



Statistical Analysis Plan

Study Code D419CC00002

Edition Number 4.0

Date 30JUL2021

A Randomized, Open-label, Multi-center Phase III Study of Durvalumab and Tremelimumab as First-line Treatment in Patients with Advanced Hepatocellular Carcinoma (HIMALAYA)

A Randomized, Open-label, Multi-center Phase III Study of Durvalumab and Tremelimumab as First-line Treatment in Patients with Advanced Hepatocellular Carcinoma (HIMALAYA)

Study Statistician

PPD



Aug 6, 2021

Date

A Randomized, Open-label, Multi-center Phase III Study of Durvalumab and Tremelimumab as First-line Treatment in Patients with Advanced Hepatocellular Carcinoma (HIMALAYA)

Global Product Statistician

PPD



Aug 6, 2021

Date

TABLE OF CONTENTS	PAGE
TITLE PAGE.....	1
SIGNATURE OF STUDY STATISTICIAN.....	2
SIGNATURE OF GLOBAL PRODUCT STATISTICIAN	3
TABLE OF CONTENTS	4
LIST OF TABLES	8
LIST OF FIGURES	9
LIST OF ABBREVIATIONS	10
AMENDMENT HISTORY	14
1. STUDY DETAILS.....	25
1.1 Study objectives	25
1.2 Study design.....	29
1.3 Number of Subjects.....	32
2. ANALYSIS SETS	34
2.1 Definition of analysis sets.....	34
2.1.1 Full analysis set.....	36
2.1.2 Safety analysis set	36
2.1.3 PK analysis set	36
2.1.4 ADA evaluable sets.....	36
2.2 Protocol Deviations.....	37
3. PRIMARY, SECONDARY AND EXPLORATORY VARIABLES	38
3.1 Primary endpoint variables	38
3.1.1 Overall survival (OS).....	38
3.2 Derivation of RECIST 1.1 Visit Responses.....	38
3.2.1 Target lesions (TLs).....	39
3.2.2 Non-target lesions (NTLs) and new lesions.....	44
3.2.3 Overall visit response – site investigator data.....	45
3.2.4 Blinded Independent Central Review (BICR)	46
3.2.5 Investigator RECIST 1.1-based secondary and CCI endpoints	46
3.3 Secondary variables	47
3.3.1 Progression Free Survival (PFS).....	47
3.3.2 Time to progression (TTP).....	48
3.3.3 Objective response rate (ORR)	48
3.3.4 Best objective response (BoR).....	49

3.3.5	Disease control rate (DCR)	50
3.3.6	Duration of response (DoR)	50
3.3.7	Proportion of subjects alive at 12, 18, 24 and 36 months after randomization (OS12, OS18, OS24 and OS36)	51
3.3.8	Time to Response (TTR)	51
3.3.9	Time from Randomization to First Subsequent Therapy or Death (TFST)	51
3.4	CCI [REDACTED]	51
3.4.1	[REDACTED]	52
3.4.2	[REDACTED]	52
3.4.3	[REDACTED]	52
3.5	Patient reported outcome (PRO) variables	53
3.5.1	EORTC QLQ-C30	53
3.5.1.1	Time to global health status/QoL, function or symptoms deterioration	55
3.5.1.2	Symptom improvement rate	56
3.5.1.3	Global health status /QoL or function improvement rate	57
3.5.2	EORTC QLQ-HCC18	57
3.5.2.1	Time to symptom deterioration	58
3.5.2.2	Symptom improvement rate	59
3.5.3	CCI [REDACTED]	59
3.5.4	[REDACTED]	59
3.5.5	[REDACTED]	60
3.5.6	Compliance	60
3.6	CCI [REDACTED]	60
3.7	Safety Variables	61
3.7.1	General considerations	61
3.7.2	Adverse events	63
3.7.2.1	AEs of special interest and AEs of possible interest	63
3.7.2.2	Other significant adverse events	63
3.7.3	Treatment exposure	63
3.7.4	Dose intensity	66
3.7.5	Laboratory data	67
3.7.6	ECGs	67
3.7.7	Vital signs	68
3.7.8	ECOG performance status	68
3.7.9	Child-Pugh score	68
3.7.10	Physical examinations	68
3.7.11	Other safety assessments	68
3.7.12	Prior and concomitant medications	69
3.8	Pharmacokinetic variables	69
3.8.1.1	Population pharmacokinetics and exposure-response/safety analysis	69
3.8.1.2	Pharmacokinetic non-compartmental analysis	70

3.9	Immunogenicity analysis	70
3.10	Pharmacogenetic variables	70
3.11	Biomarker variables	71
4.	ANALYSIS METHODS	71
4.1	General principles	72
4.1.1	Baseline	74
4.1.2	Visit windows for safety and PRO assessment	74
4.1.3	Study day will be calculated in relation to date of first treatment. Visit Windows for PK and ADA	75
4.2	Analysis methods	76
4.2.1	Multiplicity	76
4.2.2	Analysis of the primary variable	78
4.2.2.1	Overall survival	78
4.2.3	Analysis of the secondary variables	82
4.2.3.1	Progression Free Survival (PFS)	82
4.2.3.2	Time to progression (TTP)	82
4.2.3.3	Objective response rate (ORR)	82
4.2.3.4	Best objective response (BoR)	83
4.2.3.5	Disease control rate (DCR)	83
4.2.3.6	Duration of response (DoR)	83
4.2.3.7	OS12, OS18, OS24, and OS36	83
4.2.3.8	Time to Response (TTR)	84
4.2.3.9	Time from Randomization to First Subsequent Therapy or Death (TFST)	84
4.2.4	Patient-reported outcomes	84
4.2.4.1	EORTC QLQ-C30	84
4.2.4.2	EORTC QLQ-HCC18	86
4.2.4.3	CCI	87
4.2.4.4		87
4.2.4.5		87
4.2.5		87
4.2.6	Subgroup analyses	87
4.2.6.1	Subgroup analyses for OS	87
4.2.6.2	Subgroup analyses for secondary endpoints	88
4.2.7	Safety	89
4.2.7.1	Adverse events	89
4.2.7.2	Treatment exposure and intensity	93
4.2.7.3	Laboratory assessments	93
4.2.7.4	ECGs	95
4.2.7.5	Vital signs	95
4.2.7.6	ECOG performance status	95
4.2.7.7	Child-Pugh score	95
4.2.7.8	Physical examinations	96
4.2.7.9	Other safety assessments	96

4.2.8 Pharmacokinetic data.....	96
4.2.9 Immunogenicity data	96
4.2.10 CCI [REDACTED]	97
4.2.11 Biomarker data	97
4.2.12 CCI [REDACTED]	97
4.2.12.1 CCI [REDACTED]	97
4.2.12.2 [REDACTED]	98
4.2.12.3 [REDACTED]	98
4.2.12.4 [REDACTED]	98
4.2.13 Demographic, initial diagnostics and baseline characteristics data.....	98
5. INTERIM ANALYSES	99
5.1 Analysis methods	99
5.2 Blinding.....	100
5.3 Independent Data Monitoring Committee	100
6. CHANGES OF ANALYSIS FROM PROTOCOL	101
7. REFERENCES	102

LIST OF TABLES

Table 1	Summary of study objectives, outcome measures and analysis sets	26
Table 2	Summary of exploratory objectives, outcome measures and analysis sets	27
Table 3	Summary of randomized studies used to determine NI margin	34
Table 4	Summary of additional Phase 3 studies with a sorafenib control	34
Table 5	Summary of outcome variables and analysis sets	35
Table 6	TL Visit Responses (RECIST 1.1)	40
Table 7	NTL visit responses	44
Table 8	Overall Visit Responses	45
Table 9	Censoring rules for PFS	47
Table 10	Mean change and clinically meaningful change - EORTC QLQ-C30	53
Table 11	Best response in EORTC QLQ-C30 and EORTC QLQ-HCC18 scores: FAS	54
Table 12	Visit response for HRQoL and disease-related symptoms	57
Table 13	Formal Statistical Analyses to be Conducted and Pre-planned Sensitivity Analyses	71
Table 14	Response rate analyses conducted during the study	82

LIST OF FIGURES

Figure 1	Study design	30
Figure 2	Multiple testing strategy	78

LIST OF ABBREVIATIONS

Abbreviation or special term	Explanation
ADA	Anti-drug antibody
AE	Adverse Event
AEPI	Adverse Event of possible interest
AESI	Adverse Event of special interest
AFP	Alpha-fetoprotein
AJCC	American Joint Committee on Cancer
ALP	Alkaline Phosphatase
ALT	Alanine Aminotransferase
AST	Aspartate Aminotransferase
AZ	AstraZeneca
BCLC	Barcelona Clinic Liver Cancer
BICR	Blinded Independent Central Review
BID	Twice daily
BMI	Body Mass Index
BoR	Best Objective Response
CD	Cluster of differentiation
CI	Confidence interval
CR	Complete response
CSP	Clinical Study Protocol
CSR	Clinical Study Report
CT	Computed tomography
CTC	Common terminology criteria
CTCAE	Common terminology criteria for adverse event
CCI	
DBL	Database lock
DCO	Data cut-off
DCR	Disease control rate
DCR-16w	Disease control rate at 16 weeks
DCR-24w	Disease control rate at 24 weeks

Abbreviation or special term	Explanation
DLT	Dose Limiting Toxicity
DoR	Duration of Response
ECG	Electrocardiogram
ECOG PS	Eastern Cooperative Oncology Group Performance Status
eCRF	Electronic Case Report Form
EORTC	European Organisation for Research and Treatment of Cancer
CCI	
FA	Final analysis
FAS	Full analysis set
FAS-32w	FAS subjects with an opportunity for 32 weeks of follow-up prior to IA1 DCO
FWER	Familywise error rate
GI	Gastrointestinal
H1, H2, H3	Hypothesis 1, Hypothesis 2, Hypothesis 3
HBV	Hepatitis B virus
HCC	Hepatocellular carcinoma
HCV	Hepatitis C virus
CCI	
HR	Hazard ratio
HRQoL	Health-related quality of life
IA	Interim Analysis
IA1	First interim analysis; Interim Analysis 1
IA2	Second interim analysis; Interim Analysis 2
ICF	Informed Consent Form
ICH	International Conference on Harmonisation
IDMC	Independent Data Monitoring Committee
IHC	Immunohistochemical
ILD	Interstitial Lung Disease
IP	Investigational Product
IPD	Important Protocol Deviations

Abbreviation or special term	Explanation
CCI	
ITT	Intent-to-treat
IV	Intravenous
IWRS	Interactive Web Response System
KM	Kaplan-Meier
LD	Longest diameter
LIMS	Laboratory Information Management System
LLOQ	Lower Limit of Quantification
MedDRA	Medical Dictionary for Regulatory Activities
MMRM	Mixed-effect model for repeated measurement
mRECIST	Modified Response Evaluation Criteria in Solid Tumors
MRI	Magnetic resonance imaging
CCI	
MTP	Multiple Testing Procedure
NA	Not Applicable
NE	Not Evaluable
NI	Non-inferiority
NQ	Not Quantifiable
NTL	Non-target lesion
OAE	Other significant adverse event
ORR	Objective response rate
OS	Overall survival
OS12	Overall survival at 12 months
OS18	Overall survival at 18 months
OS24	Overall survival at 24 months
OS36	Overall survival at 36 months
PD	Progressive Disease; Protocol Deviation; Pharmacogenetic Data
PD-L1	Programmed cell death ligand 1
PFS	Progression-free survival (time to first progression)
CCI	

Abbreviation or special term	Explanation
CCI	
PK	Pharmacokinetics
PR	Partial Response
PRO	Patient-reported outcome
CCI	
Q12W	Every 12 weeks
Q4W	Every 4 weeks
Q8W	Every 8 weeks
QLQ-C30	30-item core quality of life questionnaire
QLQ-HCC18	18-item hepatocellular carcinoma health-related quality of life questionnaire
QoL	Quality of Life
QT	Q wave T wave
QTcF	QT interval corrected for using Fridericia's formula
RDI	Relative dose intensity
RECIST	Response Evaluation Criteria in Solid Tumors
RR	R wave to R wave
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan
SD	Stable Disease
TFST	Time to first subsequent therapy
TL	Target lesion
TSH	Thyroid-Stimulating Hormone
TTP	Time to progression
TTR	Time to response
UK	United Kingdom
ULN	Upper Limit of Normal
CCI	

AMENDMENT HISTORY

Date	Brief description of change
25-Oct-2017	Original version 1.0 PPD

23-Aug-2019

Version 2.0 PPD

1. Replaced the order of treatment arms and added labels (Arm A, B, C, D) to match the protocol.
2. Section 4.2.7.1: changed “Most common AEs with CTCAE grade 3 or higher” to “Most common AEs with CTCAE grade 3 or 4”; added categories: "Any SAE leading to discontinuation of study medication", "Any SAE leading to discontinuation of treatment, causally related to study medication".
3. Updated Protocol Deviations list.
4. Rephrased the paragraph about trial integrity and “sponsor-blind” by adding suggested wording from the Open Label Guidance document (version 17-Sep-2017).
5. Added a summary “Disease characteristics at screening”; added a note to repeat tables for subset of subjects experiencing re-challenge; included type of summary for concomitant medications; split summaries for allowed prior and concomitant medications; added “Alcohol use at baseline” to the list of baseline characteristics.
6. Added Section 3.7.12 to define prior and concomitant medications.
7. Added a clarification that p-value for survival will be obtained from log-rank test.
8. Section 4.2.6.1: added an explanation about excluding stratification factors.
9. Added a note about potential Hy’s law definition.
10. Added a table describing PFS censoring rules.
11. Added “infection events” summary.
12. Removed Section 3.7.13.
13. Added explanations about “Other significant adverse events”.
14. Added details about PRO plots and compliance tables.
15. Added details about analysis to be produced for IA1 and the timing.
16. Added CMH analysis as requested by FDA.
17. Update details about PK analyses.
18. Updated wording about PD-L1 analyses.
19. Removed text about listing other deviations that are not IPDs.
20. CCI
21. Corrected incomplete dates imputation rules to be compliant with AZ standards.
22. Removed paragraph about using local laboratory ranges.
23. Clarified definition of randomized patients.
24. Added PRO to Table 3 list.
25. Removed endpoints derived only for Payer analysis and not included in study endpoints list (PFS2, TTR, TFST, TSST) and updated section “Changes to the planned analyses”.

26. Added a general statement about conversion from weeks to days.
27. Moved rounding rules to Section 4.1.
28. Removed unnecessary table with repeated information from Section "Relationship between AFP and efficacy parameters".
29. Update baseline characteristics section following changes to eCRF (nicotine and alcohol use).
30. CCI [REDACTED]
31. Added a note in Section 4.1 about using only data until DCO at each analysis.
32. Added a note in Section 5.2 about Shadow team activities.
33. Updated slightly AE event rate definition to follow TA SAP.
34. Removed paragraph about qualitative interactions testing from "subgroup analyses" section.
35. Added subgroup analysis by BCLC score as required by CSP.
36. Added a note about the date of first subsequent therapy in Safety analyses section.
37. Corrected visit window for Day 29 visit to include Day 1.
38. Updated BoR definition to reflect study schedule.
39. Added a note about censoring at randomization for PRO TTR for subjects with no post-baseline assessment.
40. Removed text about summarizing previous cancer therapy.
41. Rephrased one paragraph in subgroup analysis section to follow TA SAP.
42. Added "Effect of covariates on the HR estimate" to Section 4.2.2.1 and removed related paragraph from section "Subgroup analyses for OS".
43. Edits about exposure in Sections 3.7.3, 3.7.4.
44. Updated windows in Section 4.1.2 to include screening and baseline, and to be relative to first dose date for PRO.
45. Updated baseline definition in Section 4.1.2 to be relative to first dose for PRO (Sections 3.5.1.1, 3.5.2.1).
46. Updated Section 4.1.1 to clarify baseline definitions.
47. Removed reference to maintenance phase from exposure analysis Sections (3.7.3 and 4.2.7.2).
48. Added max-combo test to OS sensitivity analyses.
49. Updated section 2.1.1 to include IA1 analysis set for efficacy.
50. Added a note to Section 4.1 about formal analyses.
51. Added Table 12 to specify ORR analyses at different timepoints.
52. Defined cut-off for PD-L1 analysis in Section 4.2.11.
53. Corrected DCR definition in Section 3.3.5 to match CSP.

Updates related to the study protocol amendments:

54. Updates to sample size calculations, MTP and study hypotheses definition.

Date	Brief description of change
	55. Addition of endpoints: DCR-16w, DCR-24w.
	56. Removing Arm B from formal comparisons and including for descriptive purpose.
	57. Addition of text about use of separate single-arm summary of subset of TLFs for Arm B. Arm B will not be included in main TLFs.
	58. Removed physical examination from Sections 1.1 and 2.1, added notes to Sections 3.7.10, 4.2.6.8.
	59. Updated Table 8 and Table 9 following changes to the corresponding tables in the protocol.
	60. CCI [REDACTED]
	61. Updated Section 4.2.2.3 about ORR analysis.
	62. CCI [REDACTED]
	63. [REDACTED]
	64. Added secondary objective to conduct RECIST 1.1 and mRECIST analyses (ORR, BoR, DoR) by BICR for the IA1 set of patients with an opportunity for 32 weeks of follow-up.
	65. Updated wording BCLC score to BCLC stage.
	66. Sections 4.2.3.3 and 5.1, and CCI [REDACTED] sections - clarified ORR will be presented by Investigator assessment (using RECIST1.1) and BICR (using RECIST1.1 and mRECIST) for IA1.

15-May-2020

Version 3.0 PPD

1. SAP author handover from PPD
2. PPD handover from PPD
3. Abbreviation added for Adverse event of possible interest (AEPI).
4. CCI
5. Changed 'patient' to 'subject' throughout.
6. Changed 'casually related' to 'possibly related' throughout.
7. Changed treatment arm descriptions for Arms B and D in Section 1.2 so that tremelimumab is the first treatment.
8. In Section 1.2, the phrase 'by treatment arm' was removed from the description of two interim analyses for the study.
9. Added ECOG 0/1 to the target patient population criteria in Section 1.2.
10. Removed sentences about patients from China enrolled in global study in Section 1.3 because no patients from China were enrolled in the global study. Specified that the China cohort will be made up of only patients in the China tail.
11. Added power calculations and design assumptions for the non-inferiority analysis of Arm A vs Arm D in Section 1.3. Table 3 added to summarize results of the studies used to determine the non-inferiority margin. Table 4 was added to summarize other Phase 3 studies in first-line advanced HCC that include non-inferiority to a sorafenib control. Supporting reference for Table 4, Cheng et al 2019, was added to the References section.
12. Clarified that ORR for both confirmed and unconfirmed responses will be analysed at IA2 and FA according to Investigators assessments per RECIST 1.1 in Section 3.3.3.
13. Clarified the definition of DCR in Section 3.3.5.
14. Clarified that DoR for both confirmed and unconfirmed responses will be analysed at IA2 and FA according to Investigators assessments per RECIST 1.1 in Sections 3.3.6 and 4.2.3.6.
15. In Section 3.3.9, for the TFST endpoint, the definition of censoring was clarified.
16. Clarified in Section 3.5.1.1 that the analysis set for EORTC QLQ-C30 time to symptom deterioration will consist of a subset of FAS patients who have a baseline symptom score ≤ 90 .
17. In Section 3.7.1, removed statement that denominator in vital signs data should include only those patients with recorded data. Change also applies to Section 3.7.7.

18. In Section 3.7.1, removed text for imputation of completely missing end dates for AEs and concomitant medications in Section 3.7.1. Added that for completely missing AE ends dates, if the subject has died and the AE stop date is missing, then the stop date of the AE will be imputed as the death date. Clarified that end dates will not be imputed for concomitant medications with start date after the last dose date. For immune-mediated adverse event summaries, an AE with outcome of unknown will be imputed as not resolved.
19. Modified Section 3.7.2.1 to include AEs of possible interest.
20. Added descriptions of exposure analyses for treatment durations, number of infusions/doses received, dose delays, infusion interruptions for Arms A, B, and C, and treatment cycles received for Arms A, B, and C in Section 3.7.3. These analyses are also referenced in Section 4.2.7.2.
21. Clarified definition of time on study in Section 3.7.3.
22. CCI 
23. In Section 4.1, replaced the listing of patients who discontinued from study treatment with a listing of discontinued subjects. Replaced the listing of patients excluded from the efficacy analysis with a listing of subjects excluded from the safety analysis.
24. Removed statement that intervals with a confidence level corresponding to adjusted significance level will be produced for endpoints included in the MTP form Section 4.2.
25. Clarified in Sections 4.1 and 4.2.6.1 that for subgroup analyses, stratification factor values collected from the eCRF will be used to define subgroups.
26. Clarified in Section 4.1 that MMRM estimates should only be summarized for visits where scores for at least 25% of patients in both treatment arms are available.
27. Clarified in Section 4.1.2 that safety and PRO visit windows should be applied until the last dose of study treatment + 90 days rather than until PD. Removed bullet for follow-up visit from list of visit windows to be applied.
28. Removed text from Section 4.1.2 stating that to prevent very large tables or plots being produced, visit data should only be summarised if the number of observations for each treatment group is greater than the minimum of 20 and $> 1/3$ of patients dosed.
29. Clarified in Section 4.2.1 that if H1 is rejected at IA2 or FA, alpha will be recycled to H2 across IA2 and FA.
30. In Section 4.2.1, added the anticipated number of events across Arms A and D, significance levels for H2 and H3 at the time of IA2 and FA, and confidence interval levels to be applied for the non-inferiority comparisons at IA2 and FA based on the anticipated number of events.

31. Added that the Kaplan-Meier landmark analysis will be repeated for the 12-month OS rate in Section 4.2.2.1.
32. Added text to Section 4.2.2.1 to describe a listing of subjects either diagnosed with COVID-19 or died due to COVID-19. CCI
[REDACTED]
33. CCI
[REDACTED]
34. Clarified that adjusted alpha levels for OS analyses will be performed using Lan and DeMets approach that approximates the O'Brien Fleming spending function for both the primary and key secondary analyses.
35. Removed PRO endpoints of EORTC QLQ C30 TTD in physical functioning, EORTC QLQ C30 TTD in fatigue, EORTC QLQ C30 TTD in appetite loss, EORTC QLQ C30 TTD in nausea, and EORTC QLQ HCC18 TTD in abdominal pain from the multiple testing plan. These details were removed from Sections 4.2.4 and 4.2.1.
36. Clarified in Section 4.2.4 that improvement rate will be analysed for PROs to align with Sections 3.5.1.2, 3.5.1.3, 3.5.2.2, 4.2.4.1, and 4.2.4.2.
37. CCI
[REDACTED]
38. Updated primary PRO measures in Section 4.2.4.1 to align with Section 4.2.4. Clarified that the primary PRO comparisons will be between immunotherapy arms (Arm A, Arm C) and the sorafenib arm.
39. Added details in Section 4.2.4.1 of how MMRM model will be estimated.
40. In Section 4.2.4.1, clarified the use of covariance structures for use in MMRM model.
41. Removed statement from Section 4.2.3.3 that analysis for ORR using BICR assessments may be performed at IA2 and FA. Statement also applies to Table 14, from which BICR RECIST 1.1 and BICR mRECIST 1.1 response rate analyses were removed for the Interim Analysis 2 and Final Analysis time points.
42. Removed statement from Section 4.2.3.6 that analysis for DoR using BICR assessments may be performed at IA2 and FA.
43. Removed statement from Section 4.2.6.1 that stratification factors from IWRS will be used for OS subgroup analyses where appropriate.
44. Added subgroup analysis for MVI = Yes and/ or EHS = Yes to Section 4.2.6.1.
45. Added subgroup analysis for macrovascular invasion (yes versus no) for secondary endpoints to Section 4.2.6.2.

