

STATISTICAL ANALYSIS PLAN

A Phase III Open-Label, Multi-Centre, Randomised Study Comparing NUC-1031 plus Cisplatin to Gemcitabine plus Cisplatin in Patients with Previously Untreated Locally Advanced or Metastatic Biliary Tract Cancer

PROTOCOL NUMBER: NuTide:121

PROTOCOL VERSION: Version 4.0, 18 December 2020

STUDY DRUG: NUC-1031 in combination with cisplatin

SAP VERSION: Version 6.0, 25 February 2022

IND NUMBER: 139058

EUDRACT NUMBER: 2019-001025-28

SPONSOR: NuCana plc
3 Lochside Way
Edinburgh, EH12 9DT
UK

CONFIDENTIALITY STATEMENT

This confidential document is the property of NuCana plc. No unpublished information contained herein may be disclosed without prior written approval from NuCana.

Access to this document must be restricted to relevant parties.

REVISION HISTORY

Version Number	Date	Reason for Revision
1	11 March 2019	N/A
2	31 July 2019	<p>Update of study design details in line with protocol update</p> <p>Addition of a modified intention-to-treat analysis population</p> <p>Update of visit window conventions</p> <p>Addition of a definition for ‘date at which patient is last known to be alive’ for the purposes of OS</p> <p>Update of analysis details for OS, ORR, PFS, and DCR</p> <p>Update of details for the interim analysis for power reassessment</p> <p>Update of analysis details for patient-report QoL</p> <p>Addition of subgroup analysis details</p> <p>Addition of Hy’s law definition</p>
3	21 November 2019	<p>Clarification of the procedure for follow-up of patients who discontinue without disease progression</p> <p>Update of details for stratification factors used in the primary analysis of ORR</p> <p>Update of PFS analysis details</p> <p>Update of details for the interim analysis for futility</p>
4	05 February 2021	<p>Update of secondary and supportive analyses for OS</p> <p>Update to Hy’s law criteria to align with FDA guidance</p> <p>Update of analyses for disposition, OS, ORR, PFS, DoR, and AEs, as well as for the futility analysis, to cover additions or modifications that are needed due to the COVID-19 pandemic</p>
5	15 December 2021	<p>Additional details on supportive analyses of OS and PFS</p> <p>Clarifications provided on analysis populations for which data collected prior to randomisation will be summarised</p>


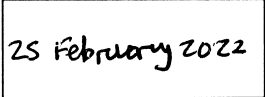
		<p>Additional details provided explicitly on the (SAS default) variance estimators to be used for analyses of ORR</p> <p>Inclusion of explicit details on how, under the CHW method for power re-assessment (Cui et al, 1999), the test statistic for OS needs to be calculated for the final analysis if under the procedure the new required number of events is increased</p> <p>Inclusion of additional safety analyses for SAEs and Deaths. Further clarifications on COVID-19 related safety summaries</p>
6	25 February 2022	<p>Inclusion of additional supportive analyses for OS, ORR, and PFS in which those patients taking particular treatment prior to randomisation are excluded.</p> <p>Inclusion of additional sensitivity analyses for OS, ORR, and PFS in which stratification factors are derived using particular information recorded by the investigator at the Screening Visit</p> <p>Updated the description of the ECG sub-study to align with that sub-study's latest SAP.</p>

STATISTICAL ANALYSIS PLAN SIGNATURE PAGE

NuTide:121 Statistical Analysis Plan version 5.0 (dated 15 December 2021)

Authored by:

Name:	Jonathan R. Smith, Ph.D.
Title:	Consultant Statistician to NuCana
Company:	Adaptive Plus, LLC

Signed:		Date:	
----------------	---	--------------	---

Approved by:

Name:	Elisabeth Oelmann, MD, PhD
Title:	SVP Medical & Clinical Development
Company:	NuCana

I certify that I have read this version of the Statistical Analysis Plan and approve its contents

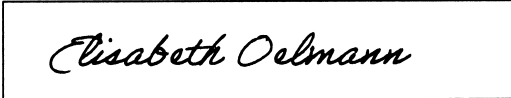
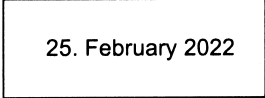
Signed:		Date:	
----------------	---	--------------	---

Table of Contents

Revision History	2
1 ADMINISTRATIVE STRUCTURE AND VERSION HISTORY	13
2 STUDY DESCRIPTION	14
2.1 Study Design.....	14
2.2 Study Treatment.....	15
2.3 Study Objectives and Endpoints	16
2.3.1 Primary Objectives.....	16
2.3.2 Secondary Objectives.....	16
2.3.3 Tertiary Objectives.....	16
2.3.4 Primary Endpoints	16
2.3.5 Secondary Endpoints	16
2.3.5.1 Key Secondary Endpoint	16
2.3.5.2 Other Secondary Endpoints	17
2.3.6 Tertiary Endpoints	18
2.4 Power and Sample Size Justification	18
2.5 Randomisation	19
3 ANALYSIS POPULATIONS	19
3.1 Intention-to-Treat (ITT) Population.....	19
3.2 Intention-to-Treat with Measurable Disease at Baseline (ITTMD) Population	19
3.3 Modified Intention-to-Treat (MITT) Population	20
3.4 Safety Population	20
3.5 Primary Analysis Populations.....	20
4 GENERAL CONVENTIONS	20

4.1	Definition of Baseline	20
4.2	Definition of Study Day	21
4.3	Visit Windows	21
4.4	Handling of Missing Data	21
4.4.1	Imputation of Partial Dates	21
4.4.2	Imputation Rules for Laboratory Values Outside of the Quantification Range	22
4.5	Safety Observation Period	22
4.6	Definition of Prior and Concomitant Medications.....	22
4.7	Software	22
4.8	Changes to Planned Analyses	22
5	STUDY POPULATION SUMMARIES	23
5.1	Disposition	23
5.2	Demographic and Baseline Characteristics.....	23
5.3	Medical History	25
5.4	Cancer History and Current Disease Status	25
5.5	Randomisation Strata compared with Baseline Current Disease Status	25
6	TREATMENTS AND MEDICATIONS	26
6.1	Prior Anti-Cancer and Radiation Therapy	26
6.2	Prior Medications, Concomitant Medications, Prohibited Medications and Prohibited Procedures.....	26
6.3	Non-Protocol Anti-Cancer Therapy (NPACT).....	27
6.4	Study Treatment Exposure and Study Treatment Modifications.....	27
6.5	Post-randomisation Surgery/Procedure, including Biliary Stenting	28
7	EFFICACY ANALYSES	28
7.1	Primary Efficacy Endpoints	28

7.1.1 Overall Survival	28
7.1.1.1 Definition of Overall Survival	28
7.1.1.2 Primary Analysis of Overall Survival.....	28
7.1.1.3 Secondary, Supportive, and Sensitivity Analyses of Overall Survival.....	29
7.1.1.4 Additional Sensitivity Analyses for Overall Survival as a Result of the COVID-19 Pandemic.....	31
7.1.2 Objective Response Rate	31
7.1.2.1 Derivation of Best Overall Response.....	31
7.1.2.2 Primary Analysis of Objective Response Rate	32
7.1.2.3 Secondary, Supportive, and Sensitivity Analyses of Objective Response Rate ...	33
7.1.2.4 Additional Sensitivity Analyses for ORR as a Result of the COVID-19 Pandemic	35
7.2 Key Secondary Efficacy Endpoint.....	35
7.2.1 Progression-Free Survival (PFS)	35
7.3 Interim Analyses	39
7.3.1 Determination of Efficacy at the Interim Analyses	39
7.3.2 Interim Analysis for Assessment of Futility	40
7.3.3 Interim Analysis for Power Re-Assessment	41
7.4 Control of Type 1 Error	42
7.5 Additional Secondary Efficacy Endpoints.....	46
7.5.1 Duration of Response.....	46
7.5.2 18-Month and 12-Month Survival	46
7.5.3 Disease Control Rate.....	46
7.6 Patient-Reported Quality of Life	46
7.6.1 EORTC-QLQ-C30 with the QLQ-BIL21 module.....	46

7.6.2 EuroQoL Health questionnaire instrument (EQ-5D-5L)	48
7.7 Health Economics	49
7.8 Pharmacokinetics	49
7.9 Efficacy Analyses by Subgroups	49
8 SAFETY SUMMARIES.....	50
8.1 Adverse Events, Serious Adverse Events, and Deaths	50
8.1.1 Treatment-Emergent COVID-19 Adverse Events, Including their Impact on the Study 52	
8.2 Laboratory Parameters	53
8.3 Vital Signs.....	54
8.4 ECOG Performance Status	54
8.5 Electrocardiogram.....	55
9 REFERENCES	56
Appendix 1: Derivation of the Objective Response Rate Assumed for the Control Arm	58
Appendix 2: Response Evaluation Criteria in Solid Tumours (RECIST) version 1.1.....	62
Appendix 3: Determination of Best Overall Response in Patients having Intermediate Timepoint Responses of NE	71
Appendix 4: Power Calculations for Subgroup Analyses of ORR	73

