ClinicalTrials.gov NCT #: NCT01556490

Study Title: Efficacy Evaluation of TheraSphere in Patients With Inoperable Liver Cancer (STOP-HCC)

Approved Document Date: 15 September 2022

Redacted Document Date: 28 April 2023

Protocol No. TS-103 (Version 7.0 22Jul2019)

A Phase III Clinical Trial of Intra-arterial TheraSphere® in the Treatment of Patients with Unresectable Hepatocellular Carcinoma (HCC)

Statistical Analysis Plan

Prepared for:
Boston Scientific Corporation
[redacted]

Final Version 6.0 Date 15Sep2022

Prepared by:
Labcorp
[redacted]

Revision History

Version	Date	Comments
Final Version 1.0	20Mar2012	Approved by Nordion
Final Version 1.1	19Jan2016	Updated Sponsor and CRO information only. Further updates required as a result of protocol amendments will be made in a subsequent version of the statistical analysis plan.
Final Version 2.0	07Nov2017	Updated for current protocol and CRF.
Final Version 2.1	16Apr2018	Clarifications and minor corrections.
Final Version 3.0	16Aug2021	1) Included the requirement of "Aggregate Data Declaration Form" being signed to maintain the integrity of study results 2) Clarified the protocol versions implemented

		3) Removed the restriction to keep randomization schedule confidential until study database is locked 4) Updated the derivation of relative days and visit windows, and the imputation of partial days; clarified the baseline definition 5) Updated the definition of best overall response (subsequently impacting ORR, DCR, duration of response and duration of control) to include assessments until the first PD 6) Added details for confirmation of ECOG>1 at two subsequent evaluations for TTSP endpoint 7) Included additional definitions for AFP response 8) Exclude assessments after the start of subsequent HCC therapy in the derivation of the following efficacy variables: best overall response, ORR, DCR, depth of response, PTTS, and AFP response 9) Updated the censoring rules for the time to event secondary and additional efficacy variables (TTP, TTUP, TTSP, TTDQoL, PFS, duration of response, duration of disease control) to censor for the start of subsequent HCC therapy and for any events that occur immediately after two or more missed visits 10) Defined the subsequent HCC therapy with detailed definition and derivations in appendix 4 11) Updated exposure summaries for TheraSphere (including definitions provided to compute the dose absorbed by perfused volume) and sorafenib; added summaries of post-progression treatments; removed the summary of duration of follow-up post treatment 12) Updated the TEAE definition and noted the changes to protocol analyses with rationale 13) Removed UADE safety summaries and added AESI safety summaries; added the calculation of event rates in TEAE overall summaries 14) [redacted] 15) Updated list of variables in the analyses of covariates

		and subgroups, and modifications to some variables 16) Updated to perform subgroup analyses only for primary and secondary efficacy endpoints and AEs; added details for efficacy subgroup analyses and forest plots to present results for time-to-event endpoints 17) Added assessment for collinearity of covariates, and clarified to include treatment group and use overall p-value in the multivariable covariate analysis 18) Updated the definition of safety analysis population and noted changes from the protocol 19) Updated the Per Protocol population to be analyzed according to the randomized treatment, and noted changes from the protocol; added details for major protocol deviations to define Per Protocol population 20) Updated baseline characteristics to be summarized 21) Clarified the definition of prior and concomitant medications 22) Added the analyses to assess the assumption of proportional hazards used to compute the HR for OS 23) Clarified details for data collection and derivation of last date known alive for OS censoring 24) Clarified details for TTP sensitivity analysis of excluding censoring at death date, and added similar sensitivity analysis for TTUP 25) Updated poolability analyses to exclude patients from sites in Germany, and to be performed only for primary and secondary efficacy endpoints 26) Added treatment comparisons for depth of response, PTTS and AFP response in additional efficacy analyses 27) Updated the CRO name from Chiltern International to Labcorp 28) Updated the list of abbreviations and abbreviations used in the text
Final Version 4.0	18Feb2022	Modified the final analysis to be performed at 417 deaths or 30April2022, whichever comes first, and adjusted the calculation of efficacy boundary for the final analysis.

		Provided other small clarifications. Updated sponsor name and address and CRO's signatory.
Final Version 5.0	24Jun2022	1) [redacted] 2) For covariate analyses, updated univariable models to include treatment group with each covariate. 3) For assessment of poolability, removed adjusting for additional factors in multivariable analysis in poolability analysis models. 4) Updated the calculation for duration of sorafenib prior to progression. 5) Updated unit to be displayed for albumin and bilirubin results. 6) In covariate analyses, subgroup analyses and baseline characteristic summaries, added number of lesions at baseline and updated tumor replacement at baseline to use blinded central review data (and both investigator assessment and blinded central review data for baseline summaries).
Final Version 6.0	15Sep2022	1) Updated categories of Child-Pugh class for covariate and subgroup analyses 2) Updated categories of duration from date of initial diagnosis of HCC to randomization for covariate and subgroup analyses, and baseline characteristics summaries 3) Updated analyses of binary endpoints of ORR, DCR, PTTS and AFP response rate to use continuity adjusted Wald approach 4) Updated unit to use grams in summaries of cumulative dose of sorafenib 5) For Per Protocol population, updated to exclude patients in the control arm receiving any Y90 treatment (including TheraSphere) prior to progression

SIGNATURE PAGE

Author:

[redacted]
Principal Biostatistic
Labcorp

Reviewed by:

[redacted]
Director, Biostatistics
Labcorp

Date

Approved by:

[redacted]
Fellow, Biostatistics
Boston Scientific

Date

Approved by:

[redacted]
Medical Safety Director
Boston Scientific

Date

Table of Contents

1 LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS 9

2 INTRODUCTION 11

3 STUDY OBJECTIVE 11

4 STUDY DESIGN 11

4.1 GENERAL DESIGN 11
4.2 METHOD OF ASSIGNMENT OF PATIENTS TO TREATMENT GROUPS 12
4.3 BLINDING 13
4.4 DETERMINATION OF SAMPLE SIZE 13

5 CHANGES IN THE CONDUCT OF THE STUDY OR PLANNED ANALYSES 13

5.1 CHANGES IN THE CONDUCT OF THE STUDY 13
5.1.1 Number of Study Centers and Subjects 14
5.1.2 Randomization Stratification 14
5.1.3 General Design 14
5.2 CHANGES IN THE PLANNED ANALYSES 14
5.2.1 Futility Stopping Rule 14
5.2.2 Secondary Endpoints 15
5.2.3 Sensitivity Analysis 15
5.2.4 [redacted] 15
5.2.5 Analysis Population 15
5.2.6 Treatment Emergent Adverse Event (TEAE) 16

6 BASELINE, EFFICACY AND SAFETY EVALUATIONS 16

6.1 SCHEDULE OF EVALUATIONS 16
6.2 TIME POINT ALGORITHMS 18
6.2.1 Relative Days 18
6.2.1.1 For Assessments except Adverse Events 18
6.2.1.2 For Adverse Events 18
6.2.1.3 Partial Dates 18
6.2.2 Windows 20
6.3 BASELINE ASSESSMENTS 20
6.4 EFFICACY VARIABLES 21
6.4.1 Primary Efficacy Variable – Overall Survival (OS) 21
6.4.2 Secondary Efficacy Variables 21
6.4.2.1 Objective Response Rate (ORR) according to RECIST v1.1 criteria by investigator determination 22
6.4.2.2 Time to Progression according to RECIST v1.1 criteria by investigator determination 24
6.4.2.3 Quality of Life Assessments (FACT-Hep) 24
6.4.2.4 Time to Untreatable Progression 25
6.4.2.5 Time to Symptomatic Progression 25
6.4.3 Additional Efficacy Variables 26
6.4.3.1 TTP according to mRECIST criteria by blinded central image review 26
6.4.3.2 Disease control rate (DCR) according to RECIST v1.1 criteria by investigator determination 26
6.4.3.3 ORR according to mRECIST criteria by blinded central image review 26

6.4.3.4 DCR according to mRECIST criteria by blinded central image review 27
6.4.3.5 PFS according to RECIST v1.1 criteria by investigator assessment 27
6.4.3.6 PFS according to mRECIST criteria by blinded central image review 27
6.4.3.7 Duration of objective response 27
6.4.3.8 Duration of Disease Control 27
6.4.3.9 Depth of response (DoR) 28
6.4.3.10 Post-Treatment Tumor Shrinkage (PTTS) 28
6.4.3.11 Tumor Marker for HCC (Alpha Fetoprotein, AFP) 28
6.5 SAFETY ASSESSMENTS 29
6.5.1 Extent of Exposure and Compliance to Study Treatment 29
6.5.1.1 Extent of Exposure to TheraSphere 29
6.5.1.2 Extent of Exposure to Sorafenib 30
6.5.1.3 Extent of Study Exposure 31
6.5.1.4 Best Available Care Post-Progression 31
6.5.2 Adverse Events 31
6.5.2.1 Serious Adverse Event (SAE) 32
6.5.2.2 Adverse Device Effect (ADE) 32
6.5.2.3 Serious Adverse Device Effect (SADE) 33
6.5.2.4 Adverse Events of Special Interest (AESI) 33
6.5.3 Clinical Laboratory Evaluations 33
6.5.4 Additional Safety Variables 33
6.5.4.1 ECOG Performance Status 33
6.5.4.2 Child-Pugh Score Status 34
6.5.4.3 Albumin-Bilirubin Score 34
6.6 [redacted] 34

7 STATISTICAL METHODS 34

7.1 GENERAL METHODOLOGY 35
7.2 ADJUSTMENTS FOR COVARIATES 35
7.3 HANDLING OF DROPOUTS OR MISSING DATA 37
7.4 INTERIM ANALYSES AND DATA MONITORING 37
7.5 MULTIPLE COMPARISONS/MULTIPLICITY 39
7.6 USE OF AN "EFFICACY SUBSET" OF PATIENTS 39
7.7 EXAMINATION OF SUBGROUPS 39

8 STATISTICAL ANALYSIS 41

8.1 DISPOSITION OF PATIENTS 41
8.2 PROTOCOL DEVIATIONS 41
8.3 ANALYSIS POPULATIONS 41
8.3.1 Modified Intent-to-Treat (mITT) Population 41
8.3.2 Safety Analysis (SA) Population 41
8.3.3 Per Protocol (PP) Population 42
8.3.4 [redacted] 43
8.4 DEMOGRAPHIC AND OTHER BASELINE CHARACTERISTICS 43
8.5 PRIOR AND CONCOMITANT THERAPY 45
8.5.1 Prior Medication 45
8.5.2 Prior Therapy for HCC 45
8.5.3 Concomitant Medication 45
8.6 ANALYSIS OF EFFICACY PARAMETERS 46
8.6.1 Analysis of Primary Efficacy Variable 46
8.6.2 Analysis of Secondary Efficacy Variables 47
8.6.2.1 Time to Progression (TTP) according to RECIST v1.1 criteria by investigator determination 48
8.6.2.2 Time to Untreatable Progression (TTUP) 49

8.6.2.3 Time to Symptomatic Progression (TTSP) 51
8.6.2.4 Objective Response Rate (ORR) according to RECIST v1.1 criteria by investigator determination 52
8.6.3 Analysis of Quality of Life Questionnaire (FACT-hep) *52*
8.6.3.1 Analysis of FACT-hep Scores 52
8.6.3.2 Analysis of Time to Deterioration in QoL (TTDQoL) 52
8.6.4 Subgroup Analyses *53*
8.6.5 Assessment of Poolability *54*
8.6.6 Additional Efficacy Analyses *55*
8.7 ANALYSIS OF SAFETY 57
8.7.1 Extent of Exposure to Study Treatment *57*
8.7.1.1 Extent of Exposure to TheraSphere 57
8.7.1.2 Sorafenib 57
8.7.1.3 Extent of Study Exposure and Follow-up 58
8.7.1.4 Best Available Care Post-Progression 58
8.7.2 Adverse Events *59*
8.7.3 Clinical Laboratory Evaluations *60*
8.8 ADDITIONAL SAFETY ANALYSES 60
8.9 [redacted]

9 COMPUTER SOFTWARE 62

10 REFERENCES 63

11 APPENDICES 64

11.1 APPENDIX 1: VARIABLE DEFINITIONS 64
11.2 APPENDIX 2: STATISTICAL ANALYSIS AND PROGRAMMING DETAILS 65
11.3 APPENDIX 3: FACT-HEP QUESTIONNAIRE SCORING RULES 66
11.4 APPENDIX 4: DEFINITION AND DERIVATION OF SUBSEQUENT HCC THERAPY 69
11.5 APPENDIX 5: STATISTICAL DETAILS OF THE ADAPTIVE DESIGN FOR PROTOCOL TS-103 STOP-HCC 70

1 LIST OF ABBREVIATIONS AND DEFINITIONS OF TERMS

Table 1: Abbreviations and Definitions of Terms

AD	Absorbed Dose
ADE	Adverse Device Effect
AE	Adverse Event
AESI	Adverse Event of Special Interest
AFP	Alpha fetoprotein
ALBI	Albumin-Bilirubin
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ATC	Anatomical Therapeutic Chemical
BCLC	Barcelona Clinic Liver Cancer
CI	Confidence Interval
CR	Complete Response
CT	Computed Tomography
CTC	Common Terminology Criteria
CTCAE	Common Toxicity Criteria for Adverse Events
DCO	Data Cut-Off
ECOG	Eastern Cooperative Oncology Group
eCRF	electronic Case Report Form
FACT –hep	Functional Assessment of Cancer Therapy – hepatobiliary
FDA	Food and Drug Administration
Gy	Gray, a measure of irradiation dose
HAP	Hepatoma Arterial-Embolization Prognostic
HCC	Hepatocellular Carcinoma
HCS	Hepatobilliary Cancer Subscale
HR	Hazard Ratio
IDMC	Independent Data Monitoring Committee
INR	International Normalized Ratio
mITT	modified Intent-To-Treat
IVRS	Interactive Voice Response System
ALBI	Albumin-Bilirubin
MedDRA	Medical Dictionary for Regulatory Activities
MR, MRI	Magnetic Resonance, Magnetic Resonance Image
mRECIST	modified Response Evaluation Criteria in Solid Tumor
NCI	National Cancer Institute
NE	Not Evaluable
NTAD	Normal Tissue Absorbed Dose
NTCP	Normal Tissue Complication Probability
ORR	Objective Response Rate

OS	Overall Survival
QoL	Quality of Life
PD	Progressive Disease
PP	Per Protocol
PR	Partial Response
PT	Prothrombin Time
PTT	Partial Thromboplastin Time
RECIST	Response Evaluation Criteria in Solid Tumor
SADE	Serious Adverse Device Effect
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
SoC	Standard of Care
^{99m}Tc-MAA	Technicium-99m Macroaggregated albumin
TAD	Tumor Absorbed Dose
TCP	Tumor Control Probability
TEAE	Treatment Emergent Adverse Event
TS	TheraSphere
TTD	Time to Deterioration
TTP	Time-to-Progression
TTSP	Time to Symptomatic Progression
TTUP	Time to Untreatable Progression
VIF	Variance Inflation Factor
WHO DE	World Health Organization Drug Enhanced

2 INTRODUCTION

According to the International Agency for Research on Cancer[1], primary liver cancer is a major health problem worldwide. Globally, it is the sixth most commonly diagnosed cancer, with more than 749,000 new cases in 2011. It is the third leading cause of cancer death in men and sixth among women. In North America and Western or Northern Europe, areas with historically low rates, the incidence of liver cancer is increasing, possibly due to increased prevalence of hepatitis C.

Based on published reports, there is extensive clinical experience demonstrating the safety of TheraSphere in the management of patients with unresectable hepatocellular carcinoma (HCC). Early reports of serious adverse events possibly associated with the use of TheraSphere included death, hepatorenal failure, liver abscess, hepatic encephalopathy, hepatic decompensation, radiation hepatitis, radiation pneumonitis, duodenal ulcer, gastrointestinal bleeding and cholecystitis. As clinical experience with TheraSphere increased, the pre-treatment high risk factors associated with these early serious events were identified, leading to improved patient selection criteria and thereby lowering the risk of these events occurring. These risk factors include infiltrative tumor type, bulk disease (tumor volume >70% or nodules too numerous to count), AST or ALT > five times the upper limit of normal, bilirubin >3 mg/dL, tumor volume >50% in the presence of albumin <3 g/dL and those in whom extra-hepatic shunting to the lungs or gastrointestinal tract cannot be managed through standard angiographic techniques.

For those patients without the pre-treatment high risk factors noted above, TheraSphere is very well tolerated, with treatment in the United States commonly administered in an outpatient setting. Hospitalization for treatment effects associated with TheraSphere administration is rare. The most commonly reported adverse events associated with TheraSphere administration are fatigue, abdominal pain, nausea/vomiting and transient laboratory values including elevated bilirubin, AST, ALT, alkaline phosphatase, decreased platelets and lymphocyte depression with no clinical sequelae.

3 STUDY OBJECTIVE

The objective of this study is to evaluate TheraSphere in the treatment of patients with unresectable HCC.

4 STUDY DESIGN

4.1 General Design

This is an open-label, prospective, multi-center, randomized, clinical trial. The primary efficacy endpoint of the trial is overall survival (OS).

Patients with unresectable HCC in whom standard of care (SoC) sorafenib therapy is planned are eligible to participate. The trial will evaluate the use of TheraSphere followed by the SoC sorafenib treatment. Up to 105 study centers will participate and

recruit patients. Participating study sites may be in the United States, Canada, Europe, and Asia. All patients will be followed prospectively from randomization to death.

Eligible patients will be randomized (1:1) to either the Control Group or the Treatment group, defined as follows:

Patients will have regular clinical study visits as long as they participate in the trial. During these visits, safety and efficacy data will be collected and recorded.

A feasibility safety assessment will be conducted after the first 20 patients in the treatment group have received TheraSphere followed by at least 2 weeks of sorafenib therapy. The IDMC will review the safety results of both the control and treatment groups.

4.2 Method of Assignment of Patients to Treatment Groups

Patients will be randomized to study treatment, either the Control group or the Treatment group in a 1:1 ratio.