46. Removed redundant safety summaries from Section 4.2.7.1 as well as the summary of dose limiting toxicities to align with the protocol.
47. Updated cutoff for the summaries of most common AEs to be 10% in Section 4.2.7.1.
48. Added description of the programmatic process for identification of imAEs using AESIs and AEPs and provided a list of summaries for imAEs to align with new imAE standards. A manual process may also be used to identify AESIs/AEPs from a list of preferred terms.
49. Removed the shift table for urinalysis (Bilirubin, Blood, Glucose, Ketones, Protein) comparing baseline CTCAE grade to maximum grade on treatment value from Section 4.2.7.3.
50. Updated the lists of summaries for patient characteristics at baseline, disease characteristics at initial diagnosis, and disease characteristics at screening in Section 4.2.13.
51. Clarified in Section 4.2.13 that PD-L1 status will be summarized at baseline and that the summary of Post IP discontinuation anti-cancer therapy is for disease related anti-cancer therapy.
52. Added maturity across Arms C and D at IA2 and FA for the OS analysis of H1 to Section 5.1.
53. Added to Section 6 that in addition to AESIs, AEPs will also be determined using the latest list of preferred terms. Added additional details of the non-inferiority analysis for Arm A vs Arm D beyond what was provided in the protocol.

July 2021

Version 4.0 (PPD

1. PPD handover from PPD
2. Study statistician handover from PPD
3. Added NI (Non-inferiority), OS36 (Overall survival at 36 months), NQ (not quantifiable) and LLOQ (lower limit of quantification) to abbreviations
4. Added abbreviation, FAS-32w. In Section 2.1.1, FAS-32w is defined as an analysis set for the subset of subjects randomized $\geq$ 32 weeks prior to IA1 DCO. This abbreviation is used in Sections 3.3.3, 3.3.6, 4.2.3.2, 4.2.3.4, 5.1 and in Tables 1, 5, and 14 in place of the text describing FAS subjects with opportunity for 32 weeks of follow-up.
5. CCI
6. Added the statistical margin for the non-inferiority comparison and clarified that 1.08 is the clinical non-inferiority margin in Section 1.3.
7. Added OS36 to efficacy data in Table 5; also add OS36 in table footnote (section 2.1)
8. Deleted the statement about analysis not being performed “A per-protocol analysis excluding subjects with specific important protocol deviations is not planned” in section 2.2.
9. Added NED as an overall visit response category to Section 3.2.4.
10. Added analyses for ORR subgroups at IA1, IA2, and FA to Section 3.3.3. Clarified that the ORR subgroup analysis at IA1 will use BICR data.
11. Added a summary to Section 3.3.4 to compare BoR by Investigator assessment to BoR by BICR assessment for the FAS-32w.
12. Added the summary of proportion of subjects live 36 months after randomization (OS36) in section 3.3.7.
13. Clarified in Section 3.5.6 that subjects who are unable to read PRO questionnaires will be excluded from compliance calculations.
14. Changed abbreviations to AESI and AEPI (removed “s”) in section 3.7.2.1
15. Added sentence to mention the Hepatic and Hemorrhage SMQ AE analysis. (section 3.7.2.1)
16. Added header “Dose delays” in section 3.7.3.
17. Added a definition of dose delays in section 3.7.3.
18. Clarified that dose delays will be summarized in Arm C for durvalumab only in section 3.7.3.
19. Added clarifications that infusion interruptions applies to only arms A, B, and C in section 3.7.3
20. Added clarifications how infusion interruption summaries will be done in section 3.7.3

21. Added a statement defining how dose reductions are calculated for sorafenib in section 3.7.3
22. Added details to Section 3.7.12 regarding review of medication data by the AZ medical team to identify disallowed medications. Disallowed medications will be summarized in a table. Summary added for concomitant medications which began prior to randomization.
23. Changed “missing” to “NQ (Not quantifiable)” for sample below LLOQ according to PK evaluation guideline and pointed the details to section 4.2.8 in section 3.8.1.2
24. Added OS 36 and the stratified test at a fixed time point to Table13 (section 4)
25. Clarified in Section 4.1 that Arm B will be summarized for descriptive purposes in all efficacy and safety tables instead of descriptive summaries for Arm B appearing in a separate set of tables.
26. Added a new section, Section 4.1.3, to describe the creation of time windows for summaries of PK and ADA data by visit.
27. Rephrased sentences about alpha recycling in section 4.2.1.
28. Added one sentence to state the alpha recycling for OS36 test in section 4.2.1.
29. Changed “PD-L1 high” and “PD-L1 low/negative” to “PD-L1 positive” and “PD-L1 negative” in Table 1 and section 4.2.2.1, 4.2.6.1, 4.2.6.2, 4.2.11, and 4.2.13.
30. Section 4.2.2.1 – Specified in Section 4.2.2.1 that the Grambsch-Therneau test may be used to assess non-proportionality and clarified the details for the max-combo test.
31. Details for summaries of duration of follow-up for prematurely censored subjects, censored subjects alive at DCO, and all subjects added to Section 4.2.2.1.
32. Clarified in Section 4.2.2.1 that the two stage-method (Latimer 2018) will be the primary method for the treatment switch analysis and that a Weibull mixture cure model will be fit to adjust for subsequent therapy initiation.
33. Added summaries by treatment arm to Section 4.2.2.1 for subsequent therapies received after discontinuation of treatment and for subjects receiving immunotherapy according to line of subsequent therapy.
34. Added sensitivity analysis to Table 13 for assessing the impact of COVID-19 on both OS. Details of the analysis are provided in Section 4.2.2.1.
35. CCI 
36. Added 36-month landmark to Kaplan Meier curve in subsection “assumption of proportionality” in section 4.2.2.1.
37. Changed the region group to (Asia (except Japan) versus Rest of World (includes Japan)) in section 4.2.2.1&4.2.6.1

38. Clarified in Section 4.2.2.3 that for IA1, descriptive summaries of ORR will be presented for all treatment arms including Arms A, B, C, and D.
 39. Added Kaplan Meier estimate of OS at 36 months and a test of OS36 in section 4.2.3.7.
 40. Added MVI = No and EHS = No as subgroup for the primary endpoint in Section 4.2.6.1.
 41. Removed summaries of adverse events of infection by infection pooled term and pooled term from Section 4.2.7.1.
 42. Changed “durvalumab” to “Study medication” in the AEs list in section 4.2.7.1
 43. Added additional sentence to mention AE results with a cutoff of 5% will be also reported (in addition to the planned 10%) in section 4.2.7.1
 44. Changed AESI to AESI/AEPI to “imAEs by AESI/AEPI category and preferred term” and “imAEs by AESI/AEPI group, preferred term, and maximum CTCAE Grade” in section 4.2.7.1
 45. Removed statement from Section 4.2.8 that PK by-visit summaries will only be produced for CSP scheduled timepoints.
 46. Added calculation details for sample value below LLOQ according to the PK evaluation guideline in section 4.2.8
 47. Added a short paragraph in section 6 to state the deviation of OS36 test
 48. References added to Section 7 for Grambsch and Therneau (1994), Karrison (2016), Latimer (2018), Lin (2020), checkmate 459 (Yau 2019), Brivanib (Johnson 2013), Linifanib (Cainap 2015), FDA guidance 2016, and EMEA guideline 2005
-

1. STUDY DETAILS

This statistical analysis plan (SAP) contains a detailed description of the analyses in the clinical study protocol (CSP) for study D419CC00002 (HIMALAYA). This SAP is based on version 6.0 (dated 20-Aug-2019) of the CSP.

This is a randomized, open-label, multi-center, global, Phase III study to assess the efficacy and safety of durvalumab monotherapy and durvalumab plus tremelimumab combination therapy versus sorafenib in the treatment of subjects with no prior systemic therapy for hepatocellular carcinoma (HCC). The subjects cannot be eligible for locoregional therapy.

The primary endpoint will be overall survival (OS) and there will be two interim analyses (IA1 and IA2) and a final analysis (FA).

1.1 Study objectives

The formal statistical analysis of OS (primary endpoint) will be performed for the following efficacy test hypotheses (alternative hypotheses):

- H1: Difference between durvalumab [redacted] plus tremelimumab [redacted] (Arm C) and sorafenib [redacted] (Arm D)
- H2: Durvalumab [redacted] monotherapy (Arm A) not inferior to sorafenib [redacted] (Arm D) with noninferiority (NI) margin of 1.08
- H3: Difference between durvalumab [redacted] monotherapy (Arm A) and sorafenib [redacted] (Arm D)

No formal efficacy analysis will be conducted for Arm B, durvalumab [redacted] plus tremelimumab [redacted] combination therapy arm, which was closed for enrolment with Amendment 4. Instead, results for Arm B will be summarized descriptively and separately from other study arms.

The primary objective is to test Hypothesis 1 (H1); i.e., whether subject's OS when randomized to receive durvalumab [redacted] plus tremelimumab [redacted] (Arm C) is significantly different than the OS of subjects randomized to receive sorafenib [redacted] (Arm D) in the full analysis set (FAS). The key secondary objectives are to first test H2; i.e., whether subject's OS when randomized to receive durvalumab [redacted] monotherapy (Arm A) is not inferior to the OS of subjects randomized to receive sorafenib [redacted] (Arm D) with NI margin of 1.08 in the FAS, and then test H3; i.e., whether subject's OS when randomized to receive durvalumab [redacted] monotherapy (Arm A) is significantly different than the OS of subjects randomized to receive sorafenib [redacted] (Arm D) in the FAS.

The study will be considered successful if the comparison defined in H1 reaches statistical significance in favour of Arm C at either IA2 or FA as outlined in Section 4.2.1.

Table 1 summarises the primary, secondary and safety objectives and the outcome measures used in the analyses of these objectives. All of these objectives will be reported in the clinical

study report (CSR). CCI

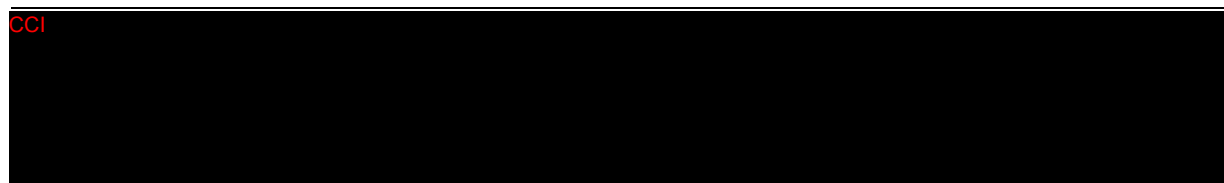
CCI

Table 1 Summary of study objectives, outcome measures and analysis sets

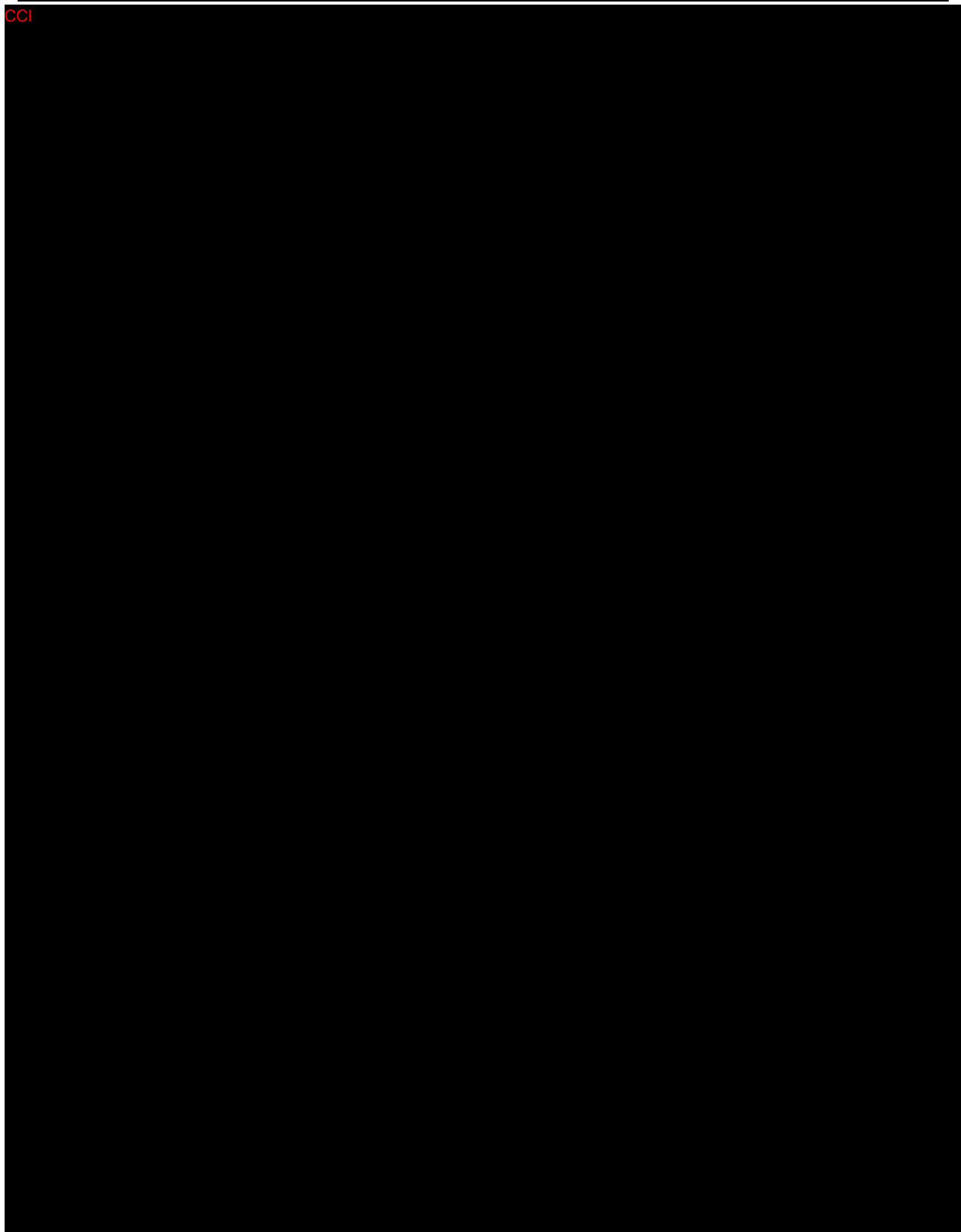
Study objectives	Outcome measure	Analysis set
Primary objective		
To assess the efficacy of Arm C vs. Arm D (for superiority)	Overall survival (OS)	FAS
Key Secondary objectives		
To assess the efficacy of Arm A vs. Arm D (for non-inferiority)	OS	FAS
To assess the efficacy of Arm A vs. Arm D (for superiority)	OS	FAS
To assess the efficacy of Arm C vs. Arm D	OS at 36 months (OS36)	FAS
Secondary objectives		
To assess the efficacy of Arm A vs. Arm D and Arm C vs. Arm D	- OS at 18 months (OS18) and OS at 24 months (OS24) - Progression-free survival (PFS), time to progression (TTP), objective response rate (ORR), disease control rate (DCR), disease control rate at 16 weeks (DCR-16w), disease control rate at 24 weeks (DCR-24w) and duration of response (DoR), according to Response Evaluation Criteria in Solid Tumors, version 1.1 (RECIST 1.1) using Investigator assessments	FAS
To assess the efficacy of Arm A and Arm C in subjects with an opportunity for 32 weeks of follow-up	ORR, best objective response (BoR), and DoR according to RECIST1.1 and modified Response Evaluation Criteria in Solid Tumors (mRECIST) by Blinded Independent Central Review (BICR)	FAS subjects with an opportunity for 32 weeks of follow-up (FAS-32w)

To assess the efficacy of Arm A vs. Arm D and Arm C vs. Arm D by PD-L1 expression	• OS • PFS, TTP, ORR, DCR, DCR-16w, DCR-24w and DoR according to RECIST 1.1 using Investigator assessments	FAS, PD-L1 positive, PD-L1 negative
To assess disease-related symptoms, impacts, and health-related quality of life (HRQoL) in Arm A vs. Arm D and Arm C vs. Arm D	• European Organisation for Research and Treatment of Cancer (EORTC) 30-item core quality of life questionnaire (QLQ-C30): Time to deterioration in global health status/QoL, functioning (physical), multi-term symptom (fatigue), single-item symptoms (appetite loss, nausea) • EORTC 18-item hepatocellular cancer health-related quality of life questionnaire (QLQ-HCC18): Time to deterioration in single-item symptoms (shoulder pain, abdominal pain, abdominal swelling)	FAS
To investigate the immunogenicity of Arm A and Arm C	Presence of anti-drug antibody (ADA) for durvalumab and tremelimumab	Safety set (listings)/ ADA evaluable set (summary)
To evaluate the population pharmacokinetics (PK) and pharmacodynamics in Arm A and Arm C	Durvalumab and tremelimumab concentrations and PK parameters in individual treatment arms	PK analysis set
Safety objective		
To assess the safety and tolerability profile across all treatment arms	Adverse events, treatment exposure, dose intensity, laboratory findings, electrocardiograms (ECGs), vital signs, ECOG PS, Child-Pugh score	Safety set

Table 2 **Summary of exploratory objectives, outcome measures and analysis sets**



CCI



CCI

1.2 Study design

This randomized Phase III study will assess the efficacy and safety of durvalumab, with or without tremelimumab, compared to sorafenib in the treatment of subjects with no prior systemic therapy for HCC that are not eligible for locoregional therapy. This will be a multi-center global study enrolling subjects from different regions including North America, South America, Asia and Europe. Therefore, the study population is expected to be representative of the demographic variation and the global distribution of HCC.

Subjects will be randomized in a 1:1:1:1 ratio to one of the following 4 arms:

- 1) Arm A: Durvalumab CCI monotherapy
- 2) Arm B: Tremelimumab CCI plus Durvalumab CCI combination therapy
- 3) Arm C: Tremelimumab CCI plus Durvalumab CCI combination therapy
- 4) Arm D: Sorafenib CCI

Protocol amendment 4 closed enrolment to Arm B. As a result of protocol amendment 4, subjects will be randomized in a 1:1:1 ratio to Arm A, Arm C and Arm D. Subjects randomized to Arm B prior to amendment 4 can remain on study as planned until discontinuation criteria are met at the discretion of the investigator.

Randomization will be stratified according to macrovascular invasion (yes versus no), etiology of liver disease (hepatitis B virus [confirmed HBV] versus hepatitis C virus [confirmed HCV] versus others), and ECOG PS (0 versus 1).

This study will use an open-label design because blinding of the treatment assignment is challenging owing to the unique safety profile differences between durvalumab and tremelimumab compared with sorafenib, and the different routes of administration causing undue burden to subjects. Moreover, the study drugs will have different administration schedules and treatment durations.

Although the study is open-label, it will be conducted “Sponsor-blind”. To maintain the integrity of the study, Sponsor access to treatment records will be restricted, and, in particular, under no circumstances will the Sponsor undertake any efficacy analysis by treatment arm during the study. A Trial Integrity Document will be generated in which nominated individuals who will be granted access to any treatment-revealing data will be pre-specified, with their reason for requiring access detailed.

The study includes two interim analyses, which will be performed by an Independent Data Monitoring Committee (IDMC). Details will be given in the IDMC charter.

Figure 1 Study design



*approximately

BCLC Barcelona Clinic Liver Cancer; BID twice daily; ECOG Eastern Cooperative Oncology Group;
HBV hepatitis B virus; HCC hepatocellular carcinoma; HCV hepatitis C virus; PS performance status;
CCI

Target patient population

The study population includes subjects 18 years of age or older with advanced HCC, Barcelona Clinic Liver Cancer stage B (not eligible for locoregional therapy) or stage C, ECOG performance score 0 or 1, and Child-Pugh class A liver disease. Subjects must not have received any prior systemic therapy for HCC.

Each subject should meet all inclusion criteria and none of the exclusion criteria for this study. Under no circumstances can there be exceptions to this rule.

Duration of treatment

Subjects in all treatment arms should, wherever possible, continue to receive their initially assigned treatment to disease progression.

At the Investigator's discretion, subjects in all treatment arms may continue receiving treatment until progressive disease (PD) by RECIST 1.1 is confirmed on a follow-up scan as per Confirmation of Radiological Progression criteria (see Section 7.2.1.3 of CSP). The follow-up scan should preferably occur at the next scheduled visit and no earlier than 4 weeks after the previous assessment of PD. Refer to Appendix B of the CSP for the criteria for confirmation of progression.

Subjects in all arms with confirmed PD who, in the Investigator's opinion, continue to receive benefit from their assigned treatment and meet the criteria for treatment in the setting of PD may continue to receive their assigned treatment. However, subjects who develop progression in target lesions (TLs) after a clear response to therapy as defined by RECIST 1.1 will not be permitted to continue therapy.

Rechallenge option for subjects in the durvalumab plus tremelimumab combination therapy arms

Subjects in the durvalumab plus tremelimumab combination therapy arms who complete the assigned dosing cycle(s) of durvalumab plus tremelimumab, and are benefiting from study drug(s) in the Investigator's opinion, and subsequently have evidence of PD with or without confirmation according to RECIST 1.1 during the durvalumab monotherapy portion of their regimen, can be rechallenged with tremelimumab, provided they meet eligibility criteria for rechallenge as described in Section 7.2.1.3 of the protocol. Subjects assigned to Arm B can be rechallenged in their assigned treatment arm, or, with tremelimumab CCI along with durvalumab with prior approval from the AstraZeneca Study Physician. Subjects in Arm C may only be rechallenged with tremelimumab CCI along with durvalumab if eligible for rechallenge.

Subjects who rechallenge with tremelimumab after PD must have a rechallenge baseline tumor assessment within 28 days of restarting treatment with tremelimumab plus durvalumab combination therapy. Using regular RECIST 1.1 baseline guidelines, the rechallenge baseline may have TLs and non-target lesions different from those at the original baseline (including pre-existing new lesions). Rechallenge follow-up scans should occur Q8W (± 1 week) for the

first 48 weeks (relative to the date of first rechallenge treatment) then Q12W thereafter until confirmed disease progression.

Tumor assessments

Tumor assessments, based on RECIST 1.1, will be performed every 8 weeks (Q8W) (± 1 week) for the first 48 weeks from the date of randomization and then Q12W (± 1 week) thereafter until RECIST 1.1-defined radiological progression is confirmed by a follow-up scan, if clinically feasible, as per Confirmation of Radiological Progression criteria (see Appendix B of the CSP). Subjects who continue treatment beyond radiological progression should continue with tumor assessments on their regular imaging schedule for the duration of their treatment.

Subject follow-up post-discontinuation of study drug

Subjects for whom AstraZeneca and the Investigator determine may not continue treatment after PD will be followed up for survival. Subjects who have discontinued treatment due to toxicity or symptomatic deterioration, or who have commenced subsequent anticancer therapy, will be followed until disease progression and for survival.

Post final data cutoff

Subjects who continue to receive benefit from their assigned treatment at the final data cutoff (DCO) and database closure may continue to receive their assigned treatment for as long as they and their physician feel they are gaining clinical benefit. For subjects continuing to receive durvalumab treatment following the final DCO and database closure, it is recommended that the subjects continue the scheduled site visits and investigators monitor the subject's safety laboratory results prior to and periodically during treatment with durvalumab in order to manage adverse events (AEs) in accordance with the durvalumab toxicity management guidelines.

In the event that a rollover or safety extension study is available at the time of the final DCO and database closure, subjects currently receiving treatment with durvalumab may be transitioned to such a study, and the current study would reach its end. The rollover or safety extension study would ensure treatment continuation with visits assessment per its protocol. Any subject that would be proposed to move to such study would be given a new Informed Consent.

1.3 Number of Subjects

This study will screen approximately 1650 subjects, with no prior systemic therapy for HCC and not eligible for locoregional therapy, in order to randomize approximately 1310 subjects. (This includes approximately 1155 subjects randomized to Arms A, C, D with approximately 385 per arm; and approximately 155 subjects in Arm B, randomized prior to the closure of this arm. Once global enrolment has completed, recruitment into an expansion cohort will continue in China (i.e. China Tail) until up to a total of 180 Chinese subjects have been randomized. The China cohort will be made up of the subjects in the China Tail. Details of the China cohort and analysis plan will be outlined in a China specific amendment and SAP.