List of Tables

Table 1	Censoring scheme used in the Primary Analysis of PFS.....	37
Table 2:	Analyses Planned, Primary Endpoints Evaluated, Approximate Timing, and Drivers of Timing 40	
Table 3:	Efficacy boundaries and properties for ORR.....	44
Table 4:	Efficacy boundaries and properties for OS.....	45
Table 5:	QLQ-C30 domain minimal clinically important differences (MCIDs).....	47
Table 6:	QLQ-BIL21 domain minimal clinically important differences (MCIDs)	48

List of Figures

Figure 1: endpoint	Type 1 error recycling between the two primary endpoints, and the key secondary 43
-----------------------	---

LIST OF ABBREVIATIONS

AE	Adverse event
ALT	Alanine aminotransferase
AST	Aspartate aminotransferase
ATA	Adequate tumour assessment
ATC	Anatomical therapeutic chemical
BICR	Blinded independent central review
BOR	Best overall response
BMI	Body mass index
BTC	Biliary tract cancer
C1D1	Cycle 1 Day 1
CHW	Cui, Hung, Wang
CI	Confidence interval
CMH	Cochran-Mantel-Haenszel
CNAR	Censoring not at random
CR	Complete response
CT	Computed tomography
CTCAE	Common terminology criteria for adverse events
DATE _{LUR}	Date of the latest unconfirmed response
DCR	Disease control rate
DILI	Drug-induced liver injury
DoR	Duration of response
ECG	Electrocardiogram
ECOG	Eastern Cooperative Oncology Group
EMA	European Medicines Agency
EORTC	European Organisation for Research and Treatment of Cancer
eCRF	Electronic case report form
EQ-5D-5L	5-level EuroQol five-dimension scale
FDA	Food and Drug Administration
HR	Hazard ratio
ICH	International Council for Harmonisation
IDMC	Independent Data Monitoring Committee

INR	International Normalised Ratio
ITT	Intention-to-treat
ITTMD	Intention-to-treat with measurable disease at baseline
ITTMD28	Subset of the ITTMD population consisting of patients randomized ≥ 28 weeks prior to the data cut-off
IxRS	Interactive voice or web-based response system
LLQ	Lower limit of quantitation
MCID	Minimal clinically important difference
MedDRA	Medical Dictionary for Regulatory Activities
MITT	Modified intention-to-treat
MRI	Magnetic resonance imaging
NCI	National Cancer Institute
NE	Not evaluable
N_{M28}	Number of patients randomised in the measurable disease stratum with the opportunity for ≥ 28 weeks of follow-up
NPACT	Non-protocol anti-cancer treatment
ORR	Objective response rate
OS	Overall survival
OS12	Proportion of patients alive at 12 months
OS18	Proportion of patients alive at 18 months
PD	Progressive disease
PET	Positron emission tomography
PFS	Progression-free survival
PK	Pharmacokinetic
PR	Partial response
PS	Performance status
QALY	Quality-adjusted life-year
QLQ-BIL21	EORTC questionnaire module for patients with cholangiocarcinoma and gallbladder cancer
QLQ-C30	EORTC questionnaire to assess the quality of life cancer patients
QoL	Quality of life
RDI	Relative dose intensity
RECIST	Response Evaluation Criteria in Solid Tumours (version 1.1)
RMST	Restricted mean survival time

SAE	Serious adverse event
SAP	Statistical analysis plan
SD	Stable Disease
SOC	System organ class
TE	Treatment-emergent
TEAE	Treatment-emergent adverse event
ULQ	Upper limit of quantitation
VAS	Visual analogue scale
WBC	White blood cell
WHO-DD	World Health Organisation drug dictionary

1 ADMINISTRATIVE STRUCTURE AND VERSION HISTORY

This study is being conducted under the sponsorship of NuCana plc.

Statistical programming and analyses are being conducted under contract by IQVIA.

This version of the NuTide:121 Statistical Analysis Plan (SAP) is based on Version 4.0 of the NuTide:121 protocol, dated 18 December 2020.

2 STUDY DESCRIPTION

2.1 Study Design

This is an open-label, randomised Phase III study of NUC-1031 in combination with cisplatin (Arm A) compared to gemcitabine in combination with cisplatin (Arm B), administered on Days 1 and 8 of 21-day cycles, in previously untreated patients with locally advanced or metastatic biliary tract cancer (BTC). A total of 828 patients will be randomised in a 1:1 ratio to Arm A or Arm B, and may continue to receive study treatment until documentation of disease progression, evidence of unacceptable treatment-related adverse events (AEs) despite optimal medical management and/or dose modification, or withdrawal of consent.

Patients will be recruited from approximately 130 sites in North America, Europe, and Asia Pacific (Asia and Australasia). Patients are eligible to participate in the study if all of the inclusion criteria are met and none of the exclusion criteria apply.

Eligible patients must have histologically- or cytologically-proven biliary adenocarcinoma, including cholangiocarcinoma (intra- and extra-hepatic biliary ducts), gallbladder or ampullary cancer, that is not amenable to surgical resection and will have had no prior systemic chemotherapy for treatment of locally advanced or metastatic disease. Patients with measurable or non-measurable disease are eligible for study participation in accordance with predefined randomisation strata; however, non-measurable disease is limited to 82 patients. All screening activities must be performed within 21 days prior to Cycle 1 Day 1 (C1D1), including the baseline radiographic assessment which must also be conducted within 21 days prior to C1D1.

In accordance with local standards of care as appropriate, patients may have received prior chemotherapy in the adjuvant setting and/or radiotherapy with or without radio-sensitising low-dose chemotherapy for localised disease if completed at least 6 months prior to randomisation. Patients must have adequate biliary drainage without evidence of biliary obstruction. Patients with disease characterised as combined or mixed hepatocellular/cholangiocarcinoma are not eligible for this study.

Objective disease assessment will be performed by radiologic evaluation and assessed according to Response Evaluation Criteria in Solid Tumours (RECIST) v1.1 criteria. All known or suspected disease sites must be assessed at baseline by either computed tomography (CT), magnetic resonance imaging (MRI) or positron emission tomography-CT (PET-CT) scan. For each patient, the same radiological method, including the use or not of contrast, used at baseline must be used for disease assessment throughout the duration of the patient's participation in the study.

Tumour measurements and disease response assessments are to be performed every 9 weeks (± 7 days) (approximating three cycles) from C1D1 until disease progression. Responses (Complete Response [CR] or Partial Response [PR]) are scheduled to be confirmed at least 28 days and not more than 42 days after the first observation of response. If a patient is found to have a response at the confirmatory scan which has improved from PR to CR, then the improved response must also be confirmed within a further 28-42 day window.

Treatment and study continuation decisions based on radiologic assessments will be made by the treating Investigator. Independent radiologic disease assessments will be performed through a Blinded Independent Central Review (BICR). In addition, assessment of measurable or non-measurable disease as baseline for the purposes of determining the appropriate randomisation

stratum as well as for use in determining whether the 82 patient limit on non-measurable patients has been reached will be based on BICR assessment. Further details on BICR are described in Section 11.2 of the Study Protocol and in the separate BICR charter.

Patients in either treatment group who stop treatment with no evidence of disease progression as defined by RECIST v1.1 criteria will continue to receive scans at regular intervals (at least every 12 weeks [± 14 days]) until disease progression in order to determine duration of overall response and progression-free survival (PFS). This follow-up for both treatment groups should continue (regardless of whether any subsequent anti-cancer medication has been prescribed) until disease progression, or death. Patients in either treatment group who no longer attend clinic every 12 weeks should continue to be followed for progression status (unless progression has already been documented), for details on any new anti-cancer treatments, and for survival status. Patients who stop treatment following an unconfirmed response should still have a confirmatory scan within the 28- to 42-day window, provided that they have not yet started subsequent anti-cancer therapy. If a patient is found to have a response at the confirmatory scan which has improved from PR to CR, then the improved response must also be confirmed within a further 28- to 42-day window, provided that they have not yet started subsequent anti-cancer therapy.

Survival and initiation of new treatments will be assessed approximately every 12 weeks (± 14 days) until death. To ensure current and complete survival data at the time of database locks, updated survival status may be requested prior to but not limited to Independent Data Monitoring Committee (IDMC) review, interim analyses, and the final analysis.

The study will continue until 637 deaths have occurred, unless the results for overall survival (OS) meet the pre-specified criterion at an interim analysis to stop for early demonstration of efficacy (as described in [Sections 7.3.1](#) and [7.4](#)), or unless terminated early on the recommendation of the IDMC.

The study design, schedule of the visits, assessments and conduct are further described within the Study Protocol.

Robust electrocardiogram (ECG) monitoring will be implemented in a subset of patients at a subset of sites to collect QT/QTc data to allow a primary analysis of the QT effect of NUC-1031 in combination with cisplatin relative to the effect of gemcitabine in combination with cisplatin, based on a by-time point analysis. A total of 64 patients (approximately 32 patients in each treatment group) will be enrolled and will have standard 12-lead ECG monitoring conducted during the first 2 cycles of treatment. This will constitute a sub-study and will be reported separately in accordance with a separate SAP.