At study enrollment, each patient will be assigned a subject identity code (e.g. T030103-001) consisting of the protocol number (T03), the country number (e.g. 01), the site number (e.g. 01), and patient number (e.g. 001).

If a patient is determined to be eligible to participate in the trial, the study site will contact the central randomization office when randomization will be determined using assignment by a computer-generated randomization scheme. Upon randomization, each patient will be assigned a 4 digits randomization number with the first digit indicating which combination of the 3 stratification factors the patient has.

A centralized randomization schedule will be generated by a statistician in the Labcorp, Biometrics department who is not associated with the conduct or analysis of the study, using a validated system. The randomization will be stratified by the following factors:

- Region (North America and Europe vs Asia)
- ECOG performance status (0 vs 1)
- Presence or absence of branch portal vein thrombosis (PVT)

In order to ensure that the study treatment groups are balanced, the schedule will have randomization numbers assigned to the 2 study treatments in blocks of 4 within each combination of the 3 stratification factors to achieve a 1:1 ratio of study treatment (i.e. an equal number of patients in each treatment group). The randomization will be performed using IVRS by Perceptive, Inc. Each eligible patient will be assigned to the next sequential randomization number within the specified stratification combination and will receive the corresponding study treatment.

Patients randomized to the Control group or the Treatment group who are unable to receive their planned study treatment will continue to be followed under the study group to which they were randomized for the purpose of the modified intent-to-treat (mITT) analysis (see Section 8.3.1).

4.3 Blinding

This is an open label study and there is no blinding.

To maintain the integrity of the study results in this open label study, the following personnel who had access to the study data before database lock, were required to sign an "Aggregate Data Declaration Form" documenting their agreement to not produce or review aggregate summaries of efficacy and death data, including AEs with an outcome of death, separated by treatment arm:

- CRO (i.e. Labcorp) personnel who were not involved in preparing data summaries for the independent data monitoring committee (IDMC) meetings, and
- Sponsor personnel.

4.4 Determination of Sample Size

This study is an adaptive trial using a group sequential design with OS as the primary efficacy endpoint. The study is designed to detect a 3.5 months increase in median OS time, from 10.7 months in the sorafenib arm to 14.2 months in the TheraSphere arm (i.e. hazard ratio [HR]= 0.754), using a log rank test. Due to uncertainty in the expected treatment effect, a sample size re-estimation is planned, which would allow the sample size to increase in order to detect a 3.0 month increase in median OS time, from 10.7 months in sorafenib arm to 13.7 months in the TheraSphere arm (i.e. HR = 0.781).

A maximum of 417 deaths will yield 80% power to detect the target difference in median OS (i.e. HR = 0.754) with a two-sided alpha of 0.05 using a group sequential design with 2 interim analyses. It is estimated that a maximum of 520 patients will need to be recruited over 60 months, with an 18 months additional follow-up period. This includes an adjustment to take account of an assumed 5% of patients who will be lost to follow-up and for whom a date of death is not recorded, and an assumed additional 5% of patients who will erroneously be randomized because they did not meet the eligibility criteria at randomization.

5 CHANGES IN THE CONDUCT OF THE STUDY OR PLANNED ANALYSES

5.1 Changes in the Conduct of the Study

The following protocol versions have been implemented for this study:

- Version 2.0 dated 24Jan2012
- Version 3.0 dated 06Sep2013
- Version 4.1 dated 20May2014
- Version 5.0 dated 08Jan2016
- Version 6.1 dated 29Nov2016

- Version 7.0 dated 22Jul2019

A separate protocol (version 4.2 dated 11Dec2014) was implemented for sites in Germany, where different eligibility criteria were used, mainly related to limits on liver function tests.

5.1.1 Number of Study Centers and Subjects

In Version 4.1 of the protocol, dated May 20th 2014, the number of study centers increased from 40 to 105 and the number of patients changed from 400 to 390 with up to a maximum of 600 based on a sample size re-estimation.

In the current protocol (Version 6.1, dated November 29th 2016) the number of patients changed from 390 to 520 with the maximum based on a sample size re-estimation increased from 600 to 700 patients. Also, the analysis population for efficacy endpoints was changed from an intention to treat population, defined as all randomized patients, to a mITT population, defined as randomized patients who met the study eligibility criteria at randomization. This protocol amendment was implemented before the first interim analysis was conducted.

5.1.2 Randomization Stratification

In Version 4.1 of the protocol dated May 20th2014, the stratification factors to be used for randomization were updated to remove HCC status (unilobar vs bilobar) and replace it with Region (North America and Europe vs. Asia). Patients randomized prior to this change would not be stratified by region. 74 patients were randomized under the HCC status stratification. Details of both randomizations can be found in each Randomization Plan Document.

5.1.3 General Design

In Version 4.1 of the protocol dated May 20th 2014, the design of the trial was amended to be an adaptive trial using a group sequential design with 2 interim analyses including a sample size re-estimation.

5.2 Changes in the Planned Analyses

5.2.1 Futility Stopping Rule

An assessment of futility at the two planned interim analyses, based on conditional power, was included in the study design. However, it was decided by the Sponsor, before the first interim analysis was performed, that the futility assessment would not be performed. This was primarily because patient recruitment was faster than expected towards the latter part of the study, such that all 520 patients for the original sample size had already been randomized before the first interim analysis was performed.

5.2.2 Secondary Endpoints

In Version 5.0 of the protocol dated Jan 8th 2016, a sequential hierarchical approach was added to control the study-wise Type I error rate. Also, in Version 5.0 supportive analyses were added using the Cox regression model to evaluate the effect of multiple covariates, including stratification factors, on the secondary efficacy time-to-event endpoints.

In the current protocol, section 10.2.3.2 states that time to progression (TTP) will be calculated as the interval between the randomization date and the date of first disease progression, including death for any cause. However, death will not be considered as a TTP event. Section 6.4.2.2 of this SAP explains how TTP will be analyzed.

In the current protocol, sections 10.2.3.1 and 10.2.5 state that tumor response rate per RECIST 1.1 by investigator determination and per mRECIST by blinded centralized independent imaging assessment will be compared between treatment arms using the continuity adjusted Newcombe-Wilson test. However, the continuity adjusted Newcombe-Wilson approach only provides confidence intervals and not p-values. Hence, the continuity adjusted Wald approach will be used instead, which provides both confidence intervals and p-values.

5.2.3 Sensitivity Analysis

In Version 5.0 of the protocol dated Jan 8th2016, a sensitivity analysis on the primary endpoint, OS, was added to address the poolability of data across regions, study sites, and gender.

[redacted]

5.2.5 Analysis Population

For the Per Protocol population, the protocol states that patients in this population will be analyzed according to the treatment actually received. Since patients who did not receive the treatment they were randomized to receive prior to progression by investigator assessment will be excluded from the Per Protocol population, the treatment actually received will be identical to the treatment randomized to receive. Hence, patients in this population will be analyzed according to the treatment group to which they were randomized.

For the safety analysis population, the protocol defines this population as all randomized patients who received at least one administration of study treatments and will be analyzed according to the treatment actually received. However, for patients with progression by investigator assessment, only study treatments received prior to progression will be used in the derivation of this population. Since both TheraSphere and sorafenib are allowed as best available care post progression, there are patients who received part of their randomized study treatment only after progression.

5.2.6 Treatment Emergent Adverse Event (TEAE)

The protocol defines a TEAE as an event that was not present at baseline or worsened in severity following the start of treatment. According to the protocol, AEs will only be collected until 30 days from discontinuation of sorafenib and after this period only AEs related to TheraSphere will be collected. However, to make the treatment groups comparable and to avoid the impact of subsequent HCC therapy on the evaluation of the AE profile, in Section 6.5.2, an AE with an onset date or a pre-existing AE worsening beyond 30 days after the end of the last study treatment initiated prior to progression date by investigator assessment, or the start date of subsequent HCC therapy, whichever comes first, will not be considered a TEAE.

6 BASELINE, EFFICACY AND SAFETY EVALUATIONS

6.1 Schedule of Evaluations

The assessments to be conducted at each scheduled visit are displayed in the following table

Table 2 Assessments Conducted at each Scheduled Visit

Evaluation/Test	**Screen**	**Rand-omize**	**Sorafenib for the Control Group**	**1st TS work up & Admin-istration**	**2nd TS work up & Admin-istration**	**Sorafenib for the Treatment Group**	**TS work up & re-treatment[1]**	**Follow Up Until Death**	
Timing of Visit(s)	**Days -14 to 0**	**Study Day 0**	**Weeks 1-4 initiate Weeks 5 & thereafter continue therapy**	**Weeks 1-4**	**Weeks 5-8**	**>2 & <6 weeks after TS – initiate & thereafter continue therapy**	**After hepatic progression**	**Q8 weeks ± 14 days[8]**	
								Prior to PD[9]	**Post PD**
Informed Consent	X								
Demographics	X								
Medical History	X								
Physical Exam	X								
ECOG Performance	X			X	X		X	X	

Status									
Medication & Prior Treatment History	X								
Review Eligibility Criteria	X								
Hematology: WBC, HgB, Hct, Platelets	X		X[7]	X[7]			X	X	
Coagulation: PT, PTT, INR	X		X[7]	X[7]			X	X	
Chemistry panel, liver function tests	X		X[7]	X[7]			X	X	
Serum Pregnancy[2]	X						X		
Tumor markers for HCC (AFP)	X		X	X				X	
Liver Volume/Mass Calculation	X			X			X		
Randomize Patient		X							
Hepatic Angiogram, ^{99m}Tc-MAA scan, TS Dose Calculation[3]				X	X		X		
Order TS[3]				X	X		X		
Administer TS[3, 4]				X	X		X		
Administer Sorafenib[5]			X			X	X[6]		
QOL questionnaire	X							X	
Triple Phase MRI/Spiral CT of abdomen	X							X	
Child-Pugh score	X							X	
Spiral CT of chest and pelvis	X							X	
Assess/Report Adverse Events	X		X	X	X		X	X	
Review/Record Concurrent Medication	X		X	X	X		X	X	
Final Endpoint Efficacy/Safety documentation & exit patient								X	X

1 Additional TS work up & Administration in lesions amenable to further TS treatment
2 Female patients of childbearing potential only
3 TS patients only
4 Additional TS treatments may be administered only after progression if lesions are amenable to treatment
5 According to package insert at Weeks 1-4 for Control group patients and after all initial TS administrations for Treatment group patients only

6 Sorafenib to be stopped 7 days before subsequent TS administration on disease progression and restarted 2 weeks after TS is administered
7 If treatment commences within 14 days of randomization the clinical laboratory assessments are not required to be repeated
8 The follow-up visits should be scheduled from the day of randomization. A window of ± 14 days is permissible from the scheduled date
9 Progression of disease resulting in termination of further treatment

6.2 Time Point Algorithms

6.2.1 Relative Days

6.2.1.1 For Assessments except Adverse Events

For all assessments except adverse events, the following relative day calculation will be used.

The date of randomization will be considered relative day 1, and the day before the randomization will be relative day -1. Relative days will be calculated as follows:

For days on or after randomization:
Date of Assessment – Date of Randomization + 1

For days before randomization:
Date of Assessment – Date of Randomization

6.2.1.2 For Adverse Events

For adverse events, the following relative day calculation will be used.

The start date of sorafenib and the date of first angiogram (whichever occurs first) will be considered relative day 1, and the day before the start date of sorafenib and the date of first angiogram (whichever occurs first) will be relative day -1. Relative days will be calculated as follows :

For days on or after the start date of sorafenib and the date of first angiogram (whichever occurs first):
Date of Assessment – Start date of sorafenib and date of first angiogram (whichever occurs first) + 1

For days before the start date of sorafenib and the date of first angiogram (whichever occurs first):
Date of Assessment – Start date of sorafenib and date of first angiogram (whichever occurs first)

6.2.1.3 Partial Dates

Partial dates with day or day and month missing will be imputed as follows:

- The missing day of onset of an adverse event (AE) will conservatively be set to the earlier of:
    - First day of the month of the AE start month, if the month of the start date of sorafenib and the date of first angiogram (whichever occurs first) is not the same as the AE start month,
    - One day after the start date of sorafenib and the date of first angiogram (whichever occurs first), if the month of the start date of sorafenib and the date of first angiogram (whichever occurs first) is the same as the AE start month.
- The missing day of resolution of an AE will be set to the last day of the month of the AE end month.
- For other variables, including date of tumor response, and progression, partial dates that need to be imputed will use the 15th of the month to replace the missing day.
- A missing day of death will be replaced by the 15th of the month if there are no other assessments after the 15th of the month for that patient. Otherwise the last day of the month will be used to replace the missing day of death.
- If the onset date of an AE is missing both day and month, it will be set to:
    - January 1 of the year of AE start year, if the year of the start date of sorafenib and the date of first angiogram (whichever occurs first) is not the same as the AE start year,
    - One day after the start date of sorafenib and the date of first angiogram (whichever occurs first), if the year of the start date of sorafenib and the date of first angiogram (whichever occurs first) is the same as the AE start year.
- If the resolution date of an AE is missing both day and month, it will be set to December 31 of the AE end year.
- For the date of initial diagnosis of HCC, a missing day will be set to the first day of the month, and a missing day and month will be set to January 1 of the year.
- For the start date of medications recorded on the "Prior and Concurrent Medications" electronic case report form (eCRF) page, a missing day will be set to the first day of the month, and a missing day and month will be set to January 1 of the year. If month/year of the start date (if only missing day) or year of the start date (if missing month/day) is the same as the month/year or year of the randomization date, respectively, the start date will be set to one day after the randomization date.
- For the end date of medications recorded on the "Prior and Concurrent Medications" eCRF page, a missing day will be set to the last day of the month, and a missing day and month will be set to December 31 of the year.
- For the start date of medications recorded on the Best Available Care - Post Treatment Discontinuation – Medication eCRF page and the start date of

procedures recorded on the Additional Procedures eCRF page, a missing day will be set to:

- the first day of the month if the month/year of progression (as determined by the investigator) is not the same as the month/year of the start of best available care medication or additional procedure,
- One day after the date of progression (as determined by the investigator) if the month/year of progression is the same the month/year of the start of best available care medication or additional procedure.

- For the end date of medications recorded on the Best Available Care - Post Treatment Discontinuation – Medication eCRF page and the end date of procedures recorded on the Additional Procedures eCRF page, a missing day will be set to the last day of the month.

6.2.2 Windows

For the purpose of statistical analysis, time windows will need defining for presentations that summarize values by visit. The windows for the visits following baseline will be constructed in such a way that the upper limit of the interval falls half way between the two visits (the lower limit of the first post-baseline visit will be Day 2), as shown below. The assignment of data to visit windows will use the relative day defined in Section 6.2.1.1.

Table 3 Analysis Windows for Assessments Performed at Eight Week Intervals

Week	Scheduled Day	Visit Window for Analysis (Days)
Week 8	57	2 – 85
Week 16	113	86 – 141
Week 24	169	142 – 197
…		
End of Study		Latest assessment available*

*Last assessment should be assigned to a Week X based on the visit windows as well as to the end of study evaluation.

If a patient has more than 1 assessment occurring in the same visit window, the data from the assessment closest to the scheduled day will be used. If 2 assessments have the same distance from the scheduled day, the data of the assessment after the scheduled day will be used. Note that the visit windows will not be used for time-to-event endpoints and subject-level tumor response endpoints (i.e. objective response rate and disease control rate).

6.3 Baseline Assessments

Baseline will be defined as the last non-missing assessment performed on or before the day of randomization.

According to the protocol, the following assessments will be conducted prior to randomization:

- Informed Consent
- Inclusion/ Exclusion Criteria
- Demographics (age, gender, race, ethnicity)
- Medical history
- Physical examination
- Vital signs (heart rate, respiration rate, blood pressure, oral temperature, height and weight)
- Disease and treatment history
- ECOG performance status
- Laboratory tests (hematology, coagulation, chemistry)
- Serum pregnancy test
- Child Pugh Score
- HCC tumor biomarkers
- Liver volume/mass and tumor burden
- FACT-hep QOL
- CT/MRI of chest, abdomen, and pelvis
- Stratification factors: region (North America and Europe vs Asia), ECOG performance status (0 vs 1), presence or absence of branch PVT

Time from diagnosis of HCC will be calculated as follows:
Time from diagnosis of HCC (in months) = (Date of Randomization – Date of Diagnosis)/30.4375.

6.4 Efficacy Variables

For all efficacy evaluations, the baseline measurement is defined as the last measurement on or prior to the date of randomization. Any tumor assessments performed within 6 weeks of randomization are less accurate for tumor response assessment, so will not be included in the analysis of imaging related efficacy endpoints.

6.4.1 Primary Efficacy Variable – Overall Survival (OS)

The primary study endpoint is OS, which is defined as the time from date of randomization until date of death due to any cause.

6.4.2 Secondary Efficacy Variables

The secondary efficacy endpoints for this study are:

- Tumor Response, defined as Objective Response Rate (ORR) according to RECIST v1.1 criteria by investigator determination
- Time to Progression (TTP) according to RECIST v1.1 criteria by investigator determination
- Quality of Life Assessments
- Time to Untreatable Progression (TTUP)

- Time to Symptomatic Progression (TTSP)

6.4.2.1 Objective Response Rate (ORR) according to RECIST v1.1 criteria by investigator determination

Tumor Response is based on the radiological tumor assessment performed at specified time points. The post baseline assessments are compared to the baseline assessment and the overall response based on investigator assessment according to RECIST criteria v1.1 is recorded at each efficacy visit. The tumor response for target lesions is categorized as Complete Response (CR), Partial Response (PR), Stable Disease (SD), Progressive Disease (PD) or Not all Evaluated (NE) according to the RECIST criteria v1.1[2] as shown in Table 4. Table 5 shows the responses for non-target lesions.