The study is sized to characterize the OS benefit of Arm C vs. Arm D.

The sample size estimation assumes an exponentially distributed OS and a 2-month delay in separation of the OS curves for Arm C vs. Arm D, hence the use of average HR (0.70 for Arm C vs. Arm D). A non-uniform accrual of subjects with a duration of 22 months is assumed when estimating the analysis times with a follow-up duration of 15.5 months and a total duration of 37.5 months. No adjustment has been included for dropouts.

For the efficacy comparisons, the median OS for sorafenib (Arm D) is assumed to be 11.5 months (Llovet 2008b, Cheng 2013, Kudo 2018).

Durvalumab [REDACTED] plus tremelimumab [REDACTED] (Arm C) versus sorafenib [REDACTED] (Arm D) (OS in FAS [ITT])

The assumed OS treatment effect is an average HR of 0.70 for Arm C vs. Arm D. This translates to an increase in median OS from 11.5 months to 16.5 months and in the 18-month OS rate from 33.8% to 46.8% in Arm C vs. Arm D.

At the time of IA2, the analysis of OS will be performed when approximately 404 OS events in Arm C and Arm D combined (~52% maturity) have occurred, approximately 30 months after the first subject is randomized. This number of OS events will provide at least 85% power to demonstrate a statistically significant difference in OS at a 2 sided 2.22% significance level.

At the time of FA, the analysis of OS will be performed when approximately 515 events in Arm C and Arm D combined (~67% maturity) have occurred, approximately 37.5 months after the first subject is randomized. This number of OS events will provide at least 97% power to demonstrate a statistically significant difference in OS at a 2 sided 4.25% significance level. The smallest treatment difference that could be observed as statistically significant at the final analysis is an average HR of 0.84 (an increase in median OS from 11.5 months to approximately 13.7 months in Arm C versus Arm D).

Durvalumab [REDACTED] monotherapy (Arm A) versus sorafenib [REDACTED] (Arm D) (OS in FAS [ITT])

It is estimated that approximately 453 and 560 events can be observed at the time of the interim and final analysis respectively. Assuming a target HR of 0.84, the power of the NI test at margin of 1.08 is approximately 84% at final analysis.

- The statistical margin of 1.21 (M_1) and clinical noninferiority margin (M_2) of HR 1.08 are determined using 95%-95% fixed margin approach (FDA Guidance 2016; EMEA Guideline 2005) based on two phase 3 trials of sorafenib (Llovet 2008a and Cheng 2009) in first-line HCC and assuming conservative 60% retention. Multiple historical trials in the same indication were also designed as NI including Brivanib (Johnson 2013), Sunitinib (Cheng 2013), Linifanib (Cainap 2015), and Lenvatinib (Kudo 2018) the last supporting global registration of Lenvatinib.

Table 3 Summary of randomized studies used to determine NI margin

Trial	Sample Size	Events (maturity)	HR (95% CI)
SHARP study (Llovet JM et al. 2008a)	602	321 (53%)	0.69 (0.55, 0.87)
Asian-Pacific study (Cheng AL et al. 2009)	226	168 (74%)	0.68 (0.50, 0.93)
Overall (random effect model)	-	--	0.686 (0.571, 0.825)

Non-inferiority for the comparison of Arm A vs Arm D will be declared if the upper limit of the two-sided alpha adjusted CI for HR is less than the NI margin of 1.08. The analysis follows the intent-to-treat principle. Assuming a HR of 0.84, a total of approximately 560 events will be observed at final analysis in the durvalumab monotherapy arm and the sorafenib arm. A target of 560 events will provide 84% power for the NI test with a NI margin of 1.08. The assumed HR is based on CheckMate-459 results for nivolumab vs sorafenib in the same population (Yau T, 2019). Other studies with sorafenib as a treatment arm in first-line advanced HCC were also reviewed. A total of four Phase 3 studies were designed with non-inferiority to a sorafenib control.

Table 4 Summary of additional Phase 3 studies with a sorafenib control

Description	Median OS Sorafenib	HR (95%CI)	NI Margin
Brivanib Trial, N = 1155 (Johnson PJ, 2013)	9.9 months	1.06 (0.94, 1.23)	1.08
Sunitinib Trial, N=1074 (Cheng A, 2013)	10.2 months	1.30 (1.13, 1.50)	N/A
Linifanib Trial, N = 1035 (Cainap C, 2015)	9.8 months	1.05 (0.90, 1.22)	1.0491
Lenvatinib Trial, N=954 (Kudo M, 2018)	12.3 months	0.92 (0.79, 1.06)	1.08
Atezolizumab Trial, N = 501 (Cheng AL, 2019, p. LBA3)	13.2 months	0.58 (0.42, 0.79)	N/A

2. ANALYSIS SETS

2.1 Definition of analysis sets

Definitions of the analysis sets for each outcome variable are provided in Table 5.

Table 5 Summary of outcome variables and analysis sets

Outcome variable	Analysis set
Efficacy data	
OS	Full analysis set (ITT)
ORR, BoR, DoR, DCR, DCR-16w, DCR-24w, PFS, TTP, OS18, OS24, OS36, PROs, CCI	Full analysis set (ITT) FAS-32w
PRO data	
PRO data	Full analysis set (ITT)
Study Population/ Demography Data	
Demography	Full analysis set (ITT)
Baseline, disease characteristics	Full analysis set (ITT)
Analysis sets	Full analysis set (ITT)
Important deviations	Full analysis set (ITT)
Medical/Surgical history	Full analysis set (ITT)
Prior and concomitant medications/procedures	Full analysis set (ITT)
Previous/current radiotherapy	Full analysis set (ITT)
Subsequent cancer therapy	Full analysis set (ITT)
PK Data	
PK data	PK analysis set
Immunogenicity Data	
Immunogenicity data	Safety analysis set / ADA evaluable sets
Pharmacogenetic Data	
Pharmacogenetic data	Full analysis set (ITT)
Biomarker Data	
Biomarker data	Full analysis set (ITT)
Safety data	
Exposure	Safety analysis set
AEs	Safety analysis set
Laboratory measurements	Safety analysis set
Vital signs	Safety analysis set
Dose intensity	Safety analysis set

ECG	Safety analysis set
ECOG performance status	Safety analysis set
Child-Pugh score	Safety analysis set

ADA Anti-drug antibody; AE Adverse event; BoR Best objective response; DCR Disease control rate; DCR-16w Disease control rate at 16 weeks; DCR-24w Disease control rate at 24 weeks; DoR Duration of response; ECG Electrocardiogram; IA1 Interim Analysis 1; ITT Intent-to-treat; ORR Objective response rate; OS Overall survival; OS18 Overall survival at 18 months ; OS24 Overall survival at 24 months; OS36 Overall survival at 36 months; PFS Progression-free survival; CCI

PK Pharmacokinetics; PRO Patient-reported outcomes; TTP Time to progression.

2.1.1 Full analysis set

The full analysis set (FAS) will include all randomized subjects (date of randomization and randomization code available in Inclusion/ Exclusion criteria CRF page), including subjects who were randomized in error. The FAS will be used for all formal efficacy analyses (including PROs). Treatment arms will be compared on the basis of randomized study drug(s), regardless of the study drug(s) actually received. Subjects who were randomized but did not subsequently go on to receive study drug(s) are included in the analysis in the treatment arm to which they were randomized.

For IA1 an additional analysis set will be defined: FAS subjects with an opportunity for 32 weeks of follow up at the time of IA1 (FAS-32w, i.e., randomized $\geq$ 32 weeks prior to IA1 DCO).

2.1.2 Safety analysis set

The safety analysis set will consist of all subjects who received any amount of study treatment (durvalumab, tremelimumab or sorafenib), including subjects who were randomized in error or not randomized and still started on treatment. Safety data will not be formally analysed but summarized using the safety analysis set according to the treatment received. If a subject receives any amount of an experimental therapy, they will be summarized in the treatment group corresponding to the first experimental treatment they received. If a subject only receives therapy from the control arm, they will be summarized in the control treatment group.

2.1.3 PK analysis set

The PK analysis set will consist of all subjects who receive at least 1 dose of study drug(s) per the protocol for whom any PK post-dose data are available (at least one non-missing post-dose PK result).

2.1.4 ADA evaluable sets

The Durvalumab ADA evaluable set will consist of all subjects in the safety analysis set who have a non-missing baseline durvalumab ADA and at least one non-missing post-baseline durvalumab ADA result.

The Tremelimumab ADA evaluable set will consist of all subjects in the safety analysis set who have a non-missing baseline tremelimumab ADA and at least one non-missing post-baseline tremelimumab ADA result.

All major ADA analyses will be based on these two ADA evaluable sets.

2.2 Protocol Deviations

The following general categories will be considered important protocol deviations. These will be listed and discussed in the CSR as appropriate:

1. Subjects randomized but who did not receive study treatment.
2. Subjects who deviate from key entry criteria per the Clinical Study Protocol (CSP).
 - a) Inclusion criteria: 7, 8, 9.
 - b) Exclusion criteria: 11, 13, 14, 17, 18, 19.
3. Baseline RECIST scan > 42 days before randomization.
4. No baseline RECIST 1.1 assessment on or before date of randomization.
5. Received prohibited systemic anti-cancer agents. Please refer to the CSP section 7.7 for the systemic anti-cancer agents that are detailed as being ‘excluded’ from permitted use during the study. This will be used as a guiding principle for the physician review of all medications prior to database lock.
6. Subjects randomized who received their randomized study treatment at an incorrect dose or received an alternative study treatment to that which they were randomized.
7. Did not have the intended disease or indication 1L HCC. (“Subjects have confirmed HCC based on histopathological findings from tumor tissues and must not have received prior systemic therapy for HCC.”)

Subjects who receive the wrong treatment at any time will be included in the safety analysis set as described in Section 2.1. During the study, decisions on how to handle errors in treatment dispensing (with regard to continuation/discontinuation of study treatment or, if applicable, analytically) will be made on an individual basis with written instruction from the study team leader and/or statistician.

The important protocol deviations will be listed and summarised by randomized treatment group. Deviation 1 will lead to exclusion from the safety analysis set. None of the other deviations will lead to subjects being excluded from the analysis sets described in Section 2.1.

3. PRIMARY, SECONDARY AND EXPLORATORY VARIABLES

3.1 Primary endpoint variables

3.1.1 Overall survival (OS)

The primary endpoint for this trial is OS; defined as the time from the date of randomization until death due to any cause regardless of whether the subject withdraws from randomized therapy or receives another anti-cancer therapy (i.e. date of death or censoring – date of randomization + 1). Any subject not known to have died at the time of analysis will be censored based on the last recorded date on which the subject was known to be alive (SUR_DAT, recorded within the SURVIVE module of the eCRF).

For any analyses or DCOs planned, which include OS analyses, a full survival sweep is planned.

Note: Survival calls will be made following the date of data cut-off (DCO) for the analysis (these contacts should generally occur within 7 days of the DCO). If subjects are confirmed to be alive or if the death date is post the DCO date, these subjects will be censored at the date of DCO. Death dates may be found by checking publicly available death registries. The status of ongoing, withdrawn (from the study) and “lost to follow-up” subjects at the time of the final OS analysis should be obtained by the site personnel by checking the subject’s notes, hospital records, contacting the subject’s general practitioner and checking publicly-available death registries. In the event that the subject has actively withdrawn consent to the processing of their personal data, the vital status of the subject can be obtained by site personnel from publicly available resources where it is possible to do so under applicable local laws.

If a subject is known to have died where only a partial death date is available, then the date of death will be imputed as the latest of the last date known to be alive +1 from the database and the death date using the available information provided:

- a. For Missing day only – using the 1st of the month
- b. For Missing day and Month – using the 1st of January

If there is evidence of death but the date is entirely missing, it will be treated as missing, i.e. censored at the last known alive date.

Subgroup analyses will be performed for OS as indicated in Section 4.2.6.

3.2 Derivation of RECIST 1.1 Visit Responses

For all subjects, the RECIST tumor response data will be used to determine each subject’s visit response according to RECIST version 1.1 (see further Appendix B of the CSP). It will also be used to determine if and when a subject has progressed in accordance with RECIST and their best objective response to study treatment.

Baseline radiological tumor assessments are to be performed no more than 28 days before the date of randomization and ideally as close as possible to randomization (Tables 2 and 3 of CSP).

Follow-up assessments will be performed every 8 weeks (± 1 week) for the first 48 weeks (relative to the date of randomization) and then every 12 weeks (± 1 week) as indicated in the schedule of procedures presented in the protocol (Tables 2 and 3 of CSP) until disease progression. The imaging schedule must be followed regardless of any delays in dosing.

If an unscheduled assessment was performed and the subject has not progressed, every attempt should be made to perform the subsequent assessments at their next scheduled visits. This schedule is to be followed in order to minimise any unintentional bias caused by some subjects being assessed at a different frequency than other subjects.

From the Investigator's review of the imaging scans, the RECIST tumor response data will be used to determine each subject's visit response according to RECIST version 1.1. At each visit, subjects will be programmatically assigned a RECIST 1.1 visit response of CR, PR, SD or PD, using the information from target lesions (TLs), non-target lesions (NTLs) and new lesions and depending on the status of their disease compared with baseline and previous assessments. If a subject has had a tumor assessment that cannot be evaluated, then the subject will be assigned a visit response of not evaluable (NE, unless there is objective disease progression according to RECIST 1.1 in which case the response will be assigned as PD).

Please refer to Section 3.2.3 for the definitions of CR, PR, SD and PD.

RECIST outcomes (i.e. PFS, TTP, ORR etc.) will be calculated programmatically for the site Investigator data (see Section 3.3) from the overall visit responses.

3.2.1 Target lesions (TLs)

Measurable disease is defined as having at least one measurable lesion, not previously irradiated, which is ≥ 10 mm in the longest diameter (except lymph nodes which must have short axis ≥ 15 mm) with computed tomography (CT) or magnetic resonance imaging (MRI) and which is suitable for accurate repeated measurements.

A subject can have a maximum of five measurable lesions recorded at baseline with a maximum of two lesions per organ (representative of all lesions involved and suitable for accurate repeated measurement) and these are referred to as target lesions (TLs). Lymph nodes are collectively considered as a single organ (regardless of designation of 'local/regional' or 'distant'). If more than one baseline scan is recorded, then measurements from the one that is closest and prior to randomization will be used to define the baseline sum of TLs. It may be the case that, on occasion, the largest lesion does not lend itself to reproducible measurement. In which circumstance the next largest lesion, which can be measured reproducibly should be selected.

All other lesions (or sites of disease) not recorded as TL should be identified as non-target lesions (NTLs) at baseline. Measurements are not required for these lesions, but their status should be followed at subsequent visits.

Measurable disease (i.e. at least one TL) is one of the entry criteria for the study. However, if a subject with non-measurable disease is enrolled in the study (i.e. no TLs), the evaluation of

overall visit responses will be based on the overall NTL assessment and the absence/presence of new lesions (see Section 3.2.2 for further details). If a subject does not have measurable disease at baseline (i.e. no TLs and no NTLs), then the TL visit response will be not applicable (NA).

For subjects with no disease at baseline (i.e. no TLs and no NTLs), evaluation of overall visit responses will be based on absence/presence of new lesions. If no TLs and no NTLs are recorded at a visit, both the TL and NTL visit response will be recorded as NA and the overall visit response will be no evidence of disease (NED). If a new lesion is observed, then the overall visit response will be PD.

Table 6 TL Visit Responses (RECIST 1.1)

Visit Responses	Description
Complete Response (CR)	Disappearance of all TLs since baseline. Any pathological lymph nodes selected as TLs must have a reduction in short axis to <10 mm.
Partial Response (PR)	At least a 30% decrease in the sum of diameters of TLs, taking as reference the baseline sum of diameters as long as criteria for PD are not met.
Progressive Disease (PD)	A $\geq 20\%$ increase in the sum of diameters of TLs and an absolute increase of $\geq 5\text{mm}$, taking as reference the smallest sum of diameters since treatment started including the baseline sum of diameters.
Stable Disease (SD)	Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD.
Not Evaluable (NE)	Only relevant in certain situations (i.e. if any of the TLs were not assessed or not evaluable or had a lesion intervention at this visit; and scaling up could not be performed for lesions with interventions). Note: If the sum of diameters meets the progressive disease criteria, progressive disease overrides not evaluable as a TL response.
Not Applicable (NA)	No target lesions are recorded at baseline.

Rounding of TL data

For calculation of PD and PR for TLs, percentage changes from baseline and previous minimum should be rounded to 1 decimal place before assigning a TL response. For example, 19.95% should be rounded to 20.0% but 19.94% should be rounded to 19.9%.

Missing TL data

For a visit to be evaluable, all TL measurements should be recorded. However, a visit response of PD should still be assigned if any of the following occurred:

- A new lesion is recorded.
- A NTL visit response of PD is recorded.
- The sum of TLs is sufficiently increased to result in at least a 20% increase, and an absolute increase of $\geq 5\text{mm}$, from nadir.

Note: the nadir can only be taken from assessments where all the TLs had a lesion diameter recorded.

If there is at least one TL measurement missing and a visit response of PD cannot be assigned, the visit response is NE.

If all TL measurements are missing then the TL visit response is NE. Overall visit response will also be NE, unless there is a progression of non-TLs or new lesions, in which case the response will be PD.

Lymph nodes

For lymph nodes, if the size reduces to $< 10\text{mm}$ then these are considered non-pathological. However, a size will still be given and this size should still be used to determine the TL visit response as normal. In the special case where all lymph nodes are $< 10\text{mm}$ and all other TLs are 0mm then although the sum may be $> 0\text{mm}$ the calculation of TL response should be over-written as a CR.

TL visit responses subsequent to CR

Only CR, PD or NE can follow a CR. If a CR has occurred, then the following rules at the subsequent visits must be applied:

- Step 1: If all lesions meet the CR criteria (i.e. 0 mm for non-nodal TLs or $< 10\text{ mm}$ for lymph node TLs) then response will be set to CR irrespective of whether the criteria for PD of TL is also met i.e. if a lymph node longest diameter increases by 20% but remains $< 10\text{ mm}$.
- Step 2: If some lesion measurements are assigned “NE”, but all other lesions meet the CR criteria (i.e. 0 mm or $< 10\text{ mm}$ for lymph nodes) then response will be set to NE irrespective of whether when referencing the sum of TL diameters the criteria for PD is also met.
- Step 3: If not all lesions meet the CR criteria and the sum of lesions meets the criteria for PD then response will be set to PD

Step 4: If after steps 1 – 3 a response can still not be determined the response will be set to remain as CR.

TL too big to measure

If a TL becomes too big to measure this should be indicated in the database and a size ('x') above which it cannot be accurately measured should be recorded. If using a value of x in the calculation of TL response would not give an overall visit response of PD, then this will be flagged and reviewed by the study team. It is expected that a visit response of PD will remain in the vast majority of cases.

TL too small to measure

If a TL becomes too small to measure, then this will be indicated as such on the case report form and a value of 5mm will be entered into the database and used in TL calculations. However, a smaller value may be used if the radiologist has not indicated 'too small to measure' on the case report form and has entered a smaller value that can be reliably measured. If a TL response of PD results (at a subsequent visit) then this will be reviewed by the study team blinded to treatment assignment.

Irradiated lesions/lesion intervention

Previously irradiated lesions (i.e. lesion irradiated prior to entry into the study) should be recorded as NTLs and should not form part of the TL assessment.

Any TL (including lymph nodes), which has had intervention during the study (for example, radiotherapy / surgery / embolization), should be handled in the following way and once a lesion has had intervention then it should be treated as having intervention for the remainder of the study noting that an intervention will most likely shrink the size of tumors:

- Step 1: The diameters of the TLs (including the lesions that have had intervention) will be summed and the calculation will be performed in the usual manner. If the visit response is PD this will remain as a valid response category.
- Step 2: If there was no evidence of progression after step 1, treat the lesion diameter (for those lesions with intervention) as missing and if $\leq 1/3$ of the TLs have missing measurements then scale up as described in the 'Scaling' section below. If the scaling results in a visit response of PD then the subject would be assigned a TL response of PD.
- Step 3: If after both steps PD has not been assigned, then, if appropriate (i.e. if $\leq 1/3$ of the TLs have missing measurements), the scaled sum of diameters calculated in step 2 should be used, and PR or SD then assigned as the visit response. Subjects with intervention are evaluable for CR as long as all non-intervened lesions are 0 (or <10mm for lymph nodes) and the lesions that have been subject to intervention have a value of 0 (or <10mm for lymph nodes) recorded. If scaling up is not appropriate due to too few non-missing measurements then the visit response will be set as NE.

At subsequent visits the above steps will be repeated to determine the TL and overall visit response. When calculating the previous minimum, lesions with intervention should be treated as missing and scaled up where appropriate (as per step 2 above).

Scaling (applicable only for lesion intervention)

If $> 1/3$ of TL measurements are missing (because of intervention) then target lesion response will be NE, unless the sum of diameters of non-missing TL would result in PD (i.e. if using a value of 0 for missing lesions, the sum of diameters has still increased by $> 20\%$ or more compared to nadir and the sum of target lesions has increased by 5 mm from nadir).

If $\leq 1/3$ of the target lesion measurements are treated as missing (because of intervention) then the results will be scaled up (based on the sizes at the nadir visit to give an estimated sum of diameters) and this will be used in calculations; this is equivalent to comparing the visit sum of diameters of the non-missing lesions to the nadir sum of diameters excluding the lesions with missing measurements.

Example of scaling

Lesion 5 is missing at the follow-up visit; the nadir TL sum including lesions 1-5 was 74 mm.

The sum of lesions 1-4 at the follow-up is 68 mm. The sum of the corresponding lesions at the nadir visit is 62 mm.

Scale up as follows to give an estimated TL sum of 81 mm:

$$68 \times 74 / 62 = 81 \text{ mm}$$

CR will not be allowed as a TL response for visits where there is missing data. Only PR, SD or PD (or NE) could be assigned as the TL visit response in these cases. However, for visits with $\leq 1/3$ lesion assessments not recorded, the scaled up sum of TLs diameters will be included when defining the nadir value for the assessment of progression.

If there was a TL intervention, then TL sum of diameters of a visit cannot be used as nadir.

Lesions that split in two

If a TL splits in two, then the LDs of the split lesions should be summed and reported as the LD for the lesion that split.

Lesions that merge

If two TLs merge, then the LD of the merged lesion should be recorded for one of the TL sizes and the other TL size should be recorded as 0 cm.

Change in method of assessment of TLs

CT, MRI and clinical examination are the only methods of assessment that can be used within the trial, with CT and MRI being the preferred methods and clinical examination only used in

special cases. If a change in method of assessment occurs between CT and MRI, this will be considered acceptable and no adjustment within the programming is needed.

If a change in method involves clinical examination (e.g. CT changes to clinical examination or vice versa), any affected lesions should be treated as missing.

3.2.2 Non-target lesions (NTLs) and new lesions

At each visit, the Investigator should record an overall assessment of the NTL response. This section provides the definitions of the criteria used to determine and record overall response for NTL at the investigational site at each visit.

NTL response will be derived based on the Investigator's overall assessment of NTLs as follows:

Table 7 NTL visit responses

Visit Responses	Description
Complete Response (CR)	Disappearance of all NTLs present at baseline with all lymph nodes non-pathological in size (<10 mm short axis).
Progressive Disease (PD)	Unequivocal progression of existing NTLs. Unequivocal progression may be due to an important progression in one lesion only or in several lesions. In all cases the progression MUST be clinically significant for the physician to consider changing (or stopping) therapy.
Non CR/Non PD	Persistence of one or more NTLs-with no evidence of progression.
Not Evaluable (NE)	Only relevant when one or some of the NTLs were not assessed and, in the Investigator's opinion, they are not able to provide an evaluable overall NTL assessment at this visit. Note: For subjects without TLs at baseline, this is relevant if any of the NTLs were not assessed at this visit and the progression criteria have not been met.
Not Applicable (NA)	Only relevant if there are no NTLs at baseline

To achieve 'unequivocal progression' on the basis of NTLs, there must be an overall level of substantial worsening in non-target disease such that, even in the presence of SD or PR in TLs, the overall tumor burden has increased sufficiently to merit discontinuation of therapy. A modest 'increase' in the size of one or more NTLs is usually not sufficient to qualify for unequivocal progression status.