2.2 Study Treatment

Eligible consenting patients will be randomised on C1D1 in a 1:1 ratio to receive treatment in Arm A (cisplatin at 25 mg/m² followed by NUC-1031 at 725 mg/m²) or Arm B (cisplatin at 25 mg/m² followed by gemcitabine at 1000 mg/m²) on Days 1 and 8 of 21-day cycles.

Criteria for inter-cycle and intra-cycle dose delay and dose modification are specified in the protocol. The reasons for dose delay or dose modification must be captured as AEs in the patient medical record and noted on the electronic case report form (eCRF).

Patients may continue to receive study treatment until documentation of objective progressive disease (PD), evidence of unacceptable treatment-related AEs despite optimal medical management and/or dose modification, or withdrawal of consent. Reasons for treatment discontinuation must be captured in the patient medical record and on the Treatment Discontinuation page of the eCRF.

A patient who is receiving clinical benefit but experiencing toxicity related to the cisplatin component may continue on study receiving single agent NUC-1031 (Arm A) or gemcitabine (Arm B).

2.3 Study Objectives and Endpoints

2.3.1 Primary Objectives

- OS
- Objective response rate (ORR) based on BICR in patients with measurable disease at baseline

2.3.2 Secondary Objectives

- PFS based on BICR
- Duration of response (DoR) based on BICR
- 18- and 12-month survival
- Disease Control Rate (DCR) based on BICR
- Safety
- Pharmacokinetics (PK) of NUC-1031
- Patient-reported Quality of Life (QoL)

2.3.3 Tertiary Objectives

- Health economics
- Assessment of archival tumour sample characteristics that may further an understanding of the mechanism(s) through which the clinical activity of NUC-1031 is achieved

2.3.4 Primary Endpoints

- OS, defined as the time from randomisation to the time of death from any cause
- ORR, defined as the percentage of patients achieving a confirmed complete or partial response to treatment as assessed by BICR according to RECIST v1.1 criteria. This will be assessed only in patients with measurable disease at baseline.

2.3.5 Secondary Endpoints

2.3.5.1 Key Secondary Endpoint

- PFS, based on BICR according to RECIST v1.1 criteria, defined as the time from randomisation to the first observation of objective tumour progression or death from any cause

2.3.5.2 Other Secondary Endpoints

Efficacy

- DoR, as assessed by BICR, defined as the time from initial clinical response (PR or CR that is subsequently confirmed) to the first observation of tumour progression or death from any cause
- 18-month survival
- 12-month survival
- DCR, based on BICR according to RECIST v1.1 criteria, defined as the percentage of patients demonstrating a Best Overall Response (BOR) of CR, PR, or Stable Disease (SD)

Objective disease assessment will be performed radiologically and assessed according to RECIST v1.1 criteria.

Safety

Safety and tolerability will be assessed by evaluation of the following:

- Treatment-emergent AEs (TEAEs), including TEAEs by severity grade using Common Terminology Criteria for Adverse Events (CTCAE) version 5.0 (or higher)
- Serious TEAEs (SAEs)
- Deaths due to TEAEs
- Treatment discontinuations due to TEAEs
- Clinically-significant changes in laboratory parameters
- Changes in Eastern Cooperative Oncology Group (ECOG) status, Physical Exam, ECGs and Vital Signs

A sub-study will be carried out to assess the effect of the NUC-1031 + cisplatin combination on cardiac repolarisation in a subset of patients.

Pharmacokinetics of NUC-1031

Sparse PK sampling will be taken on Cycle 1 Day 1 at the end of infusion, 2 hours after the end of infusion, and 6 hours after the end of infusion, to capture C_{trough} and C_{max} plasma levels.

Patient-Reported QoL

Patient-reported QoL will be assessed using the European Organisation for Research and Treatment (EORTC) QoL Questionnaire (QLQ-C30) with the QLQ-BIL21 module, and the 5-level EuroQoL five-dimension scale (EQ-5D-5L).

2.3.6 Tertiary Endpoints

Health economics

Health economics will be assessed through collection of core health resource use information, using case report forms (eCRFs) to capture procedure codes, days in hospital, and outpatient visits. Health outcomes will be quantified using quality-adjusted life years (QALYs) and a cost-utility analysis will be conducted by creating incremental cost-utility ratios for each of the treatment groups.

Biomarkers

Phenotypic, genotypic, and/or pharmacodynamic characteristics of the tumour cell that may further delineate the mechanism(s) through which NUC-1031 acts.

2.4 Power and Sample Size Justification

For OS, a hazard ratio (HR) of 0.76 has been assumed. With 3 looks (at 67%, 85%, and 100% of the required number of OS events as described in [Table 4](#) within [Section 7.4](#)), use of the Lan-DeMets O'Brien-Fleming-like α -spending function ([Lan & DeMets, 1983](#)), an overall $\alpha=0.020$ one-sided, and 1:1 randomisation, then a total of 637 OS events gives 90.9% power (after allowing for the small power loss from having the futility boundary). As described in [Section 7.4](#), initially $\alpha=0.020$ one-sided is assigned to OS and $\alpha=0.005$ one-sided is assigned to ORR.

A thirty-month duration of enrolment is assumed with gradual ramp-up in enrolment rate over the first 12 months. OS events are assumed to follow an exponential distribution, and a 11.7-month OS median has been assumed for the control arm as seen in the gemcitabine in combination with cisplatin arm from the ABC-02 study ([Valle et al, 2010](#)). The HR of 0.76 then gives a median OS of approximately 15.4 months in the NUC-1031 in combination with cisplatin arm. If the rate of discontinuation of treatment and the rate of discontinuation from the study are both assumed to be comparable to the gemcitabine in combination with cisplatin arm from ABC-02, then 811 patients would result in the last of the 637 events occurring at approximately 48 months. It is also assumed that 2% of patients will be lost to follow-up for OS (with unknown status of dead/alive) and so 828 patients will be randomised.

The gradual ramp-up in enrolment assumes the following percentages of patients are randomised in each of the first 12 months: 0.16%, 0.23%, 0.55%, 1.10%, 1.84%, 2.52%, 3.16%, 3.52%, 3.74%, 3.81%, 3.87%, and 3.90%. Then, for Months 13-30 enrolment is assumed to continue at a constant rate of 3.98% of patients per month.

The target number of OS events at the final analysis may be increased based on the effect size at Interim Analysis 3 using the Cui, Hung, & Wang (CHW) method ([Cui et al, 1999](#)). Further details are provided in [Section 7.3.3](#).

For ORR, a 19% rate is assumed for the control arm. The derivation of this rate from the gemcitabine in combination with cisplatin arms within ABC-02 ([Valle et al, 2010](#)), BT-22 ([Okusaka et al, 2010](#)), and ABC-03 ([Valle et al, 2015](#)) (allowing now for the requirement of confirmation, based on Performance Status (PS) 0 or 1 patients only, including all randomised patients in the denominator, excluding patients with non-measurable disease at baseline, and adjusting for use of BICR rather than Investigator assessment) is provided in [Appendix 1](#). For the NUC-1031 in combination with cisplatin arm, a 31% ORR is assumed, which gives an assumed true odds ratio of 1.92.

With 2 looks for ORR (at 65% and 100% as described in [Table 3](#) within [Section 7.4](#)), use of the Lan-DeMets O'Brien-Fleming-like α -spending function ([Lan & DeMets, 1983](#)), and with an overall $\alpha=0.005$ one-sided, then a total of 644 patients with measurable disease at baseline (together with 418 at the interim analysis) gives 80% power. The two looks will take place 28 weeks (corresponding to three scheduled post-baseline radiographic scans plus a one week visit window) after the last of these required numbers of patients have been randomised.

Additional power calculations for subgroup analyses are summarised in [Section 7.9](#).

2.5 Randomisation

This is a randomised, open-label, controlled study of NUC-1031 in combination with cisplatin versus gemcitabine in combination with cisplatin.

Patients will be randomised by an independent interactive voice or web-based response system (IxRS). The IxRS will receive information prior to C1D1 for each patient on whether or not the patient has measurable disease at baseline as determined by BICR, where that determination will have been made as described in the Imaging Review Charter. At C1D1, the Investigator will enter into the IxRS their site information, whether the patient has metastatic disease (yes or no), and the primary tumour location (gallbladder, intra-hepatic, extra-hepatic or ampullary). The IxRS will then indicate to which treatment group the patient has been randomised, and the study site will also obtain the patient's identification number from the IxRS.

The randomisation will be in a 1:1 ratio to receive treatment in Arm A (NUC-1031 plus cisplatin) or Arm B (gemcitabine plus cisplatin). Randomisation will be stratified by the following 4 factors:

- Measurable disease at baseline (yes, no) as determined by BICR
- Metastatic disease (yes, no)
- Tumour location (gallbladder, intra-hepatic, extra-hepatic/ampullary)
- Region (Asia, non-Asia)

The number of patients with non-measurable disease at baseline as determined by BICR is capped at 82.