Table 4: Target lesion response categories

Response	Definition
Complete Response (CR)	Disappearance of all target lesions
Partial Response (PR)	≥30% decrease in the sum of the diameters of target lesions
Stable Disease (SD)	Neither CR nor PR nor PD
Progressive Disease (PD)	≥20% increase in the sum of the longest diameter of target lesions from smallest value on study The sum of diameters must also demonstrate an absolute increase of at least 5 mm, e.g. two lesions increasing from 2 mm to 3 mm does not qualify
Not all Evaluated (NE)[a]	When imaging/measurement is not done at all at a particular time point, the patient is not evaluable (NE) at that time point. If only a subset of lesion measurements is made at an assessment, usually the case is also considered NE at that time point, unless a convincing argument can be made that the contribution of the individual missing lesion(s) would not change the assigned time point response.

Source: Protocol version 7.0, 22-Jul-2019; [a] Source: Eisenhauer et al (2009)[2]

Table 5: Non-target lesion response categories

Response	Definition

CR	Disappearance of all non-target lesions All non-target lymph nodes must be non-pathological in size (<10 mm short axis)
Non-CR/Non-PD	Persistence of 1 or more non-target lesions
PD[a]	Unequivocal progression of existing non-target lesions
Not Evaluable (NE)[a]	When no imaging/measurement is done at all at a particular time point, the patient is not evaluable (NE) at that time point. If only a subset of lesion measurements are made at an assessment, usually the case is also considered NE at that time point, unless a convincing argument can be made that the contribution of the individual missing lesion(s) would not change the assigned time point response.
[a] According to Eisenhauer et al (2009)[2] "to achieve 'unequivocal progression' on the basis of the non-target disease, there must be an overall level of substantial worsening in non-target disease such that that, even in presence of SD or PR in target disease, the overall tumor burden has increased sufficiently to merit discontinuation of therapy. A modest 'increase' in the size of one or more non-target lesions is usually not sufficient to qualify for unequivocal progression status."	

Source: Eisenhauer et al (2009)[2]

Table 6 provides a summary of the overall response status calculation at each timepoint.

Table 6: Timepoint response

Target lesions	Non-target lesions	New lesions	Timepoint response
CR	None	No	CR
PR	None	No	PR
SD	None	No	SD
PD	None	No	PD
Any	None	Yes	PD
CR	CR	No	CR
CR	Non-CR/Non-PD	No	PR
CR	Not evaluated	No	PR
PR	Non-PD or not all evaluated	No	PR
SD	Non-PD or not all evaluated	No	SD
Not all evaluated	Non-PD	No	NE
PD	Any	Yes or No	PD
Any	PD	Yes or No	PD
Any	Any	Yes	PD

None	CR	No	CR
None	Non-CR/Non-PD	No	Non-CR/Non-PD
None	Not at all evaluated	No	NE
None	Unequivocal PD	Yes or No	PD
None	Any	Yes	PD

Source: eCRF

The best overall response is based on the overall responses from each imaging assessment according to RECIST 1.1. It is the best response a patient has had following randomization, but up to and including the first PD or the last valid post baseline imaging assessment in the absence of the first PD.

If a patient received a subsequent systemic anticancer treatment (excluding sorafenib) and/or non-protocol liver directed therapy (excluding ablation/surgery) (henceforth referred to as "subsequent HCC therapy" in this document for ease of reference; see Appendix 4 for details), tumor response assessments after the start date of the subsequent HCC therapy will be excluded from the calculation of best overall response.

The ORR is defined as the proportion of randomized patients achieving a best overall response of CR or PR.

Patients who do not have any post baseline tumor assessments for any reason on or prior to the start date of the subsequent HCC therapy, are considered non-responders and are included in the denominator when calculating the ORR.

6.4.2.2 Time to Progression according to RECIST v1.1 criteria by investigator determination

This secondary endpoint is time to progression. TTP is defined as the time from date of randomization until date of radiological progression according to RECIST v1.1 criteria by investigator determination.

6.4.2.3 Quality of Life Assessments (FACT-Hep)

6.4.2.3.1 FACT-Hep Scores

The total score of the FACT-Hep QoL instrument will be calculated, the scores of each domain (Physical Well-Being, Social/Family Well-Being, Emotional Well-Being, Functional Well-Being), Hepatobiliary Cancer subscale (HCS), FACT-hep trial outcome index and each question at each time-point and their differences from baseline will be determined for each treatment group.

The scoring algorithm is in Section 11.3 Appendix 3.

6.4.2.3.2 Time to Deterioration in QoL (TTDQoL)

The time to deterioration in QoL is defined as the time from date of randomization to the assessment date when the change from baseline in FACT-Hep Total Score is ≤-7 points (i.e., a 7 point or greater decline in the total score) or date of death, whichever occurs first.

6.4.2.4 Time to Untreatable Progression

This secondary endpoint, time to untreatable progression, is defined as the time from date of randomization to date of untreatable progression, where untreatable progression is defined as one of the following events:

- Intolerance to sorafenib
- Occurrence of specific contraindications to sorafenib
- Assessment of progression in the target lesions, occurrence of new lesions after treatment, or death due to progression and, for patients randomized to the treatment group, a maximum of 2 re-treatments with TheraSphere
- Occurrence of specific contraindications to TheraSphere and or appearance of lung/intestinal shunts or anatomical constraints not correctable by radiological procedures for the Treatment group
- Confirmed extra-hepatic metastases
- Deterioration of liver function (Child Pugh score >B7)
- Clinical progression to ECOG performance status >1. Such deterioration in performance status should be observed at two subsequent evaluations at 8 week intervals.

The investigator will determine whether the patient met any of the protocol specified conditions. A response of yes to the question of "Is this progression considered untreatable according to the protocol definition?" on the Determination of Response or Progression CRF page will be used to indicate untreatable progression.

6.4.2.5 Time to Symptomatic Progression

Time to symptomatic progression (TTSP) is defined as the time from date of randomization to date of assessment of ECOG performance status >1 with or without tumor progression based on investigator assessment according to RECIST criteria v1.1 The symptomatic progression is confirmed at the first two subsequent evaluations at least 8 and 16 weeks later, respectively. The date of the first ECOG performance status >1 will be used as the event date in the TTSP analysis (assuming ECOG performance status >1 at the first two subsequent evaluations at least 8 and 16 weeks later, respectively). Illustrative examples are provided in Table 7.

Table 7 Examples of Symptomatic Progression

Week X	Week X+8	Week X+16	Week X+24	Week X+32	TTSP Event
ECOG>1	ECOG>1	ECOG>1	ECOG≤1	ECOG≤1	Yes, event at Week X

ECOG>1	ECOG>1	Missing	ECOG>1	ECOG≤1	Yes, event at Week X
ECOG>1	Missing	ECOG>1	ECOG>1	Missing	Yes, event at Week X
ECOG>1	ECOG≤1	ECOG>1	ECOG>1	Missing	No
ECOG>1	No further ECOG assessments				No
ECOG>1	ECOG>1	No further ECOG assessments			No

Week X denotes that week of the first occurrence of ECOG >1

6.4.3 Additional Efficacy Variables

6.4.3.1 TTP according to mRECIST criteria by blinded central image review

TTP by blinded central review is defined as the time from date of randomization until date of radiological progression by the blinded central image review, according to mRECIST.

6.4.3.2 Disease control rate (DCR) according to RECIST v1.1 criteria by investigator determination

DCR by investigator assessment is defined as the proportion of randomized patients achieving a best overall response of CR, PR, or SD as defined by RECIST v 1.1, as determined by the investigator.

If a patient received subsequent HCC therapy, imaging assessments after the start date of the subsequent HCC therapy will be excluded from the calculation of best overall response and DCR. Patients who do not have any post baseline tumor assessments for any reason on or prior to the start date of the subsequent HCC therapy, are considered non-responders and are included in the denominator when calculating the DCR.

6.4.3.3 ORR according to mRECIST criteria by blinded central image review

ORR by blinded central review is defined as the proportion of randomized patients achieving a best overall response of CR or PR, as defined by mRECIST, as determined by blinded central image review.

If a patient received subsequent HCC therapy, imaging assessments after the start date of the subsequent HCC therapy will be excluded from the calculation of best overall response and ORR. Patients who do not have any post baseline tumor assessments for any reason on or prior to the start date of the subsequent HCC therapy, are considered non-responders and are included in the denominator when calculating the ORR.

6.4.3.4 DCR according to mRECIST criteria by blinded central image review

DCR by blinded central review is defined as the proportion of randomized patients achieving a best overall response of CR, PR or SD, as defined by mRECIST, as determined by blinded central image review.

If a patient received subsequent HCC therapy, imaging assessments after the start date of the subsequent HCC therapy will be excluded from the calculation of best overall response and DCR. Patients who do not have any post baseline tumor assessments for any reason on or prior to the start date of the subsequent HCC therapy, are considered non-responders and are included in the denominator when calculating the DCR.

6.4.3.5 PFS according to RECIST v1.1 criteria by investigator assessment

PFS by investigator assessment is defined as the time from date of randomization until date of progression determined by the investigator, according to RECIST v1.1, or death due to any cause, whichever occurs first.

6.4.3.6 PFS according to mRECIST criteria by blinded central image review

PFS by blinded central image review is defined as the time from date of randomization until date of progression determined by the blinded central image review, according to mRECIST, or death due to any cause, whichever occurs first.

6.4.3.7 Duration of objective response

The duration of objective response will be determined for patients who have a best overall response of CR or PR. Duration of objective response is defined as the time from first date of overall response of CR or PR until date of PD, or death due to any cause, whichever occurs first. If a patient did not die or progress then the date of the last radiological assessment will be used in the calculation.

Duration of objective response will be assessed separately by investigator assessment by RECIST 1.1 and by blinded central image review by mRECIST.

6.4.3.8 Duration of Disease Control

The duration of disease control will be determined for patients who have a best overall response of CR, PR or SD. Duration of disease control is defined as the time from first date of overall response of CR, PR or SD until date of PD, or death due to any cause, whichever occurs first.

Duration of disease control will be assessed separately by investigator assessment by RECIST 1.1 and by blinded central image review by mRECIST.

6.4.3.9 Depth of response (DoR)

DoR is defined as the percentage change from baseline to nadir in the sum of the longest diameters of target lesions. If a patient received a subsequent HCC therapy, tumor assessments after the start date of the subsequent HCC therapy will be excluded from the calculation of DoR.

DoR will be assessed by investigator assessment by RECIST 1.1, and also by blinded central image review by mRECIST. In addition, DoR will be assessed for the following subgroups based on tumor replacement (%) at baseline:

- >10% tumor replacement by blinded central review
- >20% tumor replacement by blinded central review

6.4.3.10 Post-Treatment Tumor Shrinkage (PTTS)

PTTS is defined as the proportion of randomized patients achieving a ≥20% decrease in the sum of the longest diameters of target lesions, separately at the Week 8, 16, and 24 analysis visits (as defined in Table 3). PTTS will be assessed separately by investigator assessment by RECIST 1.1 and by blinded central image review by mRECIST.

If a patient received a subsequent HCC therapy, tumor assessments obtained after the start date of the subsequent HCC therapy will be excluded from the determination of achieving the threshold of PTTS at each analysis visit. Patients without post baseline tumor assessments on or prior to the start date of the subsequent HCC therapy are considered non-responders and are included in the denominator when calculating the PTTS.

6.4.3.11 Tumor Marker for HCC (Alpha Fetoprotein, AFP)

AFP will be collected along with laboratory data and will be presented similarly. Change from baseline will be calculated.

AFP response, defined as a ≥50% decrease in AFP levels for patients with a baseline AFP level of ≥200 ng/mL, will also be calculated. The following additional definitions of AFP response will also be calculated:

- a ≥50% decrease in AFP levels for patients with a baseline AFP level of ≥400 ng/mL
- a ≥20% decrease in AFP levels for patients with a baseline AFP level of ≥20 ng/mL

For each definition, the AFP response is derived per patient, and the percentage of patients achieving an AFP response is calculated based on the number of patients meeting the condition of the baseline AFP level in the definition. If a patient received a

subsequent HCC therapy, AFP assessments obtained after the start date of the subsequent HCC therapy will be excluded from the determination of AFP responses.

6.5 Safety Assessments

6.5.1 Extent of Exposure and Compliance to Study Treatment

6.5.1.1 Extent of Exposure to TheraSphere

TheraSphere exposure will be presented as described below for the Treatment arm. This includes summaries presented separately for TheraSphere administered prior to progression (i.e. prior to date of progression) and post progression (i.e. on or after date of progression) as assessed by investigator according to RECIST 1.1.

- Number of patients who received TheraSphere during the study
- Number of patients who received TheraSphere prior to progression and post progression
- Reasons for not receiving TheraSphere prior to progression
- Reasons for not receiving a second TheraSphere administration prior to progression for patients with bilobar disease
    - Note that bilobar disease is from the stratification factor on the Randomization eCRF page, with any incorrect values at randomization replaced with the corrected value from the eCRF, or values from eCRF for patients enrolled after the changes in stratification factor; this rule also applies to the derivation of unilobar or bilobar disease in later bullets
- Number of patients with bilobar disease at baseline who received TheraSphere prior to progression
    - to both lobes or to the whole liver (i.e. TheraSphere administered to both lobes in a non-lobar approach, for example through the common hepatic artery) on the same day,
    - to both lobes on different days,
        - to both lobes ≥28 days apart
        - to both lobes <28 days apart
    - to one lobe
- Number of patients with unilobar disease at baseline who received TheraSphere prior to progression
    - to both lobes or to the whole liver on the same day,
    - to both lobes on different days,
        - to both lobes ≥28 days apart
        - to both lobes <28 days apart
    - to one lobe
- Number of patients who received first TheraSphere administration post progression

- Patients with at least one TheraSphere administration not completed as planned prior to progression and separately post progression
- Time to the first angiogram (days) prior to progression, defined as (first angiography date prior to progression – randomization date + 1)
- Time to the first TheraSphere treatment (days) prior to progression, defined as (treatment date of first TheraSphere administration prior to progression - randomization date + 1)
- TheraSphere dose absorbed by perfused volume prior to progression and post progression
- TheraSphere dose delivered to lungs prior to progression and post progression

TheraSphere dose absorbed by perfused volume will be calculated as follows.

- Dose absorbed by perfused volume within a lobe (left lobe or right lobe) is defined using data for the corresponding lobe, as the sum of doses delivered by each vial if multiple vials are used to treat same target tissue, or as the weighted average of doses delivered by each vial (weights are target tissue masses) if multiple vials are not used to treat same target tissue.
- Dose absorbed by perfused volume within the liver is defined as the weighted average of doses delivered to each lobe (weights are the sum of target tissue masses in each lobe) for patients who had both lobes treated, and as the single dose delivered for patients who received whole liver dosing.
- Dose absorbed by perfused volume is defined as the dose absorbed by the perfused volume within the treated lobe for patients who had one lobe treated, and dose absorbed by the perfused volume within the liver for patients who had both lobes treated or who received whole liver dosing.

TheraSphere dose delivered to lungs will be calculated as the sum of doses delivered to lungs across all TheraSphere administrations.

6.5.1.2 Extent of Exposure to Sorafenib

Sorafenib exposure prior to progression (i.e. prior to date of progression) as assessed by investigator according to RECIST 1.1, will be presented as described below by treatment group.

- Number of patients who received sorafenib during the study
- Number of patients who received sorafenib prior to progression
- Reasons for not receiving sorafenib prior to progression
- Time to the start of sorafenib (days) prior to progression, defined as (start date of first sorafenib administration prior to progression - randomization date + 1)
    - This summary will also be produced separately for patients with unilobar and bilobar disease at baseline (Note that unilobar or bilobar disease is

from the stratification factor on the Randomization eCRF page, with any incorrect values at randomization replaced with the corrected value from the eCRF, or values from eCRF for patients enrolled after the changes in stratification factor)

- Cumulative dose (g) of sorafenib prior to progression
- Dose intensity (mg/day) of sorafenib prior to progression
- Relative dose intensity (%) of sorafenib prior to progression
- Duration of treatment (weeks) of sorafenib prior to progression
- Reason for Dose Delays and Changes for sorafenib prior to progression

6.5.1.3 Extent of Study Exposure

The duration on study will be determined.

6.5.1.4 Best Available Care Post-Progression

Systemic treatments and liver directed therapies received post-progression according to RECIST v1.1 by investigator assessment will be summarized as follows:

- Systemic treatments received post-progression
  - from the "Best Available Care - Post Treatment Discontinuation – Medication" eCRF page with start date on or after date of progression, by preferred terms (classification based on World Health Organization Drug Enhanced [WHO DE] March 2011)
  - from the "Sorafenib Administration" eCRF page with start date on or after date of progression
- Liver directed therapies received post-progression
  - from "Additional Procedures" eCRF page with start date on or after date of progression, categorized according to a manual review of the free text entered in the "Procedure term" eCRF field
  - from "TheraSphere Doses Administered" eCRF page with treatment date on or after date of progression

6.5.2 Adverse Events

All adverse events (AEs) will be documented from the date of randomization until study exit. For patients who permanently discontinue sorafenib in either arm of the study, AEs will be documented for 30 days from the date of discontinuation. After this period, only AEs considered to be related to TheraSphere will be collected.

A treatment emergent AE (TEAE) is defined as an event that was not present at baseline or worsened in severity following the start of treatment through to 30 days after the end of the last study treatment (sorafenib and/or TheraSphere) initiated prior to progression

date by investigator assessment, or start date of subsequent HCC therapy, whichever comes first.

The start of treatment will be defined as the start date of sorafenib and the date of first angiogram, prior to progression date by investigator assessment, whichever occurs first. The last study treatment will be sorafenib for patients who received only sorafenib or both sorafenib and TheraSphere prior to progression, and will be TheraSphere for patients who only received TheraSphere prior to progression. Partial AE start and stop dates will be imputed as described in Section 6.2.1.3.

The investigator's verbatim term of an AE will be mapped to a system organ class and preferred term using the MedDRA Version 14.0 dictionary (Medical Dictionary for Regulatory Activities). The investigator will use the NCI Common Toxicity Criteria for Adverse Events [CTCAE] (version 4.0) or the protocol specific criteria when no NCI CTC criteria are available for the AE to determine the severity of the AE.