Details of any new lesions will also be recorded with the date of assessment. The presence of one or more new lesions is assessed as progression.

A lesion identified at a follow-up assessment in an anatomical location that was not scanned at baseline is considered a new lesion and will indicate disease progression.

The finding of a new lesion should be unequivocal: i.e., not attributable to differences in scanning technique, change in imaging modality or findings thought to represent something other than tumor.

New lesions will be identified via a Yes/No tick box. The absence and presence of new lesions at each visit should be listed alongside the TL and NTL visit responses.

A new lesion indicates progression so the overall visit response will be PD irrespective of the TL and NTL response.

If the question ‘Any new lesions since baseline’ has not been answered with Yes or No and the new lesion details are blank this is not evidence that no new lesions are present but should not overtly affect the derivation.

Symptomatic progression is not a descriptor for progression of NTLs: it is a reason for stopping study therapy and will not be included in any assessment of NTLs.

Subjects with ‘symptomatic progression’ requiring discontinuation of treatment without objective evidence of disease progression at that time should continue to undergo tumor assessments where possible until objective disease progression is observed.

3.2.3 Overall visit response – site investigator data

Table 8 defines how the previously defined TL and NTL visit responses will be combined with new lesion information to give an overall visit response.

Table 8 Overall Visit Responses

Target Lesions	Non-target lesions	New Lesions	Overall Response
CR	CR or NA	No (or NE)	CR
CR	Non-CR/Non-PD or NE	No (or NE)	PR
PR	Non-PD or NE or NA	No (or NE)	PR
SD	Non-PD or NE or NA	No (or NE)	SD
PD	Any	Any	PD
Any	PD	Any	PD
Any	Any	Yes	PD
NE	Non-PD or NE or NA	No (or NE)	NE
NA	CR	No (or NE)	CR
NA	Non-CR/Non-PD	No (or NE)	SD

Table 8 Overall Visit Responses

Target Lesions	Non-target lesions	New Lesions	Overall Response
NA	NE	No (or NE)	NE
NA	NA	No (or NE)	NED

CR Complete response, PR Partial response, SD Stable disease, PD Progression of disease, NE Not evaluable, NA Not applicable (only relevant if there were no TL/NTL at baseline).

3.2.4 Blinded Independent Central Review (BICR)

A BICR of radiological scans will be performed for the subjects for first interim analysis, i.e. when approximately 100 subjects per treatment arm have had the opportunity for at least 32 weeks of follow-up and not prior to the last subject enrolled. Only the subset of subjects with the opportunity for at least 32 weeks of follow-up at the time of the DCO will be included in the BICR analysis. The imaging scans will be reviewed by 2 primary radiologist reviewers using RECIST 1.1. If the overall timepoint assessments differ at any timepoint between the 2 primary reviewers, the case will be adjudicated by a third radiologist who must choose all the overall timepoint assessments from the primary reviewer with which they more agree. If the overall timepoint assessments are identical between the 2 primary reviewers, the timepoint responses from the reviewer who completed their assessment of baseline scans first will be used for our analyses. For each subject, the BICR will define the overall visit response data (CR, PR, SD, PD, NED (only relevant for subjects with no disease identified at baseline), or NE) and the relevant scan dates for each time point (i.e., for visits where response or progression is/is not identified).

At IA1 BICR data will be used to analyse ORR and DoR.

3.2.5 Investigator RECIST 1.1-based secondary and CCI endpoints

Analysis of the secondary endpoints PFS, TTP, ORR, BoR, DCR, DCR-16w, DCR-24w, DoR CCI will be based on the Investigator assessments using RECIST 1.1.

All RECIST 1.1 assessments, whether scheduled or unscheduled, will be included in the calculations. This is also regardless of whether a subject discontinues study drug(s) or receives another anticancer therapy.

At each visit, subjects will be programmatically assigned a RECIST 1.1 visit response of CR, PR, SD, or PD depending on the status of their disease compared with baseline and previous assessments. Baseline will be assessed within the 28 days prior to randomization. If a subject has had a tumor assessment that cannot be evaluated, then the subject will be assigned a visit response of not evaluable (NE; unless there is evidence of progression, in which case the response will be assigned as PD).

3.3 Secondary variables

3.3.1 Progression Free Survival (PFS)

PFS (per RECIST 1.1 using Investigator assessments) will be defined as the time from the date of randomization until the date of objective disease progression or death (by any cause in the absence of progression) regardless of whether the subject withdraws from therapy or receives another anticancer therapy prior to progression (i.e. date of PFS event or censoring – date of randomization + 1). Subjects who have not progressed or died at the time of analysis will be censored at the time of the latest date of assessment from their last evaluable RECIST 1.1 assessment. However, if the subject progresses or dies after 2 or more missed visits, the subject will be censored at the time of the latest evaluable RECIST 1.1 assessment prior to the 2 missed visits (Note: NE visit is not considered as missed visit).

Given the scheduled visit assessment scheme (i.e. eight-weekly for the first 48 weeks then twelve-weekly thereafter) the definition of 2 missed visits will change. If the previous RECIST assessment is less than study day 274 (i.e. week 39) then two missing visits will equate to 18 weeks since the previous RECIST assessment, allowing for early and late visits (i.e. $2 \times 8 \text{ weeks} + 1 \text{ week for an early assessment} + 1 \text{ week for a late assessment} = 18 \text{ weeks}$). If the two missed visits occur over the period when the scheduled frequency of RECIST assessments changes from eight-weekly to twelve-weekly this will equate to 22 weeks (i.e. take the average of 8 and 12 weeks which gives 10 weeks and then apply same rationale, hence $2 \times 10 \text{ weeks} + 1 \text{ week for an early assessment} + 1 \text{ week for a late assessment} = 22 \text{ weeks}$). The time period for the previous RECIST assessment will be from study days 274 to 344 (i.e. week 39 to week 49). From week 49 onwards (when the scheduling changes to twelve-weekly assessments), two missing visits will equate to 26 weeks (i.e. $2 \times 12 \text{ weeks} + 1 \text{ week for an early assessment} + 1 \text{ week for a late assessment} = 26 \text{ weeks}$).

If the subject has no evaluable visits or does not have baseline data, they will be censored at randomization date unless they die within two visits of baseline (16 weeks plus 1 week allowing for a late assessment within the visit window), then they will be treated as an event with date of death as the event date.

The PFS time will always be derived based on scan/assessment dates and not visit dates.

Table 9 Censoring rules for PFS

Assessment	Outcome	Date of Progression or Censoring
No baseline assessments or no evaluable response visits (excluding deaths within 2 visits of baseline)	Censored	Randomization date
No baseline or evaluable tumor assessments and death within 2 visits of baseline	Progressed	Date of death
Progression documented between scheduled visits	Progressed	Date of assessment of progression

Assessment	Outcome	Date of Progression or Censoring
No progression (or death) at time of analysis	Censored	Date of last evaluable tumor assessment
Death between assessment visits	Progressed	Date of death
Death or progression after 2 or more missed visits	Censored	Date of last evaluable tumor assessment prior to the 2 missed visits

PFS Progression-free survival.

RECIST 1.1 assessments/scans contributing toward a particular visit may be performed on different dates. The following rules will be applied:

- For Investigator assessments, the date of progression will be determined based on the earliest of the RECIST 1.1 assessment/scan dates of the component that indicates progression.
- When censoring a subject for PFS, the subject will be censored at the latest of the scan dates contributing to a particular overall visit assessment.

Note: for TLs only the latest scan date is recorded out of all scans performed at that assessment for the TLs and similarly for NTLs only the latest scan date is recorded out of all scans performed at that assessment for the NTLs.

3.3.2 Time to progression (TTP)

TTP (per RECIST 1.1 using Investigator assessment) will be defined as the time from randomization until objective tumor progression in the absence of death. TTP is defined as per PFS however if subjects died without tumor progression, they will be censored at the time of death.

3.3.3 Objective response rate (ORR)

ORR (per RECIST 1.1 as assessed by the Investigator, or per mRECIST and RECIST 1.1 by BICR) will be defined as the percentage of subjects with at least one unconfirmed visit response of CR or PR. Data obtained up until progression, or the last evaluable assessment in the absence of progression, will be included in the assessment of ORR. Subjects who go off treatment without progression, receive a subsequent therapy, and then respond will not be included as responders in the ORR.

ORR based on at least one confirmed response will also be derived and reported in CSR. A confirmed response of CR/PR means that a response of CR/PR is recorded at 1 visit and confirmed by repeat imaging not less than 4 weeks after the visit when the response was first

observed with no evidence of progression between the initial and CR/PR confirmation visit. In the case where a subject has two non-consecutive visit responses of PR, then, as long as the time between the 2 visits of PR is greater than 4 weeks and there is no PD between the PR visits, the subject will be defined as a responder. Similarly, if a subject has visit responses of CR, NE, CR, then, as long as the time between the 2 visits of CR is greater than 4 weeks, then a best response of CR will be assigned.

For IA1 ORR (for both confirmed and unconfirmed responses) will be calculated in FAS-32w, both according to Investigator and BICR assessment per RECIST 1.1 and BICR assessment per mRECIST.

For IA2 and FA, ORR (for both confirmed and unconfirmed responses) will be calculated for the FAS according to Investigator assessments (per RECIST1.1).

ORR subgroup analyses will also be conducted for IA1, IA2, and FA, using the subgroups as identified in Section 4.2.6.2. For IA1, the subgroup analysis will use the BICR data for FAS-32w.

3.3.4 Best objective response (BoR)

Best objective response (BoR) is calculated based on the overall visit responses from each RECIST assessment, described in Section 3.2.3. It is the best response a subject has had following randomization, but prior to starting any subsequent cancer therapy and up to and including RECIST 1.1 progression or the last evaluable assessment in the absence of RECIST 1.1 progression. Categorisation of BoR will be based on RECIST using the following response categories: CR, PR, SD, NED (applies only to those subjects entering the study with no disease at baseline), PD and NE.

CR or PR must be confirmed. For determination of a best response of SD, the earliest of the dates contributing towards a particular overall visit assessment will be used. SD should be recorded at least 8 weeks minus 1 week (to allow for an early assessment within the assessment window), after randomization. For CR/PR, the initial overall visit assessment that showed a response will use the latest of the dates contributing towards a particular overall visit assessment.

BoR will be determined programmatically based on RECIST 1.1 from the overall visit response using all site Investigator data up until the first progression event. The denominator will be consistent with those used in the ORR analysis.

For subjects, whose progression event is death, BoR will be calculated based upon all evaluable RECIST 1.1 assessments prior to death.

For subjects who die with no evaluable RECIST 1.1 assessments, if the death occurs ≤ 9 weeks (i.e. 8 weeks + 1 week to allow for a late assessment within the assessment window) after randomization, then BoR will be assigned to the progression (PD) category. For subjects who

die with no evaluable RECIST assessments, if the death occurs >9 weeks after randomization then BoR will be assigned to the NE category.

A subject will be classified as a responder if the RECIST 1.1 for a CR or PR are satisfied at any time following randomization, prior to RECIST 1.1 progression and prior to starting any subsequent cancer therapy.

Subjects who achieve CR or PR as determined by RECIST 1.1 will be included for the analysis of duration of response (DoR).

BoR will be derived similarly using RECIST 1.1 and mRECIST by BICR at IA1. Comparison of BoR by Investigator assessment and BoR by BICR assessment will be summarised by treatment group in subjects in the FAS-32w.

3.3.5 Disease control rate (DCR)

Disease control rate (DCR), per RECIST 1.1 using Investigator assessment, will be defined as the proportion of subjects with a Best Objective Response (BoR) of CR, PR, or SD.

DCR-16w, per RECIST 1.1 using Investigator assessment, is defined as the percentage of subjects who have a best objective response of CR or PR or who have SD for at least 16 weeks (+/-7 days), following the start of study treatment.

DCR-24w, per RECIST 1.1 using Investigator assessment, is defined as the percentage of subjects who have a best objective response of CR or PR or who have SD for at least 24 weeks (+/-7days), following the start of treatment.

3.3.6 Duration of response (DoR)

DoR (per RECIST 1.1 using Investigator assessment, or per RECIST 1.1 and mRECIST by BICR) will be defined as the time from the date of first documented response until the first date of documented progression or death in the absence of disease progression (i.e., date of PFS event or censoring – date of first response +1). It will be calculated in days and analysed in months. The end of response should coincide with the date of progression or death from any cause used for the RECIST 1.1 PFS endpoint. The time of the initial response will be defined as the latest of the dates contributing towards the first visit response of PR or CR.

If a subject does not progress following a response, then their DoR will use the PFS censoring time.

For IA1 DoR will be calculated in the FAS-32w, both according to Investigator and BICR assessment, per RECIST 1.1 criteria.

For IA2 and FA, DoR will be calculated for the FAS according to Investigator assessments (per RECIST1.1), for responses with or without confirmation.

3.3.7 Proportion of subjects alive at 12, 18, 24 and 36 months after randomization (OS12, OS18, OS24 and OS36)

The proportion of subjects alive at 12 months (OS12), 18 months (OS18), 24 months (OS24), and 36 months (OS36) following randomization will be defined as the Kaplan-Meier estimate of OS at 12 months, 18 months, 24 months, and 36 months after randomization.

3.3.8 Time to Response (TTR)

Time to response (per RECIST 1.1 as assessed by the site Investigator) is defined as the time from the date of randomization until the date of first documented response (i.e. date of response – date of randomization + 1). The date of first documented response should coincide with that used for the RECIST 1.1 DoR endpoint.

Time to response will not be defined for those subjects who do not have a documented response.

TTR will only be calculated to support the Payer Analysis. It will not be reported in CSR.

3.3.9 Time from Randomization to First Subsequent Therapy or Death (TFST)

TFST is defined as the time from randomization to the earlier of first subsequent cancer therapy start date following study treatment discontinuation, or death (i.e. date of first subsequent cancer therapy/death or censoring – date of randomization + 1). Any subject not known to have died at the time of the analysis and not known to have had a further intervention of this type will be censored at the last known time to have not received subsequent cancer therapy, i.e. the last follow-up visit where this was confirmed. If this is not available (e.g. if the subject has not yet attended a survival follow up visit), then subjects who are ongoing on study treatment at DCO will be censored at DCO, and subjects who discontinued study treatment before DCO will be censored at date of discontinuation. Subjects who were randomized but did not receive any study treatment would have TFST calculated in the same way, i.e. time from date of randomization to the earliest of first subsequent therapy or death; however, if no information is available regarding first subsequent therapy or death, the subject will be censored at randomization.

TFST will only be calculated to support the Payer Analysis. It will not be reported in CSR.

3.4

CCI

CCI

3.4.1

CCI

CCI

ORR, BoR and DoR using mRECIST by BICR will be analyzed as secondary endpoints.

3.4.2

CCI

CCI

3.4.3

CCI

CCI

CCI

3.5 Patient reported outcome (PRO) variables

All items/questionnaires will be scored according to published scoring guidelines. All PRO analyses will be based on FAS.

Compliance rates summarizing questionnaire completion at each visit will be tabulated.

3.5.1 EORTC QLQ-C30

The EORTC QLQ-C30 consists of 30 questions that can be combined to produce 5 functional scales (physical, role, cognitive, emotional, and social), 3 multi-item symptom scales (fatigue, pain, and nausea/vomiting), 6 single-item symptom scales and global health status/QoL scale. The EORTC QLQ-C30 will be scored according to the EORTC QLQ-C30 Scoring Manual (Fayers et al 2001). An outcome variable consisting of a score from 0 to 100 will be derived for each of the symptom scales, each of the functional scales, and the global measure of health status scale in the EORTC QLQ-C30 according to the EORTC QLQ-C30 Scoring Manual. Higher scores on the global measure of health status and functional scales indicate better health status/function, but higher scores on symptom scales represent greater symptom severity. For each subscale, if <50% of the subscale items are missing, then the subscale score will be divided by the number of non-missing items and multiplied by the total number of items on the subscales (Fayers et al 2001). If at least 50% of the items are missing, then that subscale will be treated as missing. Missing single items are treated as missing. The reason for any missing questionnaire will be identified and recorded.

Definition of clinically meaningful changes - Visit Response and Best Overall Response

Definition of clinically meaningful changes in score compared with baseline will be evaluated. A clinically meaningful change is defined as an absolute change in the score from baseline of ≥ 10 for scales from the EORTC QLQ-C30 (Table 10). For example, a clinically meaningful improvement in physical function (as assessed by EORTC QLQ-C30) is defined as an increase in the score from baseline of ≥ 10 , whereas a clinically meaningful deterioration is defined as a decrease in the score from baseline of ≥ 10 . At each postbaseline assessment, the change in global health status/QoL, symptoms, and functioning score from baseline will be categorized as improvement, no change, or deterioration as shown in Table 10.

Table 10 Mean change and clinically meaningful change - EORTC QLQ-C30

Score	Change from baseline	Visit response
	$\geq +10$ (increase of at least 10)	Improvement

Score	Change from baseline	Visit response
EORTC QLQ-C30 global health status score	≥ -10 (decrease of at least 10) or “Subject too sick to complete the questionnaires (disease under investigation)”	Deterioration
	Otherwise	No change
EORTC QLQ-C30 symptom scales	$\geq +10$ (increase of at least 10) or “Subject too sick to complete the questionnaires (disease under investigation)”	Deterioration
	≥ -10 (decrease of at least 10)	Improvement
	Otherwise	No change
EORTC QLQ-C30 functional scales	$\geq +10$ (increase of at least 10)	Improvement
	≥ -10 (decrease of at least 10) or “Subject too sick to complete the questionnaires (disease under investigation)”	Deterioration
	Otherwise	No change

EORTC European Organisation for Research and Treatment of Cancer; QLQ-C30 30-item core quality of life questionnaire.

A subject’s best overall response in symptoms, function, or global health status/QoL will be derived as the best response the subject achieved, based on evaluable PRO data collected during the study period. The criteria in Table 11 will be used to assign a best response in symptoms or function or global health status/QoL.

Table 11 Best response in EORTC QLQ-C30 and EORTC QLQ-HCC18 scores: FAS

Overall response score	Criteria
Missing	Subject has no evaluable baseline or post-baseline PRO assessment.
Improved	Subject meets one of the following criteria: - Has 2 consecutive visit responses of “improvement” at least 21 days apart. - Has 1 visit response of “improvement” and no further assessments, and did not die within 2 PRO assessment visits.
No Change	Subject does not qualify for an overall score response of “improved” and meets 1 of the following criteria: - Has 2 consecutive visit responses of “no change” at least 21 days apart. - Has 1 visit response of “no change” with no further assessments, and did not die within 2 PRO assessment visits.

Overall response score	Criteria
Deterioration	Subject does not qualify for an overall score response of “improved” or “no change” and meets 1 of the following criteria: 1. Has 2 consecutive visit responses of “deterioration” at least 21 days apart. 2. Has 1 visit response of “deterioration” and no further assessments. 3. Has 1 visit response of “improvement” or “no change” followed by death within 2 PRO assessment visits.
Other	Does not qualify for one of the above (improved, no change or deterioration).

EORTC European Organisation for Research and Treatment of Cancer; FAS full analysis set; QLQ C30 30-item core quality of life questionnaire; QLQ-HCC18 18-item HCC specific quality of life questionnaire.

Visit responses are not considered consecutive if, according to the visit schedule, there are one or more missed (a scheduled visits record missing completely, or a missing assessment) assessments between the two evaluable assessments.

3.5.1.1 Time to global health status/QoL, function or symptoms deterioration

Time to deterioration in global health status/QoL, function or symptoms will be defined as the time from the date of randomization until the date of the first clinically meaningful deterioration that is confirmed at a subsequent visit (except if it was the subject’s last available assessment) or death (by any cause) in the absence of a clinically meaningful deterioration, regardless of whether the subject discontinues study drug(s) or receives another anticancer therapy prior to global health status/QoL, function or symptoms deterioration. Death will be included as an event only if it occurs within 2 PRO assessment visits from the last available PRO assessment.

Subjects whose global health status/QoL, function or symptoms (as measured by EORTC QLQ-C30) has not shown a clinically meaningful deterioration and who are alive at the time of the analysis will be censored at the time of their last PRO assessment where the global health status/QoL, function or symptoms could be evaluated. Subjects with no post-baseline assessment will be censored at date of randomization. Also, if global health status/QoL, function or symptoms deteriorates, or the subject dies after 2 or more missed PRO assessment visits, the subject will be censored at the time of the last PRO assessment where global health status/QoL, function or symptoms could be evaluated prior to the 2 missed visits.

The set for the analysis of time to deterioration will include a subset of the FAS who have baseline scores of ≥ 10 . The set for the analysis of time to symptom deterioration will consist of a subset of the FAS subjects who have a baseline symptom score ≤ 90 .

A death within 2 PRO assessment visits, or after 2 or more missed PRO assessment visits will be identified by comparing the visit assigned to the study day of death (relative to first dose date) using visits windows defined in Section 4.1.2 with the assigned PRO assessment visits.

The following time to deterioration analyses should be produced:

1. Time to global health status/QoL deterioration
2. Time to function deterioration:
 - a. Time to physical deterioration
 - b. Time to role deterioration
 - c. Time to cognitive deterioration
 - d. Time to emotional deterioration
 - e. Time to social deterioration
3. Time to symptoms deterioration:
 - a. Time to fatigue deterioration
 - b. Time to pain deterioration
 - c. Time to nausea/vomiting deterioration
 - d. Time to nausea deterioration (single item #14)
 - e. Time to dyspnoea deterioration
 - f. Time to insomnia deterioration
 - g. Time to appetite loss deterioration
 - h. Time to constipation deterioration
 - i. Time to diarrhoea deterioration

3.5.1.2 Symptom improvement rate

Responses in symptoms for each visit (improvement, deterioration, and no change based on Table 10) as well as the best overall response will be presented by treatment arm. The symptom improvement rate will be defined as the number (%) of subjects with a best overall score response of “improved” in symptoms.

The denominator will consist of a subset of the FAS subjects who have a baseline symptom score ≥ 10 .

The following symptom improvement rate analysis should be produced:

- a) Fatigue improvement rate
- b) Pain improvement rate
- c) Nausea/vomiting improvement rate
- d) Dyspnoea improvement rate
- e) Insomnia improvement rate
- f) Appetite loss improvement rate
- g) Constipation improvement rate
- h) Diarrhoea improvement rate

3.5.1.3 Global health status /QoL or function improvement rate

The global health status/QoL or function improvement rate will be defined as the number (%) of subjects with best overall response of “improved” in global health status/QoL or function.

The denominator will consist of a subset of the FAS subjects who have a baseline global health status/QoL or function score ≤ 90 .

The following function improvement rate analysis should be produced:

- a. Physical improvement rate
- b. Role improvement rate
- c. Cognitive improvement rate
- d. Emotional improvement rate
- e. Social improvement rate

3.5.2 EORTC QLQ-HCC18

The EORTC QLQ-HCC18 is a hepatocellular cancer-specific module from the EORTC comprising 18 questions to assess HCC symptoms. The module includes 6 multi-item domain scales and 2 single-item scales. For all items and scales, high scores indicate increased symptomatology/more problems.

The scoring approach for the QLQ-HCC18 is identical in principle to that for the symptom scales/single items of the EORTC QLQ-C30. Similar to the symptom scales of the EORTC QLQ-C30, higher scores represent greater symptom severity.

Definition of clinically meaningful changes – visit response and best overall response

Changes in score compared with baseline will be evaluated. A clinically meaningful change is defined as an absolute change in the score from baseline of ≥ 10 for scales/items from QLQ-HCC18. For example, a clinically meaningful deterioration or worsening in pain (as assessed by QLQ-HCC18) is defined as an increase in the score from baseline of ≥ 10 . At each postbaseline assessment, the change in symptoms score from baseline will be categorized as improved, no change, or deterioration, as shown in Table 12. A subject’s best overall response in symptoms will be derived as the best response the subject achieved, based on evaluable PRO data collected during the study period. The criteria in Table 12 will be used to assign a best response in symptom score.

