for ITT-pSTAT3(unknown) will be summarized.
- Tissues sample collected from EDC are summarized: tumor tissue sample type (block, core and slides, and slides); sample acquisition type (resection, biopsy), site (primary and metastatic). When multiple records reported for one patient, the record was evaluated for biomarker status from vendor will be used in the summary.

7.2.1.4. Protocol Deviation

A summary of the total number of major protocol deviation events and the number and percentage of patients in each type of major protocol deviation will be provided by treatment group and overall, in ITT-GP and/or ITT-pSTAT3(+) and/or ITT-pSTAT3(-).

All major or minor Protocol Deviation will be listed and PD due to COVID-19 pandemic will be flagged.

7.2.1.5. Study Treatment

An overall summary of drug exposure will be presented including number of maximum treated cycles, numbers and percentages of patients who had 1, 2, etc. Maximum treated cycles by treatment arm for SAS-GP and/or SAS-pSTAT3(+). A treated cycle for a specific drug is defined as a cycle in which the patient received any amount of any study drug (napabucasin and or FOLFIRI).

Duration of exposure =Earliest date of {Latest date of {(start date of last cycle +13 days), last dose date of napabucasin}, death date, end of study date} – Earliest date of {first dose date of the FOLFIRI and/or napabucasin} +1.

The following items will be summarized for BBI-608, FOLFIRI, and bevacizumab where appropriate as below: dose reduction, dose interruption and dose modification.

7.2.1.6. Exposure of BBI-608

The details of BBI-608 dose modification in protocol 7.1.7 as below.

Table 8: BBI-608 Dose Modification Table

Dose Level	Dose
Full dose	240 mg twice daily (q12h)
Modification Level-1	80 mg twice daily (q12h), up-titrate as tolerated**
Modification Level-2	80 mg once daily*, up-titrate as tolerated**
* If 80 mg once daily is not tolerated, a dose interruption of 1-3 days followed by re-challenge at 80 mg once daily is recommended. ** Morning and evening doses can be increased in 80 mg increments every 3-7 days or slower as tolerated, up to 240mg 2 times daily.	

Several measures of drug administration will be summarized.

Cumulative dose, actual dose intensity and relative dose intensity of BBI-608 will be summarized for Arm 1 by overall, run-in and first 28 days.

Treatment duration in days:

- Overall intended treatment duration = earliest date of {latest date of {start date of last cycle +13 days, last dose date of napabucasin}, death date, end of study date} – earliest date of {first dose date of the FOLFIRI and/or napabucasin} +1.
- Run-in treatment duration = actual number of days in run-in period where run-in period is defined as napabucasin treatment duration prior to first dose of FOLFIRI.
- First 28 days intended treatment duration = earliest date of {start date of 1st cycle + 27 days, death date, end of study date}– start date of 1st cycle +1.

Cumulative dose (mg) in overall or run in or first 28 days is defined as the total dose (mg) that a patient received for napabucasin in overall, run-in or first 28 days, respectively.

Actual dose intensity (DI) in mg/day:

- Overall actual DI = overall accumulative dose in mg / overall intended treatment duration in days.
- Run-in actual DI = actual total dose in mg in run-in / actual treatment duration in days in run-in.
- First 28 days actual DI = actual total dose in mg in first 28 days / intended treatment duration in days in first 28 days.

Relative dose intensity (%) in overall, run-in or first 28 days is defined as the actual DI in (mg/day) in overall, or run-in or first 28 days / 480 (mg/day) * 100.

Dose Changes

The details of BBI-608 dose change are outlined below:

- Dose Held - 2 or more consecutive doses held (AM-PM or PM-AM).
- Dose Held Due to Non-AE - 2 or more consecutive doses held (AM-PM or PM-AM) due to reasons other than an adverse event (eg, subject forgot, subject decision or other reason).
- Dose Held Due to AE - 2 or more consecutive doses held (AM-PM or PM-AM) due to adverse event.
    - Restarted at the same dose - Restarted at a dose same as the previous non-zero dose at the time of dose held.
    - Restarted as reduced dose (dose-decrease) - Restarted at a dose decreased from the previous non-zero dose at the time of dose held.
        - Dose increase to partial dose following reduction - An increase to <240 mg for 2 or more consecutive doses (AM-PM or PM-AM) following restart at reduced dose.
        - Dose increase to full dose following reduction - An increase to 240 mg for 2 or more consecutive doses (AM-PM or PM-AM) following restart at reduced dose.

7.2.1.7. Permanently Discontinued Due to AE: based on drug administration form. Exposure of FOLFIRI

FOLFIRI Dose Modification is detailed in following table (Protocol Section: 7.1.5)

Table 9: FOLFIRI Dose Modification Table

Dose Level*	Irinotecan**	5-Fluorouracil Bolus	5-Fluorouracil Infusion
Full dose	180 (mg/m^2)	400 (mg/m^2)	1200 (mg/m^2/day) (total 2400 mg/m^2)
Modification Level-1	150 (mg/m^2)	200 (mg/m^2)	900 (mg/m^2/day) (total 1800 mg/m^2)
Modification Level-2	120 (mg/m^2)	N/A	600 (mg/m^2/day) (total 1200 mg/m^2)
Modification Level-3	100 (mg/m^2)	N/A	N/A

* If the dose of any component of FOLFIRI is reduced because of potentially-related AEs, subsequent dose increases are not permitted.
** If irinotecan dose is required to be reduced below 100 mg/m^2, lower doses may be used at the discretion of the Investigator.

The FOLFIRI regimen includes the three components of irinotecan, leucovorin and 5-FU. Component specific exposure will be calculated for irinotecan, 5-FU Bolus and 5-FU Infusion separately.

Cumulative dose, actual dose intensity and relative dose intensity for the specific component will be summarized for ARM 1 and ARM 2 overall and first 28 days.

Intended treatment duration in weeks:

- Overall intended treatment duration = (earliest date of {latest date of {start date of last cycle +13 days, last dose date of napabucasin}, death date, end of study date} – earliest date of {first dose date of the FOLFIRI} +1)/ 7.
- First 28 days intended treatment duration = (earliest date of {start date of 1st cycle + 27 days, start date of FOLFIRI within Day 28 ±3 window -1, start date of 1st cycle + 13 days if patient died or end of study before 2nd dose of FOLFIRI}– start date of 1st cycle +1)/ 7.

Cumulative dose (mg/m2) in first 28 days or overall is defined as the total dose (mg/m2) that the patient received for the specific component (Irinotecan, 5-FU Bolus and 5-FU Infusion, Leucovorin/Levoleucovorin) in first 28 days or overall, respectively.

Actual dose intensity (mg/m2/week) in first 28 days or overall = cumulative dose for the specific component in mg/m2 in first 28 days or overall / intended treatment duration in weeks in first 28 days or overall, for the specific component. Relative Dose Intensity (%) in first 28 days or overall = actual dose intensity in first 28 days or overall / targeted dose in weeks for the specific component * 100. Targeted dose is 90 mg/m2/week for Irinotecan, 1400 (mg/m2/week) for 5-FU Bolus and 5-FU Infusion, 200 mg/m2/week for Leucovorin, 100 mg/m2/week for Levoleucovorin.

In order to calculate dose intensity for Leucovorin/Levoleucovorin together, the Levoleucovorin dose will be converting to the equivalent leucovorin by multiplying of 2.

The BSA to be used for calculating each dose of specific component will be calculated using the Dubois & Dubois formula (see below) based on the last available weight and height prior to each infusion.

BSA [m2] = (Weight [kg]$^{0.425}$ * Height [cm]$^{0.725}$) * 0.007184

7.2.1.8. Exposure of Bevacizumab

Exposure of bevacizumab will be summarized for the patients who are determined to receive bevacizumab by the investigator.

Cumulative dose, actual dose intensity and relative dose intensity for bevacizumab will be summarized for ARM 1 and ARM 2 overall.

Overall intended treatment duration in weeks = (Start date of the last cycle of bevacizumab + 13 days - First dose date of bevacizumab + 1) / 7. Overall cumulative dose (mg/kg) = total dose (mg/kg) that the patient received for bevacizumab.

Actual overall dose intensity (mg/kg/week) = overall cumulative dose in mg/kg / overall intended treatment duration in weeks.

Relative dose Intensity (%) overall = actual overall dose intensity in mg/kg/week / targeted dose (2.5 mg/kg/week.) for bevacizumab * 100.

The weight in kg to be used for calculating RDI is derived on the last available weight prior to each infusion.

Cumulative dose, actual dose intensity and relative dose intensity for bevacizumab will be summarized for ARM 1 and ARM 2 overall.

7.2.2. Primary Analyses of Primary Endpoints

Overall survival will be analyzed in ITT-GP and/or ITT-pSTAT3(+). For ITT-GP, difference in OS between Arm 1 and Arm 2 will be analyzed by the log-rank test stratified by actual stratification factors (Section 3.4.1). And 1-sided p-value from stratified log-rank test will be reported. If any actual stratification factor proves to have an inadequate sample size, it will be dropped from the primary analysis or pooling will be applied. Details will be determined in a blinded fashion prior to any planned unblinded analysis or data base lock.

Note that the sample size by stratification factor was reviewed and the following pooling strategy was decided prior to interim analysis in the blinded fashion. Tumor RAS status, Time to Progression from start of 1st Therapy, and Bevacizumab as Part of Study Protocol Treatment will be the stratification factors in the primary analysis in the general population. Geographic Region and Location of the Primary Tumor (left vs right) will be dropped from the primary analysis. For pSTAT3(+) subpopulation, the primary analysis will be based on unstratified analysis (for log rank test or cox regression model). This is because the stratified randomization is not conducted in this subpopulation. Stratified analysis including three factors of RAS status, Time to PD, and Bevacizumab as part of study treatment will serve as sensitivity analysis.

For ITT-pSTAT3(+), the 3 stage-wise p-values under the combination function approach will be produced from three parts of data (Subset 1 Stage 1, Subset 1 Stage 2 and Subset 2). Since Subset 2 will likely contain some sparse cells of its stratification factor combination due to a smaller sample size, the primary analysis for ITT-pSTAT3(+) will be based on unstratified analysis: difference in OS between Arm 1 and Arm 2 will be analyzed by the unstratified log-rank test. And 1-sided p-value from unstratified log-rank test will be reported.

Estimates of the OS curves obtained from the Kaplan-Meier method will be presented and displayed graphically. The median survival time and corresponding 2-sided 95% CI will be provided for each treatment arm.