Adverse events related to sorafenib, device, and angiographic procedures are defined as those events recorded on the eCRF with relationship of possibly, probably, or definitely relationship. Relation to TheraSphere (device) is not appropriate for the Control group.

The incidence of TEAEs will be the number of patients who had the AE (counted only once) divided by the number of patients in the safety population and represented as a percentage. For subgroup summaries of AEs, the percentage will be represented as the number of patients with the AE event divided by the number of patients of that group in the safety population. The incidence of AEs will be the number of times an event occurs, counting worsening events only once. For worsening events, the AE end date of the earlier AE will be the same as the start date of the same AE with a higher severity.

Adverse events reported for the Treatment group patients will be split further into three groups, all TEAEs, TEAEs after angiogram before start of sorafenib, and TEAEs after start of sorafenib.

6.5.2.1 Serious Adverse Event (SAE)

A Serious Adverse Event is any untoward medical occurrence that at any dose:

- Results in death;
- Is life-threatening ("life-threatening" refers to an event in which the subject was at risk of death at the time of the event; it does not refer to an event which hypothetically might have caused death if it were more severe);
- Requires inpatient hospitalization or prolongation of existing hospitalization;
- Results in a persistent or significant disability/incapacity; or
- Is a congenital anomaly/birth defect.

6.5.2.2 Adverse Device Effect (ADE)

An Adverse Device Effect is an AE related to a medical device and includes any event resulting from insufficiencies or inadequacies in the instructions for use or the

deployment, implantation, installation or malfunction of the device; any event that is the result of user error; or any potential ADE which might have occurred if suitable action had not been taken or intervention had not been made or if circumstances had been less fortunate. All AEs with a relationship to device of possibly, probably, or definitely will be considered to be ADEs.

6.5.2.3 Serious Adverse Device Effect (SADE)
A Serious Adverse Device Effect is an ADE that has resulted in any of the consequences characteristic of a SAE or might have led to any of these consequences if suitable action had not been taken; intervention had not been made or circumstances had been less fortunate. All SAEs with a relationship to device of possibly, probably, or definitely will be considered to be SADEs.

6.5.2.4 Adverse Events of Special Interest (AESI)
Some clinical concepts (including some selected individual preferred terms) have been considered as "AEs of special interest" (AESI). These AESIs will identify as a list of categories provided by the clinical team, and the study physician will review the AEs of interest and identify which preferred terms contribute to each AESI. A final review will take place prior to the database hard lock to ensure all applicable terms are captured within the categories. Preferred terms used to identify AESIs will be listed and documented prior to the database hard lock.

6.5.3 Clinical Laboratory Evaluations
A list of the specific clinical laboratory assessments completed at each study visit during the trial is listed in the protocol.

Clinical laboratory results will be converted to SI units. Change from baseline to each visit assessed and end of study will be defined using the windowing method specified in Section 6.2.2, as the visit value minus the baseline visit. Laboratory values will also be classified as normal (if value is within normal reference range) or lower/higher than normal (if value is either below or above the normal reference range).

Applicable laboratory values will also be classified using NCI CTCAE v4.0.

6.5.4 Additional Safety Variables

6.5.4.1 ECOG Performance Status
The ECOG Performance Status will be assessed according to the following categories:

Score	Characteristics
0	Asymptomatic and fully active
1	Symptomatic; fully ambulatory; restricted in physically strenuous activity
2	Symptomatic; ambulatory; capable of self-care; more than 50% of

	waking hours are spent out of bed
3	Symptomatic; limited self-care; more than 50% of waking hours are spent in bed
4	Completely disabled; no self-care; bedridden

6.5.4.2 Child-Pugh Score Status

Severity of liver disease will be assessed according to the Child-Pugh classification of Severity of Liver Disease at screening and at every 8 weeks visit.

6.5.4.3 Albumin-Bilirubin Score

Albumin-Bilirubin (ALBI) score will be assessed according to three ALBI grades in relationship to a patient's linear prediction at screening and at every 8 weeks visit. These grades are categorized as the following:

ALBI Grade	**Classifier**
1	Linear predictor ≤ -2.60
2	-2.60 < Linear predictor ≤ -1.39
3	Linear predictor > -1.39

The linear predictor used to compute the ALBI score for each patient is:

$(\log_{10}$ bilirubin x 0.66) + (albumin x -0.085),

where bilirubin is in µmol/L and albumin is in g/L.

[redacted]

7 STATISTICAL METHODS

7.1 General Methodology

All statistical tests will be two-sided with a significance level of α=0.05, unless specified otherwise, and will be performed using SAS® Version 9.1.3 or higher. Data will be summarized using descriptive statistics (number of patients, mean, standard deviation, median, minimum, and maximum) for continuous variables and using frequency and percentage for discrete variables.

Patient listings of all data from the eCRF as well as any derived variables will be presented.

7.2 Adjustments for Covariates

The following covariates will be included, one at a time together with the treatment group, in 'univariable' Cox regression analysis of time-to-event efficacy endpoints, including OS.

- Stratification factors
  - HCC status (unilobar vs. bilobar disease)
  - Region (North America and Europe vs Asia)
  - ECOG performance status (0 vs 1) at baseline
  - Presence or absence of branch PVT at baseline

  Notes:
  Stratification factors according to the master file, with any incorrect values at randomization replaced with the corrected value from the eCRF, will be used.

  Patients randomized prior to the stratification change to replace HCC status with Region will have their Region covariate assigned programmatically based on site. Patients randomized after the stratification change to replace HCC status with Region will have their HCC status covariate assigned based on the eCRF collected HCC status.

- Age group (≥18 to <65 years, ≥65 to <75 years, and ≥75 years)
- Gender
- Race (White or Caucasian, Black or African American, Asian, Other [Native Hawaiian or other Pacific Islander or Native American or Alaska Native or Other])
- US and non-US region
- Duration from date of initial diagnosis of HCC to randomization (<6 months, ≥6 months)
- Tumor replacement (as percentage of total liver volume) at baseline by blinded central review (<20% or ≥20%)
- Extrahepatic disease at baseline (yes or no)
  - A patient has extrahepatic disease if the patient has extrahepatic target and/or non-target lesions (identified from target lesions and non-target lesions eCRF pages at baseline, based on a manual review of the free text

entered in the "other location, specify" and "extrahepatic location, specify" eCRF fields)

- Child-Pugh class (A5, A6, B7) at baseline
- Barcelona clinic liver cancer (BCLC) stage (B or C) at baseline. This will be derived as follows:
  - BCLC B: patients with ECOG 0, and no extrahepatic lesions, and no PVT at baseline
  - BCLC C: patients with ECOG 1, or, with extrahepatic lesions, or PVT at baseline
- HCC etiology, two categorizations will be considered for this variable
  - Categorization #1: alcoholism, hepatitis B, hepatitis C, afloxatin-contaminated food, non-alcoholic fatty liver disease, other (including unknown), based on the etiology pre-specified categories in the eCRF
  - Categorization #2: alcoholism vs. non-alcoholism; note that patients with more than one etiology are recorded under "other" and will be considered as alcoholism if one of the etiologies is alcoholism (for example, a patient with "alcoholism and hepatitis B" entered in the "other, specify" free text field will be considered into alcoholism)
- Prior oncologic treatment for HCC (yes or no)
- Bilirubin (<1 mg/dL or ≥1 mg/dL) at baseline
- ALBI score (1 vs 2 and 3) at baseline
- AFP (<200 ng/mL or ≥200 ng/mL) at baseline
- Maximum lesion size at baseline, defined as the longest diameter of largest target lesion at baseline according to RECIST 1.1 by investigator assessment (<7 cm or ≥7 cm)
- Number of lesions at baseline (<3 lesions, 3-5 lesions, 6-10 lesions, >10 lesions) by blinded central review

These covariates will also be included, one at a time together with the treatment group, in a 'univariable' logistic regression analysis of binary efficacy endpoints.

All factors in the univariable models with a two-sided p-value <0.15 and treatment group will be included in a multivariable analysis to determine the impact of these factors. For both univariable and multivariable analyses, the overall p-value will be used for factors with >2 levels (i.e. the p-value corresponding to the Type 3 Wald chi-square statistic) rather than the p-values corresponding to each level of the factor.

For the multivariable analysis, collinearity of covariates will be assessed by the variance inflation factor (VIF)[9], and further action will be taken if any covariate has a VIF value >10. Highly correlated covariates (i.e. with VIF >10) will be removed, one at a time, based on the descending order of their univariable p-values or clinical justification, until VIF values are ≤10 for all covariates remaining in the multivariable model.

7.3 Handling of Dropouts or Missing Data

Dropout patients will not be replaced in this study. The handling of missing data will be discussed throughout Section 8, where relevant. Censoring for the efficacy endpoints is discussed throughout Section 8, where applicable.

7.4 Interim Analyses and Data Monitoring

An IDMC will be established to oversee the conduct of the study. The IDMC will meet periodically during the study to review enrollment, protocol deviations and safety events for the study. In addition, the IDMC will conduct and review the interim efficacy results and will make formal recommendations to the study Sponsor at the time of the interim analysis and during the conduct of the study.

After the first 20 patients in the treatment group have received TheraSphere followed by at least 2 weeks of sorafenib therapy, a feasibility safety assessment will be conducted. The IDMC will review the safety results of both the control and treatment groups in an unblinded fashion. The IDMC will take into consideration the established safety profiles of TheraSphere and sorafenib as described in the package inserts for each product as well as the published literature. The expected high rates of AEs events and death that are associated with disease progression in patients with HCC will be considered.

A consideration for stopping further enrollment to the trial may be made if there is a pattern of serious toxicity clearly related to the sequential administration of TS followed by sorafenib. Such a toxicity pattern must be clearly different from, or more severe than, what might be expected from independent administration of the products. The potential adverse impact of any such pattern of toxicity on the survival or well being of the patient should be considered in the context of the safety and outcome expectations of patients with advanced HCC.

This study uses an adaptive group sequential design with two interim analyses and one final analysis. The efficacy stopping boundaries are based on the rho family error spending function with the parameter value rho=1.5. The first interim analysis is planned at approximately, but no less than, 188 deaths, with a two-sided p-value $\leq$0.0151 allowing the study to be stopped early for efficacy. A second interim analysis is planned at approximately, but no less than, 250 deaths, with a two-sided p-value $\leq$0.0151 allowing the study to be stopped early for efficacy. If the interim analyses do not occur at exactly 188 or 250 deaths, the corresponding efficacy boundaries were to be calculated using the rho family spending function with rho=1.5.

Sample size modification is considered at the second interim analysis following the promising zone approach described in Mehta & Pocock (2011)[4] which employs an un-weighted test statistic at the final analysis as recommended by Burman & Sonneson (2006)[5]. The conditional probability boundaries for the decision rules at the second interim analysis are as follows:

- Unfavorable zone ($CP_2 < 0.42$): study size will remain at 417 deaths

- Promising zone ($0.42 \leq CP_2 < 0.8$): study size will be increased to 564 deaths
- Favorable zone ($CP_2 \geq 0.8$): study size will remain at 417 deaths

where CP_2 is defined as the conditional probability of rejecting the null hypothesis at the final analysis, given the results at the second interim analysis. Further details, including both mathematical and simulation based demonstration of type I error control, are provided in Appendix 5.

There was no sample size modification at the second interim analysis. The final analysis was originally planned when approximately, but no less than, 417 deaths have occurred. At the time of preparing this version (i.e., version 4.0) of the statistical analysis plan (SAP), 406 deaths had occurred, however, it became clear, based on the status of patients who had not yet died, that it would be unlikely to reach 417 deaths, because all patients still in follow-up on the study had already been followed up for over 40 months. Therefore, it is now planned to perform the final analysis when at least 417 deaths have occurred or on 30 April 2022, whichever comes first. If 417 deaths have not occurred on 30 April 2022 then the final analysis will be performed with the number of deaths at that time.

The boundary for the final analysis will be adjusted based on the first and second interim analyses and the actual number of deaths at the final analysis. The actual number of deaths for the first and second interim analyses were 205 and 257, respectively. However, at the first interim analysis, the efficacy boundary was inadvertently not recalculated based on the actual number of 205 deaths, and instead the p-value scale boundary stated in the SAP, based on the planned number of 188 deaths, was used. The p-value scale boundary at the second interim analysis was based on the planned number of 188 deaths at the first interim analysis, the actual number of 257 deaths at the second interim analysis and the planned final number of 417 deaths. Therefore, for the final analysis, the efficacy boundary (i.e., α_f) will be calculated using the rho family spending function with rho=1.5, based on the planned number of 188 deaths at the first interim analysis, the actual number of 257 deaths at the second interim analysis and actual number of deaths at the final analysis. This will result in a more conservative efficacy boundary than the one using actual numbers of deaths for all scenarios for the actual number of deaths at the final analysis, as shown in Table 8. The efficacy boundary α_f will be fixed and documented prior to database hard lock for the final analysis. A two-sided p-value $\leq \alpha_f$ will be required to declare a statistically significant improvement in OS at the final analysis.

Table 8 Scenarios of Efficacy Boundary at Final Analysis

	Two-Sided P-value Scale Efficacy Boundary at Final Analysis	
Actual Number of Deaths at Final Analysis	Based on 188 Planned Deaths at IA1, 257 Actual Deaths at IA2 and Actual Number of Deaths at Final Analysis	Based on 205 Actual Deaths at IA1, 257 Actual Deaths at IA2 and Actual Number of Deaths at Final Analysis

Actual Number of Deaths at Final Analysis	Two-Sided P-value Scale Efficacy Boundary at Final Analysis	
	Based on 188 Planned Deaths at IA1, 257 Actual Deaths at IA2 and Actual Number of Deaths at Final Analysis	Based on 205 Actual Deaths at IA1, 257 Actual Deaths at IA2 and Actual Number of Deaths at Final Analysis
406	0.0354	0.0355
407	0.0354	0.0355
408	0.0355	0.0356
409	0.0355	0.0356
410	0.0356	0.0357
411	0.0356	0.0357
412	0.0356	0.0358
413	0.0357	0.0358
414	0.0357	0.0358
415	0.0358	0.0359
416	0.0358	0.0359
417	0.0358	0.0360

IA = Interim Analysis. Efficacy boundaries computed using EAST version 6.5.

7.5 Multiple Comparisons/Multiplicity

For the primary analysis the Type I error is controlled at $\alpha = 0.05$ (2-sided) over the 2 planned interim analyses and final analysis. Mathematical and simulation-based demonstration of type I error control is provided in Appendix 5. For the secondary endpoints the study-wise Type I error will be controlled using a sequential hierarchical approach as explained in Section 8.6.

7.6 Use of an "Efficacy Subset" of Patients

Patients in the mITT population who do not have major protocol deviations that may affect the efficacy evaluation, will form the Per Protocol (PP) Population. Major protocol deviations resulting in a patient being excluded from the PP population are defined in Section 8.3.3

Excluding patients who have major protocol deviations will likely decrease the variability in treatment response.

7.7 Examination of Subgroups

Primary and secondary efficacy endpoints will be summarized by the following subgroups:

- Stratification factors
    - HCC status (unilobar vs. bilobar disease)

  - Region (North America and Europe vs Asia)
  - ECOG performance status (0 vs 1) at baseline
  - Presence or absence of branch PVT at baseline

  Note: stratification factors according to the master file with any incorrect values at randomization replaced with the corrected value from the eCRF will be used.

  Patients randomized prior to the stratification change to replace HCC status with Region will have their Region covariate assigned programmatically based on site. Patients randomized after the stratification change to replace HCC status with Region will have their HCC status covariate assigned based on the eCRF collected HCC status.

- Age group (≥18 to <65 years, ≥65 to <75 years, and ≥75 years)
- Gender
- Race (White or Caucasian, Black or African American, Asian, Other [Native Hawaiian or other Pacific Islander or Native American or Alaska Native or Other])
- US and non-US region
- Duration from date of initial diagnosis of HCC to randomization (<6 months, ≥6 months)
- Tumor replacement (as percentage of total liver volume) at baseline by blinded central review (<20% and ≥20%)
- Extrahepatic disease at baseline (yes or no)
- Child-Pugh class (A5, A6, B7) at baseline
- BCLC stage (B or C) at baseline
- HCC etiology, separately for two categorizations defined in Section 7.2
- Prior oncologic treatment for HCC (yes or no)
- Bilirubin (<1 mg/dL or ≥1 mg/dL) at baseline
- ALBI score (1 vs 2 and 3) at baseline
- AFP (<200 ng/mL or ≥200 ng/mL) at baseline
- Maximum lesion size at baseline, defined as the longest diameter of largest target lesion at baseline according to RECIST 1.1 by investigator assessment (<7 cm or ≥ 7 cm)
- Number of lesions at baseline (<3 lesions, 3-5 lesions, 6-10 lesions, >10 lesions) by blinded central review

AEs will be summarized by the following subgroups:

- Age group (≥18 to <65 years, ≥65 to <75 years, and ≥75 years)
- Gender

- Race (White or Caucasian, Black or African American, Asian, Other [Native Hawaiian or other Pacific Islander or Native American or Alaska Native or Other])
- Region (North America, Europe, Asia)
- US and non-US region
- BCLC stage (B or C) at baseline
- Number of lesions at baseline (<3 lesions, 3-5 lesions, 6-10 lesions, >10 lesions) by blinded central review

8 STATISTICAL ANALYSIS

8.1 Disposition of Patients

The number of patients enrolled will be summarized by region (North America, Europe, Asia), country, and site. The number of patients randomized, and the number of patients treated with sorafenib only, TheraSphere and sorafenib, and TheraSphere only, prior to progression (i.e. prior to date of progression) as assessed by investigator according to RECIST 1.1, will be summarized. The number of untreated patients who discontinued from study will be summarized. The number of treated patients who discontinued from the study (treated and untreated) and the reasons for discontinuing from the study will also be summarized.

8.2 Protocol Deviations

Protocol deviations/violations will not be entered into the database. However, protocol deviations/violations will be identified and summarized within Labcorp's Clinical Department from which the Sponsor can make determinations. All protocol deviations/violations determinations will be made before the database is locked for statistical analysis.