3 ANALYSIS POPULATIONS

3.1 Intention-to-Treat (ITT) Population

The ITT population will consist of all patients who are randomised, regardless of whether any study medication was received. Patients will be summarised on the basis of the treatment group to which they were randomised.

3.2 Intention-to-Treat with Measurable Disease at Baseline (ITTMD) Population

The ITTMD population will consist of all patients who are randomised to the stratum corresponding to having measurable disease at baseline (as assessed by BICR), regardless of whether any study medication was received. Patients will be summarised on the basis of the treatment group to which they were randomised.

The sub-population of the ITTMD population consisting of patients randomised ≥ 28 weeks before the data cut-off will be denoted by the ITTMD28 population.

3.3 Modified Intention-to-Treat (MITT) Population

The MITT population will consist of all patients who are randomised and received any study medication. Patients will be summarised on the basis of the treatment group to which they were randomised.

3.4 Safety Population

The Safety population will consist of all patients who are randomised and receive any study medication. Patients will be summarised on the basis of the actual study medication received, *i.e.*, NUC-1031 in combination with cisplatin (Arm A), or gemcitabine in combination with cisplatin (Arm B). Any patients receiving study medication from both arms will be summarised under Arm A.

3.5 Primary Analysis Populations

The ITTMD28 will be the primary analysis population for evaluating ORR and DCR. DoR will be analysed in the subset of ITTMD28 patients who have confirmed response. For evaluating all other efficacy endpoints, the primary analysis population will be the ITT population. The MITT population will be used only for a secondary analysis of the OS primary endpoint. The Safety population will be the primary analysis population for evaluating all safety endpoints.

4 GENERAL CONVENTIONS

The statistical principles applied in the design and planned analyses of this study are consistent with [International Council for Harmonisation \(ICH\) E9](#) and [Food and Drug Administration \(FDA\) Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics \(2018\)](#).

Continuous data will be summarised using descriptive statistics (number of observations, mean, standard deviation, median, 25th and 75th percentiles, minimum, and maximum). Frequencies and percentages will be used for summarising categorical (discrete) data.

P-values will be presented as two-sided in all cases. In addition, p-values for the primary analyses of OS, ORR, and PFS will also be presented as one-sided for consistency with the method for type 1 error control described in Section 7.4.

Confidence intervals (CIs), where presented, will be provided at the 95% level, with the exception of certain analyses of the primary efficacy endpoints and the key secondary endpoint (as described in [Section 7.1](#) and [Section 7.2.1](#)).

A month is operationally defined to be 30.4375 days (=365.25/12).

All summaries will be presented by treatment group unless otherwise specified.

4.1 Definition of Baseline

In general, for efficacy endpoints the last observed measurement prior to randomisation will be considered the baseline measurement.

For safety endpoints the last observation before first dose of study treatment will be considered the baseline measurement unless otherwise specified. 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 first dose where neither time nor a nominal pre-dose indicator are captured will be considered prior to first dose if such procedures are required by the protocol to be conducted before first dose.

4.2 Definition of Study Day

For the purpose of efficacy data summaries, study day is defined with respect to the randomisation date. For visits (or events) that occur on or after randomisation, study day is defined as (date of visit [event] – date of randomisation + 1). For visits (or events) that occur prior to randomisation, study day is defined as (date of visit [event] – date of randomisation).

For the purpose of safety data summary, Day 1 is defined as the date of first dose of study treatment (referred to in the protocol as C1D1). For visits (or events) that occur on or after first dose, dose day is defined as (date of visit [event] – date of first dose of study treatment + 1). For visits (or events) that occur prior to first dose, dose day is defined as (date of visit [event] – date of first dose of study treatment).

Data listings for safety will include day number in numerical format and in CxDy format (referring to Cycle x Day y).

4.3 Visit Windows

The planned analyses do not require the calculation of visit windows.

4.4 Handling of Missing Data

In general, other than for partial dates, missing data will not be imputed and will be treated as missing with the exceptions specified for certain efficacy variables.

4.4.1 Imputation of Partial Dates

Concomitant medication and adverse events start dates

- If the date is completely missing or if year is missing, then do not impute
- If (year is present and month and day are missing) or (year and day are present and month is missing), impute as January 1st
- If year and month are present and day is missing, impute day as first day of the month

Concomitant medication and adverse events end dates

- If the date is completely missing or if year is missing, then do not impute
- If (year is present and month and day are missing) or (year and day are present and month is missing), impute as December 31st, unless this is after the date of death in which case date of death will be used instead
- If year and month are present and day is missing, impute day as last day of the month, unless this is after the date of death in which case date of death will be used instead

In addition, for AEs and concomitant medications if, for a partial start date, the start date could (when also considering the end date) potentially be on the first study medication date, the start date will be imputed with the first study medication date to assume a “worst case” scenario; *e.g.*, AE from UNK-Feb-2014 to 23-Mar-2014 with first study medication date 21-Feb-2014, then the AE start date will be imputed to 21-Feb-2014.

4.4.2 Imputation Rules for Laboratory Values Outside of the Quantification Range

Lab values below the lower limit of quantification (LLQ) that are reported as “<LLQ” or “≤LLQ” in the database will be imputed by $LLQ \times 0.99$ for analysis purposes. The original value will be listed.

Lab values above the upper level of quantification (ULQ) that are reported as “>ULQ” or “≥ULQ” in the database will be imputed by $ULQ \times 1.01$ for analysis purposes. The original value will be listed.

4.5 Safety Observation Period

The safety observation period is defined as time from first dose date of study treatment to the earlier of: the date of the decision to permanently discontinue study treatment +30 days, or date patient withdrew consent, or date of death, or data cut-off date.

Generally, only the safety data (including AEs, laboratory results, vital signs, ECG, ECOG PS, concomitant medications, *etc*) reported during the safety observation period will be analysed and summarised, unless otherwise specified in this plan.

4.6 Definition of Prior and Concomitant Medications

For the purpose of inclusion in summary tables, incomplete medication or radiation start and stop dates will be imputed as detailed in [Section 4.4](#). Based on imputed start and stop dates:

- Prior medications/radiation therapies are defined as medications or radiation therapies with a start date occurring before the date of first dose of study treatment
- Concomitant medications/radiation therapies are defined as medications or radiation therapies that are already being taken or received on the date of first dose day, or which start after the date of first dose day where this start date is before the end of safety observation period

Medications that have a start date before the date of the first dose of study treatment and that have a stop date after the date of the first day of study treatment will be included under both prior medications/radiation therapies and concomitant medications/radiation therapies.

4.7 Software

All analyses will be conducted using SAS Version 9.4 or higher.

4.8 Changes to Planned Analyses

Version 3 of this SAP was finalised prior to randomising the first patient, version 5 of this SAP was finalised prior to the data cut-off for the first interim analysis, and version 6 of this SAP was finalised prior to the IDMC meeting for the first interim analysis. None of the changes to the SAP made in versions 4, 5, or 6 (after randomising the first patient) were substantive, because these do not change the primary analyses of OS, ORR, or PFS, and no changes were made to methods of type 1 error control or to interim analysis boundaries. However, any future changes to the analyses described in the protocol or in approved versions of this plan that are substantive will be fully documented in an addendum to this plan that will also be approved by the Sponsor prior to conducting the analyses. Any such substantive changes will also be documented in the clinical study report.

5 STUDY POPULATION SUMMARIES

5.1 Disposition

The number of patients screened will be summarised; for screen failures, counts for the reasons why they were not randomised will also be provided.

For the ITT population, counts and percentages will be provided by treatment group for each of the following: treated or untreated; treatment ongoing or treatment ended; primary reason for end of treatment as recorded on the eCRF; and whether the patient discontinued the study overall together with by primary reason for discontinuation from the study as recorded on the eCRF. For patients who withdrew consent as the primary reason for discontinuation from the study, separate counts will also be provided for the subset who withdraw consent to be followed for survival status, and the subset who agreed to continue to be followed for survival status only.

For the ITT population, counts and percentages will also be provided for patients randomised before approval of a COVID-19 vaccine in their country, and for patients randomised after approval of a COVID-19 vaccine in their country. Appendix 5 provide details on the date that a COVID-19 vaccine was first approved for emergency use for each country that randomized patients in the NuTide:121 trial.

For each treatment group, the number of patients in each analysis population will be summarised.

Summaries will be provided of any patients randomised that did not satisfy one or more inclusion criterion or who were randomised but satisfied one or more exclusion criterion.

Major protocol deviations will be defined in a separate "Protocol Deviation Management Plan". They will be summarised by reason and overall.

5.2 Demographic and Baseline Characteristics

For demographic and background characteristics at baseline summary statistics will be provided for continuous variables and counts with percentages will be provided for categorical variables. These analyses will be presented for patients in the ITT, ITTMD28, and Safety populations.