Cox PH model stratified by actual stratification factors will be fitted for ITT-GP. Cox PH model unstratified will be fitted for ITT-pSTAT3(+). The estimated hazard ratio and 2-sided 95% CI will be provided.

The 3 months, 6 months, 9 months, 12 months year, 18 months, 24 months, 30 months and 36 months survival probability will be estimated using the Kaplan-Meier method and a 2-sided 95% CI for the log(-log(survival probability)) will be calculated using a normal approximation and then back transformed to give a CI for survival probability itself.

7.2.3. Sensitivity Analyses of the Primary Endpoints

7.2.3.1. Sensitivity Analysis for Overall Survival in ITT-GP and/or ITT-pSTAT3(+)

Sensitivity analysis for OS will be performed for ITT-GP and/or ITT-pSTAT3(+). Sensitivity Analysis of OS includes:

The analyses of OS in the per-protocol population will be performed. An un-stratified log-rank test and Cox PH model may be used as sensitivity analyses for OS for ITT-GP. A stratified log-rank test/stratified Cox PH model may be performed based on randomization stratification factors for ITT-GP.

Stratified log-rank test and stratified Cox PH model based on actual stratification factor may be performed as sensitivity analyses for OS for ITT-pSTAT3(+) if possible.

The potential influence of the stratification factors and other baseline characteristics may be evaluated by sensitivity analysis.

Evaluation of the impact on OS of new anti-cancer therapy may be performed. The patients who receive subsequent (new) anti-cancer therapy may be excluded from the OS analysis or survival times may be censored at the time of new anti-cancer therapy. If applicable, other sensitivity analyses including the time-dependent Cox PH model may be used to adjust for the impact of new anti-cancer therapy.

Due to the open label nature of this study, the early terminated rate (early termination rate defined as patients who terminated the study before taking any dose among all randomized patients) in Arm 1 and Arm 2 may be differentially influenced by the knowledge of the treatment. Early terminated rate and reasons for premature discontinuation will be summarized between Arm 1 and Arm 2. If applicable, appropriate sensitivity analysis may be conducted to assess the impact on OS by imbalance of early termination between the two treatment arms.

Subgroup analysis will be performed considering the subgroup specified in Section 3.4.1 and 3.4.2 or other subgroups where appropriate.

For each subgroup level, the median OS (or other quartiles) and a 2-sided 95% CI will be provided for each arm. The difference in OS between Arm 1 and Arm 2 may be analyzed by unstratified log-rank test. The nominal 2-sided p-value may be provided in an exploratory basis. The estimated hazard ratio and 2-sided 95% CI may be provided based on un-stratified Cox PH model.

Forest plots of hazard ratios from the primary analysis or other subgroup analyses will be provided where appropriate.

For pSTAT3(+) Subpopulation, sensitivity analysis using overall survival data from Subset 1 patients only (i.e., data within CSS window of 6 months) will be also performed at the FA.

7.2.3.2. Analysis for Overall Survival in ITT-pSTAT3(-) and ITT-pSTAT3(Unknown)

At the IA, overall survival for ITT-pSTAT3(-) and ITT-pSTAT3(Unknown) will be performed using similar method for primary analysis for primary endpoints (Section 7.2.2) and other sensitivity analysis as mentioned in Section 7.2.3.1 will be applied if needed. Depending on the

decision from the IA, similar analysis as Section 7.2.2 and Section 7.2.3.1 may be performed at the FA.

7.2.4. Analyses of Key Secondary Endpoints

PFS will be summarized in ITT-GP and/or ITT-pSTAT3(+). For ITT-GP, difference in PFS between treatment arms will be analyzed by the log-rank test stratified by actual stratification factors (Section 3.4.1). And 1-sided p-value from stratified log-rank test will be reported. The strategy of pooled or dropping is the same as the primary analysis of OS. For ITT-pSTAT3(+), similar to the primary analysis of OS, primary analysis for PFS will be based on unstratified log-rank test. And 1-sided p-value from unstratified log-rank test will be reported. Estimates of the PFS curves obtained from the Kaplan-Meier method will be presented. PFS curves will be displayed graphically. The median event time (and other quartiles) and corresponding 2-sided 95% CI will be provided for each treatment arm. The Cox PH model stratified by actual stratification factors will be fitted for ITT-GP. The Cox PH model unstratified will be fitted for ITT-pSTAT3(+). The estimated hazard ratio and 2-sided 95% CI will be provided.

Sensitivity analysis for PFS will be performed for ITT-GP and/or ITT-pSTAT3(+).

Sensitivity Analysis of PFS includes:

The analyses of PFS in the per-protocol population will be performed for both general population and pSTAT3(+) subpopulation. An un-stratified log-rank test and Cox PH model will be used as sensitivity analyses for PFS for ITT-GP. A stratified log-rank test/stratified Cox PH model will be performed based on randomization stratification factors for ITT-GP.

Stratified log-rank test and stratified Cox PH model based on actual stratification factor may be performed as sensitivity analyses for PFS for ITT-pSTAT3(+) if possible.

Subgroup analysis will be performed considering the subgroup specified in Section 3.4.2 or other subgroups where appropriate. Similar analysis methods as subgroup analysis for OS will be applied.

Forest plots of hazard ratios from the primary analysis or other subgroup analyses will be provided where appropriate.

For pSTAT3(+) Subpopulation, sensitivity analysis using PFS data from Subset 1 patients only (i.e., data within CSS window of 6 months) will be also performed at the FA.ORR will be summarized in ODAS-GP and/or ODAS-pSTAT3(+) by treatment arm. The ORR will be summarized for each treatment arm along with the corresponding exact 2-sided 95% CI using Clopper Pearson method. For ODAS-GP, differences of ORR between the two arms will be compared using a 1-sided Cochran-Mantel-Haenszel test stratified by actual stratification factors. Treatment difference of ORR and its 95% CI based on harmonic mean method adjusting actual stratification factors will be provided. For ODAS-pSTAT3(+), differences of ORR between the two arms will be compared using a 1-sided Z-test via normal approximation. Treatment difference of ORR and its 95% CI based on normal approximation (unstratified) will be provided. Best Overall Response will be summarized in ODAS-GP and/or ODAS-pSTAT3(+) by treatment arm. The best response of stable disease (SD) can be assigned if SD criteria were met at least once after randomization at a minimum interval of 8 weeks – 5 days window from randomization.

DCR will be analyzed similar as described for ORR for in ODAS-GP and/or ODAS-pSTAT3(+). At IA, PFS for ITT-pSTAT3(-) and DCR, ORR and BOR in ODAS-pSTAT3(-) will be analyzed using similar methods as above. Depending on the decision from the IA, similar analysis may be performed at the FA.

Sensitivity analysis for ORR/DCR will be performed for ODAS-GP and/or ODAS-pSTAT3(+) including:

For ODAS-GP, sensitivity analysis of ORR/DCR comparison between arms based on randomization stratification factors will be performed. The differences of ORR/DCR between the two arms will be performed unstratified (without adjusting stratification factors).

For ODAS-pSTAT3(+), sensitivity analysis of ORR/DCR comparison between arms based on actual stratification factors will be performed.

Subgroup analysis will be performed considering the subgroup specified in Section Section 3.4.1 and Section 3.4.2 or other subgroups where appropriate:

For each subgroup level, differences of ORR/DCR between the two arms will be compared using a 1-sided Z-test via normal approximation. Treatment difference of ORR and its 95% CI based on normal approximation (unstratified) will be provided.

Forest plots of risk differences from ORR/DCR for different subgroups will be provided where appropriate.

For pSTAT3(+) Subpopulation, sensitivity analysis using PFS/DCR/ORR data from Subset 1 patients only (i.e., data within CSS of 6 months) will be also performed at the FA.

7.2.5. Analyses of Other Secondary Endpoints

Quality of Life (QoL) Endpoints will be summarized by treatment arm and by timepoint in QoL-GP and QoL-pSTAT3(+).

The endpoints in QoL analysis are change from baseline at time 2 (Cycle 5 Day 1) and time 4 (Cycle 9 Day 1) for the physical function, global health status/quality of life subscale scores of the EORTC QLQ-C30 QoL.

The mean scores and change from baseline for all QoL scales and time point will be summarized descriptively for EORTC QLQ-C30.

7.2.6. Statistical Analysis for Safety Endpoints

Safety data will be analyzed in SAS-GP and/or SAS-pSTAT3(+). Similar analysis will be performed at IA for SAS-pSTAT3(-). Depending on the decision from the IA, similar analysis maybe performed at the FA.

7.2.6.1. Adverse Event

Overall Summary of AEs

An AE will be regarded as treatment-emergent, if

- It occurs for the first time on or after the first dose date of either BBI-608 or FOLFIRI and up to 30 days after the last dose of study treatment; or

- It occurs prior to the first dose date of either BBI-608 or FOLFIRI but worsens in severity during therapy or for up to 30 days after the last dose of study treatment (or up to any time if a serious AE and considered related to study drug.

The emphasis for AE analysis will be based on treatment-emergent AE (TEAE), however all AEs will be listed regardless of TEAE or not.

All AEs will be coded using Medical Dictionary for Regulatory Activities (MedDRA version 19.0) and summarized by MedDRA System Organ Class and Preferred Term. The severity of all AEs will be graded by the Investigator using NCI CTCAE version 4.0.

An overall summary of TEAEs will be provided. The number and percentage of patients with following will be summarized by ARM 1 and ARM2:

- Patients with at least 1 TEAE
- Patients with TEAE of CTCAE Grade 3 or higher
- Patients with serious TEAE
- Patients with BBI-608 related TEAE
- Patients with BBI-608 related TEAE of CTCAE Grade 3 or higher
- Patients with FOLFIRI related TEAE
- Patients with FOLFIRI related TEAE of CTCAE Grade 3 or higher
- Patients with either BBI-608 or FOLFIRI related TEAE
- Patients with either BBI-608 or FOLFIRI related TEAE of CTCAE Grade 3 or higher
- Patients with TEAE leading to BBI-608 dose interruption, reduction, permanent discontinuation
- Patients with TEAE leading to FOLFIRI interruption, reduction, permanent discontinuation

Summary of AEs by SOC and Preferred Term

The number and percentage of patients who experience any TEAE will be summarized by SOC and PT. A summary of TEAEs by PT and maximum CTCAE grade will be presented. TEAE preferred terms with missing grade will be reported in the Unknown category if no other grade was found for the same PT. The most commonly reported TEAEs using a different cutoff (e.g. 2% or 10% or more of patients in either arm) may also be summarized by PT as needed. Meaningful differences between treatment arms (reported an incidence of ≥5% of patients in either arm or difference by >=2%) will be compared.