8.3 Analysis Populations

8.3.1 Modified Intent-to-Treat (mITT) Population

All randomized patients who met the study eligibility criteria at randomization will form the modified Intent-to-Treat (mITT) Population and will be analyzed according to the treatment group to which they were randomized. This population will be used to analyze all efficacy endpoints.

8.3.2 Safety Analysis (SA) Population

All randomized patients who received study treatments at least once, prior to progression (i.e. prior to date of progression) as assessed by investigator according to RECIST 1.1, will form the Safety Population and will be analyzed according to the treatment actually

received prior to progression. Note that patients who received only TheraSphere (no sorafenib) prior to progression do not receive full treatment of either protocol-defined treatment group, so will be analyzed as a separate group in the safety population. This population will be used in all safety reporting and analysis.

8.3.3 Per Protocol (PP) Population

The Per Protocol population is the subset of the mITT population excluding patients with major protocol deviations which may affect the efficacy evaluation. Patients in the PP population will be analyzed according to the treatment group to which they were randomized.

Major protocol deviations resulting in a patient being excluded from the PP population will include, but may not be limited to, the following:

- Baseline imaging assessment not performed
- Baseline imaging assessment performed >42 days prior to date of randomization (note that although the screening period for baseline imaging assessment was 28 days, an additional 14-day window is being applied so that only baseline imaging assessments >42 days before randomization will be deemed to be a major protocol deviation that may affect the efficacy evaluation)
- Post-randomization imaging assessments not performed for 3 consecutively planned timepoints (i.e. the number of days between imaging assessments is >224 days), defined as
    - the first post-randomization imaging assessment, prior to progression (i.e. prior to date of progression) as assessed by investigator according to RECIST 1.1, occurs at >32 weeks (i.e. >224 days) after randomization, or
    - any post-randomization imaging assessment, prior to progression as assessed by investigator according to RECIST 1.1, occurs at >32 weeks (i.e. >224 days) after the previous post-randomization imaging assessment
- Randomized study treatment not received (TheraSphere and/or sorafenib) prior to progression as assessed by investigator according to RECIST 1.1
- Y90 (including TheraSphere) received by patients in the control arm prior to progression assessed by investigator according to RECIST 1.1
- TheraSphere dose absorbed by perfused volume (as defined in Section 6.5.1.1) lower than the protocol stated range of 120 Gy – 10% (i.e. <108 Gy) prior to progression as assessed by investigator according to RECIST 1.1
- Bilobar disease at baseline (as defined in Section 6.5.1.1) but only one lobe treated with TheraSphere prior to progression assessed by investigator according to RECIST 1.1
- For patients enrolled under all protocol versions implemented:
    - For the control arm:
        - Start of sorafenib >28 days after randomization

- For the treatment arm:
  - First administration of TheraSphere prior to progression >35 days after randomization, or
  - Start of sorafenib >42 days after the last TheraSphere administration prior to progression

The deviations listed above will be programmatically determined. In addition, monitoring notes or data listings will be reviewed to determine any major deviations that are not identifiable via programming, and to check that those identified via programming are correctly classified. The final classification of major protocol deviations and decisions to exclude patients from the Per Protocol population will be made at the time between the database soft close and hard lock.

[redacted]

8.4 Demographic and Other Baseline Characteristics

All demographic and baseline summaries will be displayed for the mITT Population, Safety Population and PP Population.

Gender, race, ethnicity, and female childbearing potential will be summarized using counts and percentages. Age, height, and weight will be summarized with descriptive statistics (number of patients, mean, standard deviation, median, minimum, and maximum). Age group (≥18 to <65 years, ≥65 to <75 years, and ≥75 years) will be summarized using counts and percentages.

The number and percentage of patients with abnormal physical examination findings at screening will be summarized. The number and percentage of patients with medical history events will be summarized. Vital signs collected at baseline (blood pressure, heart rate, respiratory rate, and body temperature) will be summarized with descriptive statistics (number of patients, mean, standard deviation, median, minimum, and maximum).

Baseline characteristics of HCC will be summarized using counts and percentages as follows:

- Child-Pugh class (A [A5 or A6] or B)
- BCLC stage (B or C)
- HCC etiology (alcoholism, hepatitis B, hepatitis C, afloxatin-contaminated food, non-alcoholic fatty liver disease, other (including unknown)); note that patients with more than one etiology are recorded under “other” so two summaries will be provided:

  - based on data directly from these etiology pre-specified categories on CRF
  - patients with more than one etiology recorded under "other" will be summarized once under each corresponding category (for example, "alcoholism and hepatitis B" entered in the "other, specify" free text field will be summarized under the pre-defined categories of alcoholism and hepatitis B separately); only patients with etiologies that do not fall into the pre-defined categories will be summarized under "other" (including patients with unknown etiology)

- HCC etiology (alcoholism, non-alcoholism) as defined in Section 7.2
- Extrahepatic disease at baseline (yes or no)
- Duration from date of initial diagnosis of HCC to randomization (<6 months, ≥6 months; <12 months, ≥12 months)
- Tumor replacement at baseline separately by investigator assessment and by blinded central review (<20% or ≥20%)
- Prior oncologic treatment for HCC (yes or no, as well as summarizing separately for liver directed therapy, prior resection, or systemic treatment)
- Hepatoma Arterial-Embolization Prognostic (HAP) score at baseline (1, 2, 3, or 4)
  - Patients are assigned one point for each of the following:
    - Bilirubin at baseline ≥1 mg/dL
    - Albumin at baseline <3.6 g/dL
    - AFP at baseline >400 ng/mL
    - Longest diameter of largest target lesion at baseline according to RECIST 1.1 by investigator assessment >7 cm
- Bilirubin at baseline (<1 mg/dL or ≥1 mg/dL)
- Albumin at baseline (<3.6 g/dL or ≥3.6 g/dL)
- ALBI score at baseline (1, 2 or 3)
- AFP at baseline (<200 ng/mL or ≥200 ng/mL; <400 ng/mL or ≥400 ng/mL)
- Maximum lesion size at baseline, defined as the longest diameter of largest target lesion at baseline according to RECIST 1.1 by investigator assessment (<7 cm or ≥7 cm)
- HCC related stratification factors
  - HCC status (unilobar vs. bilobar disease)
  - ECOG performance status (0 vs 1) at baseline
  - Presence or absence of branch PVT at baseline

  Note: stratification factors according to the master file, with any incorrect values at randomization replaced with the corrected value from the eCRF, will be used. Patients randomized after the stratification change to replace HCC status with Region will have their HCC status assigned based on the eCRF collected HCC status.
- Baseline number of lesions (<3 lesions, 3-5 lesions, 6-10 lesions, >10 lesions) by blinded central review

The time from diagnosis of HCC to randomization and tumor replacement at baseline (separately by investigator assessment and by blinded central review) will also be summarized descriptively.

Baseline tumor replacement (%) will be summarized with descriptive statistics (n, mean, SD, median, min, and max) by treatment group, and also as the number and percentage of patients with tumor replacement <20% and ≥20%.

Pregnancy test results will be summarized by number and percentage.

8.5 Prior and Concomitant Therapy

The WHO DE March 2011 dictionary will be used to classify medications by preferred term and WHO Anatomical Therapeutic Chemical (ATC) classification of ingredients.

The following applies to all data collected on the prior and concomitant eCRF page and will be reported by each category separately.

Where a medication start date is missing, this medication will be assumed to be concomitant for reporting purposes, unless the end date is prior to the date of randomization. Partial dates will be imputed as detailed in Section 6.2.1.3.

Frequency counts and percentages will be provided to summarize the use of prior and concomitant medications by WHO ATC classification of ingredients and by preferred term.

8.5.1 Prior Medication

A prior medication is defined as any medication stopped prior to the date of randomization.

The number and percentage of patients who had at least one prior medication will be tabulated as well as the number and percentage of patients with each medication. Patients will only be counted once for each medication.

8.5.2 Prior Therapy for HCC

Prior HCC treatment type and treatment will be summarized from the treatment type and treatment recorded on the Medical History of HCC eCRF page.

8.5.3 Concomitant Medication

A concomitant medication is defined as any medication given prior to the patient being randomized and continuing after randomization, or any medication that is initiated on or after randomization. Medications are considered concomitant through to the end of the study.

The number and percentage of patients who had at least one concomitant medication will be tabulated as well as the number and percentage of patients with each medication. Patients will only be counted once for each medication.

8.6 Analysis of Efficacy Parameters

8.6.1 Analysis of Primary Efficacy Variable

The primary efficacy analysis is of OS. OS rates will be derived from the Kaplan-Meier estimates for each treatment group; 6, 12, 18 and 24 month OS rates will be presented, together with 95% confidence intervals (CIs). Quartiles will be presented and 95% CIs will also be calculated on the quartiles for each treatment group. A log-rank (two-sided) test, converted to a z-score, will be used to compare OS between the two treatment groups at an overall two-sided alpha level of 0.05. The two-sided alpha level of 0.05 will be adjusted over the 2 planned interim analyses and final analysis (as described in Sections 7.4 and 7.5). The HR and corresponding two-sided 95% CI for the treatment effect will be computed from a Cox proportional hazards model. Plots of the Kaplan-Meier curves will be provided for each treatment group. This analysis will be performed on the mITT population (primary analysis) and PP population (secondary analysis).

The assumption of proportional hazards used to compute the HR for OS will be assessed. Firstly, a plot of log[-log (estimated probability of event-free survival)] versus log(time) will be examined, with nonparallel curves for the 2 treatment groups indicating non-proportional hazards. Also, a time-dependent covariate Cox regression model (i.e. adding a treatment group by time interaction) will be fitted and if the time-dependent covariate has two-sided p-value <0.15 piecewise HRs over distinct time periods will be calculated.

For each patient that has not known to have died, OS will be censored at the time of last contact date known to be alive.

Note that any patient who is still on the study and has not withdrawn or died at the data cut-off (DCO) date for the final analysis will have an additional survival contact on the day of DCO or within 2 weeks of DCO date, with Patient Contact/Overall Survival log and/or Confirmation of Death eCRF page to be completed accordingly. If patients are confirmed to be alive or if the death date is after the DCO date these patients will be censored at the date of DCO.

All patients should be followed for OS, however, if a patient has not known to have died and not been followed for OS for any reason (i.e. no contact date known alive on Patient Contact/Overall Survival log eCRF page), censoring date will be defined as the latest among the following dates recorded on the eCRF:

- Adverse event start and end dates

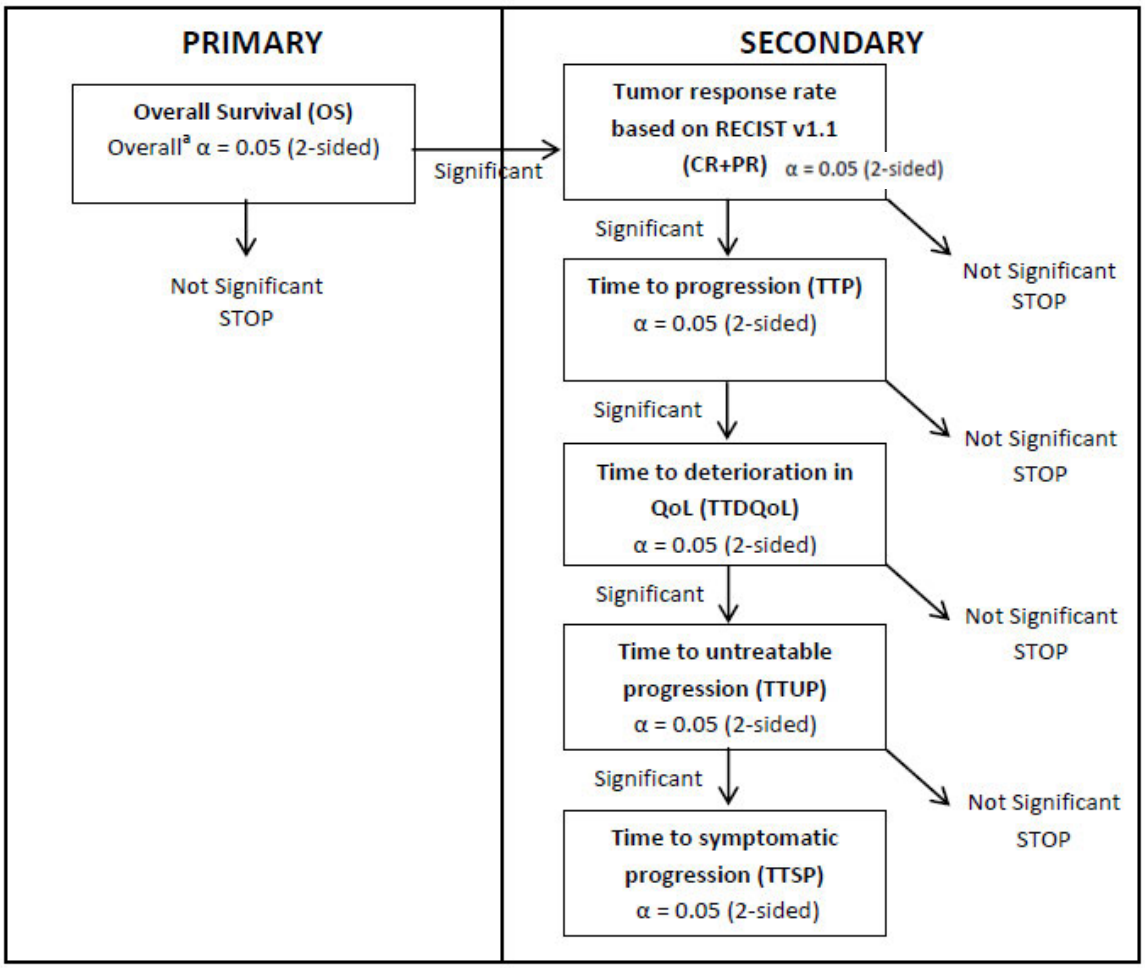
- Dates of collection for laboratory tests (hematology, coagulation, chemistry)
- Date of collection for AFP
- ECOG assessment date
- FACT-Hep assessment date
- Child Pugh Score assessment date
- TheraSphere administration date
- Sorafenib administration start and stop dates
- Date of imaging on Determination of Response eCRF page
- Date of curative treatment on Resection / HCC Stage Migration eCRF page
- Best available care start and end dates on Best Available Care - Post Treatment Discontinuation - Medication eCRF page
- Procedure start and stop dates on Additional Procedures eCRF page
- Date of study exit on Study Exit eCRF page

Sensitivity analysis will be performed to address the poolability of data (See Section 8.6.5).

8.6.2 Analysis of Secondary Efficacy Variables

For secondary efficacy endpoints, each comparison between treatment groups will be conducted at α=0.05 (two-sided). Secondary study endpoints will be analyzed only at the final analysis to determine the statistical significance, if any, between the treatment groups. Study-wise Type I error will be controlled using a sequential hierarchical approach, as shown in the figure below. That is, if the primary comparison is statistically significant, the secondary endpoints will be analyzed as secondary endpoints in order of the list below and will continue as long as the obtained 2-sided probability is equal to or less than 0.05. If a probability of greater than 0.05 is obtained, the inferential analysis of secondary endpoints will stop and not proceed further down the ordered list. In this manner the overall study alpha is protected and no further adjustment for multiplicity of analyses is required. If a probability of greater than 0.05 is obtained for an endpoint then the analysis of that endpoint and the endpoints further down the ordered list will still be presented but will be considered as exploratory endpoints rather than secondary endpoints.

Hierarchical approach to control study-wise Type I error of primary and secondary efficacy endpoints



[a] Type I error is controlled at α=0.05 (two-sided) over the 2 planned interim analyses and final analysis.

8.6.2.1 Time to Progression (TTP) according to RECIST v1.1 criteria by investigator determination

TTP rates will be derived from the Kaplan-Meier estimates for each treatment group; 3, 6, 9, and 12-month TTP rates will be presented, together with 95% CIs. Quartiles will be presented and 95% CIs will be calculated on the quartiles for each treatment group. A log-rank (two-sided) test will be used to compare TTP between the two treatment groups at a 0.05 significance level. The HR and corresponding 95% CI will be computed from a Cox proportional hazards model. Plots of the Kaplan-Meier curves will be provided for each treatment group. This analysis will be performed on both the mITT and PP populations.

TTP (Months) = (Date of event/censor – Date of Randomization +1)/ 30.4375.

The censoring is performed in the following order:

1) If a patient does not have a baseline tumor assessment, then the TTP time will be censored at the randomization date, regardless of whether or not radiological disease progression (i.e. PD) has been observed.
2) If a patient received a subsequent HCC therapy before PD or in the absence of PD, the TTP time will be censored at the last valid (i.e. evaluable) post baseline radiological tumor assessment on or before the start date of the subsequent HCC therapy. If the patient has no valid post-baseline radiological tumor assessments on or before the start date of the subsequent HCC therapy, the patient will be censored at the randomization date.
3) If a patient is known not to have PD and did not receive subsequent HCC therapy, the TTP time will be censored at the date of death* or last valid post baseline radiological tumor assessment date in the absence of death, or at the randomization date if the patient does not have any valid post-baseline radiological tumor assessments or death.
4) If a patient had PD immediately after two or more missed visits (Note: a response of NE at a visit is not considered as a missed visit), the patient will be censored at the time of the last valid post baseline radiological tumor assessment date that occurred before the missed visits. If the patient has no valid post-baseline radiological tumor assessments before the missed visits, the patient will be censored at the randomization date.
    - For example, if a patient had a tumor assessment at Week 8, but did not have tumor assessments at Weeks 16 and 24, and then had PD at the Week 32 assessment (i.e. after 2 missed visits), then the TTP time would be censored at the date of the Week 8 assessment.
    - Given the scheduled visit scheme of tumor assessments (i.e. every 8 weeks), the definition of 2 missed visits will equate to 16 weeks since the previous tumor assessment, or since the randomization date if no previous post-baseline tumor assessment.

* Note: Censoring at the date of death, in the above censoring rules, is according to US Food and Drug Administration (FDA) guidance[8]. However, a sensitivity analysis of TTP, excluding censoring at the date of death (i.e. censoring at the last valid post baseline radiological tumor assessment date instead) with all other censoring rules described above still used, will also be performed.