Table 12 Visit response for HRQoL and disease-related symptoms

Score	Change from baseline	Visit response
QLQ-HCC18 symptom scales and items	$\geq +10$ (increase of at least 10) or “Subject too sick to complete the questionnaires (disease under investigation)”	Deterioration

Score	Change from baseline	Visit response
	≥ -10 (decrease of at least 10)	Improved
	Otherwise	No change

HRQoL Health-related quality of life; QLQ-HCC18 18-item HCC quality of life questionnaire.

3.5.2.1 Time to symptom deterioration

For each of the symptom scales/items in the QLQ-HCC18, time to symptom deterioration will be defined as the time from randomization until the date of the first clinically meaningful symptom deterioration that is confirmed at a subsequent visit (except if it was the subject's last available assessment) or death (by any cause) in the absence of a clinically meaningful symptom deterioration, regardless of whether the subject discontinues study drug(s) or receives another anticancer therapy prior to symptom deterioration. Only deaths occurring within 2 PRO assessment visits from the last available PRO assessment will be included as events.

Subjects whose symptoms (as measured by the QLQ-HCC18) have not shown a clinically meaningful deterioration and who are alive at the time of the analysis will be censored at the time of their last PRO assessment where the symptom could be evaluated. Subjects with no post-baseline assessment will be censored at date of randomization. Also, if symptoms progress or the subject dies after 2 or more missed PRO assessment visits, the subject will be censored at the time of the last PRO assessment where the symptom could be evaluated prior to the 2 missed visits.

The set for the analysis of time to symptom deterioration will include a subset of the FAS who have baseline scores ≤ 90 .

A death within 2 PRO assessment visits, or after 2 or more missed PRO assessment visits will be identified by comparing the visit assigned to the study day of death (relative to the first dose date) using visits windows defined in Section 4.1.2 with the assigned PRO assessment visits.

Time to deterioration analyses should be produced for the following EORTC QLQ-HCC18 subscales and items:

1. Muscle loss (single item #33)
2. Abdominal swelling (single item #34)
3. Shoulder pain (single item #38)
4. Abdominal pain (single item #39)
5. Early satiety (full up too quickly) (single item #43)
6. Weight loss (single item #44)
7. Jaundice
8. Pain
9. Nutrition
10. Fatigue
11. Fever

3.5.2.2 Symptom improvement rate

Responses in symptoms for each visit (improvement, deterioration, and no change based on Table 10) as well as the best overall response will be presented by treatment arm. The symptom improvement rate will be defined as the number (%) of subjects with a best overall score response of “improved” in symptoms.

The denominator will consist of a subset of the FAS subjects who have a baseline symptom score ≥ 10 .

The following symptom improvement rate analysis should be produced:

1. Muscle loss (single item #33)
2. Abdominal swelling (single item #34)
3. Shoulder pain (single item #38)
4. Abdominal pain (single item #39)
5. Early satiety (full up too quickly) (single item #43)
6. Weight loss (single item #44)
7. Jaundice
8. Pain
9. Nutrition
10. Fatigue
11. Fever

3.5.3

CCI

CCI

3.5.4

CCI

CCI

3.5.5

CCI

CCI

3.5.6 Compliance

Summary measures of overall compliance and compliance over time will be derived for each PRO, respectively. These will be based upon:

- Received questionnaire = a questionnaire that has been received and has a completion date and at least one individual item completed.
- Expected questionnaire = a questionnaire that is expected to be completed at a scheduled assessment time e.g. a questionnaire from a subject who has not withdrawn from the study at the scheduled assessment time but excluding subjects in countries with no available translation as well as subjects who are unable to read the questionnaire (eg, subject is blind or illiterate). For subjects that have progressed, the latest of progression and safety follow-up will be used to assess whether the subject is still under PRO follow-up at the specified assessment time. Date of study discontinuation will be mapped to the nearest visit date to define the number of expected forms.
- Evaluable questionnaire = a questionnaire with a completion date and at least one subscale that is non-missing.
- Overall PRO compliance rate is defined as: Total number of evaluable questionnaires across all time points, divided by total number of questionnaires expected to be received across all time points multiplied by 100.
- Overall subject compliance rate is defined for each randomized treatment group as: Total number of subjects with an evaluable baseline and at least one evaluable follow-up questionnaire (as defined above), divided by the total number of subjects expected to have completed at least a baseline questionnaire multiplied by 100.

Compliance over time will be calculated separately for each visit, including baseline, as the number of subjects with an evaluable questionnaire at the time point (as defined above), divided by number of subjects still expected to complete questionnaires. Similarly, the evaluability rate over time will be calculated separately for each visit, including baseline, as the number of evaluable questionnaires (per definition above), divided by the number of received questionnaires.

3.6

CCI

CCI

CCI

3.7 Safety Variables

Safety and tolerability will be assessed in terms of AEs (including SAEs), deaths, laboratory data, vital signs, ECGs, physical examination, exposure, dose intensity, ECOG performance status and Child-Pugh score. These will be collected for all subjects. Data from all cycles of treatment will be combined in the presentation of safety data.

“On treatment” will be defined as assessments between date of start dose and 90 days following the date of the last dose of study drug(s) (i.e., the last dose of durvalumab, tremelimumab, or sorafenib) (including re-treatment), or up to the date of initiation of the first subsequent therapy (whichever occurs first). This definition applies to all safety reporting, unless otherwise specified.

For AEs, on treatment (or treatment-emergent AEs) will be defined as any AEs with an onset date on or after the date of first dose or pre-treatment AEs that increase in severity on or after the date of first dose and up to and including 90 days following the date of last dose of study drug(s) (i.e., the last dose of durvalumab, tremelimumab, or sorafenib) (including re-treatment), or up to the date of initiation of the first subsequent therapy (whichever occurs first).

3.7.1 General considerations

Safety and tolerability data will be presented by treatment arm using the safety analysis set.

All summary statistics by visit tables (actual and change from baseline) should use windows defined in Section 4.1.2.

Missing safety data will generally not be imputed. However, safety assessment values of the form of “< x” (i.e. below the lower limit of quantification) or “> x” (i.e. above the upper limit of quantification) will be imputed as “x” in the calculation of summary statistics but displayed as “< x” or “> x” in the listings. Note that 0 should not be used as an imputed value in case the endpoint requires a log transformation. Additionally, adverse events that have missing causality (after data querying) will be assumed to be related to study drug.

The denominator used in laboratory summaries will only include evaluable subjects, i.e., those who had sufficient data to have the possibility of an abnormality.

For example:

- If a CTCAE criterion involves a change from baseline, evaluable subjects would have both a predose and at least 1 postdose value recorded.
- If a CTCAE criterion does not consider changes from baseline to be evaluable, the subject need only have 1 postdose value recorded.

For handling missing and incomplete dates the following rules should be followed:

- For missing diagnostic dates, if day and/or month are missing use 01 and/or Jan. If year is missing, put the complete date to missing.
- For missing start AE/medication dates, the following will be applied:
 - Missing day - Impute the 1st of the month unless month is same as month of first dose of study drug then impute first dose date.
 - Missing day and month – impute 1st January unless year is the same as first dose date then impute first dose date.
 - Completely missing – impute first dose date unless the end date suggests it could have started prior to this in which case impute the 1st January of the same year as the end date.
 - When imputing a start date ensure that the new imputed date is sensible i.e. is prior to the end date of the AE or med.
- For missing end AE/medication dates, the following will be applied:
 - Missing day - Impute the last day of the month unless month is the same as month of the first dose of study drug then impute last dose date.
 - Missing day and month – impute 31st December unless year is the same as first dose date then impute last dose date.

For completely missing AE stop dates, if the subject has died and the AE stop date is missing, then the stop date of the AE will be imputed as the death date. End dates will not be imputed for concomitant medications with a start date after the last dose date.

For immune-mediated adverse event summaries, an AE with outcome of unknown will be imputed as not resolved.

3.7.2 Adverse events

Definitions of Adverse Events (AEs) and serious adverse events (SAEs) can be found in the study protocol Section 6.1 and 6.2. The Medical Dictionary for Regulatory Activities (MedDRA) (using the latest or current MedDRA version) will be used to code the AEs. AEs will be graded according to the National Cancer Institute Common Terminology Criteria for AEs (using the CTCAE version referenced in the Clinical Study Protocol).

TEAEs will be used for summary tables. A separate data listing of AEs occurring more than 90 days after discontinuation of study drug(s) will be produced. These events will not be included in AE summaries.

3.7.2.1 AEs of special interest and AEs of possible interest

An AstraZeneca medically qualified expert, after consultation with the Global Patient Safety Physician, has reviewed the AEs of special interest and AEs of possible interest and identified which preferred terms contribute to each adverse events of special interest (AESI) and adverse events of possible interest (AEPI), the list can be found in Section 6.5 in the CSP. A further review will take place prior to database lock to ensure new terms not already included in the older MedDRA version are captured within the categories for the new higher MedDRA version. The list will be provided by AZ prior to database lock.

AEs of special interest will be summarized by treatment group.

3.7.2.2 Other significant adverse events

During the evaluation of the AE data, an AstraZeneca medically qualified expert will review the list of AEs that were not reported as SAEs and AEs leading to discontinuation. Based on the expert's judgment, significant AEs of particular clinical importance may, after consultation with the Global Patient Safety Physician, be considered other significant AEs (OAEs) and reported as such in the CSR, for example other significant AEs based on a subset of PTs from the Hepatic SMQ and Haemorrhages SMQ. A similar review of laboratory, vital signs, and ECG data will be performed for identification of OAEs.

Examples of these are marked hematological and other laboratory abnormalities, and certain events that lead to intervention (other than those already classified as serious) or significant additional treatment.

3.7.3 Treatment exposure

Exposure to study drug(s), time on study, treatment durations, number of infusions/doses received, dose delays(all arms), infusion interruptions (Arms A, B, and C), treatment cycles received (Arms A, B, and C) and dose reductions (sorafenib CCI arm) will also be summarized.

Exposure will be defined as follows.

Total (or intended) exposure of study medication:

- Total (or intended) durvalumab exposure = $\min(\text{last durvalumab dose date where dose} > 0 + 27 \text{ days, date of death, date of DCO}) - \text{first dose date} + 1$
- Total (or intended) tremelimumab exposure = $\min(\text{last tremelimumab dose date where dose} > 0 + 27 \text{ days, date of death, date of DCO}) - \text{first dose date} + 1$
- Total (or intended) durvalumab and tremelimumab exposure = $\min(\max(\text{last durvalumab or tremelimumab dose date where dose} > 0) + 27 \text{ days, date of death, date of DCO}) - \text{first dose date} + 1$
- Total (or intended) sorafenib exposure = $\min(\text{last sorafenib dose date where dose} > 0, \text{date of death, date of DCO}) - \text{first dose date} + 1$

Actual exposure:

- Actual exposure = intended exposure – total duration of dose delays/interruptions, where intended exposure will be calculated as above, and a dose interruption is defined as any length of time where the subject has not taken any of the planned daily dose.

The actual exposure calculation makes no adjustment for any dose reductions that may have occurred.

Infusion interruptions (Arms A, B, and C)

For durvalumab and tremelimumab, a dose interruption is an infusion interruption that occurs during the infusion. To count as an infusion interruption, the total dose received must be >0 . The drug can be restarted after the interruption and so it is possible for an infusion interruption to occur and the whole dose to be administered. If the same infusion was interrupted multiple times, then this would just be captured as one infusion interruption. For Arms B and C, the number of subjects with infusion interruption(s) of durvalumab and tremelimumab will be summarized separately, along with the number of subjects with infusion interruption(s) of either drug.

Number of infusions and doses

Number of infusions and number of cycles of durvalumab and tremelimumab:

Exposure will also be summarized by the number of infusions received. Cycles of treatment with durvalumab or tremelimumab are of 28 days duration with a **CCI** of each cycle. If a cycle is prolonged due to toxicity, this should be counted as one cycle. A cycle will be counted if treatment is started even if the full dose is not delivered.

Number of doses of sorafenib:

Exposure will also be summarized by the number of doses received. Sorfaenib is administered CCI The number of doses received will be determined the number of days a dose was administered.

Dose delays

A treatment cycle is started when >0 dose of durvalumab or tremelimumab is administered. As such, a dose delay for durvalumab or tremelimumab occurs when the start of a cycle is started at a later date than planned.

Durvalumab and tremelimumab

Since subject will receive drug via IV infusions CCI until confirmed PD the duration of dose delays will be calculated as:

Total duration of dose delays= Sum of (Date of the dose start - Date of previous dose end – 28 days)

For Arm C, since dosing of tremelimumab only occurs on CCI dose delays will be summarized for Durvalumab only.

Sorafenib

Since subject will receive drug CCI until confirmed PD the duration of dose delays will be calculated as:

Total duration of dose delays= Sum of (Date of the dose start - Date of previous dose end – 1 day)

Dose reduction

For sorafenib, a dose reduction is counted once for each time the dose is reduced.

Time on Study

Time on study should be defined for all treatment arms as follows:

time on study = (death date or data cut off or date of study withdrawal, whichever occurs earlier - randomization date + 1) / (365.25/12). Total exposure, actual exposure and time on study expressed in months will be summarized in tables. The duration in months will be calculated as follows:

- Duration in days / (365.25/12)
Exposure should be calculated separately for the following three segments: initial phase, re-challenge phase, total study. For Durvalumab monotherapy and Sorafenib arms the initial phase and total study will be the same. For combination therapy arms, the initial phase includes data before re-challenge, and the re-challenge phase includes

data from re-challenge onwards. The total study period includes all data from initial and re-challenge phase.

3.7.4 Dose intensity

Relative dose intensity (RDI) is the percentage of the actual dose delivered relative to the intended dose through to treatment discontinuation. It should be calculated for each study drug separately.

RDI will be defined as follows:

- $RDI = 100\% * d/D$, where d is the actual cumulative dose delivered up the actual last day of dosing and D is the intended cumulative dose (mg) up to the actual last day of dosing. D is the total dose (mg) that would be delivered, if there were no modification to dose or schedule.

Intended cumulative dose will be calculated as follows:

- For durvalumab/tremelimumab: number of cycles received * intended dose per cycle;
- For sorafenib: (min(last Sorafenib dose date where dose>0, date of death, date of DCO) – first dose date +1) * intended daily dose.

Intended dose during treatment (excluding rechallenge period) will be assigned as follows:

- Arm A: CCI Durvalumab (if a subject's weight decreases to ≤ 30 kg, the subject should receive weight-based dosing of durvalumab CCI)
- Arm B: CCI Durvalumab, CCI Tremelimumab (if a subject's weight decreases to 30 kg or below (≤ 30 kg), the subject should receive weight-based dosing of durvalumab CCI and tremelimumab CCI)
- Arm C: CCI Durvalumab, CCI Tremelimumab (if a subject's weight decreases to 30 kg or below (≤ 30 kg), the subject should receive weight-based dosing of durvalumab CCI and tremelimumab CCI)
- Arm D: CCI Sorafenib.

Intended dose during rechallenge period will be assigned as follows:

- Arm A: CCI Durvalumab (if a subject's weight decreases to ≤ 30 kg, the subject should receive weight-based dosing of durvalumab CCI)
- Arm B: CCI Durvalumab, subject can be rechallenged with either Tremelimumab CCI so intended Tremelimumab dose will be CCI based on actual dosing (if a subject's weight decreases to 30 kg or below (≤ 30 kg), the subject should receive weight-based dosing of durvalumab CCI and tremelimumab CCI)
- Arm C: CCI Durvalumab, CCI Tremelimumab (if a subject's weight decreases to 30 kg or below (≤ 30 kg), the subject should receive weight-based dosing of durvalumab CCI and tremelimumab CCI)
- Arm D: CCI Sorafenib.

Details of weight-based dosing calculations are provided in CSP Appendix F and G.

3.7.5 Laboratory data

On-treatment laboratory data will be used for summaries.

Laboratory data will be collected throughout the study, from screening to the follow-up visits as described in Tables 2, 3, 4 and 5 of the CSP. Blood and urine samples for determination of haematology, clinical chemistry, and urinalysis will be collected as described in Section 5.2.1 of the CSP. For derivation of baseline and post baseline visit values considering visit window and how to handle multiple records, derivation rules as described in Section 4.1.2 will be used.

Change from baseline in haematology and clinical chemistry variables will be calculated for each post-dose visit on treatment. CTC grades will be defined at each visit according to the CTC grade criteria using local or project ranges as required, after conversion of lab result to corresponding SI units. The following parameters have CTC grades defined for both high and low values: Potassium, Sodium, Magnesium, Glucose and Corrected calcium, so high and low CTC grades will be calculated.

Corrected Calcium will be derived during creation of the reporting database using the following formulas:

$$\text{Corrected calcium (mmol/L)} = \text{Total calcium (mmol/L)} + ([40 - \text{Albumin (g/L)}] \times 0.02)$$

Absolute values will be compared to the project reference range and classified as low (below range), normal (within range or on limits of range) and high (above range).

The maximum or minimum on treatment value (depending on the direction of an adverse effect) will be defined for each laboratory parameter as the maximum (or minimum) post-dose value at any time.

For example:

- If a CTCAE criterion involves a change from baseline, evaluable subjects would have both a pre-dose and at least 1 post-dose value recorded.
- If a CTCAE criterion does not consider changes from baseline, to be evaluable the subject needs only to have 1 post dose-value recorded.

3.7.6 ECGs

On-treatment ECG data will be used for summaries.

For the derivation of post baseline visit values considering visit window and to handle multiple records present in any visit window, derivation rules as described in Section 4.1.2 will be used.

At each time point the Investigator's assessment of the ECG will be collected locally.

For triplicate ECGs, the mean of the three ECG assessments will be used to determine the value at that time point.

QTcF (QT interval corrected for using Fridericia's formula) will be derived during creation of the reporting database using the reported ECG values (RR and QT) using the following formula:

$QTcF = QT/RR^{(1/3)}$ where RR is in seconds

3.7.7 Vital signs

On-treatment vital signs data will be used for summaries. Change from baseline in vital signs variables will be calculated for each post-dose visit on treatment. For derivation of post baseline visit values considering visit window and to handle multiple records, derivation rules as described in Section 4.1.2 will be used.

3.7.8 ECOG performance status

Performance status as determined by the ECOG Scale will be recorded in the eCRF as per the schedules defined in CSP and will be summarized by visit.

3.7.9 Child-Pugh score

Cirrhosis severity, as determined by the Child-Pugh score (Pugh et al 1973), will be recorded in the eCRF as specified in the assessment schedules (see Table 2, Table 3, and Table 4 of CSP). The modified Child-Pugh classification of liver disease severity according to the degree of ascites, serum concentrations of bilirubin and albumin, prothrombin time, and degree of encephalopathy is shown in Table 9 of CSP. The severity of cirrhosis is classified as follows:

- Child-Pugh class A (well-compensated disease): score of 5 to 6
- Child-Pugh class B (significant functional compromise): score of 7 to 9
- Child-Pugh class C (decompensated disease): score of 10 to 15

Child-Pugh classification and the total score will be summarized by visit.

3.7.10 Physical examinations

Physical examinations will be performed according to the schedule of assessments (see Tables 2,3 and 4 of CSP). Full physical examinations will include assessments of the head, eyes, ears, nose, and throat and the respiratory, cardiovascular, GI, urogenital, musculoskeletal, neurological, dermatological, hematologic/lymphatic, and endocrine systems. Results of these measurements will not be collected in eCRF. Height will be measured at screening only.

3.7.11 Other safety assessments

If new or worsening pulmonary symptoms (e.g., dyspnea) or radiological abnormality suggestive of pneumonitis/ILD is observed, toxicity management will be applied and all related safety data will be listed.

3.7.12 Prior and concomitant medications

All allowed medications (other than study drugs) with non-zero dose should be classified as prior and/or concomitant medications. In case both medication start and stop dates are available and complete, prior and concomitant medications will be defined as follows:

- Prior medication – any allowed medication with medication stop date before first study drug intake.
- Concomitant medication – any allowed medication with medication start date on or after first study drug intake date, and on or before last study drug intake date.

For all incomplete medication start and stop dates, worst case scenarios will be applied, like below:

- If both medication start and stop date are missing, the medication will be counted as both prior and concomitant medication.
- In case a medication started prior to first drug intake and no medication stop date is available (treatment is ongoing or stop date is missing), it will be counted as both prior and concomitant medication.
- If medication start date is incomplete and it cannot be determined if the medication started before or after first study drug intake (e.g. only year is available in medication start date, and it is equal to the year of first study drug intake), the medication will be counted as both prior and concomitant medication.
- If medication start date is incomplete and it cannot be determined if the medication started before or after last study drug intake (e.g. only year is available in medication start date, and it is equal to the year of last study drug intake), the medication will be counted as concomitant medication.

An AstraZeneca medically qualified expert will review medications received by subjects during the study to identify disallowed medications, which will be summarized in a separate table. A separate summary table will be produced for concomitant medications which began prior to randomization.

3.8 Pharmacokinetic variables

3.8.1.1 Population pharmacokinetics and exposure-response/safety analysis

A population PK model will be developed using a non-linear, mixed-effects modelling approach. The impact of physiologically relevant subject characteristics (covariates) and disease on PK will be evaluated. The relationship between the PK exposure and the effect on safety and efficacy endpoints will be evaluated. The results of such an analysis will be reported in a separate report. CCI

[REDACTED]

[REDACTED]

[REDACTED]

3.8.1.2 Pharmacokinetic non-compartmental analysis

Serum concentration of durvalumab and tremelimumab will be listed and summarized by descriptive statistics. Individual and mean (SD) serum concentration-time profiles will be generated. Non-compartmental analysis will not be conducted due to sparse sampling scheme. Samples below the lower limit of quantification will be reported as NQ (Not Quantifiable) in the descriptive statistics (see details in section 4.2.8).

3.9 Immunogenicity analysis

Serum samples for ADA assessments will be conducted utilizing a tiered approach (screen, confirm, titer), and ADA data will be collected at scheduled visits shown in the CSP. ADA result from each sample will be reported as either positive or negative. If the sample is positive, the ADA titer will be reported as well. In addition, the presence of neutralizing antibody (nAb) will be tested for all ADA positive samples using a ligand-binding assay. The nAb results will be reported as positive or negative. A subject is defined as being ADA-positive if a positive ADA result is available at any time, including baseline and all post-baseline measurements; other ADA negative.

The number of ADA evaluable subjects in the following ADA categories in each of the treatment group will be determined:

- ADA positive at any visit, at baseline and/or post-baseline;
- ADA positive post-baseline and positive at baseline;
- ADA positive post-baseline and not detected at baseline (treatment-induced ADA);
- ADA not detected post-baseline and positive at baseline;
- Baseline ADA titer that was boosted by ≥ 4 -fold following drug administration (treatment-boosted ADA);
- Treatment-emergent ADA positive, defined as the sum of treatment-induced ADA and treatment-boosted ADA;
- Persistently positive ADA, defined as having at least 2 post-baseline ADA positive measurements with at least 16 weeks (112 days) between the first and last positive measurements, or an ADA positive result at the last available assessment;
- Transiently positive ADA, defined as having at least one post-baseline ADA positive measurement and not fulfilling the conditions for persistently positive;
- nAb positive at any visit (at baseline and/or post-baseline).

3.10

CCI

CCI

CCI

3.11 Biomarker variables

PL-L1 expression, as defined in the secondary objectives, will be assessed for evaluable subjects in each cohort according to prespecified criteria.

4. ANALYSIS METHODS

The formal statistical analysis of OS will be performed for the following efficacy test hypotheses (alternative hypotheses):

- H1: Difference between durvalumab CCI plus tremelimumab CCI (Arm C) and sorafenib CCI (Arm D)
- H2: Durvalumab CCI monotherapy (Arm A) not inferior to sorafenib CCI (Arm D) with noninferiority margin of 1.08
- H3: Difference between durvalumab CCI monotherapy (Arm A) and sorafenib CCI (Arm D)

Table 13 details which endpoints are to be analysed, together with pre-planned sensitivity analyses indicating which analysis is regarded as primary for that endpoint.