TEAEs associated with permanent discontinuation /dose reduction/drug interruption of either BBI-608 and/or FOLFIRI will be summarized by SOC and PT and maximum CTCAE Grade (taking into consideration the action taken from CRF AE page).

Treatment Related Adverse Events

Treatment-related TEAEs are those judged by the investigator to be at least possibly related to the study drugs (BBI-608 and/or FOLFIRI) or for which relatedness is recorded as "unknown" by the investigator. Treatment related will include "possible", "probable" or "definite" causality assessments. Missing relationship will be considered as "Related" to all drugs received by the patient. Relationship to FOLFIRI is obtained from the CRF which includes "related to any of the components". Relationship to each component may be summarized, if necessary. Similar summaries as noted for all causality TEAEs will be provided for treatment related TEAEs.

Serious Adverse Events and Death

Treatment-emergent SAEs and Treatment-related SAEs will be summarized by MedDRA SOC and PT and maximum CTCAE grade.

TE AE leading to death will be summarized by MedDRA SOC and Preferred Term.

Deaths will be summarized on study and on treatment (within 30 days of last dose of study medication). The number and percentage of patients who died on study and on treatment, as well as primary cause of death will be summarized. Deaths may also be categorized according to time of occurrence after first dose.

A listing of death data will also be provided and will include all deaths that occurred from the signing of the informed consent to the end of the follow up period. The listing will include primary cause of death and the number of days relative to the administration of first and last dose.

Adverse Events of COVID-19.

Preferred Terms corona virus infection and coronavirus test positive will be used to identified COVID-19 positive cases in adverse event.

Clustered Adverse Events

Specific AEs, termed AE clusters, have been selected based upon the known and/or potential association with napabucasin administration. MedDRA (version 19.0) standard medical queries (SMQs) and/or Sponsor selected Preferred terms (PTs) have been identified to review safety data from napabucasin trials and assess for incidence, severity and causality. The selected MedDRA SMQ or sponsor derived terms are:

- Noninfectious diarrhea (narrow SMQ)
- Gastrointestinal nonspecific inflammation (narrow SMQ) plus Enterocolitis
- Gastrointestinal Hemorrhage (narrow SMQ)
- Abdominal Pain (Sponsor derived event cluster)
- Fatigue (Sponsor derived event cluster)
- Acute renal failure (narrow SMQ)
- Hypotension (Sponsor derived event cluster)
- Gastrointestinal Perforation (narrow SMQ)

- Drug related hepatic disorders [Drug related hepatic disorders - severe events only (narrow SMQ) plus Liver related investigations, signs and symptoms (narrow SMQ)]
- Haematopoietic leukopenia (broad SMQ)
- Haematopoietic thrombocytopenia (broad SMQ)
- Torsade de Pointes (Torsade de pointes/QT Prolongation (broad SMQ) plus PT Seizure)

Similar to the overall summary of TEAEs, an overview of clustered TEAEs will be provided. The clustered TEAEs, clustered TEAEs related to study drug, and clustered treatment emergent SAE will be summarized by PT in SAS-GP and/or SAS-pSTAT3(+).

7.2.6.2. Clinical Laboratory Assessments

Lab parameters collected in CRF will be summarized.

All records on or after first dose (planned and unplanned) will be considered as post baseline. The results of laboratory parameters will be graded according to NCI CTCAE v4.0. The CTCAE grading will be performed based on observed value without considering any clinical symptoms or findings. CTCAE grade at baseline and maximum CTCAE grade post baseline will be summarized. A shift summary of baseline grade to maximum CTCAE grade post baseline will be summarized. Unplanned laboratory test results will be included in maximum CTCAE grade post baseline and the shift tables.

Table 10: CTCAE Terms for Grading

Hematology	Biochemistry
White blood cell decreased Leukocytosis Anemia Platelet count decreased Neutrophil count decreased Lymphocyte count decreased Lymphocyte count increased	Creatinine increased Blood bilirubin increased Alanine aminotransferase increased Alkaline phosphatase increased Aspartate aminotransferase increased Hypoalbuminemia Hyperkalemia Hypokalemia Hypermagnesemia Hypomagnesemia Hypophosphatemia Lactate dehydrogenase

Additional CTCAE terms may be added if data are available.

A listing of all lab parameters including hematology, biochemistry, and urinalysis will be provided, including the test result, units, normal range (H and L), change from baseline, and CTCAE grades if graded.

Incidence of patients with liver function test results satisfying the drug-induced liver injury (DILI) criterion defined as (> 3x upper limit of normal [ULN] for ALT/AST, >2xULN for total bilirubin and < 2xULN or missing for alkaline phosphatase within a day) will be presented.

In addition, incidence of elevated liver function test results will be presented by elevation criterion. Elevation criteria are given as follows:

- ALT (> 3xULN, > 5xULN, >8xULN, > 10xULN, > 20xULN)
- ALT >3xULN and Baseline ALT <=3xULN (or Baseline ALT missing)
- AST (> 3xULN, > 5xULN, >8xULN, > 10xULN, > 20xULN)
- AST > 3xULN and Baseline AST <= 3xULN (or Baseline AST missing)
- Total Bilirubin >2xULN
- Total Bilirubin >2xULN and Baseline Total Bilirubin <= 2xULN or missing
- ALP (> 1.5xULN, >= 3xULN)
- ALP >= 3xULN and Baseline ALP < 3xULN or missing

7.2.6.3. Concomitant Medication/Post-Treatment Anti-Cancer Therapy

Prior and concomitant medications will be coded to ATC (Anatomical Therapeutic Chemical) classification and Drug Name using WHO Drug Dictionary (WHO-DD-Enhanced B2 MAR 16). Prior medication is defined as medication taken any time prior to the date of the first dose of study medication. Concomitant medication is defined as medications taken any time during treatment period (i.e., first dose date to last dose date of study medication, inclusive).

Summaries of concomitant medications will be provided by level 3 ATC classification, preferred term and level 4 ATC classification using frequencies and percentages. A listing of all prior and concomitant medications will be provided.

New anti-cancer therapy will be summarized by ingredient and preferred term using frequencies and percentages.

7.2.6.4. Vital Signs

For weight, heart rate, systolic blood pressure (BP), and diastolic BP, summary statistics for baseline values and maximum change (maximum increase, maximum decrease. and no change) from baseline will be presented.

Changes from baseline to on-treatment weight assessments will be presented by treatment and time point considering the frequency of patients with changes falling in the following categories: >= 5% -< 10% gain, >=10% -< 20% gain, >= 20% gain, >= 5% -< 10% loss, >=10% -< 20% loss, >= 20% loss.

Summaries of markedly abnormal vital sign parameters, including BP and pulse, will be presented by treatment group. For systolic, diastolic blood pressure and BMI, shift from baseline will be provided by for the worst value on the study using the following categories (see Table 11).

Values for vital signs for all patients will be presented in a listing, and patients with markedly abnormal values will be flagged.

Markedly abnormal ranges for vital sign parameters are given in the table below.

Table 11: Markedly Abnormal Ranges for Vital Sign

Vital Sign Parameter	Markedly Abnormal (Low)	Markedly Abnormal (High)
Systolic BP	Absolute value ≤ 90 mmHg, or a decrease from baseline ≥20 mmHg	Absolute value ≥ 180 mmHg, or an increase from baseline ≥20 mmHg
Diastolic BP	Absolute value ≤ 50 mmHg, or a decrease from baseline ≥15 mmHg	Absolute value ≥ 105 mmHg, or an increase from baseline ≥ 15 mmHg
Pulse	Absolute value ≤ 50 bpm, or a decrease from baseline ≥15 bpm	Absolute value ≥ 120 bpm, or an increase from baseline ≥15bpm
BMI	Absolute value ≤ 18 kg/m^2	Absolute value ≥ 25 kg/m^2

7.2.6.5. ECG Evaluation

12-lead ECG with categorical results (Normal, Abnormal Not Clinically Significant, Abnormal Clinically Significant) will be summarized by treatment and visit. The shift from baseline to worst post baseline will be produced. A patient listing will also be provided.

7.2.6.6. Physical Exam

Physical examination abnormalities will be summarized for each visit by body system and by treatment group. Patients with clinically significant abnormal findings will be flagged in the data listing.

7.2.6.7. ECOG

The number and percentage of patients in each category of ECOG will be listed for baseline and worst post baseline value. In addition, ECOG will be summarized in a shift table from baseline to worst post baseline for safety analysis set.

7.2.7. Subgroup Analysis for PMDA Submission

To support the PMDA submission in Japan, selected analyses will be repeated for patients enrolled in Japan in this study. The following analyses may be included:

- Standard analyses as described in Section 7.2.1;
- Analysis for primary endpoint as in Section 7.2.2, sensitivity analysis for the primary endpoint if appropriate for Japan in Section 7.2.3;
- Other secondary endpoints analyses as described in Section 7.2.4 and Section 7.2.5;
- Safety analyses as described in Section 7.2.6.

7.2.8. PK Analysis

A listing of sparse PK sample collection includes actual sampling time relative to the BBI-608 dose administration on the day of the PK sample collection will be provided.

The procedures for the population PK modeling including model evaluation, will be described in a separate Population-PK analysis plan. The corresponding results will be reported in a separate document.

7.2.9. Sensitivity Analysis on Overall Survival in pSTAT3(+) w.r.t. Baseline Imbalance

As randomization is not stratified based on the pSTAT3 status, there could be some numerical imbalance in baseline covariates including the stratification factors in the pSTAT3+ Subpopulation. Sensitivity analyses on OS in the pSTAT3(+) Subpopulation as described in this section may be performed at the IA and/or the FA. Baseline covariates including, but not limited to, the baseline demographic, baseline disease characteristics, baseline stratification factors and baseline subgroups will be summarized descriptively by treatment arms in the pSTAT3(+) Subpopulation.

Propensity Score Methods [Rosenbaum and Rubin, 1983, Yue L, 2007] will be applied to evaluate the potential imbalance between the two treatment arms in the pSTAT3(+) Subpopulation. Propensity score analysis is a versatile statistical method for improving treatment comparisons by adjusting for a relatively large number of potentially confounding covariates.