8.6.2.2 Time to Untreatable Progression (TTUP)

TTUP rates will be derived from the Kaplan-Meier estimates for each treatment group; 3, 6, 9, and 12 month TTUP rates will be presented, together with 95% CIs. Quartiles will

be presented and 95% CIs interval will be calculated on the quartiles for each treatment group. A log-rank (two-sided) test will be used to compare TTUP between the two treatment groups at a 0.05 significance level. The HR and corresponding 95% CI will be computed from a Cox proportional hazards model. Plots of the Kaplan-Meier curves will be provided for each treatment group. This analysis will be performed on both the mITT and PP populations.

TTUP (Months) = (Date of event/censor – Date of Randomization +1)/ 30.4375.

The censoring is performed in the following order:

1) If a patient does not have a baseline tumor assessment, then the TTUP time will be censored at the randomization date, regardless of whether or not untreatable PD has been observed.
2) If a patient received a subsequent HCC therapy before untreatable PD or in the absence of untreatable PD, the TTUP time will be censored at the last valid (i.e. evaluable) post baseline radiological tumor assessment on or before the start date of the subsequent HCC therapy. If the patient has no valid post-baseline radiological tumor assessments on or before the start date of the subsequent HCC therapy, the patient will be censored at the randomization date.
3) If a patient is known not to have untreatable PD and did not receive subsequent HCC therapy, the TTUP time will be censored at the date of death or last valid post baseline radiological tumor assessment date in the absence of death, or at the randomization date if the patient does not have any valid post-baseline radiological tumor assessments or death.
4) If a patient had untreatable PD immediately after two or more missed visits (Note: a response of NE for a visit is not considered as a missed visit), the patient will be censored at the time of the last valid post baseline radiological tumor assessment date that occurred before the missed visits. If the patient has no valid post-baseline radiological tumor assessments before the missed visits, the patient will be censored at the randomization date.
    - Given the scheduled visit scheme of tumor assessments (i.e. every 8 weeks), the definition of 2 missed visits will equate to 16 weeks since the previous tumor assessment, or since the randomization date if no previous post-baseline tumor assessment.

A sensitivity analysis of TTUP, excluding censoring at the date of death for any reason other than progression of disease (i.e. censoring at the last valid post baseline radiological tumor assessment date instead) with all other censoring rules described above still used, will also be performed.

8.6.2.3 Time to Symptomatic Progression (TTSP)

TTSP rates will be derived from the Kaplan-Meier estimates for each treatment group; 3, 6, 9, and 12 month TTSP rates will be presented, together with 95% CIs. Quartiles will be presented and 95% CIs will be calculated on the quartiles for each treatment group. A log-rank (two-sided) test will be used to compare TTSP between the two treatment groups at a 0.05 significance level. The HR and corresponding 95% CI will be computed from a Cox proportional hazards model. Plots of the Kaplan-Meier curves will be provided for each treatment group. This analysis will be performed on both the mITT and PP populations.

TTSP (Months) = (Date of first ECOG >1/censor – Date of Randomization + 1)/ 30.4375.

The censoring is performed as follows:

1) If a patient does not have a post baseline ECOG assessment, then the TTSP will be censored at the randomization date, regardless of whether or not symptomatic progression has been observed.
2) If a patient received a subsequent HCC therapy before symptomatic progression or in the absence of symptomatic progression (note that for this scenario a subsequent HCC therapy may occur before the assessment of ECOG>1 or before either confirmation of ECOG>1 at the first two subsequent assessments at least 8 and 16 weeks later, respectively), the TTSP time will be censored at the last valid post baseline ECOG assessment on or before the start date of the subsequent HCC therapy. If the patient has no post-baseline ECOG assessments on or before the start date of the subsequent HCC therapy, the patient will be censored at the randomization date.
3) If a patient did not have a symptomatic progression and did not receive subsequent HCC therapy, the TTSP time will be censored at the last post baseline ECOG assessment date or at the randomization date if the patient does not have any post-baseline ECOG assessments.
4) If a patient had symptomatic progression immediately after two or more missed visits (note that for this scenario two or more missed visits may occur before the assessment of ECOG>1 or before either confirmation of ECOG>1 at the first two subsequent assessments at least 8 and 16 weeks later), the patient will be censored at the last post baseline ECOG assessment before the missed visits. If the patient has no post-baseline ECOG assessments before the missed visits, the patient will be censored at the randomization date.
    - Given the scheduled visit scheme of tumor assessments (i.e. every 8 weeks), the definition of 2 missed visits will equate to 16 weeks since the previous tumor assessment, or since the randomization date if no previous post-baseline tumor assessment.

8.6.2.4 Objective Response Rate (ORR) according to RECIST v1.1 criteria by investigator determination

ORR will be computed for the two treatment groups as proportion of CR+PR over the total number of patients in the specified population. The 95% CIs for the ORR for each of the treatment groups will be computed according to Wilson (1927).

ORR, as determined by the investigator using RECIST 1.1, will be compared between treatment groups using the continuity adjusted Wald test, and the corresponding 95% CI for the difference in ORRs between the two treatment groups will be calculated. This analysis will be performed for each time point and the best overall response on both the mITT and PP populations.

8.6.3 Analysis of Quality of Life Questionnaire (FACT-hep)

8.6.3.1 Analysis of FACT-hep Scores

The total, domain, and individual question scores of the FACT-hep QoL instrument and their differences from baseline will be summarized at each time point by treatment group. The two treatment groups will be compared by applying a mixed linear model repeated measures analysis using a residual maximum likelihood estimation with the treatment, visit and the interaction between treatment and visit as factors, and the baseline score as a covariate. If the interaction term has a 2-sided p-value $\geq$0.15 then the model will be re-fitted without the interaction term. The Kenward-Roger approximation will be used to estimate the degrees of freedom. An unstructured covariance approach will be applied. If the fit of the unstructured covariance structure fails to converge, the following covariance structures will be tried in order until convergence is reached: Toeplitz with heterogeneity, autoregressive with heterogeneity, Toeplitz, and autoregressive. Means and least squares mean difference between treatment groups, along with a two-sided 95% CI and p-value for the difference between treatments will also be provided. This analysis will be performed on the mITT and PP populations.

8.6.3.2 Analysis of Time to Deterioration in QoL (TTDQoL)

A deterioration in QoL is defined as a $\geq$7-point decline in the total FACT-hep score (i.e., a change from baseline in the total score of $\leq$-7 points) or death whichever occurs first. TTDQoL rates will be derived from the Kaplan-Meier estimates for each treatment group; 3, 6, 9, and 12-month TTDQoL rates will be presented, together with 95% CIs. Quartiles will be presented and 95% CIs will be calculated on the quartiles for each treatment group. A log-rank (two-sided) test will be used to compare TTDQoL between the two treatment groups at a 0.05 significance level. The HR and corresponding 95% CI will be computed from a Cox proportional hazards model. Plots of the Kaplan-Meier curves will be provided for each treatment group. This analysis will be performed on the mITT and PP populations.

TTDQoL (months) = ((Date of change from baseline in total FACT-Hep score ≤-7 or death/Censor) – Date of Randomization +1)/ 30.4375.

The censoring is performed as follows:

- If a patient does not have a baseline total FACT-hep score, then the TTDQoL time will be censored at the randomization date, regardless of whether or not TTDQoL has been observed
- If a patient received a subsequent HCC therapy before a deterioration in QoL or in the absence of a deterioration in QoL, the TTDQoL time will be censored at the last post baseline FACT-hep assessment on or before the start date of the subsequent HCC therapy where total FACT-hep score could be evaluated. If the patient has no post baseline total FACT-hep scores on or before the start date of the subsequent HCC therapy, the patient will be censored at the randomization date.
- If a patient did not have a deterioration in QoL and did not receive subsequent HCC therapy, the TTDQoL time will be censored at the last post baseline FACT-hep assessment where total FACT-hep score could be evaluated, or at the randomization date if the patient does not have any post-baseline total FACT-hep scores.
- If a patient had a deterioration in QoL immediately after two or more visits where total FACT-hep score could not be evaluated (e.g. missed visits or visits missing items to derive total FACT-hep score), the patient will be censored at the last post baseline FACT-hep assessment before the missed visits where total FACT-hep score could be evaluated. If the patient has no post baseline total FACT-hep scores before the missed visits, the patient will be censored at the randomization date.
    - Given the scheduled visit scheme of FACT-hep assessments (i.e. every 8 weeks), the definition of 2 missed visits will equate to 16 weeks since the previous FACT-hep assessment where total FACT-hep score could be evaluated, or since the randomization date if no previous post-baseline total FACT-hep scores.

8.6.4 Subgroup Analyses

Subgroup analyses of primary and secondary efficacy endpoints will be performed for the mITT population according to the subgroups listed in Section 7.7.

For each subgroup, a Cox proportional hazards model for time-to-event endpoints or a logistic regression for binary endpoints, will be fitted with treatment as the only covariate for each level of the subgroup separately.

For time-to-event endpoints, the HRs and associated 95% CIs will be summarized and presented on a forest plot, along with results of the overall analysis. If an endpoint has <10 events available at a subgroup level then the relationship between that subgroup level and endpoint will not be formally analyzed, since it is unlikely to be a meaningful analysis, and only descriptive summaries will be provided.

8.6.5 Assessment of Poolability

Since this is a multi-center study, analysis will be performed by pooling data across study sites. The clinical study will be conducted under a common protocol for each investigational site, except for sites in Germany, where a separate protocol was used with different eligibility criteria, mainly related to the limits on liver function tests. It is expected that approximately 1% of the total number of patients randomized in the study will be from sites in Germany.

In the event that there are small sample sizes at some sites, sites may be grouped using the following procedure to create "analysis-sites" for analysis purposes. These analysis-sites will be created for North America, Europe, and Asia independently to preserve the ability to differentiate between regions. Patients from sites in Germany will not be included in this grouping mechanism. Analysis-sites are based on a target size of at least 5 patients per treatment group at each site. If investigative sites have at least 5 mITT patients per treatment group, they will retain their identities in the analysis. All investigative sites with fewer than 5 mITT patients per treatment group will be rank ordered by size and sorted secondarily by site identification number to break ties. Starting with the smallest investigative site, patients will be combined site by site by treatment group, until the first time the resulting analysis-site has at least 5 mITT patients in each treatment group. The process continues until all investigative sites are accounted for. If the last analysis-site has fewer than 5 mITT patients per treatment group, it will be combined with the most recently created analysis-site.

To assess the poolability of data across sites, a Cox regression analysis of the primary efficacy endpoint, OS, and all secondary time-to-event endpoints (i.e., TTP, TTUP, TTSP, TTDQoL) will be conducted including factors of treatment group, analysis-site, and treatment group by analysis-site interaction. Estimates of treatment effect and 95% CIs will be calculated separately by analysis-site.

Similarly, to assess the poolability of data across regions a Cox regression analysis will be conducted with analysis-site replaced by region. NB: region and study site will not be included simultaneously in the same model due to collinearity. Also, to assess the poolability of data across genders, a Cox regression analysis will be conducted with analysis-site replaced by gender. Patients from sites in Germany will not be included in these analyses.

Logistic regression of binary secondary endpoints (i.e., ORR and DCR) will be conducted using the same methodology described above for the Cox regression.

These analyses will be performed on the mITT population.

If, in the above analyses, the treatment group by analysis-site interaction, treatment group by gender interaction or treatment group by region interaction is statistically significant at a two-sided level of 0.15, the reasons for the observed differential treatment effect, such as patient demographic or clinical characteristics, will be investigated and reported. If the poolability of results is in direct question as a result of this sensitivity analysis, the endpoint(s) will also be analyzed separately by site, region, and/or gender.

8.6.6 Additional Efficacy Analyses

All the additional analyses will be performed on both the mITT and PP populations.

DCR according to RECIST 1.1 by investigator assessment will be analyzed in the same way as ORR.

PFS according to RECIST 1.1 by investigator assessment is defined as the time from date of randomization until date of PD determined by investigator assessment, according to RECIST 1.1, or death due to any cause, whichever is earlier.

Progression-Free Survival (months) = (Date of event/censor – Date of Randomization +1) /30.4375.

The censoring of PFS is performed in the following order:

1) If a patient does not have a baseline tumor assessment, then the PFS time will be censored at the randomization date, regardless of whether or not PD or death has been observed.
2) If a patient received a subsequent HCC therapy before PD or death or in the absence of PD or death, the PFS time will be censored at the last valid (i.e. evaluable) post baseline radiological tumor assessment on or before the start date of the subsequent HCC therapy. If the patient has no valid post-baseline radiological tumor assessments on or before the start date of the subsequent HCC therapy, the patient will be censored at the randomization date.
3) If a patient is known not to have died or have PD and did not receive subsequent HCC therapy, the PFS time will be censored at the last valid post baseline radiological tumor assessment date or at the randomization date if the patient does not have any valid post-baseline radiological tumor assessments.
4) If a patient had PD immediately after two or more missed visits (Note: a response of NE for a visit is not considered as a missed visit), the patient will be censored at the time of the last valid post baseline radiological tumor assessment date that occurred before the missed visits. If the patient has no valid post-baseline radiological tumor assessments before the missed visits, the patient will be censored at the randomization date.

- Given the scheduled visit scheme of tumor assessments (i.e. every 8 weeks), the definition of 2 missed visits will equate to 16 weeks since the previous tumor assessment, or since the randomization date if no previous post-baseline tumor assessment.

PFS rates will be derived from the Kaplan-Meier estimates for each treatment group; 3, 6, 9, and 12-month TTP rates will be presented, together with 95% CIs. Quartiles will be presented and 95% CIs will be calculated on the quartiles for each treatment group. A log-rank (two-sided) test will be used to compare TTP between the two treatment groups at a 0.05 significance level. The HR and corresponding 95% CI will be computed from a Cox proportional hazards model. Plots of the Kaplan-Meier curves will be provided for each treatment group.

TTP, PFS, ORR and DCR, according to mRECIST criteria by blinded central image review, will be analyzed in the same way as the corresponding analysis of investigator assessed data described above.

Duration of objective response and duration of disease control by both blinded central image review and investigator assessment will be summarized by descriptive statistics (n, mean, SD, median, min, and max) by treatment group. Also, Kaplan-Meier analyses of duration of objective response and duration of disease control will be conducted.

Duration of response (months) = (Date of PD or death in absence of PD/Censor – Date of first overall response of CR or PR +1) /30.4375.

Duration of disease control rate (months) = (Date of PD or death in absence of PD/Censor – Date of first overall response of CR, PR or SD +1) /30.4375.

The censoring for duration of response and duration of disease control will be performed as follows:

1) If a patient received a subsequent HCC therapy before PD or death, or in the absence of PD or death, the duration of response and duration of disease control will be censored at the last valid post baseline radiological tumor assessment on or before the start date of the subsequent HCC therapy.
2) If a patient is known not to have died or have PD and did not receive subsequent HCC therapy, the duration of response and duration of disease control will be censored at the last valid post baseline radiological tumor assessment date.
3) If a patient had PD or died immediately after two or more missed visits (Note: a response of NE for a visit is not considered as a missed visit), the duration of response and duration of disease control will be censored at the time of the last valid post baseline radiological tumor assessment date that occurred before the missed visits. Given the scheduled visit scheme of tumor assessments (i.e. every 8 weeks), the definition of 2 missed visits will equate to 16 weeks since the previous tumor assessment.

DoR by both blinded central image review and investigator assessment will be summarized by descriptive statistics (number of patients, mean, standard deviation, median, minimum, and maximum) by treatment group. The number and percentage of patients achieving DoR with >10% tumor replacement by blinded central review and >20% tumor replacement by blinded central review will be summarized by treatment group. A 2-sample t-test will be performed to compare the mean DoR between treatment groups, and the corresponding 95% CI for the mean difference between the two treatment groups will be calculated.

PTTS by both blinded central image review and investigator assessment will be summarized by number and percentage of patients by treatment group at Week 8, 16, and 24. The 95% CIs for the PTTS rate for each of the treatment groups will be computed according to Wilson approach[7]. The PTTS rates will be compared between treatment groups using the continuity adjusted Wald test, and the corresponding 95% CI for the difference in PTTS rates between the two treatment groups will be calculated.

The percentage and absolute change from baseline in the sum of the longest diameters of target lesions on or before the start date of subsequent HCC therapy will be summarized separately at Week 8, 16, and 24 analysis visits (as defined in Table 3). A waterfall plot of the best percentage change in the sum of longest diameters for the best overall response will be presented.

The tumor marker for HCC (AFP) will be summarized with descriptive statistics (number of patients, mean, standard deviation, median, minimum, and maximum) for each time-point and change from baseline by treatment group. The number and percentage of patients achieving an AFP response, defined in Section 6.4.3.11, will also be summarized and compared between treatment groups using the continuity adjusted Wald test.

8.7 Analysis of Safety

All safety analyses will be performed on the Safety Population.

8.7.1 Extent of Exposure to Study Treatment

Analyses of the extent of exposure to study treatment will be performed on the mITT, PP and Safety populations.

8.7.1.1 Extent of Exposure to TheraSphere

The extent of patient exposure to TheraSphere as defined in Section 6.5.1.1 will be summarized using descriptive statistics (as appropriate, including number of patients, mean, median, standard deviation, minimum, and maximum, or counts and percentages).

8.7.1.2 Sorafenib

The extent of patient exposure to sorafenib as defined in Section 6.5.1.2 will be summarized by treatment group using descriptive statistics (as appropriate, including

number of patients, mean, median, standard deviation, minimum, and maximum, or counts and percentages).

The cumulative dose of sorafenib (g) prior to progression according to RECIST 1.1 by investigator assessment will be the sum of all administered doses prior to progression date per patient.