5.2.1 Demographic Characteristics

- Age (continuous)
- Age category 1
 - <65 years
 - ≥65 years
- Age category 2
 - <65 years
 - 65 to <75 years
 - 75 to <85 years
 - ≥85 years
- Sex
 - Male
 - Female

- Not reported
- Ethnicity
 - Hispanic or Latino
 - Not Hispanic or Latino
 - Not reported
- Race
 - American Indian or Alaska Native
 - Asian
 - Black or African American
 - Native Hawaiian or Other Pacific Islander
 - White
 - Not reported
 - Other
- Geographic Region
 - Asia
 - Non-Asia
 - North America (Canada/USA)
 - Europe
 - Western Europe
 - Central/Eastern Europe
 - Australia/New Zealand
 - Rest of World

5.2.2 Background Characteristics

- Height in metres
- Weight in kg
- Body mass index (BMI) in kg/m^2 , calculated as $(\text{weight in kg}) / (\text{Height in m})^2$ - continuous
- BMI category
 - $< 25 \text{ kg}/\text{m}^2$
 - $25 - < 30 \text{ kg}/\text{m}^2$
 - $\geq 30 \text{ kg}/\text{m}^2$
- Smoking history
 - Current
 - Former
 - Never

5.3 Medical History

General medical history data will be coded per Medical Dictionary for Regulatory Authorities (MedDRA) version 20.0 (or higher), and summarised for the ITT population.

5.4 Cancer History and Current Disease Status

For the ITT population as well as for the ITTMD28 and Safety populations, cancer history and current disease characteristics data collected related to cancer history will be summarised categorically or with descriptive statistics as appropriate. The following summaries are planned:

- Diagnosis of BTC by histology or cytology (yes, no)
- Time in years to randomisation since diagnosis of BTC as identified by histology or cytology
- Primary tumour location (gallbladder, intra-hepatic, extra-hepatic, ampullary)
- Extent of Disease (locally advanced, metastatic)
- Measurable Disease (yes, no)
 - As assessed by the Investigator
 - As assessed by BICR
- Number of target lesions as identified by the Investigator
 - Overall
 - By organ site
- Number of target lesions as identified by BICR
 - Overall
 - By organ site
- ECOG PS (0, 1, missing)

5.5 Randomisation Strata compared with Baseline Current Disease Status

For the ITT population the occurrence of any cases of mis-stratifications in the randomisation will be summarised by:

- Comparing primary tumour location (gallbladder, intra-hepatic, extra-hepatic/ampullary) as recorded on the eCRF with primary tumour location as used for stratification within the IxRS
- Comparing extent of disease (locally advanced, metastatic) as recorded on the eCRF with extent of disease as used for stratification within the IxRS
- Comparing measurable disease (yes, no) as assessed by BICR with measurable disease (yes, no) as used for stratification within the IxRS

Patient counts, separately by treatment group, will also be provided for the complete cross-classification of the four stratification factors as used within the IxRS.

6 TREATMENTS AND MEDICATIONS

6.1 Prior Anti-Cancer and Radiation Therapy

Prior anti-cancer therapies will be coded per World Health Organisation drug dictionary (WHO-DD).

The following will be summarised categorically or with descriptive statistics as appropriate for all patients in the ITT, ITTMD28, and Safety populations:

- Prior adjuvant chemotherapy or low dose adjuvant chemotherapy with radiotherapy completed at least 6 months prior to study enrolment (yes, no)
- Prior surgery for BTC (yes and completely resected, yes and incompletely resected, no)
- Prior surgery for BTC with subsequent evidence of non-resectable disease that requires systemic chemotherapy (yes, no)
- Prior radiotherapy for local BTC disease with subsequent evidence of non-resectable disease that requires systemic chemotherapy (yes, no)
- Prior photodynamic therapy for local BTC disease with subsequent evidence of disease progression that requires systemic chemotherapy (yes, no)
- Prior photodynamic therapy for localised disease to relieve biliary obstruction with subsequent evidence of disease progression that requires systemic chemotherapy (yes, no)
- Prior surgery to implant a biliary stent (yes, no)
- Stent currently in place (yes, no)
- Type of stent (plastic, metal)
- Length of time that stent has been in place (months)

All prior non-radiation anti-cancer agents will be summarised categorically by Anatomical Therapeutic Chemical (ATC) Class Text and WHO-DD base substance preferred name by treatment group. Listing of these prior non-radiation anti-cancer agents will also be provided.

6.2 Prior Medications, Concomitant Medications, Prohibited Medications and Prohibited Procedures

6.2.1 Prior and Concomitant Medications

Medications recorded on the eCRFs will be coded using WHO-DD. Prior medications and concomitant medications will each be summarised by treatment group in the Safety population by ATC and WHO-DD base substance preferred name.

6.2.2 Prohibited Medications and Prohibited Procedures

An incidence table and a listing-type table will be provided for those Safety population patients who took prohibited medications. In addition, a listing-type table will be provided for Safety population patients who underwent prohibited procedures after randomisation.

An IQVIA clinician will review all concomitant medication and procedure records via an Excel extract from the database. This regular review will be on a case-by-case basis and should a medication meet any of the prohibited criteria detailed in section 10.3.1 or appendix 3 of the study protocol, this will be indicated via a Prohibited Medication flag (Yes/No) that the clinician will add to the Excel file. This prohibited medication flag will be incorporated during the production

of the SDTM datasets, by importing and then merging this external Excel file with the respective raw data panels.

Note: Post-randomisation procedures allowed by the protocol are summarized within Section 6.5.

6.3 Non-Protocol Anti-Cancer Therapy (NPACT)

NPACT will be summarised for the ITT, ITTMD28, and Safety populations by ATC Class Text and WHO-DD base substance preferred name. This will include:

- NPACT given in the period from the time of initial dosing of study medication on C1D1 until the day of last dosing of study medication
- Anti-cancer medications taken after the day of last dosing of study medication (subsequent anti-cancer medications)

The date of first starting (where applicable) any subsequent anti-cancer medication will be listed for all patients in the ITT population.

6.4 Study Treatment Exposure and Study Treatment Modifications

Study treatment exposure will be summarised categorically or with descriptive statistics, as appropriate, in the Safety population and in the ITTMD28 population. The following will be derived for each patient and summarised by treatment group:

- Number of treatment cycles received overall and separately by:
 - Treatment cycles with cisplatin infused and NUC-1031 (Arm A) or gemcitabine (Arm B) infused on each day of the cycle on which infusions are received
 - Treatment cycles with NUC-1031 (Arm A) or gemcitabine (Arm B) infused on each day on which infusions are received, but with no cisplatin infusion during the cycle
 - Treatment cycles with NUC-1031 (Arm A) or gemcitabine (Arm B) infused on each of the two days (within the cycle) on which infusions are received, but with cisplatin infusion given on only one out of those two days
- Number of days with infusions received overall and separately by:
 - Number of days with infusion received containing cisplatin (Arm A and Arm B) and infusion also received containing NUC-1031 (Arm A), or containing gemcitabine (Arm B)
 - Number of days with infusion received containing NUC-1031 (Arm A) or containing gemcitabine (Arm B), but on which day no infusion of cisplatin was given
- Number of infusions delayed separately by:
 - Number of cisplatin infusions delayed
 - Number of NUC-1031 (Arm A) or gemcitabine (Arm B) infusions delayed
- Total dose given during the study
 - Total dose given during the study of NUC-1031 (Arm A), or gemcitabine (Arm B)
 - Total dose of cisplatin given during the study
- Relative dose intensity (RDI)
 - RDI for NUC-1031 (Arm A), or gemcitabine (Arm B)
 - RDI for cisplatin

RDI is calculated as $100\% \times (d/D)$, where d is the actual cumulative dose delivered up to the actual last day of dosing and D is the intended cumulative dose to be delivered up to the actual last day of dosing.

Treatment modifications (interruptions, temporary modifications, permanent reductions) in study drug dosing due to AE/toxicity will be summarised in the Safety population. In this summary patients will be included in the counts for each treatment modification that applies. Counts will also be given for patients with none of these treatment modifications. Such modifications will be summarised separately for: (i) the cisplatin component; (ii) the NUC-1031 (Arm A) or gemcitabine (Arm B) component; and (iii) for either component of the treatment regimen. In addition, permanent discontinuations of cisplatin (where the patient continues on NUC-1031 [Arm A] or gemcitabine [Arm B]) will also be summarised.

6.5 Post-randomisation Surgery/Procedure, including Biliary Stenting

Post-randomisation surgery/procedures that impacted the tumour lesion (Yes, No, Unknown) will be summarised by treatment group for patients in the ITT and ITTMD28 populations.

7 EFFICACY ANALYSES

7.1 Primary Efficacy Endpoints

7.1.1 Overall Survival

7.1.1.1 Definition of Overall Survival

Duration of OS is defined as the time from randomisation to the time of death due to any cause. For patients who are alive at the time of a data cut-off or are permanently lost to follow-up, duration of OS will be censored at the date at which they were last known to be alive.