Let $T = 1$ indicate patients with the treatment ARM1, $T = 0$ indicate patients with the treatment ARM2, x indicates a vector of the pre-specified covariates, and $e(x)$ denotes the probability of being assigned to ARM1 given the covariates x,

$$e(x) = P(T = 1|x)$$

Here $e(x)$ is the propensity score, which will be estimated for each patient given their covariates

The clinically relevant baseline factors to be included in the propensity score analysis are pre-specified and listed in Table 12.

Table 12: Clinically Relevant Baseline Factors

Stratification Factors (actual)	Geographical Region	Japan/Korea vs ROW vs WE/AUS/NA
	Time to progression on first line therapy	< 6 Months vs >= 6 Months
	RAS mutation status	Mutant vs Wild Type
	Bevacizumab as part of study protocol treatment	Yes vs No
	Location of the primary tumor	Right vs Left
Subgroup	Age	>= 65 years vs < 65 years
	Sex	Male vs Female
	Presence of liver metastases	Yes vs No
	ECOG Performance Status	0 vs 1
	Primary Tumor Site	Colon vs Rectum
	Prior Bevacizumab	Yes vs No
Baseline Disease Characteristic	BMI(kg/m2)	<= 25 vs >25

	Primary Tumor	Present vs Resected
	Initial TMN Staging	I/II vs III vs IV
	Any Prior Cancer Surgery	Yes vs No
	Any Priory Radiotherapy	Yes vs No
Baseline lab	Baseline LDH	Low/Normal vs High

The following steps will be implemented to perform the sensitivity analyses.

Step 1: The baseline factors will be summarized by two arms.

Step 2: Propensity score calculation. Multiple logistic regression models will be fitted with the treatment assignment indicator T as the response variable. The variables in Table 12 will be included as covariates.

Step 3: Evaluate the distribution of propensity score calculation from two arms. Propensity scores (estimated probability) from the multiple logistic regression models will be summarized between two treatment groups descriptively and graphically by box plots. The propensity score is a summary score, which may provide an indicator how the baseline covariatesin the propensity score calculation is equally (or unequally) distributed between the 2 treatment arms.

Step 4: Sensitivity analysis for an outcome adjusting the propensity score. Sensitivity analysis for an outcome (e.g., OS) considering propensity scores adjustment may provide an assessment of the outcome adjusted by any potential imbalances in the baseline covariates.

The Propensity Score may be used in 2 ways in the sensitivity analysis of OS data in the pSTAT3(+) Subpopulation:

- As a continuous covariate in the Cox PH model models for OS. The models will include terms for treatment and propensity score.
- As a stratification factor in the Cox PH model models for OS where the subjects are stratified by the quintiles of the propensity scores. Here a simple stratification scheme will be employed: propensity score <= q20, > q20 < -<= q40, > q40 -≤ q60, > q60 -≤ q80 and > q80, where q20, q40, q60 and q80 are the 20th, 40th, 60th and 80th percentiles of propensity scores. Patients in the strata defined by propensity scores have approximately the same probability of treatment assignment and hence are comparable within each stratum. The Cox PH model models will include the terms for treatment stratified by propensity score quintiles. Within-strata subgroup analyses may also be conducted.

The sample SAS code are provided in Appendix V: Sample SAS codes for Propensity Score and Multiple Imputation .

7.2.10. Sensitivity Analysis on Missing Biomarker Data

Depending on the proportion of missing biomarker data, sensitivity analysis described in this section may be performed at the time of the IA and/or FA.

7.2.10.1. Comparison Baseline Covariates between Biomarker Analysis Set and Patients Missing Biomarker Data

Baseline covariates, including but not limited to the baseline demographic, baseline disease characteristics, baseline stratification factors and baseline subgroups will be summarized descriptively by the 2 groups (Biomarker Analysis Set and ITT-pSTAT3(Unknown)).

Since the pSTAT3 sample collection is independent from the baseline covariates, including demographics and baseline disease characteristics, it is expected that the pSTAT3 status would be missing at random, and that baseline covariates should be approximately balanced between Biomarker Analysis Set and ITT-pSTAT3(Unknown).

If deemed necessary, propensity scores may be calculated and summarized using the similar procedure as described in Section 7.2.9 where replacing $T = 1$ indicate patients in Biomarker Analysis Set and $T = 0$ indicate ITT-pSTAT3(Unknown).

7.2.10.2. Comparison of Overall Survival of Biomarker Analysis Set and Patients Missing Biomarker Data

Estimates of the overall survival obtained from the Kaplan-Meier method will be presented for Biomarker Analysis Set and ITT-pSTAT3(Unknown) by ARMs and overall. The median survival time and corresponding 2-sided CI will be provided for each treatment arm for Biomarker Analysis Set and ITT-pSTAT3(Unknown). An unstratified Cox PH model will be used to obtain the Hazard Ratio of ARM1 vs ARM2 in Biomarker Analysis Set and ITT-pSTAT3(Unknown), respectively. Additional covariates, such as propensity score from the last step may be included in multivariate Cox PH model either as continuous variable or categorical variable of propensity score quintiles.

7.2.10.3. Handling Missing Biomarker Status

To evaluate the impact of missing biomarker status on OS endpoint and robustness of conclusion from the pSTAT3(+) Subpopulation, the following methods may be applied where appropriate.

Multiple Imputation of Missing Biomarker Status

Using the following steps for multiple imputation of missing biomarker status.

Assuming the biomarker status is missing at random, the multiple imputation method [Rubin 1978, Rubin 1987, Yan 2009, Campbell 2011] is deemed appropriate to address the impact of missing pSTAT3 data on OS, and is described in the following 3 steps. Step 1 describes the missing data imputation procedure, Steps 2 and 3 illustrate the subsequent analysis strategy.

Step 1: Missing Data Imputation based on the Selected Covariates (M = 100 imputations):

Missing biomarker status will be imputed by SAS PROC MI with the all baseline factors specified in Table 12 using logistic regression to generate M complete data sets. The seed will be set up as 3789 in PROC MI procedure.

Step 2: Statistical Analysis for Each Complete Dataset:

Within the pSTAT3(+) group, the unstratified Cox PH model will be applied to each of the complete datasets. The Observed Hazard Ratio and its corresponding statistics will be computed from the model. Similar analysis may be performed using the pSTAT3(-) group if needed.

Step 3: Combine the Results from M Datasets:

PROC MIANALYZE will be applied to combine the analysis results (HR and its corresponding statistics) from the M datasets. The HR and p-value from the combination of results will be reported as a sensitivity analysis result compared with the primary analysis.

In the case that missing at random is questionable, other methods including the tipping point method [Yan, 2009, Campbell, 2011] will be used to examine the sensitivity of the conclusions of missing data and aid clinical reviewers in making a judgment regarding the treatment effect in the study.

Tipping Point Analysis

Another strategy to handle missing data, particularly under worse case scenarios, is the tipping point analysis, which does not rely on the assumption of missingness at random. Let X1 and X2 denote the number of patients to be imputed as pSTAT3(+) in the FOLFIRI Arm and the NAPABUCUSIN+ FOLFIRI arm from the pSTAT3(Unknown) Subpopulation. X1 and X2 will be determined based on the prevalence of pSTAT3(+) estimated from the Biomarker Analysis Set. For example, if the prevalence of pSTAT3(+) in the Biomarker Analysis Set is 50% and there are 100 patients in the FOLFIRI arm and 100 patients in NAPABUCUSIN+ FOLFIRI arm with unknown pSTAT3 status (total unknown is 200), then X1=X2=50.

The following tipping points analysis will be applied to evaluate the impact of MNAR (missing not at random) of biomarker status may have on the primary analysis, starting with the worst-case scenario.

Step 1: from pSTAT3(Unknown) Subpopulation of FOLFIRI Arm, randomly select X1 patients to be pSTAT3(+);

Step 2: from pSTAT3(Unknown) Subpopulation of napabucasin + FOLFIRI Arm, list all uncensored survival times ordered from the shortest to the longest and append to a second list of all censored survival time ordered from the shortest to the longest (for convenience, we will call this list – the ranked outcome list). Assign patients occupying the first X2 positions to be pSTAT3(+);

Step 3: combine data from Step 1 and Step 2 with the observed pSTAT3(+) data to form a complete pSTAT3(+) dataset. The log-rank test and Cox PH model will be performed in the same way as in the primary analysis for the pSTAT3(+) Group;

Step 4: Repeat Step 1 to Step3 10000 times to obtain an average p-value and HR estimate. This would be considered as the worst-case scenario (in the worst-case scenario, Step 2 is the same for each of the 10000 runs);

Step 5: Repeat Step 1 to Step 3 except that in Step 2, randomly assign 95% of the patients occupying the first X2 positions on the ranked outcome list to be pSTAT3(+). Randomly select X2*5% patients from those occupying the remaining positions on the ranked outcome list to be pSTAT3(+) also. Compute p-value and HR for the resulting dataset;

Step 6: Repeat Step 5 10000 times to obtain an average p-value and HR estimate for the 95% scenario;

Step 7: Repeat Step 5 to Step 6, while changing the percentage from 95% to p% (p=90 to 5, by a decrement of 5 each time). Randomly assign p% of patients occupying the first X2 positions on the ranked outcome list to be pSTAT3(+). Randomly select X2*(100-p)% patients from those occupying the remaining positions on the ranked outcome list to be pSTAT3(+) also;

Step 8 A graph of p-values/HR versus corresponding to different proportions (p%) as outlined above will be presented to illustrate results of tipping point analysis.