Duration of sorafenib in weeks prior to progression according to RECIST 1.1 by investigator assessment will be calculated as:

- Duration of Treatment (weeks) prior to progression by investigator assessment = sum of all sorafenib administration periods (stop date – start date) for sorafenib started prior to progression date /7
- Dose intensity of sorafenib (mg/day) prior to progression by investigator assessment per patient will be calculated as:
- Dose Intensity (mg/day) prior to progression by investigator assessment = Cumulative dose of sorafenib (mg) started prior to progression / (Duration of treatment (weeks) prior to progression *7)
- Relative dose intensity of sorafenib (%) prior to progression by investigator assessment will be the dose intensity (mg/day) prior to progression divided by the one-day equivalent dose (2 * 400 mg = 800 mg/day) *100.

8.7.1.3 Extent of Study Exposure and Follow-up

The duration of study will be summarized by treatment group using descriptive statistics (number of patients, mean, median, standard deviation, minimum, maximum).

Duration on study (months) = (earlier of study exit date and death date – randomization date + 1) / 30.4375

In addition, an alternative method for determining duration of follow-up will be performed using the reverse Kaplan-Meier method for the mITT population. The censored values for OS will be reversed so that 1's will be 0's and 0's will be 1's. The median follow-up time will be derived from the Kaplan-Meier method using the OS values of overall survival and the reversed censoring values.

8.7.1.4 Best Available Care Post-Progression

The number and percentage of patients who received each post-progression treatment (systemic treatments will be presented by preferred terms) will be summarized by treatment group.

8.7.2 Adverse Events

The investigator's verbatim term of each AE will be mapped to system organ class and preferred term using the MedDRA Version 14.0 dictionary.

Adverse events will be summarized by system organ class and preferred term; a patient will only be counted once per system organ class and once per preferred term within a treatment. Patient counts and percentages and event counts will be presented for each treatment group for the following summaries:

1. Overall summary of TEAEs
2. All TEAEs (also presented by preferred term only in descending order).
3. All TEAEs with CTCAE grade ≥3 (also presented by preferred term only in descending order).
4. All TEAEs considered related to sorafenib.
5. All TEAEs considered related to device (ADE).
6. All CTCAE grade ≥3 TEAEs considered related to device.
7. All TEAEs related to angiographic procedure.
8. All TEAEs with fatal outcome.
9. All treatment emergent serious adverse events (SAE).
10. All treatment emergent serious adverse device events (SADE)
11. All TEAEs leading to sorafenib discontinuation
12. All TEAEs of special interest (presented by AESI category and preferred term)
13. All TEAEs of special interest with CTCAE grade ≥3 (presented by AESI category and preferred term)

Event rate (per 100 patient years) will also be presented by treatment group for each of the above TEAE categories. For each category, the event rate is defined as the number of patients with TEAEs in that category divided by the total duration of patients at risk for TEAEs and then multiplied by 365.25 x 100 to present in terms of per 100 patient years.

For the summary of TEAEs by CTCAE grade, if a patient has multiple events occurring in the same body system or same preferred term, then the event with the highest CTCAE grade will be counted. For TEAEs by relationship to study device, if a patient has multiple events occurring in the same body system or same preferred term, the event with the highest association to study device will be summarized (unknown is considered a higher association to study device than not related, but less of a relationship than possibly, probably, and definitely). Adverse events related to sorafenib, device and angiographic procedure are defined as a subset of AEs with a relationship of either possibly, probably, definitely, or unknown.

No statistical inference between the treatments will be performed on AEs.

Listings will be presented by patient for all AEs as well as for SAEs including SADE, AEs with fatal outcome, and AEs leading to discontinuation of sorafenib.

8.7.3 Clinical Laboratory Evaluations

Clinical laboratory results will be converted to SI units (except albumin will use the unit of g/dL and bilirubin will use the unit of mg/dL). Change from baseline to each visit assessed and end of study will be defined using the windowing method specified in Section 6.2.2, as the visit value minus the baseline visit. Laboratory test values at each assessment and for change from baseline to each assessment will be displayed using summary statistics (number of patients, mean, median, and standard deviation). Hematology, chemistry, and coagulation results will each be summarized separately.

All clinical laboratory data will be presented in listings. Within each listing, laboratory values outside the normal ranges will be flagged as either high or low. In addition, shift tables will be presented to display the shift in the normal range categories (low, normal, high) from baseline to the final evaluation. Baseline is defined as the latest non-missing value prior to randomization.

A shift table of baseline to each assessment by CTC grade (NCI CTCAE v4.0), and a table of laboratory parameters of CTCAE Grade 3 or higher that worsened from baseline will be summarized.

A shift table comparing the baseline ALBI score to the ALBI score at each time-point will also be summarized by treatment group.

8.8 Additional Safety Analyses

A shift table comparing the baseline ECOG score to the ECOG score at each time-point will be summarized for the mITT population. This will be used to ascertain the number of patients with an ECOG score that worsens after baseline and any difference between the treatment groups.

Child-Pugh scores will be summarized for the mITT population in the same way as ECOG scores.

[redacted]

[redacted]

8.9.2 Association between ^{99m}Tc-MAA and Y-90 absorbed doses

The relationship between the pre-treatment and post-treatment ADs will be assessed separately for each AD endpoint listed in Section 6.6, using Bland-Altman analysis. The Bland-Altman analysis will calculate the 95% limit of agreement and will be performed by plotting the difference in AD (pre-treatment minus post-treatment) against the post-treatment AD. The figure will also present the number of patients.

In addition, a linear regression of the pre-treatment and post-treatment ADs (for each AD endpoint listed in Section 6.6) will be performed and Pearson's correlation coefficient with be calculated and displayed, along with the number of patients, on scatter plots.

8.9.3 Determination of Dose Volume Histogram (DVH) for tumoral and normal liver tissue volumes

The DVH values for tumor and normal tissue volumes will be determined. Summary statistics will be provided for both ^{99m}Tc-MAA and Y-90 values.

Additionally, the pre and post DVH value of D70 for tumor volumes and V30 for normal tissue volumes will be assessed for the same efficacy and safety endpoints as described in Section 8.9.1 in the same manner using logistic and Cox regressions.

9 COMPUTER SOFTWARE

All analyses will be performed by Labcorp using Version 9.1.3 or later of SAS® software. All summary tables and data listings will be prepared utilizing SAS® software.

For continuous variables, descriptive statistics (number of patients, mean, standard deviation, median, minimum, and maximum) will be generated. For discrete/categorical variables, the number and proportion of patients will be generated. The standard operating procedures of Labcorp will be followed in the creation and quality control of all data displays and analyses.

10 REFERENCES

1 http://www.dep.iarc.fr/ accessed July, 2015.

2 Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009 Jan; 45(2):228-47.

3 Lencioni R, Llovet JM (2010), Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. Semin Liver Dis. 2010 Feb;30(1):52-60

4 Mehta, C. R. and Pocock, S. J. (2011), Adaptive increase in sample size when interim results are promising: A practical guide with examples. Statist. Med., 30: 3267-3284. Doi: 10.1002/sim.4102

5 Burman and Sonesson (2006), Are flexible designs sound? Biometrics., 62: 664-683

6 Proschan MA, Lan KKG, Wittes JT (2006), Statistical Monitoring of Clinical Trials: A Unified Approach. 1st edn. Springer: USA

7 Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," Journal of the American Statistical Association, 22, 209–212.

8 Food and Drug Administration (May 2007), Guidance for Industry Clinical Trial Endpoints for Approval of Cancer Drugs and Biologics

9 Belsley DA, Kuh E, Welsch RE. Regression Diagnostics. New York: John Wiley & Sons (1980).

11 APPENDICES

11.1 APPENDIX 1: VARIABLE DEFINITIONS

Age will be calculated as the date of birth subtracted from the randomization date, divided by 365.25 [Age= (Randomization Date-DOB)/365.25]. Only the integer part of the result will be taken.

Weight will be displayed in kilograms, height will be displayed in centimeters, and temperature will be displayed in Celsius. Weights, heights, or temperatures recorded in alternate units will be converted to the units being displayed using standard conversion formulas.

11.2 APPENDIX 2: STATISTICAL ANALYSIS AND PROGRAMMING DETAILS

The SAS procedure LIFETEST will be used in the Kaplan-Meier analyses. Patients who did not have an event will be censored.

The following code will be used:

```
proc lifetest data=all method=km alpha=0.05 outsurv=interval;
        time aval*censr(1);
        strata trtp;
        id usubjid;
run;
```

The SAS procedure PHREG will be used in the Cox regression analysis of time-to-event endpoints. Patients who did not have an event will be censored. The SAS method of discrete will be used to handle ties.

The SAS procedure MIXED will be used for mixed modeling. The following code will be used:

```
proc mixed method = reml;
class BASE TRT VISIT SUBJID;
model CH= BASE TRT VISIT  TRT*VISIT /s ddfm=kr;
repeated VISIT/type=UN subject=SUBJID;
lsmeans TRT*VISIT /slice=VISIT diff alpha=0.05 cl;
run;
```

where BASE is the baseline score, TRT is the assigned treatment, VISIT is the visit based on the window mapping, CH is the change from baseline.

11.3 APPENDIX 3: FACT-hep Questionnaire Scoring Rules

FACT-hep Scoring Guidelines (Version 4)

Instructions:* 1. Record answers in "item response" column. If missing, mark with an X
2. Perform reversals as indicated, and sum individual items to obtain a score.
3. Multiply the sum of the item scores by the number of items in the subscale, then divide by the number of items answered. This produces the subscale score.
4. Add subscale scores to derive total scores (TOI & FACT-hep).
5. **The higher the score, the better the QOL.**

Subscale	Item Code	Reverse item?		Item response	Item Score
PHYSICAL WELL-BEING (PWB)	GP1	4	-	_____	=_____
	GP2	4	-	_____	=_____
	GP3	4	-	_____	=_____
	GP4	4	-	_____	=_____
Score range: 0-28	GP5	4	-	_____	=_____
	GP6	4	-	_____	=_____
	GP7	4	-	_____	=_____

Sum individual item scores: _____
Multiply by 7: _____
Divide by number of items answered: _____
=**PWB subscale score**

SOCIAL/FAMILY WELL-BEING (SWB)	GS1	0	+	_____	=_____
	GS2	0	+	_____	=_____
	GS3	0	+	_____	=_____
	GS4	0	+	_____	=_____
Score range: 0-28	GS5	0	+	_____	=_____
	GS6	0	+	_____	=_____
	GS7	0	+	_____	=_____

Sum individual item scores: _____
Multiply by 7: _____
Divide by number of items answered: _____
=**SWB subscale score**

EMOTIONAL WELL-BEING (EWB)	GE1	4	-	_____	=_____
	GE2	0	+	_____	=_____
	GE3	4	-	_____	=_____
	GE4	4	-	_____	=_____
Score range: 0-24	GE5	4	-	_____	=_____
	GE6	4	-	_____	=_____

Sum individual item scores: _____
Multiply by 6: _____
Divide by number of items answered: _____
=EWB subscale score

FUNCTIONAL WELL-BEING (FWB)	GF1	0	+	_____	=_____
	GF2	0	+	_____	=_____
	GF3	0	+	_____	=_____
	GF4	0	+	_____	=_____
Score range: 0-28	GF5	0	+	_____	=_____
	GF6	0	+	_____	=_____
	GF7	0	+	_____	=_____

Sum individual item scores: _____
Multiply by 7: _____
Divide by number of items answered: _____
=FWB subscale score

Subscale	**Item Code**	**Reverse item?**		**Item response**	**Item Score**
HEPATOBILIARY CANCER SUBSCALE (HCS)	C1	4	-	_____	=_____
	C2	4	-	_____	=_____
	C3	0	+	_____	=_____
	C4	0	+	_____	=_____
	C5	4	-	_____	=_____
Score range: 0-72	C6	0	+	_____	=_____
	Hep1	4	-	_____	=_____
	Cns7	4	-	_____	=_____
	Cx6	4	-	_____	=_____
	HI7	4	-	_____	=_____
	An7	0	+	_____	=_____
	Hep2	4	-	_____	=_____
	Hep3	4	-	_____	=_____
	Hep4	4	-	_____	=_____
	Hep5	4	-	_____	=_____
	Hep6	4	-	_____	=_____
	HN2	4	-	_____	=_____
	Hep8	4	-	_____	=_____

Sum individual item scores:_____

Multiply by 18: _____
Divide by number of items answered: _____
=HC Subscale score

To derive a FACT-hep Trial Outcome Index (TOI):
Score range: 0-128

_____ + _____ + _____ =_____=**FACT-hep TOI**
(**PWB** score) (**FWB** score) (**HCS** score)

To Derive a FACT-hep total score:
Score range: 0-180

_____ + _____ + _____ + _____ + _____ =_____=**FACT-hep Total score**
(**PWB** score) (**SWB** score) (**EWB** score) (**FWB** score) (**HCS** score)

*For guidelines on handling missing data and scoring options, please refer to the Administration and Scoring Guidelines in the manual or on-line at www.facit.org.

11.4 APPENDIX 4: Definition and Derivation of Subsequent HCC Therapy

A patient is considered to have received 'subsequent HCC therapy', after the protocol required treatments, if they had

- a subsequent systemic anti-cancer treatment (excluding sorafenib), and/or
- a non-protocol liver directed therapy (excluding ablation/surgery)

Subsequent systemic anti-cancer treatment will be identified using all non-sorafenib records from the "Best Available Care - Post Treatment Discontinuation – Medication" eCRF page. Note that all records on the "Best Available Care - Post Treatment Discontinuation – Medication" eCRF page are expected to be systemic anti-cancer treatments for HCC after data cleaning has been completed.

Non-protocol liver directed therapy will be identified using all records (except ablation/surgery) from "Additional Procedures" eCRF page. This eCRF page captures any procedures relating to liver directed therapies outside the study protocol, such as liver ablation, liver surgery, liver external beam radiation therapy, SIRT with rhenium, SIRT with Y90 (including TheraSphere for patients randomized to control group), TACE, and transarterial embolization (TAE). The free text field "procedure term" will be reviewed and categorized by physician, which will be finalized prior to the database hard lock. Procedures identified as liver ablation and liver surgery by physician's review will not be considered as subsequent HCC therapy.

Notes:

- Treatments recorded on the "Resection / HCC Stage Migration" eCRF page will not be considered as subsequent HCC therapy, since these treatments are delivered to patients who had a downstaging of disease and represents a population possible to have better outcomes and important to follow-up.
- TheraSphere treatment/retreatment for patients randomized to TheraSphere group, recorded on the "TheraSphere Doses Administered" eCRF page, are study treatment per protocol, so will not be considered as subsequent HCC therapy.

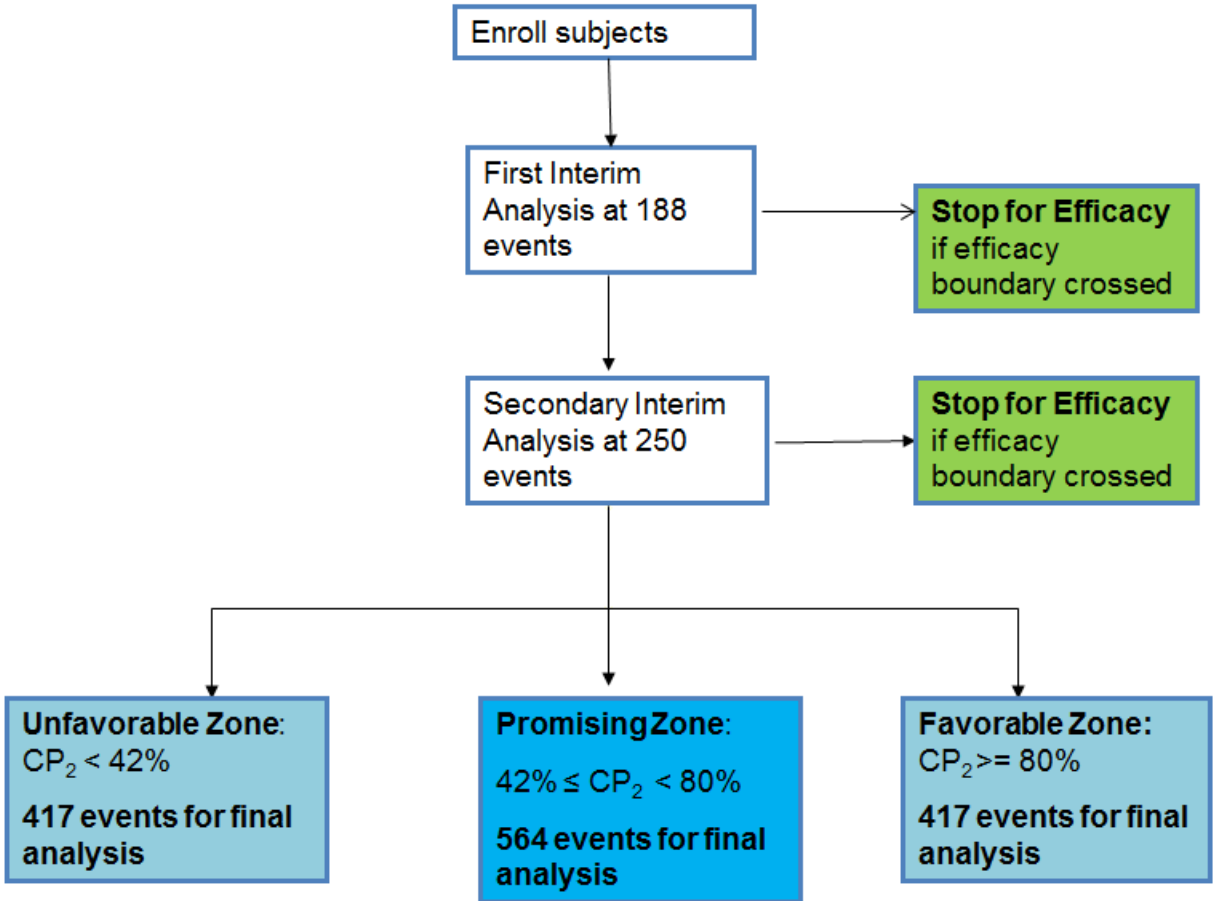
11.5 APPENDIX 5: Statistical Details of the Adaptive Design for Protocol TS-103 STOP-HCC

Statistical Details of the Adaptive Design

1 Design Overview

This study uses an adaptive group sequential design with two interim analyses and one final analysis. The efficacy stopping boundaries will be based on the rho family error spending function with the parameter value $\rho = 1.5$. An assessment of futility at the two interim analyses, based on conditional power, was included in the study design. However, it was decided by the Sponsor before the first interim analysis was performed, that the futility assessment would not be performed, primarily because patient recruitment was faster than expected towards the latter part of the study, such that all patients for the original sample size had already been randomized. A total of 417 events are planned based on providing more than 80% power to detect a hazard ratio of 0.754 with the group sequential design. The first interim analysis is planned at 45% (188 events) information time and the second interim analysis is planned at 60% (250 events) information time. However if the true hazard ratio is 0.781 which is larger than 0.754 but is still considered as a clinically meaningful improvement, the study will be at risk of being underpowered. Therefore a promising zone approach as in Mehta and Pocock (2011) will be employed at the second interim analysis to increase the total number of events. The design is depicted by the following flow chart. More specifically, the first interim analysis is planned to be performed when 188 overal survival (OS) events are observed at which the trial may be stopped for overwhelming efficacy at the first interim analysis time if the observed Z-score statistic crosses the efficacy boundary 2.430. If at the first interim look the study is not stopped for efficacy, the trial will continue and the second interim analysis is planned to be performed when 250 OS events are observed. At the second interim analysis, the trial will be terminated for efficacy (if efficacy boundary 2.429 is crossed). If the trial is not stopped for efficacy, the conditional power given the observed trend will be calculated. If the conditional power of the trial at the second interim analysis falls in the unfavorable zone with conditional power less than 42% or in the favorable zone with conditional power larger than 80%, the trial will continue as planned and the final analysis will be performed when 417 OS events are observed. If the trial falls in the promising zone with conditional power between 42% and 80%, the total number of OS events required will be increased and the final analysis will occur when 564 OS events are observed. The final analysis will be performed by comparing the Z-score statistic to the original group sequential boundary 2.093. The efficacy boundary would be re-calculated using the rho family spending function (rho=1.5) if the actual interim look did not occur at exactly the planned number of events.