The date at which the patient is last known to be alive is defined as the latest date of: (i) last site visit; (ii) last date at which the patient had a radiographic scan; and (iii) last date at which the patient, the study investigator, their other physicians, or a family member confirmed that the patient was alive.

When expressed in months:

$$OS = (\text{earliest date of death or censoring} - \text{date of randomisation} + 1) \times [12/365.25]$$

7.1.1.2 Primary Analysis of Overall Survival

The primary analysis of OS will be performed using the ITT population. As described further in [Section 7.3.1](#) (with summary in Table 2) as well as in [Section 7.4](#) (as illustrated there in Table 4), OS will be assessed for demonstration of efficacy on 3 occasions, *i.e.*, after approximately 425 OS events (67%), after 541 OS events (85%), and after 637 OS events (100%). At each of these 3 looks the treatment effect for OS will be assessed by the stratified log-rank test. The one-sided p-value will be compared with the critical values as derived in [Section 7.4](#) (and illustrated in Table 4) to determine if statistical significance has been obtained at that time.

To determine the stratification factors (primary tumour location, extent of disease, measurable disease at baseline, and region) to be included within the stratified log-rank test the following rule is used: consider the 24 combinations of the stratification factor levels (as used within the IxRS at randomisation), *i.e.*, primary tumour location (gallbladder, intra-hepatic, extra-hepatic/ampullary), extent of disease (locally advanced, metastatic), measurable disease at baseline (yes, no), and

region (Asia, non-Asia). If for any of the 24 combinations (cells) there are less than 10 OS events, then the factor (when considered separately) that has the category which contains the smallest number of OS events will be removed from the stratified analysis. The removal of stratification factors will continue until all cells contain at least 10 OS events. The rule will be implemented at Interim Analysis 2 which is the first look at which OS determination of efficacy is to be made. The same resultant set of stratification factors will also be used (without change) for analyses of OS at Interim Analysis 3 and at the Final Analysis. It will also be used for the analysis of PFS.

A stratified Cox proportional hazards model with Efron's method of handling ties will be used to estimate the HR for OS. The model will include a term for treatment and the same set of stratification factors as used within the stratified log-rank test. The 100 (1-2 α) % two-sided CI for the HR will also be derived where " α " is the "Critical p-value (1-sided)" for OS derived as described in [Section 7.4](#) and illustrated there in Table 4.

Overall counts of deaths will also be summarized together with counts for censored patients subdivided into those still in survival follow-up, patients who are lost to follow up, and patients who have withdrawn consent to continue to be followed for survival status. The non-parametric Kaplan-Meier method will also be used to estimate the survival curves, within which Greenwood's formula will be used to derive the standard error. The number of patients at risk at the start of each 3-month timepoint after randomisation will also be displayed. The median survival time will be estimated for each treatment group, together with its 95% CI using the [Brookmeyer & Crowley \(1982\)](#) method. Estimates of the 25th and 75th percentiles will also be derived together with their corresponding CIs.

7.1.1.3 Secondary, Supportive, and Sensitivity Analyses of Overall Survival

A secondary analysis of OS based on a stratified Cox proportional hazards model will be carried out in the same manner as described in [Section 7.1.1.2](#), but with terms for treatment and ECOG status at baseline (0 or 1), as well as including terms for any of the design stratification factors which (based on the rule given above) are not able to be included as analysis stratification factors.

Subgroup analyses of OS for each of the randomisation stratification factors as well as for other important factors will be provided as described in [Section 7.9](#).

The analyses of OS described in [Section 7.1.1.2](#) will also be carried out based on the MITT population (which is the subset of the ITT population including only those patients that received any study medication), and these analyses will be viewed as supportive.

In addition, the analyses described in [Section 7.1.1.2](#) will be carried out on a subset of the ITT population which excludes those patients for whom either of the following conditions related to particular prior treatment were identified at Screening: (i) Prior treatment with NUC-1031, gemcitabine, cisplatin, or other platinum-based agents (Condition 1); or (ii) Prior systemic therapy for advanced or metastatic biliary tract cancer other than as allowed in the adjuvant setting as specified in the Protocol's wording of exclusion criterion 2 (Condition 2). These analyses will also be viewed as supportive.

Sensitivity analyses for the stratified log-rank test and for the stratified Cox proportional hazards model will be carried out for OS based on the ITT population, as described in [Section 7.1.1.2](#), but where the stratification factors (for primary tumour location and for extent of disease) are derived using information recorded by the investigator at the Screening Visit (rather than using the information that the investigator entered into the IxRS at randomisation). To determine the

stratification factors to be used in the stratified log-rank test and in the stratified Cox proportional hazards model an analogous procedure to that described in [Section 7.1.1.2](#) will be used but now starting with the 24 combinations of: (i) primary tumour location as recorded by the investigator at the Screening Visit; (ii) extent of disease as recorded by the investigator at the Screening Visit; (iii) measurable disease (yes, no) at baseline as entered by the BICR into the IxRS; and (iv) Region.

The methods described in [Section 7.1.1.2](#) assume censoring is at random. Since censoring may depend upon progression, and progression is expected to be affected by treatment group, it is possible that censoring may be informative, *i.e.*, may vary systematically between treatment groups. Therefore, methods which assume censoring not at random (CNAR) will be used to assess the robustness of the inference from the primary analysis for OS.

To allow straightforward comparison with the planned Cox proportional hazards regression analysis, multiple imputation of time to event using the Cox model will be used to implement the CNAR analyses ([Lipkovich *et al*, 2016](#)). Full details of the methodology and its assumptions are presented in the paper cited. The analysis will use the SAS macro publicly available at missingdata.org.uk under the “DIA working group” tab, at the “Imputation approaches” option under “Multiple imputation for time to event data under Kaplan-Meier, Cox or piecewise-exponential frameworks – SAS macros”, downloadable as `Package_Release_V3 Final.zip`.

A tipping point approach will be used. For all patients censored in the primary analysis of OS for whom a progression has occurred, irrespective of treatment group, this analysis will as a first step assume censoring at random but will then impose a multiple of the estimated hazard of death – say *delta* – before imputing censored times to event. In a tipping point analysis, the quantity *delta* is successively increased in a sequence of analyses, *i.e.*, the assumption about post-censoring hazard is made more severe. Note that the patients thus CNAR in the tipping point approach would be only those patients who are censored in the primary analysis (excluding patients administratively censored due to reaching the data cut-off date) and who progressed at any time whether the progression occurred while on study treatment, or if it occurred after discontinuing study treatment, and even if it occurred after starting another cancer treatment. Depending upon the patterns of censoring there may be a *delta* large enough such that inference made by the primary analysis no longer holds – this would be the “tipping point”. If the “tipping point” *delta* exists and is judged unrealistically high, then this provides evidence to help assess the robustness of inference from the primary analysis to the assumption of censoring at random.

The Restricted Mean Survival Time (RMST) method may also be conducted for OS to account for any possible non-proportional hazards and to estimate the absolute benefit of experimental treatment. Here RMST would be evaluated at: (i) the minimum of the maximum observed event time of each arm (minimax event time); (ii) 24 months; (iii) 18 months; and (iv) 12 months. Treatment groups would then be compared in each case using the SAS RMSTREG procedure (SAS Institute Inc. 2020). In addition to a term for treatment the model will include covariables (as CLASS variables) corresponding to the variables that are identified as stratification factors in the primary analysis of OS by the method described in [Section 7.1.1.2](#).

7.1.1.4 Additional Sensitivity Analyses for Overall Survival as a Result of the COVID-19 Pandemic

All analyses for Overall Survival in [Sections 7.1.1.2](#) and [7.1.1.3](#) are based on all-cause mortality, *i.e.*, include OS events with cause of death not reported as COVID-19 as well as OS events with COVID-19 cause of death. All analyses in those two sections also include all OS events regardless of whether or not the patient discontinued treatment due to a COVID-19 infection.

A first set of COVID-19-related sensitivity analyses for OS will be carried out using each method described in [Section 7.1.1.2](#), but where in each case, patients with COVID-19 recorded as the cause of death will be censored at their date of death. Frequencies and percentages based on the ITT population will also be derived for each treatment group separately for: (i) OS with cause of death related to reasons other than COVID-19; and (ii) OS with cause of death reported as COVID-19.

Note: As the eCRF does not include a category of COVID-19 as the primary cause of death the identification of such cases is carried out using the following two-stage process: (i) cases that are potentially COVID-19 deaths are found by first searching the primary cause of death on the death details page of the eCRF for the text 'COVID', 'VIRUS', 'CORONA', or '19'; and then (ii) all cases found by the search in (i) will be manually reviewed to identify cases where 'COVID-19' is the primary cause of death.

A second set of COVID-19-related sensitivity analyses for OS will be carried out using each method described in [Section 7.1.1.2](#), but where patients discontinuing treatment due to a COVID-19 infection will be censored at the date of such treatment discontinuation. Any patients that did not discontinue treatment due to a COVID-19 infection but who died with COVID-19 cause of death will be censored at their date of death. Frequencies and percentages based on the ITT population will also be derived for each treatment group separately for: (i) OS with cause of death related to reasons other than COVID-19 in patients who did not discontinue treatment due to a COVID-19 infection; and (ii) all other deaths.