Table 13: Summary of Key Efficacy Analysis

Endpoint	Analysis Set	Statistical Methods	Missing Data	Interpretation
OS (7.2.2)	ITT-GP	Stratified log-rank test (1-sided, all S factors); KM Method (median, 95% CI); HR and 95% CI from Stratified Cox PH Model	Censor patients who didn't die or who lack any data beyond randomization	Primary analysis for primary endpoint in GP
	ITT-pSTAT3(+)	Unstratified log-rank test (1-sided); KM Method (median, 95% CI); HR and 95% CI from unstratified Cox PH Model	Censor patients who didn't die or who lack any data beyond randomization	Primary analysis for primary endpoint in pSTAT3(+)
6-month, 1 year, 18-month survival	ITT-GP ITT-pSTAT3(+)	KM method (95% CI)		
OS (7.2.3.1)	PPAS-GP	Stratified log-rank test (1-sided, all S factors); KM Method (median, 95% CI); HR and 95% CI from Stratified Cox PH Model		Sensitivity analysis of primary endpoint
	PPAS-pSTAT3(+)	Unstratified log-rank test (1-sided); KM Method (median, 95% CI). HR and 95% CI from unstratified Cox PH Model		Sensitivity analysis of primary endpoint
6-month, 1 year, 18-month survival	PPAS-GP PPAS-pSTAT3(+)	KM method (95% CI)		
OS (7.2.3)	ITT-GP	Unstratified log-rank test (1-sided), HR and 95% CI from unstratified Cox PH Model		Sensitivity analysis
	ITT-pSTAT3(+)	Stratified log-rank test (1-sided, all S factors); KM Method (median, 95% CI); HR and 95% CI from Stratified Cox PH Model		Sensitivity analysis
	ITT-pSTAT3(-)	HR and 95% CI from unstratified Cox PH Model		Sensitivity analysis
OS (7.2.3)	ITT-GP ITT-pSTAT3(+)	Stratified log-rank test (1-sided, all S factors on randomized		Sensitivity analysis

Endpoint	Analysis Set	Statistical Methods	Missing Data	Interpretation
		stratification factor); KM Method (median, 95% CI); HR and 95% CI from Stratified Cox PH Model		
OS (7.2.3)	Subgroup from ITT-GP, ITT-pSTAT3(+), and ITT-pSTAT3(-)	For each S level, KM Method (median, 95% CI) for each treatment for each S level. Stratified log-rank test(2-sided) and HR and 95% CI from stratified Cox PH Model		Sensitivity analysis
OS (7.2.3)	Subgroup from ITT-GP, ITT-pSTAT3(+), ITT-pSTAT3(-)	For each baseline factor level, KM Method (median, 95% CI); Unstratified log-rank test (2-sided); HR and 95% CI from unstratified Cox PH Model		Sensitivity analysis
OS (7.2.3)	ITT-GP ITT-pSTAT3(+) ITT-pSTAT3(-)	Multivariate Cox PH model stratified by S factors and including other baseline covariates (HR and 95% CI)		Sensitivity analysis
OS (7.2.3.2)	ITT-pSTAT3(-)	KM method (median, 95% CI); HR and 95% CI from stratified Cox PH model		Sensitivity analysis for OS at IA and/or FA
6-month, 1 year, 18-month survival	ITT-pSTAT3(-)	KM method (95% CI)		
OS (7.2.9)	ITT-pSTAT3(+)	Propensity Score Methods will be used to evaluate potential imbalance baseline covariates between ARMs		Sensitivity analysis
OS (7.2.10)	BAS and ITT-pSTAT3(Unknown)	1) OS between BAS and ITT-pSTAT3(Unknown): KM method (median, 95% CI) by ARM and Overall; 2) HR (ARM1 vs ARM2) and 95% CI from unstratified Cox PH		Sensitivity analysis

Endpoint	Analysis Set	Statistical Methods	Missing Data	Interpretation
		model for both BAS and ITT-pSTAT3(Unknown); 3) Multivariate Cox PH model using propensity score, et al.; 4) Multiple Imputation and Tipping Point analysis		
PFS (7.2.4)	ITT-GP	Stratified log-rank test (1-sided, all S factor); KM method (median, 95% CI); HR and 95% CI from stratified Cox PH Model	Censoring patients following criteria in 3.1.2.1	Primary analysis for PFS in GP
	ITT-pSTAT3(+)	Unstratified log-rank test (1-sided); KM Method (median, 95% CI); HR and 95% CI from unstratified Cox PH Model	Censoring patients following criteria in 3.1.2.1	Primary analysis for PFS in pSTAT3(+)
PFS (7.2.4)	Subgroup from ITT-GP and ITT-pSTAT3(+)	For each S level, KM method (median, 95% CI) for each treatment		Sensitivity analysis
PFS (7.2.4)	ITT-pSTAT3(-)	KM method (median, 95% CI); HR and 95% CI from unstratified Cox PH Model		Sensitivity analysis at IA and/or FA
ORR (7.2.4)	ODAS-GP	Stratified CMH (1-sided, stratified by S factors); TrtDiff and 95% CI from HM method adjusting S factors; Exact CI based on Clopper-Pearson method	Patients without on-study tumor assessment or who die, progress or drop out for any reason, or receive anti-tumor treatment prior to reaching a CR or PR as non-responders.	Primary analysis for ORR in GP
ORR (7.2.4)	ODAS-pSTAT3(+)	1-sided Z test; TrtDiff and 95% CI from normal approximation (unstratified); Exact CI based on Clopper-Pearson		Primary analysis for ORR in pSTAT3(+)

Endpoint	Analysis Set	Statistical Methods	Missing Data	Interpretation
		method		
ORR (7.2.4)	ODAS-pSTAT3(-)	TrtDiff and 95% CI from normal approximation (unstratified); Exact CI based on Clopper-Pearson method		Sensitivity Analysis at IA and/or Final analysis.
DCR (7.2.4)	ODAS-GP	Stratified CMH (1-sided, stratified by S factors); TrtDiff and 95% CI from HM method adjusting S factors; Exact CI based on Clopper-Pearson method		Primary analysis for DCR in GP
	ODAS-pSTAT3(+)	1-sided Z test; TrtDiff and 95%CI from normal approximation (unstratified); Exact CI based on Clopper-Pearson method		Primary analysis for DCR in pSTAT3(+)
DCR (7.2.4)	ODAS-pSTAT3(-)	TrtDiff and 95%CI from normal approximation (unstratified); Exact CI based on Clopper-Pearson method; Exact CI based on Clopper-Pearson method		Sensitivity analysis at IA and/or FA

OS: overall survival. PFS: progression-free survival. MD: measurable disease. GP: General Population.
HM: Harmonic Means.

8. SUMMARY OF MAJOR CHANGES IN THE PLANNED ANALYSES

No major change from planned analyses was made.

9. REFERENCES

Protocol Amendment Version 7.0 – Clean (Submitted to IND 100887, Serial 0357, 14 November 2019.

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

Benjamini, Y., Hochberg, Y. Multiple hypotheses testing with weights. Scandinavian Journal of Statistics. 1997; 24: 407-418.

Brannath, W., Posch, M., Bauer, P. Recursive combination tests. Journal of the American Statistical Association. 2002; 97: 236-244.

226969_BBI_226969_Blinding_Maintenance_Plan_20191024_v2.pdf

CanStem303C Sponsor Blinding Plan v4.0 20191030_Final

Clinical Protocol_Study CanStem303C_Amendment 6.0.pdf

Chan ISF, Zhang Z. (1999). Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics, 55:1201–1209.Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B 1972; 34:187-220.

Cui, L., Hung, H.M., Wang, S.J. Modification of sample size in group-sequential clinical trials. Biometrics. 1999; 55: 853-857.

Dmitrienko, A., Tamhane, A.C. Mixtures of multiple testing procedures for gatekeeping applications in clinical trials. Statistics in Medicine. 2011; 30: 1473-1488.

Dmitrienko, A., Tamhane, A.C. General theory of mixture procedures for gatekeeping. Biometrical Journal. 2013; 55: 402-419.

Dmitrienko, A., Kordzakhia, G., Brechenmacher, T. (2016). Mixture-based gatekeeping procedures for multiplicity problems with multiple sequences of hypotheses. Journal of Biopharmaceutical Statistics. 26, 758-780.

Dmitrienko, A., Yuan, Y. Interim data monitoring. Analysis of Clinical Trials Using SAS: A Practical Guide (Second Edition). Dmitrienko, A., Koch, G.G. (editors). SAS Press: Cary, NC, 2017.

DeMets D, Lan KKG. Interim Analysis: the alpha spending function approach. Statistics in Medicine 1994; 13:1314-1352.

Eisenhauer EA, P. Therasse, J. Bogerts et al. New Response evaluation criteria in Solid Tumours: Revised RECIST Guideline (version 1.1). European Journal of Cancer. 45: 228-247, 2009.

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

Guidance for Industry, Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics,USDHHS, FDA, May 2007

Campbell, G, Pennello, G, Yue L. Missing Data in the Regulation of Medical Devices, Journal of Biopharmaceutical Statistics, 21:2, 180-195,2011.

Hochberg Y, A sharper Bonferroni's procedure for multiple tests of significance. Biometrika. 75:800-3, 1988.

Jennison, C., Turnbull, B.W. Group Sequential Methods with Applications to Clinical Trials. Boca Raton: Chapman and Hall, 2000.Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 53:457-481, 1958.

Kordzakhia, G., Dmitrienko, A., Ishida, E. Mixture-based gatekeeping procedures in adaptive clinical trials. Journal of Biopharmaceutical Statistics. 2018; 28: 129-145.

Kordzakhia, G., Brechenmacher, T., Ishida, E., Dmitrienko, A., Zheng, W.W., Li, D.F. An enhanced mixture method for constructing gatekeeping procedures in clinical trials. Journal of Biopharmaceutical Statistics. 2018b; 28: 113-128.

Magirr D, Jaki T, Koenig F, Posch M. Sample size reassessment and hypothesis testing in adaptive survival trials. PLoS ONE 2016; 11(2): e0146465.

Mehrotra, D., Raikar, R. "Minimum Risk Weights for Comparing Treatments in Stratified Binomial Trials". Statistics in Medicine, 2000, 19, pp. 811-825.

Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. Statistics in Medicine 2011; 30:3267–3284.

Miettinen O., Nurminen M. (1985) Comparative Analysis of Two Rates. Statistics in Medicine, 4, 213-226.

Millen B, Dmitrienko A, Ruberg S, Shen, L. A statistical framework for decision making in confirmatory multipopulation tailoring clinical trials. Drug Information Journal. 2012; 46: 647-656.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979; 35(3);549-556.

Pampallona S and Tsiatis AA, Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favor of the null hypothesis. J. Statistical Planning and Inference, 1994; 42:19-35.

Rosenbaum, P. R., Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55.

Rubin, D. (1978). Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse. Imputation and Editing of Faulty or Missing Survey Data. Washington, DC: U.S. Department of Commerce.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

Sarkar, S.K. On the Simes inequality and its generalization. Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen. Balakrishnan N, Pena EA, Silvapulle MJ (eds). Institute of Mathematical Statistics: Beachwood, Ohio. 2008; 231-242.

Shen Y, Cai J. Sample size re-estimation for clinical trials with censored survival data. Journal of the American Statistical Association. 2003; 98: 418-426.