Figure 1: Design Schema



2 Justification of Type I Error

This adaptive design with sample size re-estimation described in Section 1 strongly controls type I error. The type I error control follows from a similar argument as in Mehta and Pocock (2011) or Gao et al (2008). The essential idea behind it is the so-called Muller and Shaffer principle as proposed in Muller and Shaffer (2001). More specifically, let Z_2 and Z_3 denote the Z-score based on a log rank statistic using the cumulative data available at the second look and the last look for the three-look group sequential design with the initial planned 417 OS events without sample size adaptation. Let c_3 be the efficacy boundary of the group sequential design at the last look. Let Z_3^* be the Z-score based on all cumulative data available at the final look for the adaptive design depicted by Figure 1, where the total number of events will remain at 417 if the trial at the second look lands in the unfavorable zone with conditional power (CP) <42% or in the favorable zone with CP $\geq$80%, otherwise the total number of events

will be increased in a fixed amount to 564 if the trial lands into the promising zone with conditional power between 42% to 80%. The adaptive design in Figure 1 will test the null hypothesis by comparing Z_3^* to the group sequential boundary c_3 at the final look. The Muller and Shaffer (2001) principle says that the overall type I error will be controlled if the final test in the adaptive design preserves the conditional type I error of the group sequential design. The conditional type I error for the group sequential design is $P_0(Z_3 > c_3 \mid Z_2 = z_2)$ and the conditional type I error for the adaptive design is $P_0(Z_3^* > c_3 \mid Z_2 = z_2)$. Therefore to show that the type I error for the adaptive design in Figure 1 is strongly controlled, it is sufficient to show that

$$P_0(Z_3^* > c_3 \mid Z_2 = z_2) \le P_0(Z_3 > c_3 \mid Z_2 = z_2) \quad (1)$$

for any z_2 which lands into one of the three zones: unfavorable, favorable and promising.

If the conditional power of the trial at the second interim analysis lands in the unfavorable zone ($CP_2 < 0.42$) or favorable zone ($CP_2 \ge 0.8$), the number of events is not increased which implies Z_3^* is the same as Z_3. Therefore $P_0(Z_3^* > c_3 \mid Z_2 = z_2) = P_0(Z_3 > c_3 \mid Z_2 = z_2)$. On the other hand, if the trial lands into the promising zone with conditional power between 0.42 and 0.8, the total number of events will be increased to564. Let c_3^* be the adjusted boundary such that the conditional type I error is preserved i.e. $P_0(Z_3^* > c_3^* \mid Z_2 = z_2) = P_0(Z_3 > c_3 \mid Z_2 = z_2)$. In the Appendix, we have shown that $c_3^* < c_3$ for any z_2 in the promising zone as seen from Figure 2 which implies that $P_0(Z_3^* > c_3 \mid Z_2 = z_2) \le P_0(Z_3^* > c_3^* \mid Z_2 = z_2)$. Therefore $P_0(Z_3^* > c_3 \mid Z_2 = z_2) \le P_0(Z_3 > c_3 \mid Z_2 = z_2)$.

3 Simulations

This section summarizes the operating characteristics of the adaptive design with the option to increase events and sample size (the design described in Section 1) at interim analysis time compared to the group sequential design. Two interim analyses are planned and the efficacy boundaries are derived based on the ρ family error spending function with $\rho = 1.5$. The first interim analysis is planned at 188 overall survival events and the second interim at 250 events. At the second interim analysis, the conditional power will be computed. If the conditional power is $<42\%$ or $\ge 80\%$, the trial will continue and the final analysis will be performed when 417 events are observed. On the other hand, if the conditional power of the trial is between 42% and 80%, the total number of events will be increased to 564 uniformly (as shown in Figure 1) based on providing 80% power to detect an improvement in median OS from 10.7 to 13.7 months using a log rank test with a final two-sided alpha of 0.0363. It is planned to enroll 520 subjects in 60 months with additional 18 months follow-up. This includes an adjustment to take account of an assumed 5% of patients who will be lost to follow-up and for whom a date of death is not recorded, and an assumed additional 5% of patients who will be erroneously randomized because they did not meet the eligibility criteria at randomization. If the total number of events is increased to 564, it is planned to enroll 700 subjects in 66 months with

additional 18 month follow-up. Similarly, this includes an adjustment to take account of an assumed 5% of patients who will be lost to follow-up and for whom a date of death is not recorded, and an assumed additional 5% of patients who will be erroneously randomized because they did not meet the eligibility criteria at randomization.

The simulations assumed that the effective sample size is 468 (520*0.9) if no adaptation occurs and 630 (700*0.9) in case that the total sample size is increased. The median survival time is assumed to be 10.7 months. The simulations for type I error were performed under the null hypothesis of the true hazard ratio 1. The total number of simulations is 1 million. All the simulations were conducted using East 6.4.1. Based on the simulations, 24881 trials out of 1 million are significant for the adaptive design (Adapt) and therefore the type I error level is 0.02488 which is below the nominal one-sided level 0.025. Note that the type I error for the group sequential design without sample size adaptation is 0.02533, which is slightly above the nominal one-sided level 0.025. This is due to the use of the normal approximation to the log rank statistics. Also , 0.02533 is still within three standard errors of Monte Carlo simulation. Table 1a shows the zonewise summary of the simulations under the null hypothesis for the group sequential design (GSD) and the adaptive design (Adapt). Table 1b shows the average sample size, events, accrual duration and study duration for the group sequential design and the adaptive design. The power performance were simulated for hazard ratio 0.754 and hazard ratio 0.781 and the results are summarized in Table 2a, Table 2b, Table 3a and Table 3b. The number of simulations for power performance is 100000. If the true hazard ratio is 0.754, the adaptive design with sample size increase provides 82% power compared to the group sequential design which has 80% power (Table 2a). If the trial lands in the promising zone, the power of the adaptive design is boosted to 89.6% as compared to 76.6% for the group sequential design (Table 2a). This gain of power comes with a cost of increased number of subjects and study duration. With the adaptive design, the average study duration is 93.9 months if the interim result of the trial falls in the promising zone (Table 2b). If the true hazard ratio is 0.781, then the power of the adaptive design is 71.2% which is about a 3 percentage point increase from 68.4% for the group sequential design. The conditional power given that the trial falls into the promising zone is boosted from 67.5% to 83% (Table 3a). Given that the trial lands into the promising zone, the average study duration of the adaptive design is 16.3 months longer (93.3 months) than the group sequential design (77 months) (Table 3b). If the trial falls out of the promising zone, the operating characteristcs of the adaptive design is the same as the group sequential design since no changes are made to the trial conduct.

Table 1a: Simulation for Type I Error under Hazard Ratio 1

Zone	Prob. of Entering Each Zone (%)	Power (%)	
		GSD	Adapt
Unfavorable	93.4%	0.6%	0.6%
Promising	4.2%	12.5%	10.5%
Favorable	1.2%	26.7%	27.2%
Efficacy	1.2%	100%	100%
All Trials	**100%**	**2.533%**	**2.488%**

Table 1b: Average Sample Size, Events, and Durations under Hazard Ratio 1

Zone	Average Sample Size		Average Number of Events		Average Accrual Duration (months)		Average Study Duration (months)	
	GSD	Adapt	GSD	Adapt	GSD	Adapt	GSD	Adapt
Unfavorable	467	467	417	417	59.9	59.9	73	73
Promising	467	629	417	563	59.9	80.6	73	89.8
Favorable	467	467	417	417	59.9	59.9	73	73
Efficacy	321	321	209	210	41	41	41	41
All Trials	**465**	**472**	**415**	**421**	**59.6**	**61**	**72.7**	**73.4**

Table 2a: Power under Hazard Ratio 0.754

Zone	Prob. of Entering Each Zone (%)	Power (%)	
		GSD	Adapt
Unfavorable	23.6%	39%	39%
Promising	17.7%	76.6%	89.6%
Favorable	13.3%	89%	88.8%
Efficacy	45.3%	100%	100%
All Trials	**100%**	**80%**	**82%**

Table 2b: Average Sample Size, Events, and Durations under Hazard Ratio 0.754

Zone	Average Sample Size		Average Number of Events		Average Accrual Duration (months)		Average Study Duration (months)	
	GSD	Adapt	GSD	Adapt	GSD	Adapt	GSD	Adapt
Unfavorable	467	467	417	417	59.9	59.9	77.5	77.5
Promising	467	629	417	563	59.9	80.6	77.7	93.9
Favorable	467	467	417	417	59.9	59.9	77.7	77.7
Efficacy	334	335	208	208	42.7	42.8	42.8	42.9
All Trials	**407**	**435**	**322**	**347**	**52**	**55.7**	**62**	**64.6**

Table 3a: Power under Hazard Ratio 0.781

Zone	Prob. of Entering Each Zone (%)	Power (%)	
		GSD	Adaptive
Unfavorable	33.2%	30%	29.5%
Promising	19.2%	67.5%	83%
Favorable	12.8%	84%	84%
Efficacy	34.8%	100%	100%
All Trials	**100%**	**68.4%**	**71.2%**

Table 3b: Average Sample Size, Events, and Durations under Hazard Ratio 0.764

Zone	Average Sample Size		Average Number of Events		Average Accrual Duration (months)		Average Study Duration (months)	
	GSD	Adapt	GSD	Adapt	GSD	Adapt	GSD	Adapt
Unfavorable	467	467	417	417	59.9	59.9	76.9	76.9
Promising	467	629	417	563	59.9	80.6	77	93.3
Favorable	467	467	417	417	59.9	59.9	77	77
Efficacy	334	334	209	209	42.7	42.7	42.8	42.8
All Trials	**421**	**451**	**345**	**372**	**53.9**	**57.8**	**65**	**68**

Appendix

This appendix describes how the lower bound of the promising zone is derived. Let Z_2 and Z_3 be Z-score based on the log rank statistics using the cumulative data at the second look and the last look for the three-look group sequential design with a total of 417 OS events without sample size adaptation. Let c_3 be the efficacy boundary of the group sequential design at the last look. Let Z_3^* be the Z-score based on the actual events at the final look for the adaptive design where the total number of events is increased to 564 for any observed z_2. Let c_3^* be the adjusted boundary which exactly preserves the conditional type I error. The Muller and Shaffer (2001) principle states that the overall type I error will be controlled if the final test boundary of the adaptive design is adjusted such that the final test of the adaptive design preserves the conditional type I error of the group sequential design. The conditional type I error for the group sequential design is $P_0(Z_3 > c_3 \mid Z_2 = z_2)$. The conditional type I error of the adaptive design with a fixed sample size increase to 564 for any observed z_2 is $P_0(Z_3^* > c_3^* \mid Z_2 = z_2)$. In other words, the overall type I error is controlled at one-sided nominal level $\alpha = 0.025$ (or two-sided 0.05) as long as c_3^* satisfies the following equation for any z_2.

$$P_0(Z_3 > c_3 \mid Z_2 = z_2) = P_0(Z_3^* > c_3^* \mid Z_2 = z_2) \quad (2)$$

where $P_0()$ denotes that the probability is evaluated under the null hypothesis that the hazard ratio is 1. The left hand side of (2) is the conditional type I error of the group sequential design given that $Z_2 = z_2$ is observed. By Equation (3) in Gao et al (2008), the left hand side of (2) is given by

$$P(Z_3 > c_3 \mid Z_2 = z_2) = 1 - \Phi\left(\frac{c_3\sqrt{t_3} - z_2\sqrt{t_2}}{\sqrt{t_3 - t_2}}\right)$$

and the right hand side of (2) is the conditional type given by

$$P(Z_3^* > c_3^* \mid Z_2 = z_2) = 1 - \Phi\left(\frac{c_3^*\sqrt{t_3^*} - z_2\sqrt{t_2}}{\sqrt{t_3^* - t_2}}\right)$$

where $t_2 = \frac{n_2}{4}$ and $t_3 = \frac{n_3}{4}$ are the cumulative information of the log rank statistics at the second look and the last look for the original group sequential design, $t_3^* = \frac{n_3^*}{4}$ is the cumulative information of the log rank statistic at the last look for the adaptive design. For this trial, $n_2 = 250$, $n_3 = 417$ and $n_3^* = 564$ and$c_3 = 2.093$. Therefore to preserve conditional type I error, c_3^* need to satisfy the following equation

$$\frac{c_3^*\sqrt{t_3^*} - z_2\sqrt{t_2}}{\sqrt{t_3^* - t_2}} = \frac{c_3\sqrt{t_3} - z_2\sqrt{t_2}}{\sqrt{t_3 - t_2}}$$

i.e.

$$c_3^* = \left[\frac{\sqrt{t_3^* - t_2}}{\sqrt{t_3 - t_2}} \sqrt{t_3} c_3 + z_2 \sqrt{t_2} \left(1 - \frac{\sqrt{t_3^* - t_2}}{\sqrt{t_3 - t_2}} \right) \right] \frac{1}{\sqrt{t_3^*}} \tag{3}$$

By Equation (4) in Gao et al 2008, the conditional power given observed treatment effect $\hat{\theta} = \frac{z_2}{\sqrt{t_2}}$ with the initial planned sample size of 417 is given by

$$CP_2 = \Phi\left(\hat{\theta} \sqrt{t_3 - t_2} - \frac{1}{\sqrt{t_3 - t_2}} \left(c_3 \sqrt{t_3} - \sqrt{t_2} z_2 \right) \right)$$

$$= \Phi\left(z_2 \left[\sqrt{\frac{t_3 - t_2}{t_2}} + \sqrt{\frac{t_2}{t_3 - t_2}} \right] - c_3 \sqrt{\frac{t_3}{t_3 - t_2}} \right)$$

which implies

$$z_2 = \left(\Phi^{-1}(CP_2) + c_3 \sqrt{\frac{t_3}{t_3 - t_2}} \right) \Big/ \left(\sqrt{\frac{t_3 - t_2}{t_2}} + \sqrt{\frac{t_2}{t_3 - t_2}} \right) \tag{4}$$

We can express the new boundary c_3^* as a function of the conditional power CP_2 by plugging (4) into (3) as follows

$$c_3^* = \left[\frac{\sqrt{t_3^* - t_2}}{\sqrt{t_3 - t_2}} \sqrt{t_3} c_3 + \frac{\Phi^{-1}(CP_2) + c_3 \sqrt{\frac{t_3}{t_3 - t_2}}}{\sqrt{\frac{t_3 - t_2}{t_2}} + \sqrt{\frac{t_2}{t_3 - t_2}}} \sqrt{t_2} \left(1 - \frac{\sqrt{t_3^* - t_2}}{\sqrt{t_3 - t_2}} \right) \right] \frac{1}{\sqrt{t_3^*}} \tag{5}$$

Similarly we can express CP_2 as a function of c_3^*

$$CP_2 = \Phi\left[\left(c_3^* \sqrt{t_3^*} - \frac{\sqrt{t_3^* - t_2}}{\sqrt{t_3 - t_2}} \sqrt{t_3} c_3 \right) \left(\frac{\sqrt{\frac{t_3 - t_2}{t_2}} + \sqrt{\frac{t_2}{t_3 - t_2}}}{\sqrt{t_2} \left(1 - \frac{\sqrt{t_3^* - t_2}}{\sqrt{t_3 - t_2}} \right)} \right) - c_3 \sqrt{\frac{t_3}{t_3 - t_2}} \right] \tag{6}$$

The green line in the following plot shows the behavior of the adjusted boundary c_3^* against the conditional power and z_2. Note that c_3^* is a decreasing function of the conditional power. The blue line shows the group sequential efficacy boundary $c_3 = 2.093$ at the last look. In (5) or (6), if we set $c_3^* = c_3$, we can find the conditional power such that the blue line and the green line meet each other which gives $CP_2 = 0.41578$ corresponding to $z_2 = 1.516$. Note that for conditional power > 0.41578 (or $z_2 > 1.516$), we have $c_3^* < c_3$. In particular, $c_3^* < c_3$ for conditional power between 0.41578 and 0.8 (or z_2 between 1.516 and 2.033). Thus setting CP_{min} of the promising zone to 0.41578 assures no inflation of the overall study

type 1 error after increasing the sample size of the study.