While these two sets of analyses include all statistical methods described within [Section 7.1.1.2](#), greater emphasis will be placed here on the estimated hazard ratios rather than on statistical significance. This is because the power will be lower than originally planned because fewer OS events will be able to be included in these analyses.

In addition to the sensitivity analyses for OS described above, further sensitivity analyses for OS may ultimately be needed once the nature and extent of the direct and indirect impact of COVID-19 becomes clearer, and such additional analyses will then be described within future versions of this SAP.

7.1.2 Objective Response Rate

7.1.2.1 Derivation of Best Overall Response

Radiographic disease assessment will be made by BICR using RECIST v1.1 for all randomised patients. A detailed description of how RECIST v1.1 is implemented within the current study is given in [Appendix 2](#). For all scheduled and unscheduled visits at which radiographic scans take place, these scans will be reviewed by two independent radiologists, each of whom will record all details as described within [Appendix 2](#). The two radiologists will independently derive each timepoint response categorisation (into CR, PR, SD, PD, or Not Evaluable [NE]) in accordance with [Appendix 2 Table 1](#), *i.e.*, the response obtained overall at each visit by assessing target lesions, non-target lesions, and new lesions. The BICR Charter describes the process by which the reviewer

who made what is regarded as the "definitive assessment" is identified in the case that adjudication is required. All statistical analyses based on BICR assessment will, for a given patient, be based only on the independent reviewer who provided this "definitive assessment".

Then based on the timepoint response categorisations made by this independent reviewer the BOR from BICR will be derived programmatically using the rules given in [Appendix 2 Table 3](#), taking into account:

- (i) Confirmation after at least 28 days is required for CR or PR;
- (ii) For classification as SD, the radiographic assessment on which this is based must have met the SD criteria at least once after C1D1 at a minimum of 8 weeks later (*i.e.*, allowing for the earliest possible time from the 8-10 week window for the first scheduled scan);
- (iii) In patients who discontinue treatment without progression, timepoint responses will be considered in the calculation of BOR only up until the time that the patient receives a subsequent anti-cancer therapy; and
- (iv) The rules given in [Appendix 3](#) for when patients have one or more intermediate visits with timepoint responses of NE.

From the investigator's review of the imaging scans their recorded tumour response data will be used to programmatically determine each patient's timepoint response categorisation (into CR, PR, SD, PD, or NE) in accordance with [Appendix 2 Table 1](#) for patients in the ITTMD population, or categorisation (into CR, non-CR/non-PD, PD, or NE) in accordance with [Appendix 2 Table 2](#) from for patients randomised to the non-measurable disease at baseline stratum. The BOR will then be derived programmatically using the rules given in [Appendix 2 Table 3](#), taking into account (i)-(iv) above.

Note: The independent reviewer's timepoint response categorisation is used prior to programmatically deriving BOR based on BICR assessment. However, for BOR based on investigator assessment the timepoint response is also derived programmatically.

7.1.2.2 Primary Analysis of Objective Response Rate

All analyses of ORR will be based only on patients randomised to the stratum corresponding to having measurable disease at baseline. As described further in [Section 7.3.1](#) and summarised there in [Table 2](#), ORR will be assessed for demonstration of efficacy on two occasions which are scheduled to take place 28 weeks after 418 patients in this stratum, and 28 weeks after 644 patients in this stratum have been randomised. This minimum period of follow-up allows for three scheduled post-baseline radiographic scans plus a one-week visit-window. Let N_{M28} denote the actual number of patients in the ITTMD population that were randomised ≥ 28 weeks before the data cut-off for a given interim analysis, *i.e.*, N_{M28} denotes the number of patients in the ITTMD28 population.

The primary analysis of ORR at a given look will be performed based on the N_{M28} patients in the ITTMD28 population, where ORR is then defined as the proportion of the N_{M28} patients within the treatment group that have BOR of CR or PR based on BICR assessment.

Note: For the first interim analysis all patients randomised on the day that the 418th patient in this stratum is randomized will be included in the ITTMD28 population. Likewise, for the second interim analysis all patients randomised on the day that the 644th patient in this stratum is randomized will be included in the ITTMD28 population.

Note: At the time of the data cut-off for each of the two interim analyses at which ORR is assessed, there will be some of the N_{M28} patients who have an unconfirmed response based on investigator assessment at their most recent radiographic scan (with no earlier confirmed response based on investigator assessment) but who have not yet had their confirmatory radiographic scan. The date of the latest unconfirmed response over all such patients ($DATE_{LUR}$) will be determined. Radiographic assessments (by BICR and by the investigator) from those confirmatory scans which are scheduled to take place within at most 6 weeks after $DATE_{LUR}$ will be included at the interim analyses only in the calculation of counts with confirmed CR, counts with confirmed PR, and in ORR.

At each of these two looks, the odds ratio for ORR will be assessed by the stratified Cochran-Mantel-Haenszel (CMH) test. The one-sided p-value will then be compared with the critical values as derived in [Section 7.4](#) and illustrated in [Table 3](#) to determine if statistical significance has been obtained at that time for ORR.

Measurable disease (yes/no) will not be used as a stratification factor for the CMH test because all ITTMD patients will be randomised to the measurable disease at baseline stratum. To determine the remaining stratification factors (primary tumour location, extent of disease, and region) to be included within the stratified CMH test the following rule is used: consider the 12 combinations of the stratification factor levels (as used within the IxRS at randomisation), *i.e.*, primary tumour location (gallbladder, intra-hepatic, extra-hepatic/ampullary), extent of disease (locally advanced, metastatic), and region (Asia, non-Asia). If for any of the 12 combinations (cells) there are less than 10 patients, then the factor (when considered separately) that has the category which contains the smallest number of patients will be removed from the stratified analysis. The removal of stratification factors will continue until all cells contain at least 10 patients. The rule will be implemented on the N_{M28} patients at Interim Analysis 1 which is the first look at which ORR determination of efficacy is to be made. The same resultant set of stratification factors will also be used (without change) for analysis of ORR at Interim Analysis 2.

Note: In analyses of ORR the values from the IxRS at randomisation will be used for the stratification factors, even if it is subsequently found that these values were incorrect for one or more patient.

The Mantel-Haenszel estimate of the common odds ratio will be derived, together with the 100 (1-2 α) % two-sided CI for the odds ratio, using the Robins et al (1986) variance estimate, where " α " is the "Critical p-value (1-sided)" for ORR derived as in [Section 7.4](#) and illustrated there in [Table 3](#).

Counts for patients with a BOR of CR, PR, SD, PD, or NE, based on BICR assessment, will also be presented by treatment group for the N_{M28} patients.

Clopper-Pearson exact 95% two-sided CIs for ORR based on BICR assessment will be provided separately by treatment group ([Clopper & Pearson, 1934](#)).

7.1.2.3 Secondary, Supportive, and Sensitivity Analyses of Objective Response Rate

Let N_M denote the total number of patients in the ITTMD population that were randomised before the data cut-off for a given interim analysis (regardless of how long beforehand they were randomised). As a supportive analysis, the CMH test as described in [Section 7.1.2.2](#) will also be carried out for ORR using BICR assessment, based on the N_M patients. Counts for patients with a BOR of CR, PR, SD, PD, or NE, based on BICR assessment, will also be presented by treatment group for the N_M patients.

Counts will also be provided by treatment group for the number out of the N_M patients in the ITTMD population that have tumour resection. Counts will also be provided separately by the number of such patients that: (i) previously had a confirmed PR; or (ii) did not previously have a confirmed PR.

A secondary analysis of ORR will also be provided in terms of the difference in proportions based on BICR assessment. This estimate of the difference in ORRs will be based on Mantel-Haenszel weights (using the same stratification factors derived using the methodology described in [Section 7.1.2.2](#)). The p-value and the corresponding 100 (1-2 α) % two-sided CI will be calculated for the N_{M28} patients in the ITTMD population using the Sato variance estimator (Sato, 1989).

Subgroup analyses of ORR based on BICR assessment will be provided as described in [Section 7.9](#).

The analyses of ORR based on BICR assessment as described in [Section 7.1.2.2](#) will also be carried out based on the subset of the N_{M28} patients in the ITTMD population patients that received any treatment, and these analyses will be viewed as supportive.

In addition, the analyses of ORR based on BICR assessment as described in [Section 7.1.2.2](#) will also be carried out on the subset of the ITTMD28 population which excludes those patients for whom Condition 1 or Condition 2 (related to particular prior treatment, as defined in [Section 7.1.1.3](#)) applies. These analyses will be viewed as supportive.