Sugitani, T., Bretz, F., Maurer, W. (2016). A simple and flexible graphical approach for adaptive group-sequential clinical trials. Journal of Biopharmaceutical Statistics. 26, 202-216.

Sugitani, T., Posch, M., Bretz, F., Koenig, F. (2018). Flexible alpha allocation strategies for confirmatory adaptive enrichment clinical trials with a prespecified subgroup. Statistics in Medicine. 2018; 37: 3387-3402.

Wassmer, G. Planning and analyzing adaptive group sequential survival trials. Biometrical Journal. 2006; 48:714-729.

Yan, X., Lee, S., Li, N. (2009). Missing data handling methods in medical device clinical trials. Journal of Biopharmaceutical Statistics 19(6):1085–1098.

Yue, L. Q. (2007). Statistical and regulatory issues with the application of propensity score analysis to non-randomized medical device clinical studies. Journal of Biopharmaceutical Statistics 17(1):1–13.

10. APPENDIX

Appendix I: Patient Evaluation Flow Sheet for Arm 1 (BBI-608 in Combination with FOLFIRI)

Tests & Procedures	Pre-Treatment	During Protocol Treatment				After Protocol Treatment Discontinuation (+/- 3 days)	
		Run-in [1]		Cycle 1	Additional Cycles (+/- 3 days)		
Day		1	2	1	1		
Timing	≤14 days prior to randomization					4 weeks post protocol treatment	Every 8 weeks thereafter[15]
History[2] and Physical Exam	X			X	X	X	
ECOG PS	X			X	X	X	
Weight	X			X	X	X	
Height	X						
Vital signs	X			X	X	X	
FOLFIRI Infusion[3]				X	X		
Begin BBI-608 administration		X					
Hematology[4]	X			X	X	X	
Biochemistry[4]	X			X	X	X	
Urinalysis[4]	X			X	X	X	
ECG (12-lead)	X			X		X	
Radiology and Imaging[5]	X	Every 8 weeks for 6 months and every 12 weeks thereafter until progressive disease is documented.					
Submission of representative block of diagnostic tumor tissue	before or after randomization						
Blood collection for correlative studies[6,7]	X				X		
Blood collection for sparse PK analysis					X		
Pregnancy test, serum or urine (if applicable)[8,9]	X				X	X	
Adverse Event assessment[10,11]	X	X	X[11]	X	X	X	X[14]
Quality of Life assessment (EORTC QLQ-C30)[12]	X				X	X[13]	

Tests & Procedures	Pre-Treatment	During Protocol Treatment				After Protocol Treatment Discontinuation (+/- 3 days)	
		Run-in [1]		Cycle 1	Additional Cycles (+/- 3 days)		
Day		1	2	1	1		
Assessment for survival of patient						X[16]	X[16]

1. BBI-608 administration will begin 2 days prior to the FOLFIRI infusion on day 1 of cycle 1. These two days are referred to as *run-in day 1* and *run-in day 2*. The *run-in day* period may be extended by up to 3 additional calendar days. *Run-in day 1* should occur within 2 calendar days of patient randomization on study Arm 1.
2 Medical history must include date of diagnosis including histological documentation of malignancy, documentation of *RAS* status of tumor, prior anticancer therapy and prior date(s) of disease progression.
3 FOLFIRI administration should proceed according to institutional standard practice (with respect to pre-treatment laboratory evaluation, clinical assessment, pre-medication, and monitoring during and after infusion). Addition of bevacizumab to the FOLFIRI regimen, per Investigator choice, will be permissible.
4 Laboratory investigations should be performed within 3 days prior to FOLFIRI administration on Day 1 of every 14-day Cycle of protocol treatment or within 3 days prior to each FOLFIRI infusion. Laboratory testing performed as part of standard of care prior to patient signature of the study consent will be acceptable as baseline lab work as long as testing is performed ≤ 14 days prior to randomization. Laboratory testing performed as part of pre-treatment assessments within 3 days of Cycle 1 Day 1 will be acceptable as Cycle 1 Day 1 lab work and will not need to be repeated.
5 Tumor measurement and evaluation by RECIST 1.1 criteria. The same method of assessment and the same technique should be used to identify and report each lesion at baseline and at reassessment during treatment. Tumor evaluations will continue until progressive disease is documented (as described in Protocol Section 9). For patients who remain on protocol therapy after objective disease progression has been documented, no further imaging assessments are mandated, but where these occur as a component of care, tumor measurements and assessment must be reported. Tumor assessments should be obtained within +/- 5 days of protocol specified schedule. Qualifying scans performed as part of standard of care prior to patient signature of the study informed consent will be acceptable as baseline scanning as long as scanning is performed ≤ 21 days prior to randomization.
6 Sample collections should be performed at baseline and at Day 1 of Cycle 3 after randomization.
7 A sample will be collected following protocol treatment discontinuation if discontinuation occurs prior to 4 weeks of therapy.
8 In women of childbearing potential only. The minimum sensitivity of the pregnancy test must be 25 IU/L or equivalent units of HCG. Baseline pregnancy test should be done within 5 days of randomization.
9 In women of childbearing potential only a negative pregnancy test must be demonstrated every 4 weeks until 4 weeks after the administration of the final dose of protocol therapy.
10 Adverse events will be recorded and graded according to the NCI Common Terminology Criteria for Adverse Events version 4.0 (see Protocol Appendix 4). Following permanent protocol treatment discontinuation, patients will be assessed for any protocol treatment related adverse events every 8 weeks, starting with the 4-week post-protocol treatment discontinuation visit.
11 Adverse event assessments by phone should be performed on *run-in day 2*.
12 To be completed in clinic. Questionnaires should be completed at baseline and at Day 1 of Cycle 3, Day 1 of Cycle 5, Day 1 of Cycle 7, Day 1 of Cycle 9 and Day 1 of Cycle 13 (+/- 3 days) after randomization for as long as patient remains on Protocol therapy or until deterioration to ECOG PS 4 or hospitalization for end of life care.
13 EORTC QLQ-C30 questionnaires will be collected in the post-protocol discontinuation period only if the patient discontinues protocol treatment prior to 24 weeks of therapy and has an ECOG PS of less than 4 and has not been hospitalized for end of life care.
14 These assessments may occur within a ±7-day window, and after the first visit at which the patient has been off protocol treatment for 4 weeks, patients will be assessed every 8 weeks for survival and any protocol treatment related adverse events. Medical history at post-progression follow up must include post-protocol treatment cancer therapies.

15 After the first visit at which the patient has been off protocol treatment for 4 weeks, patients will be assessed for survival every 8 weeks until deterioration to ECOG PS 4 or hospitalization for end of life care
16 In the event that the patient is unable to attend clinic, post-progression follow-up for survival may be by means of telephone contact

Appendix II: Patient Evaluation Flow Sheet for Arm 2 (FOLFIRI)

Tests & Procedures	Pre-Treatment	During Protocol Treatment		After Protocol Treatment Discontinuation (± 3 days)	
		Cycle 1	Additional Cycles (± 3 days)		
Day		1	1		
Timing	≤14 days prior to randomization			4 weeks post protocol treatment	Every 8 weeks thereafter[13]
History[1] and Physical Exam	X	X	X	X	
ECOG PS	X	X	X	X	
Weight	X	X	X	X	
Height	X				
Vital signs	X	X	X	X	
FOLFIRI Infusion[2]		X	X		
Hematology[3]	X	X	X	X	
Biochemistry[3]	X	X	X	X	
Urinalysis[3]	X	X	X	X	
ECG (12-lead)	X	X		X	
Radiology and Imaging[4]	X	Every 8 weeks for 6 months and every 12 weeks thereafter until progressive disease is documented.			
Submission of representative block of diagnostic tumor tissue	before or after randomization				
Blood collection for correlative studies[5,6]	X		X		
Pregnancy test, serum or urine (if applicable)[7,8]	X		X	X	
Adverse Event assessment[9]	X	X	X	X	X[12]
Quality of Life assessment (EORTC QLQ-C30)[10]	X		X	X[11]	
Assessment for survival of patient				X[14]	X[14]

1 Medical history must include date of diagnosis including histological documentation of malignancy, documentation of *RAS* status of tumor, prior anticancer therapy and prior date(s) of disease progression.

2 FOLFIRI administration should begin within 7 calendar days of randomization and proceed according to institutional standard practice (with respect to pre-treatment laboratory evaluation, clinical assessment, pre-medication, and monitoring during and after infusion). Addition of bevacizumab to the FOLFIRI regimen, per Investigator choice, will be permissible.

3 Laboratory investigations should be performed within 3 days prior to FOLFIRI administration on Day 1 of every 14-day study Cycle of protocol treatment or within 3 days prior to each FOLFIRI infusion. Laboratory testing

performed as part of standard of care prior to patient signature of the study consent will be acceptable as baseline lab work as long as testing is performed ≤ 14 days prior to randomization. Laboratory testing performed as part of pre-treatment assessments within 3 days of Cycle 1 Day 1 will be acceptable as Cycle 1 Day 1 lab work and will not need to be repeated.

4 Tumor measurement and evaluation by RECIST 1.1 criteria. The same method of assessment and the same technique should be used to identify and report each lesion at baseline and at reassessment during treatment. Tumor evaluations will continue until progressive disease is documented (as described in Protocol Section 9). Qualifying scans performed as part of standard of care prior to patient signature of the study informed consent will be acceptable as baseline scanning as long as scanning is performed ≤ 21 days prior to randomization.

5 Sample collections should be performed at baseline and at Day 1 of Cycle 3 after randomization.

6 A sample will be collected following protocol treatment discontinuation if discontinuation occurs prior to 4 weeks of therapy.

7 In women of childbearing potential only. The minimum sensitivity of the pregnancy test must be 25 IU/L or equivalent units of HCG. Baseline pregnancy test should be done within 5 days of randomization.

8 In women of childbearing potential only a negative pregnancy test must be demonstrated every 4 weeks until 4 weeks after the administration of the final dose of protocol therapy.

9 Adverse events will be recorded and graded according to the NCI Common Terminology Criteria for Adverse Events version 4.0 (see Protocol Appendix). Following permanent protocol treatment discontinuation, patients will be assessed for any protocol treatment related adverse events every 8 weeks, starting with the 4-week post-protocol treatment discontinuation visit.

10 To be completed in clinic. Questionnaires should be completed at baseline and At Day 1 of Cycle 3, Day 1 of Cycle 5, Day 1 of Cycle 7, Day 1 of Cycle 9 and Day 1 of Cycle 13 (+/- 3 days) after randomization for as long as patient remains on Protocol therapy or until deterioration to ECOG PS 4 or hospitalization for end of life care.