Table 13 Formal Statistical Analyses to be Conducted and Pre-planned Sensitivity Analyses

Endpoint	Analysis
Overall survival (OS)	Primary analysis: Stratified log-rank test (for p-value), HR from Cox model (with 95% CI) Sensitivity analyses: - Attrition bias. Kaplan-Meier plot of time-to-censoring where the censoring indicator of the primary analysis is reversed. - CCI - Impact of COVID19. OS analysis will be repeated but subjects who died from COVID-19 Infection will be censored at their COVID infection death date.
Progression Free Survival (PFS)	Primary analysis: Stratified log-rank test using Investigator assessments per RECIST 1.1 (for p-value), HR from Cox model (with 95% CI)
Time to progression (TTP)	Primary analysis: Stratified log-rank test using Investigator assessments per RECIST 1.1 (for p-value), HR from Cox model (with 95% CI)

Endpoint	Analysis
Objective response rate (ORR)	IA1: Exact confidence intervals; IA2 and FA: Logistic regression using Investigator assessments per RECIST 1.1 (odds ratio with 95% CI and p-value)
Best Objective Response (BoR)	Descriptive statistics
Duration of response (DoR)	Descriptive statistics including KM plot
Disease control rate (DCR, DCR-16w, DCR-24w)	Descriptive statistics
Proportion of subjects alive at 18m (OS18)	KM estimates of OS at 18 months
Proportion of subjects alive at 24m (OS24)	KM estimates of OS at 24 months
Proportion of subjects alive at 36m (OS36)	KM estimates of OS at 36 months Stratified chi-square test of difference in KM estimators at a fixed time point (36 months) (for p-value)
CCI	
Time to deterioration (EORTC QLQ-C30 and EORTC QLQ-HCC18)	Stratified log-rank test (for p-value), HR from Cox model (with 95% CI), KM plot
EORTC QLQ-C30, EORTC QLQ-HCC18	Average change from baseline using an MMRM analysis, Summary statistics
Improvement based best overall response (EORTC QLQ-C30, EORTC QLQ-HCC18)	Logistic regression with odds ratio, 95% CI and p-value
CCI	

EORTC European Organisation for Research and Treatment of Cancer; CCI
MMRM Mixed effect model repeat measurement; OS overall survival;
QLQ-C30 30-item core quality of life questionnaire; QLQ-HCC18 18-item hepatocellular cancer health-related quality of life questionnaire.

4.1 General principles

Descriptive statistics will be used for all variables, as appropriate, and will be presented by treatment arm. Continuous variables will be summarized by the number of observations, mean, standard deviation, median, minimum, and maximum. Categorical variables will be summarized by frequency counts and percentages for each category. Unless otherwise stated, percentages will be calculated out of the total for the corresponding treatment arm.

Efficacy data will be analysed on the basis of randomized study drug(s), regardless of the study drug(s) actually received. Safety data will be analysed based on the study drug(s) actually received.

Efficacy data for Arm B, which was closed for enrollment with protocol Amendment 4, will be summarized descriptively, however will not be formally analyzed. All other Arm B data will be summarized like for the other treatment arms.

All formal analysis will be limited to the three continuing arms of the study (Arm A, C, D). Arm B will be summarized for descriptive purposes in all efficacy and safety tables.

At each analysis timepoint only data until the corresponding DCO is included in analyses.

The following study data will be listed:

- Discontinued subjects
- Subjects with Important Protocol Deviations
- Subjects excluded from the safety analysis
- Demographic and baseline characteristics
- Administration of durvalumab
- Administration of tremelimumab
- Administration of sorafenib
- List of lesion assessments based on Investigator Response
- Efficacy endpoints
- Deaths
- Adverse Events with outcome of death
- Serious Adverse Events
- Adverse Events
- Individual laboratory assessments

Efficacy, PRO, pharmacogenetic and biomarker data will be summarized and analysed based on FAS (ITT). PK data will be summarized and analysed based on the PK analysis set. Safety data will be summarized on the safety analysis set. Listings of immunogenicity data will be based on the safety analysis set, and summaries will be based on ADA evaluable set.

Results of all statistical analysis will be presented using a 95% CI and 2-sided p-value, unless otherwise stated. Refer to Section 4.2.1 for example adjusted significance levels at IA2 and FA.

Where included in the analyses stratification variables will be taken from those reported via IWRS at randomization. For subgroup analyses, stratification factor values collected in eCRF will be used to define subgroups.

Time-to-event efficacy endpoints (OS, PFS, etc.) will be calculated in days. In the analysis and when interpreted they will be expressed in months.

Whenever required in the analysis, the conversion from days to months will be done by dividing the values in days by 30.4375 (365.25/12). The conversion from weeks to days will be done by multiplying the values in weeks by 7 (value in weeks * 7 = value in days).

Overall totals will be calculated for baseline summaries only. For continuous data, the mean and median will be rounded to 1 additional decimal place compared to the original data. The standard deviation will be rounded to 2 additional decimal places compared to the original data. Minimum and maximum will be displayed with the same accuracy as the original data. For categorical data, percentages will be rounded to 1 decimal place. P-values should be displayed with 4 decimal places, confidence intervals with the same precision as the corresponding statistic, and all ratios with 1 decimal place.

Each time when the formal analysis between the two treatment arms will be performed (e.g. to produce HR), only data for these arms should be included when running the statistical procedure. Generate MMRM estimates for only visits where scores for at least 25% of subjects in both treatment arms are available for analysis.

4.1.1 Baseline

In general, for efficacy the last observed measurement prior to randomization (before or on randomization date) will be considered the baseline measurement. For safety and PRO endpoints the last observation before the first dose of study treatment will be considered the baseline measurement unless otherwise specified (compare both date and time if possible). For assessments on the day of first dose where time is not captured, a nominal pre-dose indicator, if available, will serve as sufficient evidence that the assessment occurred prior to first dose. Assessments on the day of the first dose where neither time nor a nominal pre-dose indicator are captured will be considered prior to the first dose if such procedures are required by the protocol to be conducted before the first dose (see Table 2 and 3 of CSP).

If two assessments are equally eligible to assess subject status at baseline (e.g., screening and baseline assessments both on the same date prior to first dose with no washout or other intervention in the screening period), the average should be taken as a baseline value. For non-numeric laboratory tests (i.e. some of the urinalysis parameters) where taking an average is not possible then the best value would be taken as baseline as this is the most conservative. In the scenario where there are two assessments on day 1, one with time recorded and the other without time recorded, the one with time recorded would be selected as baseline.

In all summaries change from baseline variables will be calculated as the post-treatment value minus the value at baseline.

4.1.2 Visit windows for safety and PRO assessment

Time windows will need defining for any presentations that summarise safety or PRO data values by visit. The following conventions should also apply:

- The time windows should be exhaustive so that data recorded at any time point has the potential to be summarized. Inclusion within the time window should be based on the actual date and not the intended date of the visit.
- All unscheduled visit data should have the potential to be included in the summaries.

- The window 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. If an even number of days exists between two consecutive visits, then the upper limit will be taken as the midpoint value minus 1 day.

Visit windows will be defined as follows in the study:

- Screening, visit window (D-28, D-1)
- Baseline, visit window Low-D1
- Day 29 ($4*7+1$), visits window D2-D43
- Day 57 ($8*7+1$), visit window D44-D71
- Day 85 ($12*7+1$), visit window D72-D99
- Day 113 ($16*7+1$), visit window D100-D127
- ... (and continued every 4 weeks until last dose of study treatment + 90 days)

Listings will display all values contributing to a time point for a subject. Post treatment discontinuation follow up visits may also be summarized when available.

If there is more than one value per subject within a time window then the closest value to the scheduled visit date should be summarised, or the earlier, in the event the values are equidistant from the nominal visit date (date when visit occurred). The listings should highlight the value for the subject that contributed to the summary table, wherever feasible.

For summaries at a subject level, all values should be included, regardless of whether they appear in a corresponding visit based summary, when deriving a subject level statistic such as a maximum.

4.1.3 Study day will be calculated in relation to date of first treatment. Visit Windows for PK and ADA

Time windows for PK and ADA will use the same conventions defined in Section 4.1.2. Time windows will be defined as follows for presentations that summarise PK values by visit :

For Arms A and B:

- Baseline, Cycle 1: visit window Low-D1
- Post-baseline assessments during treatment including post-dose assessments on day 1:

C2-Day 29 ($4*7+1$), visits window D1 - D57;

C4-Day 85 ($12*7+1$), visit window D58 - last exposure to treatment
+ 90 days;

For Arm C:

Durvalumab PK:

- Baseline, Cycle 1: visit window Low-D1
- Post-baseline assessments during treatment including post-dose assessments on day 1:
C2-Day 29 ($4*7+1$), visits window D1 - D57

C4-Day 85 ($12*7+1$), visit window D58 - last exposure to treatment + 90 days;

Tremelimumab:

- Baseline, Cycle 1: visit window Low-D1
- Post-baseline assessments during treatment including post-dose assessments on day 1:
C2-Day 29 ($4*7+1$), visits window D1- last exposure to treatment + 90 days;

Time windows will be defined as follows for presentations that summarise ADA values by visit.

For all treatment arms:

- Baseline, Cycle 1: visit window Low-D1
- Post-baseline assessments during treatment including post-dose assessments on day 1:
C4-Day 85 visits window: D1- last exposure to treatment + 90 days

4.2 Analysis methods

4.2.1 Multiplicity

Two interim analyses and a final analysis are planned as described in Section 5.

To strongly control the familywise error rate (FWER) at the 5% level (2-sided), an alpha level of 0.1% will be spent on the interim ORR analysis (IA1) while the remaining 4.9% alpha level will be spent on all OS analyses. The primary objective of OS will be tested (H_1 : Arm C vs. Arm D) with 4.9% for this comparison.

Since two analyses of OS are planned (Interim Analysis, Final Analysis), the Lan DeMets approach (Lan and DeMets 1983) that approximates the O'Brien and Fleming spending function

will be used to maintain an overall 2-sided 4.9% type I error across the two planned analyses of OS (Interim and Final) for the primary comparison (H1: Arm C vs. Arm D)

If 78% of the target OS events for H1 (i.e. 404/515) are available at the time of the interim analysis, the 2-sided significance levels to be applied for the interim and final OS analyses would be 0.0222 and 0.0425, respectively.

If the primary comparison (H1: Arm C vs. Arm D) is statistically significant in the OS analyses at IA2 or FA, then the 4.9% alpha will be recycled to H2 across IA2 and FA. Otherwise neither H2 nor H3 will be tested. In case the primary comparison (H1: Arm C vs. Arm D) is not statistically significant at IA2, however significance is achieved at FA, then H2 will be tested at FA only. If non-inferiority is achieved in H2 testing at IA2 or FA, then the 4.9% alpha will be recycled to testing H3 at IA2 or FA respectively. Otherwise H3 will not be tested.

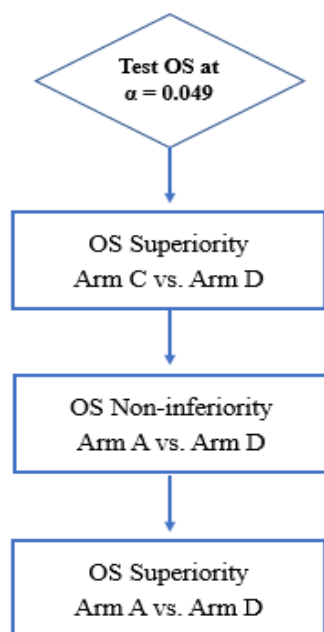
If 81% of the target OS events for H2 and H3 (ie. 453/560) are available at the time of the interim analysis, the 2-sided significance levels to be applied for the interim and final OS analyses for H2 and H3 will be 0.0248 and 0.0418 respectively; these alpha levels will be applied to the calculation of the confidence interval for the non-inferiority test for H2 so that a 97.52% and 95.82% CI will be presented for the non-inferiority comparison at the interim and final, respectively.

If all the OS analyses (H1, H2, and H3) are considered successful (superiority tests are statistically significant and non-inferiority is achieved), the 4.9% alpha level will be passed to test the difference in the three year survival rates (OS36) between Arm C and Arm D; Otherwise the test will not be conducted.

The study will be considered positive (a success) if the primary OS analysis result is statistically significant at either IA2 or FA. If significance is achieved at IA2, it does not need to be tested again at FA.

Strong control of the type I error will be applied testing endpoints as outlined in Figure 2.

Figure 2 Multiple testing strategy



4.2.2 Analysis of the primary variable

4.2.2.1 Overall survival

The primary analysis will be to compare OS for Arm C vs. Arm D (for superiority) in the FAS (ITT) analysis set.

The primary OS endpoint will be analysed using a stratified log-rank tests adjusting for etiology of liver disease (confirmed HBV versus confirmed HCV versus others), ECOG (0 versus 1), and macrovascular invasion (yes versus no) for generation of the p-value and using rank tests for association as the testing approach, which corresponds to Cox regression with the Breslow approach for handling ties (Breslow, 1974). The effect of Arm C vs. Arm D treatment will be estimated by the HR from stratified Cox proportional hazards model (with ties=*Efron* and stratification variables as listed above) together with its corresponding 95% confidence interval (CI) calculated using a profile likelihood approach. The stratification variable will use the values recorded in the randomization system (IWRS). If there is >10% discordance in stratification factors as recorded in IWRS versus the Case Report Form (CRF), then a sensitivity analysis of the primary endpoint OS will be performed using CRF based stratification factors.

The key secondary analyses are to compare OS for Arm A vs. Arm D (for non-inferiority, then superiority) in the FAS (ITT) analysis set. CCI

Secondary OS analyses will be performed using the same methodology as for primary analysis described above.

Kaplan-Meier plots of OS will be presented by treatment arm. Summaries of the number and percentage of subjects who have died, still in survival follow-up, lost to follow-up, and have withdrawn consent will be provided along with the median OS for each treatment.

The superiority boundary (i.e., adjusted alpha levels) for the HR of treatment comparison at the interim and the final for the primary OS and key secondary analyses will be derived based on the exact number of OS events using the Lan and DeMets approach that approximates the O'Brien Fleming spending function (see Section 4.2.1).

Subgroup analyses will be performed for OS as indicated in Section 4.2.6.

CCI

Assumptions of Proportionality

The assumption of proportionality of hazard will be assessed first by examining plots of complementary log-log (event times) versus log (time) and, if these raise concerns, by fitting a time-dependent covariate to assess the extent to which this represents random variation. If a lack of proportionality of hazard is evident, the variation in treatment effect will be described by presenting piecewise HR calculated over distinct time periods. In such circumstances, the HR can still be meaningfully interpreted as an average HR over time unless there is extensive crossing of the survival curves. If lack of proportionality of hazard is found, this may be a result of treatment-by-covariate interactions, which will be investigated. In addition, the Kaplan-Meier curve along with landmark analyses (e.g., 12-, 18-, 24 and 36-month OS rate) will also help in understanding the treatment benefit.

The Grambsch-Therneau test and Schoenfeld residuals may also be used to check violation of the proportional hazards assumption (Grambsch and Therneau 1994, Keele 2010).

As a lack of proportionality is expected (due to delayed effect in IO agents), a three-component stratified MaxCombo test will be used as a sensitivity analysis with the same stratification factors as the primary analysis. The MaxCombo test is the maximum of the normalized log rank test ($FH^{0,0}$) and selected Fleming-Harrington (FH) weighted log-rank tests (Fleming and Harrington 1991) ($FH^{0,1}$ and $FH^{1,1}$), i.e. $Z_{\max} = \max \{ FH^{0,0}, FH^{0,1} \text{ and } FH^{1,1} \}$, with multiplicity adjustment based on the asymptotic multivariate distribution (Karrison et al 2016). For group sequential design with formal interim analyses, the p value and rejection boundary at each analysis will be calculated based on the asymptotic multivariate normal distribution among the weighted logrank test statistics (logrank, $FH^{0,1}$ and $FH^{1,1}$) at interim and final analyses (He, Koch and Kurland 2021). The Fleming-Harrington tests of $FH^{0,1}$ and $FH^{1,1}$ assign less weight to early events and are more powerful in the scenario of delayed effect, while the log-rank test is optimal in the scenario of proportional hazards (Schoenfeld 1981). Under proportional hazards, the power loss from the Maxcombo test is usually small (Lin et al. 2020).

The Restricted Mean Survival Time (RMST) will also be analysed up to the minimum of the largest observed event time in each of the two arms and /or suitable clinically relevant timepoint, using the pseudovalues approach (Andersen et al. 2004) , to estimate RMST with standard error, for each treatment group, along with the estimate of difference in means between treatment groups, confidence interval and p-value. In addition, an area-under-the-curve approach (Kaplan-Meier method) and Royston-Parmar model (Royston and Parmar 2011, 2013) may also be used.

Sensitivity analysis for censoring patterns

A sensitivity analysis for OS will examine the censoring patterns to rule out attrition bias, achieved by a Kaplan-Meier plot of time-to-censoring where the censoring indicator of OS is reversed. This means that the status indicator will take the value of 0 for those subjects who died and the value of 1 for censored subject.

The number of subjects prematurely censored will be summarized by treatment arm. A subject would be defined as prematurely censored if their survival status was not defined at the DCO.

In addition, duration of follow-up will be summarized using medians:

- In censored subjects who are alive at data cut-off only: Time from randomization to date of censoring (date last known to be alive) by treatment arm.
- In all subjects: Time from randomization to the date of death (i.e. overall survival) or to the date of censoring for censored subjects regardless of treatment arm.

CCI

CCI

Effect of COVID-19

A sensitivity analysis will be conducted to assess for the potential impact of COVID deaths on OS. This will be assessed by repeating the OS analysis except that any subject who had a death with primary/secondary cause as COVID-19 Infection will be censored at their COVID infection death date.

A listing of all subjects diagnosed with COVID-19 or death due to COVID-19 by unique subject number identifier and investigational site will be generated along with the description of how the individual's participation was altered.

CCI

Effect of covariates on the HR estimate

Cox proportional hazards modelling will be employed to assess the effect of pre-specified covariates on the HR estimate for the primary OS treatment comparisons.

CCI

Additional covariates for this model will be:

- Sex (male versus female)
- Age at randomization (<65 versus ≥ 65 years of age)
- PD-L1 expression (positive versus negative)
- Extrahepatic spread (yes versus no; defined as Distant metastases on Pathology at screening module)
- Region (Asia (except Japan) versus Rest of World (includes Japan))
- Alpha-fetoprotein (AFP) (<400 ng/ml versus ≥ 400 ng/ml)
- BCLC stage at study entry (B versus C)

The model will include the covariates regardless of whether their inclusion significantly improves the fit of the model, providing there is enough data to make them meaningful.

For the definition of PD-L1 expression subgroup, refer to Section 4.2.11.

4.2.3 Analysis of the secondary variables

4.2.3.1 Progression Free Survival (PFS)

Analysis of PFS (time to first progression) will be performed to compare Arm C vs. Arm D and Arm A vs. Arm D using the same methodology as for OS. CCI

4.2.3.2 Time to progression (TTP)

Analysis of TTP will be performed to compare Arm C vs. Arm D and Arm A vs. Arm D using a stratified log-rank test as described for OS. CCI

4.2.3.3 Objective response rate (ORR)

The types of response rate analyses performed in the study together with their timepoints are defined in Table 14.

Table 14 Response rate analyses conducted during the study

Analysis time point	Analysis set	Confirmed/unconfirmed response	Assessment method and criteria
Interim Analysis 1	FAS-32w	Both	Investigator RECIST 1.1, BICR RECIST 1.1, BICR mRECIST
Interim Analysis 2	Full Analysis Set	Both	Investigator RECIST 1.1,
Final Analysis	Full Analysis Set	Both	Investigator RECIST 1.1,

*BICR is currently planned only for IA1, however, if it is performed for IA2/FA, these analyses will be conducted.

At IA1, only descriptive summaries of ORR including exact 95% CIs will be presented for each treatment arm (Arm A, Arm B, Arm C, and Arm D) for FAS-32w. ORR results will be presented by Investigator assessment (per RECIST1.1) and BICR (per RECIST1.1 and mRECIST).

At IA2 and FA, the ORR (per RECIST 1.1 using Investigator assessments) will be compared between Arm C vs. Arm D and Arm A vs. Arm D. Logistic regression models adjusting for the same factors as the primary endpoint (etiology of liver disease, ECOG, and macrovascular invasion) will be fitted. The results of the analysis will be presented in terms of an odds ratio together with its associated profile likelihood 95% CI (e.g. using the option 'LRCI' in SAS procedure GENMOD) and p-value (based on twice the change in log-likelihood resulting from the addition of a treatment factor to the model). This analysis will be performed in the FAS (ITT).

Additionally, at IA2 and FA a stratified Cochran Mantel–Haenszel (CMH) test will be performed using randomization stratification factors (macrovascular invasion, etiology of liver disease, and ECOG). CMH test results will include odds ratios and p-values.

4.2.3.4 Best objective response (BoR)

Summaries will be produced that present the number and percentage of subjects with a tumor-confirmed response (CR/PR). Overall visit response data will be listed for all subjects. For each treatment arm, best objective response (BoR) will be summarized by n (%) for each category (CR, PR, SD, PD, and NE). No formal statistical analyses are planned for BoR.

For IA1, BoR will be calculated in the FAS-32wp according to Investigator (per RECIST1.1) and BICR (per RECIST1.1 and mRECIST) assessments.

For IA2 and FA, BoR will be calculated for the FAS according to Investigator assessments (per RECIST1.1)..

4.2.3.5 Disease control rate (DCR)

The DCR, DCR-16w and DCR-24w will be summarized (i.e., number of subjects [%]) per treatment arm.

4.2.3.6 Duration of response (DoR)

Descriptive data will be provided for the DoR in responding subjects, including the associated Kaplan-Meier curves (without any formal comparison of treatment arms or p-value attached). This analysis will be based on FAS (ITT).

For IA1, DoR will be calculated in the FAS-32w according to Investigator (per RECIST 1.1) and BICR (per RECIST 1.1 and mRECIST) assessments.

For IA2 and FA, DoR will be calculated for the FAS according to Investigator assessments (per RECIST1.1), for responses with or without confirmation.

4.2.3.7 OS12, OS18, OS24, and OS36

OS12, OS18, OS24, and OS36 will be defined as the Kaplan-Meier estimate of OS at 12 months, 18 months, 24 months, and 36 months.

OS12, OS18, OS24, and OS36, along with their 95% CI, will be summarized (using the Kaplan-Meier curve) and presented by treatment arm.

An analysis of OS36 will be performed to compare Arm C vs. Arm D using a stratified chi-square test for the difference in KM estimators (cloglog transformed) for Arms C and D at a fixed time point (36 months). The test will be conducted using the methods described in (Klein et al., 2007), including cloglog transformation on KM estimators, with randomization stratification factors (macrovascular invasion, etiology of liver disease, and ECOG). Note that the adjustment for the stratification factors will be applied only if there are sufficient number of

events and subjects at risk available in each strata at 36 months. Otherwise, an unstratified chi-square test will be used to compare the difference in KM estimators at 36 months.

4.2.3.8 Time to Response (TTR)

TTR will be analysed using same methods like for PFS. No multiplicity adjustment will be applied, as it is a supportive endpoint calculated for the Payer Analysis. It will be reported outside CSR.

4.2.3.9 Time from Randomization to First Subsequent Therapy or Death (TFST)

TFST will be analysed using same methods like for PFS. No multiplicity adjustment will be applied, as it is a supportive endpoint calculated for the Payer Analysis. It will be reported outside CSR.

4.2.4 Patient-reported outcomes

The main PRO measures identified in the secondary objectives are global health status/QoL, physical function and fatigue scales along with single items appetite loss and nausea of the EORTC QLQ-C30; shoulder pain, abdominal pain, and abdominal distension symptom scales of the EORTC QLQ-HCC18.

All endpoints will be tested at a 5% significance level and 95% CIs will be produced.