11 EORTC QLQ-C30 questionnaires will be collected in the post-protocol discontinuation period only if the patient discontinues protocol treatment prior to 24 weeks of therapy and has an ECOG PS of less than 4 and has not been hospitalized for end of life care.

12 After the first visit at which the patient has been off protocol treatment for 4 weeks, patients will be assessed every 8 weeks for survival and any protocol treatment related adverse events. Medical history at post-progression follow up must include post-protocol treatment cancer therapies.

13 These assessments may occur within a ±7-day window, and after the first visit at which the patient has been off protocol treatment for 4 weeks, patients will be assessed for survival every 8 weeks until deterioration to ECOG PS 4 or hospitalization for end of life care.

14 In the event that the patient is unable to attend clinic, post-progression follow-up for survival may be by means of telephone contact

Appendix III Response and Evaluation Endpoints

Response and progression will be evaluated in this study using the revised international criteria (1.1) proposed by the RECIST (Response Evaluation Criteria in Solid Tumors) committee.

10.2.1 *Measurable Disease:* Measurable *tumor lesions* are defined as those that can be accurately measured in at least one dimension (longest diameter to be recorded) as ≥ 20 mm with chest x-ray and as ≥ 10 mm with CT scan, or clinical examination. Bone lesions are considered measurable only if assessed by CT scan and have an identifiable soft tissue component that meets these requirements (soft tissue component ≥ 10 mm by CT scan). *Malignant lymph nodes* must be ≥ 15mm in the short axis to be considered measurable; only the short axis will be measured and followed. All tumor measurements must be recorded in millimeters (or decimal fractions of centimeters). Previously irradiated lesions are not considered measurable unless progression has been documented in the lesion.

10.2.2 *Non-measurable Disease:* All other lesions (or sites of disease), including small lesions are considered non-measurable disease. Bone lesions without a measurable soft tissue component, leptomeningeal disease, ascites, pleural/pericardial effusions, lymphangitis cutis/pulmonis, inflammatory breast disease, lymphangitic involvement of lung or skin and abdominal masses followed by clinical examination are all non-measurable. Lesions in previously irradiated areas are non-measurable, unless progression has been demonstrated.

10.2.3 *Target Lesions:* When more than one measurable tumor lesion is present at baseline all lesions up to *a maximum of 5 lesions total* (and a maximum of *2 lesions per organ*) representative of all involved organs should be identified as target lesions and will be recorded and measured at baseline. Target lesions should be selected on the basis of their size (lesions with the longest diameter), be representative of all involved organs, but in addition should be those that lend themselves to *reproducible repeated measurements*. Note that pathological nodes must meet the criterion of a short axis of ≥ 15 mm by CT scan and only the *short* axis of these nodes will contribute to the baseline sum. All other pathological nodes (those with short axis ≥ 10 mm but < 15 mm) should be considered non-target lesions. Nodes that have a short axis < 10 mm are considered non-pathological and should not be recorded or followed (see 10.2.4). At baseline, the sum of the target lesions (longest diameter of tumor lesions plus short axis of lymph nodes: overall maximum of 5) is to be recorded.

After baseline, a value should be provided on the CRF for all identified target lesions for each assessment, even if very small. If extremely small and faint lesions cannot be accurately measured but are deemed to be present, a default value of 5 mm may be used. If lesions are too small to measure and indeed are believed to be absent, a default value of 0 mm may be used.

10.2.4 *Non-target Lesions:* All non-measurable lesions (or sites of disease) plus any measurable lesions over and above those listed as target lesions are considered *non-target lesions*.

Measurements are not required but these lesions should be noted at baseline and should be followed as "present" or "absent".

10.2.5 *Response:* All patients will have their BEST RESPONSE from the start of study treatment until the end of treatment classified as outlined below:

Complete Response (CR): disappearance of *target* and *non-target*. Pathological lymph nodes must have short axis measures < 10 mm (Note: continue to record the measurement even if < 10 mm and considered CR). Residual lesions (other than nodes < 10 mm) thought to be non-malignant should be further investigated (by cytology specialized imaging or other techniques as appropriate for individual cases *[Eisenhauer 2009]*) before CR can be accepted.

Partial Response (PR): at least a 30% decrease in the sum of measures (longest diameter for tumor lesions and short axis measure for nodes) of target lesions, taking as reference the baseline sum of diameters. Non-target lesions must be non-PD.

Stable Disease (SD): neither sufficient shrinkage to qualify for PR, nor sufficient increase to qualify for PD taking as reference the smallest sum of diameters on study.

Progressive Disease (PD): at least a 20% increase in the sum of diameters of measured lesions taking as references the smallest sum of diameters recorded on study (including baseline) AND an absolute increase of $\geq$ 5mm. Appearance of new lesions will also constitute progressive disease (including lesions in previously unassessed areas).
In exceptional circumstances, unequivocal progression of non-target disease may be accepted as evidence of disease progression, where the overall tumor burden has increased sufficiently to merit discontinuation of treatment or where the tumor burden appears to have increased by at least 73% in volume. Modest increases in the size of one or more non-target lesions are NOT considered unequivocal progression. If the evidence of PD is equivocal (target or non-target), treatment may continue until the next assessment, but if confirmed, the earlier date must be used.

Table 14: Integration of Target, non-Target and New Lesions into Response Assessment:

Target Lesions	Non-Target Lesions	New Lesions	Overall Response	Best Response for this Category also Requires
Target lesions ± non target lesions				
CR	CR	No	CR	tumor nodes < 10mm
CR	Non-CR/Non-PD	No	PR	
CR	Not all evaluated	No	PR	
PR	Non-PD/ not all evaluated	No	PR	
SD	Non-PD/ not all evaluated	No	SD	documented at least once ≥ 6 wks. from baseline
Not all evaluated	Non-PD	No	NE	
PD	Any	Any	PD	
Any	PD	Any	PD	
Any	Any	Yes	PD	
Non target lesions ONLY				
No Target	CR	No	CR	tumor nodes < 10mm
No Target	Non-CR/non-PD	No	Non-CR/non-PD	
No Target	Not all evaluated	No	NE	
No Target	Unequivocal PD	Any	PD	
No Target	Any	Yes	PD	

Note: Patients with a global deterioration of health status requiring discontinuation of treatment without radiological progression having been observed at that time should be reported as "symptomatic deterioration". This is NOT objective PD. Every effort should be made to document the objective progression even after discontinuation of treatment.

10.3 RESPONSE DURATION

Response duration will be measured from the time measurement criteria for CR/PR (whichever is first recorded) are first met until the first date that recurrent or progressive disease is objectively documented, taking as reference the smallest measurements recorded on study (including baseline).

10.4 STABLE DISEASE DURATION

Stable disease duration will be measured from the time of randomization until the criteria for progression are met, taking as reference the smallest sum on study (including baseline).

10.5 METHODS OF MEASUREMENT

The same method of assessment and the same technique should be used to characterize each identified and reported lesion at baseline and during follow-up. Assessments should be identified on a calendar schedule and should not be affected by delays in therapy. While on study, all lesions recorded at baseline should have their actual measurements recorded at each subsequent evaluation, even when very small (e.g. 2 mm). If it is the opinion of the radiologist that the lesion has likely disappeared, the measurement should be recorded as 0 mm. If the lesion is believed to be present and is faintly seen but too small to measure, a default value of 5 mm should be assigned. For lesions which fragment/split add together the longest diameters of the fragmented portions; for lesions which coalesce, measure the maximal longest diameter for the "merged lesion".

Additionally, for optimal tumor assessment scanning options are listed below in the decreasing order of preference:

Order of Preference	Scanning Option
1	Chest-Abdomen-Pelvis CT with oral and I.V. contrast
2	Chest CT without I.V. contrast PLUS MRI Abdomen-Pelvis with oral and I.V. contrast[1]
3	Chest-Abdomen-Pelvis CT with oral contrast[2]

[1]If Iodine contrast media is medically contraindicated.
[2]If Iodine contrast media is medically contraindicated and MRI cannot be performed.

10.5.1 *Clinical Lesions*. Clinical lesions will only be considered measurable when they are superficial and ≥ 10 mm as assessed using callipers (e.g. skin nodules). For the case of skin lesions, documentation by colour photography including a ruler to estimate the size of the lesion is recommended. If feasible, imaging is preferred.

10.5.2 *Chest X-ray*. Chest CT is preferred over chest X-ray, particularly when progression is an important endpoint, since CT is more sensitive than X-ray, particularly in identifying new lesions. However, lesions ≥ 20 mm on chest X-ray may be considered measurable if they are clearly defined and surrounded by aerated lung.
Appendix *Non-Permitted Treatments:*
Concurrent chemotherapy, hormonal therapy (except corticosteroids), immunotherapy, biologic therapy OR other experimental agents should not be given to study patients while on protocol treatment.

10.5.3 *CT/MRI*. CT is the best currently available and reproducible method to measure lesions selected for response assessment. This guideline has defined measurability of lesions on CT scan based on the assumption that CT slice thickness is 5 mm or less. When CT scans have slice thickness greater than 5 mm, the minimum size for a measurable lesion should be twice the slice thickness. MRI is also acceptable in certain situations (e.g. for body scans). Other specialized imaging or other techniques may also be appropriate for individual case *[Eisenhauer 2009]*. For example, while PET scans are not considered adequate to measure lesions, PET-CT scans may be used providing that the measures are obtained from the CT scan and the CT scan is of identical diagnostic quality to a diagnostic CT (with IV and oral contrast).

10.5.4 *Ultrasound*. Ultrasound is not useful in assessment of lesion size and should not be used as a method of measurement. If new lesions are identified by ultrasound in the course of the study, confirmation by CT is advised.

10.5.5 *Endoscopy/Laparoscopy*. The utilization of these techniques for objective tumor evaluation is not advised. However, they can be useful to confirm complete pathological response when biopsies are obtained or to determine relapse in trials where recurrence following complete response or surgical resection is an endpoint.

10.5.6 *Cytology/Histology*. These techniques can be used to differentiate between PR and CR in rare cases (for example, residual lesions in tumor types such as germ cell tumors, where known residual benign tumors can remain).

The cytological confirmation of the neoplastic origin of any effusion that appears or worsens during treatment when the measurable tumor has met criteria for response or stable disease is mandatory to differentiate between response or stable disease (an effusion may be a side effect of the treatment) and progressive disease.

Appendix IV: Non-Permitted Treatments

Non-Permitted Treatments:
Concurrent chemotherapy, hormonal therapy (except corticosteroids), immunotherapy, biologic therapy OR other experimental agents should not be given to study patients while on protocol treatment.

Appendix V: Sample SAS codes for Propensity Score and Multiple Imputation

```
/*Propensity Score to Check Imbalance between ARMs within BM+ cohort: Section 7.2.9 */

/* Step 1:  get propensity score calculation based on baseline covariates: within BM+ cohort: */
/* The baseline covariates included are according to Table 12: Clinically Relevant Baseline Factors in
Section 7.2.9 */
/* arm = cov1 cov2 cov3 ... */

proc logistic data = adsl_os (where = (BIOSTATN  = 1));
  class cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov9 cov10 cov11  ;
  model arm = cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov9 cov10 cov11 ;
  output out=scr predicted=ps xbeta=logit_ps;
run;

/* Step 2: Exam the propensity score overlapping between arm: continuous variable or quintiles */

proc sort data = scr ; by arm; run;
proc univariate data= scr ;
  histogram;
  class arm ;
  var ps ;
run;

proc boxplot data= scr  ;
  plot ps *arm ;
run;

proc rank data=scr out= os1  groups=5;
  var ps;
  ranks q;
run;

proc freq data = os1  ;
  table  q * arm / nopercent norow  ;
run;

 /* Step 3:  adjusted propensity score as continuous covariate */
proc phreg data = os1;
```

```
  class arm ;
  model aval * cnsr (1)  = arm ps /ties = efron ;
run;

 /*stratified propensity score quintiles */
proc phreg data = os1;
  class arm q;
  model aval * cnsr (1)  = arm  /ties = efron ;
  strata q ;
run;

/*Section 7.2.10: multiple imputation for missing biomarker status.   */

/* check if the missing pattern is monotone including all the covariates from Table 12 and BM status */
proc  mi data= adsl_os out=mi_1  nimpute=0;
  var cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov9 cov10 cov11 biostatn;
run;

/*Step 1: multiple imputation including all the baseline factors in section X*/

proc  mi data= adsl_os SEED=3789 out=mi_2 nimpute=100;
  class cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov9 cov10 cov11 biostatn ;
  var cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov9 cov10 cov11 biostatn;
/*Baseline covariate with large number of missing values will be imputed using fcs logistic. An example as
cov8 cov9 */
  fcs logistic (cov8 = cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov9 cov10 cov11 biostatn/details) ;
  fcs logistic (cov9 = cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov10 cov11 biostatn/details) ;
  fcs logistic (biostatn= cov1 cov2 cov3 cov4 cov5 cov6 cov7 cov8 cov9 cov10 cov11 /details) ;
run;

proc sort data=mi_2;
  by _imputation_ biostatn;
run;

/*Step 2:  Analysis HR for each imputation and biomarker status. */

ods output ParameterEstimates=xyz HazardRatios = HR ;
Proc phreg data=mi_2 out=outx ;
  by _imputation_ biostatn;
  class arm;
```

```
  model aval*cnsr(1)= arm / rl ties=efron;
  hazardratio "HR "  arm ;
run;

 proc sort data=xyz;
   by biostatn;
 run;

 /*Step 3:  Combined Results from 100 Imputation */

 proc mianalyze data=xyz ;
   by biostatn;
   modeleffects estimate ;
   stderr stderr;
   ods output parameterestimates=overall;
 run;

 data overall1;
   set overall;
   HR = exp(estimate) ;
   UL95 = exp(estimate + 1.96*stderr) ;
   LL95 = exp(estimate - 1.96*stderr) ;
run;

proc sort data = adsl_os; by biostatn; run;
/*compare with the original results. */
ods output ParameterEstimates=xyz1 HazardRatios = HR1 ;
Proc phreg data=adsl_os out = observed ;
  by biostatn;
  class arm ;
  model aval*cnsr(1)= arm / rl ties=efron;
  hazardratio "HR "  arm ;
run;
```

Appendix VI: ORR and DCR for 36 weeks and ORR/DCR Analysis Set with minimum 36 weeks duration

Endpoints:

DCR within 36 weeks is defined as the proportion of patients with BOR of a documented complete response, partial response, and stable disease (CR + PR + SD), based on RECIST 1.1 within 36 weeks (<=36*7+5 = 257 days) from randomization.
ORR within 36 weeks is defined as the proportion of patients with BOR of a documented complete response and partial response (CR + PR) based on RECIST 1.1 within 36 weeks (<=36*7+5 = 257 days) from randomization.
The p-values will be calculated based on DCR within 36 weeks and ORR within 36 weeks.

Analysis Sets

For ORR, the analysis sets are defined as below:
ORR Analysis Set in the General Population with minimum 36 weeks duration (ORR-GP-36 Weeks)

ORR Analysis Set in the General Population with minimum 36 weeks duration (ODAS-GP-36 Weeks) will include ODAS-GP who have adequate tumor scans after 36 weeks (>= 36*7-5 =247 days); or have the first CR/PR/PD, or die, or get new anti-cancer treatment, or end the study within 36 weeks (<=36*7+5 = 257 days) from randomization based on Final Analysis Data Cutoff.

ORR Analysis Set in the General Population with minimum 36 weeks duration Stage 1 (ORR-GP-36 Weeks- Stage 1)

ORR Analysis Set in the General Population with minimum 36 weeks duration Stage 1 (ORR-GP-36 Weeks- Stage 1) will include ODAS-GP who have adequate tumor scans after 36 weeks (>= 36*7-5 =247 days); or have the first CR/PR/PD, or die, or get new anti-cancer treatment, or end the study within 36 weeks (<=36*7+5 = 257 days) from randomization based on IA Data Cutoff of 22 February 2019.

ORR Analysis Set in the General Population with minimum 36 weeks duration Stage 2 (ORR-GP-36 Weeks- Stage 2)

ORR Analysis Set in the General Population with minimum 36 weeks duration Stage 2 (ORR-GP-36 Weeks- Stage 2) will include patients in ORR-GP-36 Weeks but not in ORR-GP-36 Weeks-Stage 1.

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration (ORR-pSTAT3(+)-36 Weeks)

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration (ORR-pSTAT3(+)-36 Weeks) will include patients in both ODAS-pSTAT3(+) and ORR-GP-36 Weeks analysis sets.

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 1 (ORR-pSTAT3(+)-36 Weeks-Subset 1 Stage 1)

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 1 (ORR-pSTAT3(+)-36 Weeks-Subset 1 Stage 1) will include patients in both ORR-GP-36 Weeks- Stage 1 and ODAS-pSTAT3(+) with specimen age up to a 6-month CSS window.

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 2 (ORR-pSTAT3(+)-36 Weeks-Subset 1 Stage 2)

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 2 (ORR-pSTAT3(+)-36 Weeks-Subset 1 Stage 2) will include patients in both ORR-GP-36 Weeks- Stage 2 and ODAS-pSTAT3(+) analysis sets with specimen age up to a 6-month CSS window but not in ORR-pSTAT3(+)-36 Weeks Subset 1 Stage 1.

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 2 Stage 3 (ORR-pSTAT3(+)-36 Weeks-Subset 2 Stage 3)

ORR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 2 Stage 3 (ORR-pSTAT3(+)-36 Weeks-Subset 2 Stage 3) will include patients ORR-pSTAT3(+)-36 Weeks with specimen age > 6 month and <= final CSS window.

For DCR, the analysis sets are defined as below

DCR Analysis Set in the General Population with minimum 36 weeks duration (DCR-GP-36 Weeks)
DCR Analysis Set in the General Population with minimum 36 weeks duration (ODAS-GP-36 Weeks) will include ODAS-GP who have adequate tumor scans after 36 weeks (>= 36*7-5 =247 days); or have the first CR/PR/PD/(SD >= 51 days from randomization), or die, or get new anti-cancer treatment, or end the study within 36 weeks (<=36*7+5 = 257 days) from randomization based on Final Analysis Data Cutoff.

DCR Analysis Set in the General Population with minimum 36 weeks duration Stage 1 (DCR-GP-36 Weeks- Stage 1)

DCR Analysis Set in the General Population with minimum 36 weeks duration Stage 1 (DCR-GP-36 Weeks- Stage 1) will include ODAS-GP who have adequate tumor scans after 36 weeks (>= 36*7-5 =247 days); or have the first CR/PR/SD/(SD >= 51 days from randomization), or die, or get new anti-cancer treatment, or end the study within 36 weeks (<=36*7+5 = 257 days) from randomization based on IA Data Cutoff of Feb 22, 2019.

DCR Analysis Set in the General Population with minimum 36 weeks duration Stage 2 (DCR-GP-36 Weeks- Stage 2)

DCR Analysis Set in the General Population with minimum 36 weeks duration Stage 2 (DCR-GP-36 Weeks- Stage 2) will include patients in DCR-GP-36 Weeks but not in DCR-GP-36 Weeks-Stage 1.

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration (DCR-pSTAT3(+)-36 Weeks)

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration (DCR-pSTAT3(+)-36 Weeks) will include patients in both ODAS-pSTAT3(+) and DCR-GP-36 Weeks analysis sets.

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 1 (DCR-pSTAT3(+)-36 Weeks-Subset 1 Stage 1)

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 1 (DCR-pSTAT3(+)-36 Weeks Subset 1 Stage 1) will include patients in both DCR-GP-36 Weeks-Stage 1 and ODAS-pSTAT3(+) with specimen age up to a 6-month CSS window.

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 2 (DCR-pSTAT3(+)-36 Weeks-Subset 1 Stage 2)

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 1 Stage 2 (DCR-pSTAT3(+)-36 Weeks-Subset 1 Stage 2) will include patients in (both DCR-GP-36 Weeks- Stage 2 and ODAS-pSTAT3(+) analysis sets) with specimen age up to a 6-month CSS window but not in DCR-pSTAT3(+)-36 Weeks Subset 1 Stage 1.

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 2 Stage 3 (DCR-pSTAT3(+)-36 Weeks-Subset 2 Stage 3)

DCR Analysis Set in the pSTAT3(+) Subpopulation with minimum 36 weeks duration Subset 2 Stage 3 (DCR-pSTAT3(+)-36 Weeks-Subset 2 Stage 3) will include patients DCR-pSTAT3(+)-36 Weeks with specimen age > 6 month and <= final CSS window.