Statistical analyses comparing treatment arms will include: visit specific and overall (across all visits) adjusted mean change from baseline scores (using mixed-effect model for repeated measurement [MMRM]), time to deterioration, improvement rate, visit response (improvement, no change, and deterioration) as well as best overall response.

Absolute and change from baseline scores for each visit will be presented as descriptive analysis. Appropriate plots and graphs will be presented. Compliance rates summarizing questionnaire completion at each visit will be tabulated.

By visits summaries will use visits windows defined in Section 4.1.2.

CCI

4.2.4.1 EORTC QLQ-C30

The key analysis of EORTC QLQ-C30 will be focused on the following scales/domains: global health status/QoL, physical function and fatigue. The primary assessment of HRQoL or symptom will focus on comparing mean change from baseline in the global health status/QoL, functions (physical, role, cognitive, social and emotional) and fatigue scores (from the EORTC QLQ-C30 questionnaire) along with single items appetite loss, nausea, and diarrhea of the EORTC QLQ-C30 between immunotherapy arms (Arm A, Arm C) and the sorafenib arms. The analysis sets for mean change in HRQoL or symptoms data will be the FAS (ITT) set and will include all randomized subjects with an evaluable baseline assessment and at least one evaluable

post baseline assessment. Change from baseline will be analysed using a MMRM analysis of all the post-baseline scores for each visit. The MMRM model will include treatment, visit, and treatment by visit interaction as explanatory variables and the baseline score and the baseline score-by-visit interaction as covariates. Restricted maximum likelihood (REML) estimation will be used. An overall adjusted mean estimate will be derived that will estimate the average treatment effect over visits given each visit equal weight. For overall treatment comparison, Adjusted mean change from baseline estimates per treatment group and corresponding 95% CIs will be presented along with an overall estimate of the treatment difference, 95% CI and p-value.

Unstructured covariance matrix will be used to model the within-subject error and the Kenward-Roger approximation will be used to estimate the degrees of freedom. 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, autoregressive and compound symmetry.

Time to deterioration will be analyzed using a stratified log-rank test as described for the primary OS endpoint. The effect of Arm C vs. Arm D and Arm A vs. Arm D will be estimated by the HR together with its corresponding CI and p-value. Kaplan-Meier plots will be presented by treatment arm. Summaries of the number and percentage of subjects who have an event as well as who were censored will be provided along with the medians for each treatment.

Summary tables of visit responses for each EORTC QLQ-C30 scale/item score (global health status/QoL, 5 functions and all symptoms) and for each visit (improvement, deterioration and no change) will be presented by treatment arm. In addition, summary tables of the best overall response will be provided for the following domains by treatment arm: global health status/QoL, functions (physical, role, cognitive, social, and emotional) and fatigue.

For the analysis of symptom and global/ health status/QoL improvement rate, symptom (fatigue), global health status/QoL and function (physical, role, cognitive, social, and emotional) improvement proportions based on best overall response will be compared between each immunotherapy arm (Arm A, Arm C) and the sorafenib arm using a logistic regression model as described for ORR. The odds ratio, p-value, and 95% CI will be presented graphically on a forest plot.

Finally, summaries of absolute and unadjusted change from baseline values of each EORTC QLQ-C30 scale/item will be reported by visit for each treatment arm.

Graphical plots of the mean of each EORTC QLQ-C30 scale/item scores (except financial difficulty scale), including change from baseline, and associated 95% CI by scheduled visits/timepoints in the study will be produced.

Empirical cumulative distribution function (eCDF) and probability density function (PDF) curves will be produced for parameters of global health status/QoL, function (physical), fatigue, and single-item symptoms (nausea and appetite loss).

The eCDF displays a continuous plot of the change from baseline in the PRO score on the horizontal axis and the cumulative percent of subjects experiencing up to that change on the vertical axis. Compared to eCDF curves, probability density function curves may provide an easier overview of the shape, dispersion, and skewness of the distribution of the change from baseline in the PRO score across treatment arms.

For each parameter listed above, the eCDF and PDF curves will be produced by treatment groups (Arm A, Arm B, Arm C, and Arm D) at weeks 8, 16, 24, and 48 (relative to first dose of study drug).

4.2.4.2 EORTC QLQ-HCC18

The primary assessment of symptoms comparing mean change from baseline using the MMRM as described for the EORTC QLQ-C30 will be repeated for shoulder pain, abdominal pain, and abdominal distention symptoms of the EORTC QLQ-HCC18. All assumptions and outputs as described for the EORTC QLQ-C30 are applicable.

Similarly, the time to deterioration as described for the EORTC QLQ-C30 will be evaluated for shoulder pain, abdominal pain, and abdominal distention symptoms of the EORTC QLQ-HCC18.

For shoulder pain, abdominal pain, and abdominal distention symptom scales of the EORTC QLQ-HCC18, time to deterioration will be presented using a Kaplan-Meier plot as well as the HR together with the corresponding 95% CI and p-values. Summaries of the number and percentage of subjects experiencing a clinically meaningful deterioration or death, and the median time to deterioration, will also be provided for each treatment arm.

Summary tables of visit responses for each EORTC QLQ-HCC18 scale/item score and for each visit (improvement, deterioration and no change) will be presented by treatment arm. In addition, summary tables of best overall response will be provided for the following symptom scales by treatment arm: shoulder pain, abdominal pain, and abdominal distention. Proportions of subjects with improvement based on best overall response will be compared between Arm C vs. Arm D and Arm A vs. Arm D using a logistic regression model as described for ORR. The odds ratio, p-value and 95% CI will be presented graphically on a forest plot.

As described for the EORTC QLQ-C30, summaries of absolute and unadjusted change from baseline values of each EORTC QLQ-HCC18 scale/item will be reported by visit for each treatment arm.

Graphical plots of the mean of each EORTC QLQ-HCC18 scale/item scores, including change from baseline, and associated 95% CI by scheduled visits/timepoints in the study will be produced.

Empirical cumulative distribution function (eCDF) and probability density function (PDF) curves will be produced for parameters of abdominal pain, shoulder pain, and abdominal distension single item symptoms by treatment groups (Arm A, Arm B, Arm C, and Arm D) at weeks 8, 16, 24, and 48 (relative to first dose of study drug).

4.2.4.3 CCI

CCI

4.2.4.4 CCI

CCI

4.2.4.5 CCI

CCI

4.2.5 CCI

CCI

4.2.6 Subgroup analyses

4.2.6.1 Subgroup analyses for OS

Subgroup analyses will be conducted comparing OS between Arm C vs. Arm D and Arm A vs. Arm D, and between Arm A and Arm C, in the following subgroups of the FAS, but not limited to:

- Sex (male versus female)

- Age at randomization (<65 versus ≥ 65 years of age)
- PD-L1 expression (positive versus negative)
- Etiology of liver disease (confirmed HBV versus confirmed HCV versus others)
- ECOG PS (0 versus 1)
- Macrovascular invasion (yes versus no)
- Extrahepatic spread (yes versus no; defined as distant metastases on Pathology at Screening module)
- Region (Asia (except Japan) versus Rest of World (includes Japan))
- Alpha-fetoprotein (AFP) (<400 ng/ml versus ≥ 400 ng/ml)
- BCLC stage at study entry (B versus C)
- MVI = Yes and/or EHS = Yes
- MVI = No and EHS = No

Values collected on the eCRF will be used to define subgroups for stratification factors. Other baseline variables may also be assessed if there is clinical justification or an imbalance is observed between the treatment arms. The purpose of the subgroup analyses is to assess the consistency of treatment effect across expected prognostic and/or predictive factors. Forest plots will be performed.

No adjustment to the significance level for testing of the subgroup and sensitivity analyses will be made since all these analyses will be considered supportive of the analysis of OS.

For each subgroup level of a factor, the HR and 95% CI will be calculated from a Cox proportional hazards model that only contains a term for treatment. The Cox models will be fitted using SAS® PROC PHREG with the Efron method to control for ties, and using a BY statement for the subgroup factor.

If there are too few events available for a meaningful analysis of a particular subgroup (it is not considered appropriate to present analyses where there are less than 20 events in a subgroup in total), the HR and CI will not be produced for that subgroup. In this case, only descriptive summaries will be provided.

For the definition of PD-L1 expression subgroup, refer to Section 4.2.11.

4.2.6.2 Subgroup analyses for secondary endpoints

Analyses described in Section 4.2.3 will be performed comparing PFS, TTP, ORR, BoR, DoR, DCR, DCR-16w and DCR-24w between Arm C vs. Arm D, and Arm A vs. Arm D in the following subgroups:

- PD-L1 expression (positive versus negative)
- Etiology of liver disease (confirmed HBV versus confirmed HCV versus others)
- Macrovascular invasion (yes versus no)

For the definition of PD-L1 expression subgroup, refer to Section 4.2.11.

4.2.7 Safety

Safety data will be summarized. No formal statistical analyses will be performed on the safety data. All safety and tolerability data will be using the safety analysis set and will be presented by treatment arm, and where indicated by-visit (using visits windows defined in Section 4.1.2)

Data from all cycles of treatment will be combined in the presentation of safety data. AEs (both in terms of Medical Dictionary for Regulatory Activities [MedDRA] preferred terms and CTCAE grade) will be listed individually by subject. The number of subjects experiencing each AE will be summarized by treatment arm and CTCAE grade. Additionally, data presentations of the rate of AEs per person-years at risk will be produced.

Other safety data will be assessed in terms of physical examination, serum chemistry, hematology, vital signs, and ECGs. Exposure to study drug(s), time on study, dose delays (all arms), and dose reductions (sorafenib CCI arm) will also be summarized. At the end of the study, appropriate summaries of all safety data will be produced.

“On treatment” will be defined as assessments between date of start dose and 90 days following the date of the last dose of study drug(s) (i.e., the last dose of durvalumab, tremelimumab, or sorafenib) (including re-treatment), or up to the date of initiation of the first subsequent therapy (whichever occurs first). This definition applies to all safety reporting, unless otherwise specified.

Date of initiation of the first subsequent therapy should be the date of the first Post IP Discontinuation Systemic Cancer Therapy.

4.2.7.1 Adverse events

All AEs, both in terms of current Medical Dictionary for Regulatory Activities (MedDRA) preferred term and Common Toxicity Criteria for Adverse Events (CTCAE) grade, will be listed and summarized descriptively by count (n) and percentage (%). The current MedDRA dictionary will be used for coding. Any AE occurring before treatment with IP will be included in the AE listings, but will not be included in the summary tables (unless otherwise stated). These will be referred to as ‘pre-treatment’.

On treatment (or treatment-emergent AEs) will be defined as any AEs with an onset date or pre-treatment AEs that increase in severity on or after the date of first dose and up to and including 90 days following the date of last dose of study drug(s) (i.e., the last dose of durvalumab, tremelimumab, or sorafenib) (including re-treatment), or up to the date of initiation of the first subsequent therapy (whichever occurs first). TEAEs will be included in all summary tables.

AEs and SAEs will be collected from the time of signature of informed consent, throughout the treatment period, and up to the follow-up period (90 days after the last dose of study drug(s) [durvalumab, tremelimumab, or sorafenib]). Any events in this period that occur after a subject has received further therapy for cancer (following discontinuation of study drug[s]) will be flagged in the data listings. A separate data listing of AEs occurring more than 90 days after

discontinuation of study drug(s) will be produced. These events will not be included in AE summaries.

All reported AEs will be listed along with the date of onset, date of resolution (if AE is resolved) and Investigator's assessment of severity and relationship to study drug. Frequencies and percentages of subjects reporting each preferred term will be presented (i.e. multiple events per subject will not be accounted for apart from on the episode level summaries).

Summary information (the number and percent of subjects by system organ class and preferred term) will be tabulated for:

- All AEs
- All AEs possibly related to study medication (as determined by the reporting Investigator)
- AEs with CTCAE grade 3 or 4
- AEs with CTCAE grade 3 or 4, possibly related to study medication (as determined by the reporting Investigator)
- Most common AEs
- Most common AEs with CTCAE grade 3 or 4
- All SAEs
- All SAEs possibly related to study medication (as determined by the reporting Investigator)
- All SAE leading to discontinuation of study medication
- All SAE leading to discontinuation of study medication, possibly related to study medication (as determined by the reporting Investigator)
- AEs leading to discontinuation of study medication
- AEs leading to discontinuation of study medication, possibly related to study medication (as determined by the reporting Investigator)
- AEs leading to dose interruption of study medication
- Infusion reaction AEs
- AEs with outcome of death by system organ class, preferred term and maximum reported CTCAE grade
- AEs with outcome of death, possibly related to durvalumab by system organ class, preferred term and maximum reported CTCAE grade
- AEs with outcome of death, possibly related to tremelimumab by system organ class, preferred term and maximum reported CTCAE grade
- AEs with outcome of death, possibly related to sorafenib by system organ class, preferred term and maximum reported CTCAE grade

An overall summary of the number and percentage of subjects in each category will be presented. In addition, a truncated AE table of most common AEs, showing all events that occur in at least 10% of subjects in any treatment arm will be summarized by preferred term, by decreasing frequency in the durvalumab + tremelimumab combination therapy arm, this table will also be produced using a 5% cut-off point. These cut-offs may be modified after review of the data. When applying the cut-offs, the raw percentage should be compared to the cut-off, no

rounding should be applied first (i.e., an AE with frequency of 9.9% will not appear because the cut-off is 10%).

Each AE event rate (per 100 years) will also be summarized by preferred term within each system organ class. For each preferred term, the event rate (defined as the number of subjects with at least one AE divided by the total time at risk, i.e. the number of days of exposure to drug summed over all subjects in each group multiplied by 100) will be presented.

The number and percentage of subjects experiencing each AE will be summarized by treatment arm and maximum CTCAE grade. Any safety summaries examining rechallenge with tremelimumab may be produced separately.

Deaths

A summary of deaths will be provided with number and percentage of subjects, categorized as:

- Total number of deaths (regardless of date of death)
- Related to disease under investigation,
- AE outcome of death only and onset date prior to initiation of subsequent anti-cancer therapy
- AE outcome of death only and onset date falling after 90 days following last dose of study medication or initiation of subsequent anti-cancer therapy (whichever is earlier)
- Both related to disease under investigation and with AE outcome of death and onset date prior to initiation of subsequent anti-cancer therapy
- Death related to disease under investigation and AE with outcome of death > 90 days after last dose of study medication or $\geq$ date of subsequent therapy, whichever occurs first
- Deaths > 90 days after last dose of study medication or $\geq$ date of subsequent therapy (whichever occurs first), unrelated to AE or disease under investigation
- Subjects with unknown reason for death.
- Other deaths.

A corresponding listing will also be produced.

Adverse events of special interest

Preferred terms used to identify adverse events of special interest will be listed before database lock (DBL) and documented in the Study Master File. Grouped summary tables of certain MedDRA preferred terms will be produced. For each 'grouped' term, the number (%) of subjects experiencing any of the specified terms will be presented by maximum CTCAE grade.

Additional summaries will include Time to onset of first CTCAE grade 3 or higher. Time to onset of first AE for each grouped term and preferred term within it will also be produced. Groupings will be based on preferred terms provided by the medical team prior to DBL, and a listing of the preferred terms in each grouping will be provided.

Additional summaries of the above-mentioned grouped AE categories will include number (%) of subjects who have:

- At least one adverse event of special interest presented by outcome
- At least one adverse event of special interest by CTCAE grade
- At least one adverse event of special interest possibly related to study medication
- At least one adverse event of special interest leading to discontinuation of IP

A summary of total duration (days) including median duration of AESI will be provided for events which have an end date and this will be supported by summaries of ongoing AESIs at death and separately at data cut-off, as well as a summary of time to resolution to grade 1 or less and time to resolution to grade 2 or less.

Additionally, there will be several summaries of AESIs requiring concomitant treatment, and particularly the relationship of AESIs to the use of immunosuppressive agents (a table of AESI leading to concomitant medications use by grouped term and preferred term).

Immune- mediated adverse events

A programmatic process will be used to identify whether adverse events of special interest (AESIs) and adverse events of possible interest (AEPIs) are immune-mediated adverse events (imAEs). The programmatic process will allow for imAE frequencies to be calculated from both AESIs and AEPIs based on applied rules and a treatment algorithm that considers interventions involving systemic steroid therapy, immunosuppressant use, and/or endocrine therapy (which, in the case of AEPI, occurs after first applying consideration of an Investigator's causality assessment of the AE to any study treatment and/or an Investigator's designation of an event as immune-mediated).

A manual adjudication process will also be applied using the latest list of AESIs/AEPIs.

The following summaries will be provided for imAEs:

- AESIs, AEPIs, and imAEs in any category
- Time to onset of AESIs, AEPIs, and imAEs
- Pneumonitis AESI and AEPI and imAE, repeated for all other categories
- AESIs, AEPIs, and imAEs by category
- imAEs by AESI/AEPI category and preferred term
- imAEs by AESI/AEPI group, preferred term, and maximum CTCAE Grade

In addition, a listing of suspected imAEs will be produced based on comparison to the latest list of AESIs/AEPs.

4.2.7.2 Treatment exposure and intensity

Exposure to study drug(s), time on study, dose delays (all arms), and dose reductions (sorafenib CCI arm) will also be summarized. At the end of the study, appropriate summaries of all safety data will be produced.

The following summaries related to durvalumab, tremelimumab and sorafenib will be produced:

- Total exposure
- Time on study in months
- Actual exposure
- Dose delays, infusion interruptions, and dose reductions
- Number of infusions, doses, and treatment cycles received
- RDI (relative dose intensity)

For subjects on study treatment at the time of the ORR and OS analysis, the DCO date will be used to calculate exposure. Summaries of exposure will also be presented for the subgroup of discontinued subjects.

The number of durvalumab, tremelimumab and sorafenib doses/ infusions and total dose received will be summarized by descriptive statistics and by frequency. Dose intensity and relative dose density of will be summarized by descriptive statistics.

All summaries should be done for three study periods: initial treatment, re-challenge (if applicable), and total study.

4.2.7.3 Laboratory assessments

On-treatment laboratory data will be used for reporting. Only laboratory tests required by CSP (Tables 5, 6, 7 and 8 in CSP) will be included in tables.

Any data post 90 days after the last dose of the study treatment or initiation of subsequent therapy will not be summarised. Data summaries will be provided in preferred units.

All available laboratory data will be summarized in by-visits tables using summary statistics.

Scatter plots (shift plots) of baseline to maximum value/minimum value (as appropriate) on treatment may be produced for certain parameters if warranted after data review.

Box-plots of absolute values by week, and box-plots of change from baseline by week, may be presented for certain parameters if warranted after data review. For continuous laboratory assessments, absolute value and change from baseline will be summarised using descriptive statistics at each scheduled assessment time by actual treatment group.

Shift tables for laboratory values by worst CTCAE grade will be produced, and for specific parameters separate shift tables indicating hyper- and hypo- directionality of change will be produced. The laboratory parameters for which CTCAE grade shift outputs will be produced are:

- Haematology: Haemoglobin; Leukocytes; Lymphocytes (count, absolute); Neutrophils (count, absolute); Platelets
- Clinical chemistry: ALT, AST, ALP, Total Bilirubin, Albumin, Magnesium – hypo and –hyper, Sodium – hypo and – hyper, Potassium – hypo and – hyper, Corrected Calcium –hypo and – hyper, Glucose – hypo and – hyper, Creatinine.

Liver Enzyme Elevations and Hy's law

Potential Hy's law cases will be defined as Aspartate aminotransferase (AST) or alanine aminotransferase (ALT) $\geq 3 \times$ ULN together with total bilirubin (TBL) $\geq 2 \times$ ULN at any point during the study following the start of study medication irrespective of an increase in alkaline phosphatase (ALP). The onset date of ALT or AST elevation should be prior to or on the date of Total Bilirubin elevation

The following summaries will include the number (%) of subjects who have:

- Elevated ALT, AST, and Total bilirubin during the study
 - ALT $\geq 3x - \leq 5x$, $> 5x - \leq 8x$, $> 8x - \leq 10x$, $> 10x - \leq 20x$, and $> 20x$ Upper Limit of Normal (ULN) during the study
 - AST $\geq 3x - \leq 5x$, $> 5x - \leq 8x$, $> 8x - \leq 10x$, $> 10x - \leq 20x$, and $> 20x$ ULN during the study
 - Total bilirubin $\geq 2x - \leq 3x$, $> 3x - \leq 5x$, $> 5x$ ULN during the study
 - ALT or AST $\geq 3x - \leq 5x$, $> 5x - \leq 8x$, $> 8x - \leq 10x$, $> 10x - \leq 20x$, $> 20x$ ULN during the study
 - ALT or AST $\geq 3x$ ULN and Total bilirubin $\geq 2x$ ULN during the study (Potential Hy's law): The onset date of ALT or AST elevation should be prior to or on the date of Total Bilirubin elevation

Liver biochemistry test results over time for subjects with elevated ALT or AST (i.e. $\geq 3x$ ULN), and elevated Total Bilirubin (i.e. $\geq 2x$ ULN) (at any time) will be plotted. Individual

subject data where ALT or AST (i.e. $\geq 3 \times \text{ULN}$) plus Total Bilirubin (i.e. $\geq 2 \times \text{ULN}$) are elevated at any time will be listed also.

Plots of ALT and AST vs. Total Bilirubin by treatment group will also be produced with reference lines at $3 \times \text{ULN}$ for ALT, AST, and $2 \times \text{ULN}$ for Total Bilirubin. In each plot, Total Bilirubin will be in the vertical axis.

Abnormal Thyroid function

Elevated TSH will be summarized per treatment group in terms of number (%) of subjects with:

- elevated high TSH (higher than the upper normal range),
- low TSH (lower than lower normal range),
- elevated high TSH post-dose and within normal range at baseline,
- low TSH post-dose and within normal range at baseline.

4.2.7.4 ECGs

On-treatment ECG data will be used for reporting.

Overall evaluation of ECG is collected at each visit in terms of normal or abnormal, and the relevance of the abnormality is termed as “clinically significant” or “not clinically significant”. A shift table of baseline evaluation to worst evaluation on-treatment will be produced.

4.2.7.5 Vital signs

On-treatment vital signs data will be included in the summary tables.

Box plots for absolute values and change from baseline by week may be presented for certain vital signs parameters if warranted after data review.

Vital signs (systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse rate, temperature, respiratory rate and weight) will be summarised over time in terms of absolute values and change from baseline at each scheduled measurement by actual treatment group.

4.2.7.6 ECOG performance status

Performance status as determined by the ECOG Scale will be recorded in the eCRF as per the schedules defined in CSP and will be summarized by visit over time.

4.2.7.7 Child-Pugh score

The severity of chronic liver disease, mainly cirrhosis, as determined by the Child-Pugh score (Pugh et al 1973), will be recorded in the eCRF as specified in the assessment schedules (see Table 2, Table 3, and Table 4 of CSP). The modified Child-Pugh classification of liver disease severity according to the degree of ascites, serum concentrations of bilirubin and albumin, prothrombin time, and degree of encephalopathy is shown in Table 9 of CSP. The severity of cirrhosis is classified as follows:

- Child-Pugh class A (well-compensated disease): score of 5 to 6
- Child-Pugh class B (significant functional compromise): score of 7 to 9
- Child-Pugh class C (decompensated disease): score of 10 to 15

Child-Pugh classification and the total score will be summarized by visit.

4.2.7.8 Physical examinations

For physical examination data only assessment date will be collected in eCRF. Therefore, no summaries will be produced.

4.2.7.9 Other safety assessments

If new or worsening pulmonary symptoms (e.g., dyspnea) or radiological abnormality suggestive of pneumonitis/ILD is observed, toxicity management will be applied and all related safety data will be listed.

Data from positive pregnancy tests will not be summarized.

4.2.8 Pharmacokinetic data

PK concentration data will be listed for each subject and each dosing day by treatment and analyte, and a summary will be provided using descriptive statistics by treatment and analyte for all evaluable subjects.

At a time point where less than or equal to 50% of the concentration values are NQ (Not Quantifiable), all NQ values will be set to the Lower Limit of Quantification (LLOQ), and all descriptive statistics will be calculated accordingly.

At a time point where more than 50% (but not all) of the values are NQ, the gmean, $\text{gmean} \pm \text{gSD}$ and gCV% will be set to NC (Not Calculated). gmean, gSD, gCV are geometric mean, geometric standard deviation, and geometric coefficient of variation. The maximum value will be reported from the individual data, and the minimum and median will be set to NQ.

If all concentrations are NQ at a time point, no descriptive statistics will be calculated for that time point. The gmean, minimum, median and maximum will be reported as NQ and the gCV% and $\text{gmean} \pm \text{gSD}$ as NC.

4.2.9 Immunogenicity data

A summary of the number and percentage of subjects in the ADA evaluable set who developed detectable ADA to duralumab or tremelimumab by ADA categories (Section 3.9) in different treatment arms will be presented and will include ADA prevalence and ADA incidence (proportion of ADA-positive and treatment-emergent ADA-positive subjects, respectively, in the ADA evaluable set). Immunogenicity results will be listed for all subjects in the Safety Analysis Set regardless of ADA-evaluable status. ADA titer and nAb data will be listed for samples confirmed positive for the presence of anti-durvalumab and/or anti-tremelimumab antibodies. AEs in ADA positive subjects by ADA positive category will be listed.

CCI

4.2.10 CCI

CCI

4.2.11 Biomarker data

The relationship of PD-L1 expression and, if applicable, of CCI to clinical outcomes (including but not restricted to) of PFS, ORR, and OS will be assessed.

PD-L1 expression determined by IHC will be reported in the CSR. CCI

The cut-off for PD-L1 expression subgroup analysis (high vs low/negative) is defined using data outside of the HIMALAYA study. The plans for determination of the cut-off value are detailed in a separate analysis plan for HCC PD-L1 Analyses.

PD-L1 expression will be determined by the analytically validated VENTANA PD-L1 (SP263) Assay using the TIP score method. The TIP score will be defined as the total percentage of the tumor area covered by tumor cells with PD-L1 membrane staining at any intensity and/or tumor-associated immune cells with any pattern of PD-L1 staining at any intensity. PD-L1 positive will be defined as PD-L1 staining of any intensity in tumor cell membranes and/or tumor-associated immune cells covering $\geq 1\%$ of tumor area. PD-L1 negative will be defined as PD-L1 staining of any intensity in tumor cell membranes and/or tumor-associated immune cells covering $< 1\%$ of tumor area.

4.2.12 CCI

CCI

4.2.12.1 CCI

CCI

4.2.12.2 CCI

CCI

4.2.12.3 CCI

CCI

4.2.12.4 CCI

CCI

4.2.13 Demographic, initial diagnostics and baseline characteristics data

The following will be summarized for all subjects in the FAS (unless otherwise specified):

- Subject disposition (including screening failures and reason for screening failure);
- Important protocol deviations;
- Inclusion in analysis sets;
- Demographics (age, age group [<65 , $\geq 65 - <75$, and ≥ 75 years], sex, race and ethnicity);
- Subject characteristics at baseline (height [cm], weight [kg], weight group [<70 , $\geq 70 - <90$, and ≥ 90 kg], Body Mass Index (BMI) and BMI group [<18.5 , $\geq 18.5 - 25.0$, $\geq 25.0 - <30.0$, ≥ 30.0 kg/m²], ECOG performance status, PD-L1, Child Pugh Classification);
- Subject recruitment by country and centre;
- Disease characteristics at initial diagnosis (primary tumor location, histology type, primary tumor grade, time from diagnosis to randomization, time from diagnosis to first dose, Fibrosis score [F0, F1] AJCC staging);
- Disease characteristics at screening (primary tumor location, histology type, primary tumor grade, Fibrosis score [F0, F1], Child Pugh Classification, alpha Fetoprotein [<400 , ≥ 400 ng/ml, missing], MVI [Yes, No, Missing], MVI = Yes and/or EHS = Yes, Virology status at screening [HBV, HCV, Other, Missing], ECOG performance status, study entry BCLC score [B, C, Missing] AJCC staging);
- Nicotine use at baseline (never, current, former);
- Alcohol use at baseline (never, current, former);

- Disallowed concomitant medications;
- Allowed concomitant medications (summary by ATC class and preferred term); allowed prior medications will only be listed;
- Disease related medical and surgical history;
- PD-L1 status at baseline (positive, negative);
- Stratification factors by IWRS and CRF [etiology of liver disease (confirmed HBV versus confirmed HCV versus others), ECOG (0 versus 1), and macrovascular invasion (yes versus no)];
- Primary tumor location and TNM classification at baseline;
- Post IP discontinuation disease-related anti-cancer therapy;
- Subsequent anti-cancer therapy (number of regimens, time to first and second subsequent therapy);

The WHO Drug Dictionary (WHO DD) will be used for concomitant medication coding.

The following tables should be repeated for the subset of subjects (if any) experiencing re-challenge: subject disposition (discontinuation reasons from “Discontinuation of Durvalumab-Rechallenge” CRF form), demographics, subject characteristics at baseline, disease characteristics at initial diagnosis, disease characteristics at screening, baseline tumor characteristics, allowed concomitant medications.

For the definition of PD-L1 expression subgroup, refer to Section 4.2.11.

5. INTERIM ANALYSES

5.1 Analysis methods

Two interim analyses and a final analysis are planned as described below:

Interim Analysis 1 (IA1): The first interim analysis will be performed after approximately 100 subjects per treatment arm have had the opportunity for 32 weeks of follow-up and not prior to the last subject enrolled. The objective is to evaluate the efficacy of Arm A and Arm C in terms of ORR and DoR. The analysis set for ORR and DoR will be the FAS-32wA BICR of radiological scans will be performed on all subjects included in IA1 who have been randomized and have had the opportunity for at least 32 weeks follow-up. Both Investigator (using RECIST 1.1) and BICR (using RECIST 1.1 and mRECIST) assessments are planned for IA1. Therefore, ORR and DoR (for both confirmed and unconfirmed responses) according to both Investigator using RECIST 1.1 and BICR using RECIST 1.1 and mRECIST will be reported for IA1.

Interim Analysis 2 (IA2): The second interim analysis will be performed when approximately 404 OS events in Arm C and Arm D combined (~52% maturity), approximately 30 months after the first subject is randomized. The goal is to evaluate the efficacy of Arm C vs. Arm D (for superiority) and then Arm A vs. Arm D (for non-inferiority, then superiority) in terms of OS. It is anticipated that approximately 453 OS events will have occurred across Arms A and D combined (~59% maturity) at the time of the DCO for IA2.

Final Analysis (FA): The final analysis is expected to be performed when approximately 515 OS events in Arm C and Arm D combined (~67% maturity), approximately 37.5 months after the first subject is randomized. The primary objective is to assess the efficacy of Arm C vs. Arm D in terms of OS for superiority. The key secondary objectives are to assess the efficacy of Arm A vs. Arm D in terms of OS (for non-inferiority, then superiority). It is anticipated that approximately 560 OS events will have occurred across Arms A and D combined (~73% maturity) at the time of the DCO for the final analysis. Efficacy data for Arm B (which was closed for enrollment with Amendment 4) will be summarized descriptively, however will not be formally analyzed.

The familywise error rate will be strongly controlled across all analyses using the strategy outlined in Section 4.2.1.

5.2 Blinding

Although the study is open-label, it will be conducted as “Sponsor-blind”. To maintain the integrity of the study, Sponsor access to treatment records will be restricted. Under no circumstance will the Sponsor perform any efficacy analysis by treatment arm during the study – any exceptions will be documented in a Trial Integrity Document (TID). The TID will pre-specify nominated individuals who will be granted access to treatment-revealing data with their reason for requiring access detailed.

The study includes 2 interim analyses (by treatment arm), which will be performed by an Independent Data Monitoring Committee (IDMC). Details will be given in the IDMC charter. Study team unblinding will not occur if IA1 is positive. A separate team will be involved in any review/submission activities as documented in the TID. At a positive interim (IA2) or final analysis, the blind may be broken for the whole study team. The timing of the unblinding of the study team will be formally documented, per SOP AZDoc0022221 – *Planned Unblinding of a Clinical Trial*.

5.3 Independent Data Monitoring Committee

An IDMC will be established to monitor data on an ongoing basis to ensure the continuing safety of subjects enrolled in this study, to ensure the integrity of the study, and to oversee the 2 planned interim analyses. The first IDMC safety review will occur when approximately 30 subjects per arm are randomized or 6 months after the first subject is dosed (whichever comes first), and will occur approximately every 6 months thereafter; the frequency of IDMC review may be adjusted by the IDMC as needed. The IDMC will be composed of individuals external to AstraZeneca. An IDMC charter will be developed which will specify the Committee’s

responsibilities, authorities, and procedures along with details of the interim analysis planning, decision-making guidance, and dissemination of the results as well as the recommendations and decisions after the interim analyses. Formal implementation and communication of IDMC recommendations will be managed by the AstraZeneca Executive Committee, which will be unrelated to the study project team.

Full details of the IDMC procedures, processes, and interim analyses can be found in the IDMC Charter.

6. CHANGES OF ANALYSIS FROM PROTOCOL

The following efficacy endpoints will be derived, which are not detailed per clinical study protocol: Time to Response (TTR), Time from Randomization to First Subsequent Therapy or Death (TFST). They will be calculated to support the payer analysis.

Section 6.5 of the protocol does not define or provide instructions for determining AEPs. The latest list of preferred terms will be used to determine both AESIs and AEPs.

Additional details of the NI approach have been provided in Sections 1.3 and 4.2.2.1 expanding up on the details of the NI margin in Section 8.2 of the protocol. The additional details in Section 1.3 include the results of the 3 studies used to determine the NI margin, clarification that the assumed HR for the Arm A vs Arm D comparison is based on Checkmate 459-results for nivolumab vs sorafenib in the same population (Yau T, 2019), and results from 4 other studies that were designed with non-inferiority to a sorafenib control in first line HCC. In Section 4.2.2.1, it is specified that for interim and final analyses of the primary OS and key secondary analyses (including NI), adjusted alpha levels will be derived based upon the exact number of OS events using the Lan and DeMets approach that approximates the O'Brien Fleming spending function. It is also clarified in the section that the primary analysis method of the log-rank test will be used to assess NI and superiority for the OS comparisons of Arms A vs Arm D.

A test of the three year overall survival rate (OS36) between Arm C and Arm D has been added. The test of OS36 will be conducted using the stratified method described in (Klein et al., 2007), using stratification factors collected at randomization (macrovascular invasion, etiology of liver disease, and ECOG). The adjustment for stratification factors will be applied only if there are sufficient number of events and number of subjects at risk available in each strata at 36 months, otherwise unstratified methods from (Klein et al., 2007) will be used.

7. REFERENCES

Alosh M et al. 2015

Statistical considerations on subgroup analysis in clinical trials. Stat Bio Res. 2015; 7 (4):286-303

Andersen PK, Hansen MG, Klein JP. 2004

Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Anal. 2004 Dec;10(4):335-50.

Berry G, Kitchin RM, Mock PA. 1991

A comparison of two simple hazard ratio estimators based on the logrank test. Stat Med. 1991; 10:749-755

Breslow N. 1974

Covariance Analysis of Censored Survival Data. Biometrics. 1974; 30:89–99

Brown LD, Cai TT, DasGupta A. 2001

Interval estimation for a binomial proportion. Statist Sci. 2001; 16:101–133

Bruix J et al. 2012

Efficacy and safety of sorafenib in patients with advanced hepatocellular carcinoma. J Hepatol 2012;57(4):821-9.

Burman CF, Sonesson C, Guilbaud O. 2009

A recycling framework for the construction of Bonferroni-based multiple tests. Stat Med. 2009; 28:739-761

Cainap C et al. 2015

Linifanib versus Sorafenib in patients with advanced hepatocellular carcinoma: results of a randomized phase III trial. Journal of Clinical Oncology 33.2 (2015): 172.

Campbell MJ, Gardner MJ. 1988

Calculating confidence intervals for some non-parametric analyses. Br Med J. 1988; 296:1454-1456

Carroll KJ. 2003

On the use and utility of the Weibull model in the analysis of survival data. Control Clin Trials. 2003; 24:682-701

Cheng AL et al. 2009

Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. Lancet Oncol 2009;10(1):25-34.

Cheng AL et al. 2012

Efficacy and safety of sorafenib in patients with advanced hepatocellular carcinoma according to baseline status: subset analyses of the phase II sorafenib Asia-Pacific trial. Eur J Cancer 2012;48(10):1452-65.

Cheng AL et al. 2013

Sunitinib versus sorafenib in advanced hepatocellular cancer: results of a randomized phase III trial. J Clin Oncol 2013; 31:4067-4075.

Cheng AL et al. 2019

IMbrave150: Efficacy and safety results from a ph III study evaluating atezolizumab (atezo) + bevacizumab (bev) vs sorafenib (Sor) as first treatment (tx) for patients (pts) with unresectable hepatocellular carcinoma (HCC). ESMO Asia 2019; 22–24 November 2019; Singapore2019. p. LBA3.

Clopper C and Pearson ES. 1934

The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934; 26:404-413

Cox DR. 1972

Regression models and life-tables. J R Stat Soc Series B Stat Methodol 1972;34(2):187-220.

Devlin et al. 2016

Valuing health-related quality of life: An EQ-5D-5L value set for England. Office of Health Economics 2016.

Dodd LE et al.. 2011

An audit strategy for progression free survival. Biometrics, 2011; 67(3): 1092-9

Duke-Margolis 2018

Oncology Clinical Trials in the Presence of Non- Proportional Hazards, The Duke- Margolis Center for Health Policy Feb. 2018.

Dunnnett C. 1955

A multiple comparison procedure for comparing several treatments with a control. J Amer Statist Assoc. 1955; 50:1096-1121

Dunnnett CW, Tamhane AC. 1992

A step-up multiple test procedure. J Amer Statist Assoc 1992;87(417):162-70.

Ellis S, Carroll K J, Pemberton K. 2008

Analysis of duration of response in oncology trials. Contemp Clin Trials. 2008; 29:456-465

EMA guideline. 2005

Guideline on the choice of the non-inferiority margin.

Fayers et al 2001

Fayers P, Aaronson NK, Bjordal K, Curran D, Groenvold M on behalf of the EORTC Quality of Life Study Group. EORTC QLQ-C30 Scoring Manual: 3rd Edition 2001. Available on request.

FDA endpoint guidance.

Guidance for industry clinical trial endpoints for the approval of cancer drugs and biologics.

FDA guidance. 2016

Guidance for industry non-inferiority clinical trials to establish effectiveness.

Fleischer F, Gaschler-Markefski B, Bluhmki E. 2011

How is retrospective independent review influenced by investigator-introduced informative censoring: A quantitative approach. Stat Med. 2011; 30(29):3373-86

Fleming TR, Harrington DP. 1991

Counting Processes and Survival Analysis. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley and Sons Inc. 1991;New York.

Gail M, Simon R. 1985

Testing for qualitative interactions between treatment effects and subject subsets. Biometrics. 1985; 41:361-72

Glimm E, Maurer W, Bretz F. 2009

Hierarchical testing of multiple endpoints in group-sequential trials. Stat Med. 2009; 29:219-228

Gooley TA et al. 1999

Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. Stat Med. 1999; 18:695-706

Grambsch PM, Therneau TM. 1994

Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994; 81:515-26

He P, Koch G, Kurland J. 2021

Robust Group Sequential Design Using Weighted Logrank Tests and Practical Considerations in Immuno-oncology Trials. 2021; *Manuscript submitted*.

Hertz-Picciotto I and Rockhill B. 1997

Validity and efficiency of approximation methods for tied survival times in Cox regression. Biometrics. 1997; 53:1151-1156

Hochberg, Y. 1988

A Sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988; 75:800-802

Holm S. 1979

A simple sequentially rejective multiple test procedure. *Scand J Stat.* 1979; 6:65-70

Hsu CH, Taylor JMG. 2009

Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Stat Med.* 2009; 28:462-475

Hung HMJ, Wang SJ, O'Neill R. 2007

Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *J Biopharm Stat.* 2007; 17:1201-1210

Jennison C, Turnbull BW. 2000

Group Sequential Methods with Application to Clinical Trials. Chapman&Hall/CRC, 2000

Johnson PJ. et al. 2013

Brivanib versus sorafenib as first-line therapy in patients with unresectable, advanced hepatocellular carcinoma: results from the randomized phase III BRISK-FL study. *J Clin Oncol* 31.28 (2013): 3517-3524.

Karrison TG. 2016

Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata J* 2016; 16: 678–90.

Keele L. 2010

Proportionally Difficult: Testing for Nonproportional Hazards in Cox Models, *Political Analysis* 2010; 18:189–205. doi:10.1093/pan/mpp044.

Klein JP et al. 2007

Klein JP, Logan B, Harhoff M, Andersen PK. Analyzing survival curves at a fixed point in time. *Stat Med* 2007;26(24):4505-19.

Kudo M et al. 2018

Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomized phase 3 non-inferiority trial. *Lancet* 2018; 391:1163-72.

Lan KKG and DeMets DL. 1983

Discrete sequential boundaries for clinical trials. *Biometrika* 1983; 70:659-63.

Latimer NR et al. 2018

Assessing methods for dealing with treatment switching in clinical trials: a follow-up simulation study. *Stat Methods Med Res.* 2018;27(3):765–84.

Lin RS et al. 2020

Alternative analysis methods for time to event endpoints under nonproportional hazards: a comparative analysis. *Stat Biopharm Res* 2020; 12: 187–98.

Llovet JM et al. 2008a

Design and endpoints of clinical trials in hepatocellular carcinoma. J Natl Cancer Inst 2008;100(10):698-711.

Llovet JM et al. 2008b

Sorafenib in advanced hepatocellular carcinoma. N Engl J Med 2008;359(4):378-90.

Miettinen O, Nurminen N. 1985

Comparative analysis of two rates. Stat Med. 1985; 4:213-226

O'Brien PC, Fleming TR. 1979

A multiple testing procedure for clinical trials. Biometrics. 1979; 35:549-556

Palta M, Amini S. 1982

Magnitude and likelihood of loss resulting from non-stratified randomisation. Stat Med. 1982; 1:267-275

Pintilie M. 2006

Competing risks: A practical perspective. Vol. 58. John Wiley & Sons, 2006.

Pocock SJ. 1977

Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64:191-199

Pugh RN et al. 1973

Transection of the oesophagus for bleeding oesophagus varices. Br J Surg 1973;60(8):646-9.

Quan H et al. 2010

Sample size considerations for Japanese patients in a multi-regional trial based on MHLW guidance. Pharm Stat. 2010; 9(2):100–112

Ricardo A et al. 2006

Robust Statistics: Theory and Methods. Wiley, 2006

Robins JM, Tsiatis AA. 1991

Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in statistics 1991;20(8):2609-31.

Robins 1993

Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proceedings of the Biopharmaceutical Section, American Statistical Association 1993;24-33.

Royston P, Parmar MK. 2011

The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. Stat Med. 2011 Aug 30;30(19):2409-21.

Royston P, Parmar MK. 2013

Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013 Dec 7;13:152.

Schoenfeld D. 1981

The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika, 68 (Dec 1981), pp. 219-316

Sellke, T, Siegmund, D. 1983

Sequential analysis of the proportional hazards model. Biometrika. 1983; 70:315-326

Stone A. 2010

The application of bespoke spending functions in group-sequential design and the effect of delayed treatment switching in survival trials. Pharm Stat. 2010; 5:151-161

Stone, A et al. 2015

Model free audit methodology for bias evaluation of tumour progression in oncology. Pharmaceutical statistics. 2015 Nov;14(6):455-63.

Sun J, Zhao Q, Zhao X. 2005

Generalized Log-Rank Tests for Interval-Censored Failure Time Data. Scand J Statist. 2005; 32:49-57

Sun X, Chen C. 2010

Comparison of Finkelstein's Method With the Conventional Approach for Interval-Censored Data Analysis. Stat Biopharm Res. 2010; 2:97-108

Tamhane AC, Mehta CR, Liu L. 2010

Testing a primary endpoint and a secondary endpoint in a group sequential design. Biometrics. 2010; 66(4):1174-1184

Van Hout B et al. 2012

Interim scoring for the EQ-5D-5L: Mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health. 2012; 15(5):708-15

Westfall PH, Young SS. 1993

Resampling-based multiple testing: Examples and methods for p-value adjustment. Wiley, New York 1993

Yau T et al. 2019

CheckMate 459: A randomized, multi-center phase III study of nivolumab (NIVO) vs sorafenib (SOR) as first-line (1L) treatment in patients (pts) with advanced hepatocellular carcinoma (aHCC). Annals of Oncology 30 (2019): v874-v875.

Zhao Q, Sun J. 2004

Generalized log-rank test for mixed interval-censored failure time data. Stat Med. 2004;
23:1621-1629

Certificate Of Completion

Envelope Id: 98D07F1D82FD45EFB46533CFDE37385A

Status: Completed

Subject: Please DocuSign: d419cc00002-sap-ed-4_30JUL2021.docx

Source Envelope:

Document Pages: 108

Signatures: 2

Envelope Originator:

Certificate Pages: 3

Initials: 0

PPD

AutoNav: Enabled

Puerta de Hierro

Envelope Stamping: Disabled

Guadalajara, Jalisco 45116

Time Zone: (UTC-04:00) Atlantic Time (Canada)

PPD

IP Address: PPD

Record Tracking

Status: Original

Holder: PPD

Location: DocuSign

8/6/2021 1:31:25 PM

PPD

Signer Events**Signature****Timestamp**

PPD

PPD

Sent: 8/6/2021 2:42:53 PM

PPD

Viewed: 8/6/2021 2:46:25 PM

Security Level: Email, Account Authentication
(None)

Signed: 8/6/2021 2:46:45 PM

Signature Adoption: Pre-selected Style

Using IP Address: PPD

Electronic Record and Signature Disclosure:

Accepted: 8/6/2021 2:46:25 PM

ID: PPD

PPD

PPD

Sent: 8/6/2021 2:42:52 PM

PPD

Viewed: 8/6/2021 2:43:36 PM

AstraZeneca

Signed: 8/6/2021 2:43:44 PM

Security Level: Email, Account Authentication
(None)

Signature Adoption: Pre-selected Style

Using IP Address: PPD

Electronic Record and Signature Disclosure:

Accepted: 6/23/2021 10:34:51 AM

ID: PPD

In Person Signer Events**Signature****Timestamp****Editor Delivery Events****Status****Timestamp****Agent Delivery Events****Status****Timestamp****Intermediary Delivery Events****Status****Timestamp****Certified Delivery Events****Status****Timestamp****Carbon Copy Events****Status****Timestamp****Witness Events****Signature****Timestamp****Notary Events****Signature****Timestamp****Envelope Summary Events****Status****Timestamps**

Envelope Sent

Hashed/Encrypted

8/6/2021 2:42:53 PM

Certified Delivered

Security Checked

8/6/2021 2:43:36 PM

Envelope Summary Events	Status	Timestamps
Signing Complete	Security Checked	8/6/2021 2:43:44 PM
Completed	Security Checked	8/6/2021 2:46:45 PM
Payment Events	Status	Timestamps
Electronic Record and Signature Disclosure		

I hereby consent to that AstraZeneca Worldwide <https://www.astrazeneca.com> may disclose personal information such as; full name, email address, and any other information you may supply on the electronic form to AstraZeneca affiliates and third party service providers throughout the world in relation to the handling and administration of the Electronic Signature Service solution. This consent relates to any electronic records or signatures associated with the electronic contract.

AstraZeneca and the third party administering this service store and process personal information that AstraZeneca collects from you for the purposes of operating the Electronic Signature Service solution.

This also applies after termination of the Agreement. Processing of your personal information will be done in accordance with applicable law.

You may request access to your personal data and withdraw agreement to this processing at any time by contacting us in writing at PPD

Personal details and electronic signatures of signatories contained in contracts cannot be removed once the contract has been executed and will remain part of such contracts until these are destroyed in accordance with applicable law and AstraZeneca internal data retention policies.