Sensitivity analyses for the stratified CMH test will be carried out for ORR based on the ITTMD28 population, as described in [Section 7.1.2.2](#), but where the stratification factors (for primary tumour location and for extent of disease) are derived using information recorded by the investigator at the Screening Visit (rather than using the information that the investigator entered into the IxRS at randomisation). To determine the stratification factors to be used in the stratified CMH test an analogous procedure to that described in [Section 7.1.2.2](#) will be used but now starting with the 12 combinations of: (i) primary tumour location as recorded by the investigator at the Screening Visit; (ii) extent of disease as recorded by the investigator at the Screening Visit; and (iii) Region.

To provide a more detailed understanding of anti-tumour activity, the maximum tumour reduction since baseline in target lesions will be derived for the subset of those N_{M28} patients from the ITTMD population who have baseline and at least one post-baseline measure. The maximum percentage tumour reduction from baseline in target lesions for each group will be displayed graphically using waterfall plots. For each patient, data from time points after the first date of any of the censoring outcomes defined for the primary analysis of PFS as described in [Section 7.2.1.2](#) will be excluded from the waterfall plots.

Further supportive analyses of ORR will be carried out based on Investigator assessment of response, within which CR, PR, SD, PD, NE, and then BOR will all be determined programmatically. All analyses described for BICR assessment (except for the subgroup analyses) will also be conducted based on Investigator assessment.

In addition, a concordance analysis will be performed in which separately for each treatment group the categories (PR/CR or SD/PD/NE) of BOR for BICR assessment will be presented cross-classified with the corresponding categories based on Investigator assessment. For this concordance analysis, only those ITTMD28 population patients that are considered by the Investigator as having measurable disease at baseline will be included, *i.e.*, the analysis will be based on patients considered by both the BICR and the Investigator to have measurable disease at baseline.

7.1.2.4 Additional Sensitivity Analyses for ORR as a Result of the COVID-19 Pandemic

For ORR, additional sensitivity analyses that account for the impact of COVID-19 will be described in a future version of this SAP, once the nature and extent of the impact becomes clearer. The focus of the additional sensitivity analyses will be on effect sizes (*e.g.*, in terms of odds ratios) rather than on statistical significance because the power will be lower than originally planned as fewer patients will be able to be included in these analyses.

Potential impacts of COVID-19 on ORR include death due to COVID-19 prior to response, treatment discontinuation or interruption due to COVID-19, delayed or missing regular scheduled radiographic scans, and delayed or missing confirmatory radiographic scans (following a PR or CR). For radiographic scans these delays could be directly related to COVID-19, such as resulting from the patient's COVID-19 infection, or the delays could be indirectly related to COVID-19, such as caused by regional lockdowns (or by other logistical reasons). These sensitivity analyses will clearly specify whether only the direct impact of COVID-19 is to be taken into account, or whether the indirect impact of COVID-19 will also be taken into account.

7.2 Key Secondary Efficacy Endpoint

7.2.1 Progression-Free Survival (PFS)

7.2.1.1 Definition of Progression-Free Survival

PFS will be calculated as the time from randomisation until objective disease progression or death from any cause, whichever occurs earlier. PFS based on BICR assessment is the single key secondary efficacy parameter.

Progression will be based on radiographic assessments only, and whether or not a patient has a progression will be determined for each scheduled visit and for any unscheduled visits. For patients with target lesions at baseline (and so having measurable disease at baseline), whether or not a patient has PD at a timepoint will be determined by [Appendix 2 Table 1](#). For patients with no target lesions at baseline (and so having non-measurable disease at baseline), whether or not a patient has PD at a timepoint will be determined by [Appendix 2 Table 2](#), where "unequivocal PD" is defined as an increase in overall disease burden that is comparable in magnitude to the increase that would be required to declare PD for measurable disease, *i.e.*, an increase in tumour burden representing an additional 73% increase in volume (which is equivalent to a 20% increase in diameter in a measurable lesion). Examples include an increase in a pleural effusion from 'trace' to 'large', an increase in lymphangitic disease from localised to widespread, or an increase sufficient to require a change in therapy.

7.2.1.2 Analyses for Progression-Free Survival

For all analyses of PFS, the recorded date of radiographic progression is the date of the tumour assessment visit at which progression is declared. If multiple scan dates are associated with a tumour assessment visit, the earliest assessment date within the set will be chosen as the progression date.

Only adequate tumour assessments (ATAs) will be considered in the determination of radiographic progression and censoring dates. Here ATA is defined as one that results in a time point assignment of: response (complete or partial), SD/(non-CR, non-PD), or progression. For the primary analysis of PFS, the censoring scheme described in [Table 1](#) will be used.

The rules within [Table 1](#) are consistent with [European Medicines Agency \(EMA\) Appendix 1 to the guideline on the evaluation of anti-cancer medicinal products in man \(2012\)](#). The rules in [Table 1](#) are also fully consistent with the rules given in [Table C2 from FDA Guidance for Industry: Clinical Trial Endpoints for the approval of non-small cell lung cancer drugs and biologics \(2015\)](#).

The primary analysis of PFS will be performed using the ITT population. As described further in [Section 7.4](#), PFS will be assessed for demonstration of efficacy only at Interim Analysis 2, at which time approximately 534 PFS events are projected to have occurred. PFS will be assessed by the stratified log-rank test, with the same set of stratification factors as identified for use with OS. If statistical significance is obtained for both ORR and OS (see [Section 7.4](#)) then PFS will be tested, at the 2.5% one-sided significance level. A stratified Cox proportional hazards model with Efron's method of handling ties will be used to estimate the HR for PFS, in which the model only includes the term for treatment. The 95% two-sided CI for the HR will also be derived for PFS from this model. If instead statistical significance is obtained for OS but not for ORR (see [Section 7.4](#)) then the above testing will be carried out at the 0.8% one sided significance level and the CI for the HR will be 98.4% two-sided.

Table 1 Censoring scheme used in the Primary Analysis of PFS

Case	Situation	Date of Progression or Censoring	Outcome
1	Incomplete or no baseline tumour assessments	Randomisation	Censored
2	Progression documented between scheduled visits	Earliest time at which progression can be declared: <ul style="list-style-type: none"> • Date of first progression assessment showing new lesion (if progression is based on new lesion); or • Date of first radiological assessment of target lesions showing a predefined increase in the sum of the target lesion measurements 	Progressed
3	No progression	Date of last progression assessment with no documented progression	Censored
4	Treatment discontinuation for undocumented progression	Date of last progression assessment with no documented progression	Censored
5	Death or Progression after treatment discontinuation for toxicity or other non-progression related reason	Date of progression or death, whichever occurs first	Progressed
6	Death or Progression after new anti-cancer treatment started	Date of progression or death, whichever occurs first	Progressed
7	Death before first scheduled PD assessment	Date of death	Progressed
8	Death after adequate progression assessment visit or death after one missed visit (where in either situation there is no previous documentation of progression)	Date of death	Progressed
9	Death or progression after more than one missed visit	Date of progression or death, whichever occurs first	Progressed

Note: Missed visit includes visits that were completely missed or visits that did not have an adequate tumour assessment (ATA), and "progression assessment" in the above censoring rules only refers to visits with ATA.

Overall counts of patients with a PFS event will be provided together with counts for RECIST progression, and counts for death in the absence of progression. For RECIST progression separate counts will also be given for progression based on target lesions, non-target lesions, and new lesions, where for patients who have progression from multiple sources, the earliest source of progression is presented. If the earliest progression event occurs from multiple sources then the progression events will be categorized using the precedence ordering: Target lesions, non-target lesions, new lesions. Counts will also be provided for censored patients subdivided into those progression-free at the time of analysis, patients who are lost to follow up, and patients who have withdrawn consent. For patients who have multiple censoring reasons the earliest of such censorings will be used.

The Kaplan-Meier method will also be used to estimate the survival curves for PFS in each treatment group, within which Greenwood's formula will be used to derive the standard error. The number of patients at risk at the start of each 3-month timepoint will also be displayed. The median will be estimated for each treatment group, together with its CI using the [Brookmeyer & Crowley \(1982\)](#) method. Estimates of the 25th and 75th percentiles will also be derived together with their corresponding CIs.

Supportive analyses will also be carried out for each of the analyses of PFS described above (using the censoring scheme as defined for the primary analysis) based on the subset of the ITT population which excludes those patients for whom Condition 1 or Condition 2 (related to particular prior treatment, as defined in Section 7.1.1.3) applies.

The first sensitivity analysis of PFS will follow the primary analysis except for the following change to Case 6:

- Patients with new anti-cancer treatment started (with no prior documented progression) will be censored at the date of their last progression assessment (prior to starting new anti-cancer treatment).

The second sensitivity analysis of PFS, which is fully consistent with Table C1 from [FDA Guidance for Industry: Clinical Trial Endpoints for the approval of non-small cell lung cancer drugs and biologics \(2015\)](#) will follow the rules from the first sensitivity analysis but will also make the following changes to Cases 5 and 9:

- Patients with treatment discontinuation (for toxicity, or for another non-progression-related reason) and no documented prior progression will be censored at the date of the last progression assessment (prior to treatment discontinuation)
- Patients that die or progress after more than one missed visit will be censored. The date of censoring will be the date of last progression assessment with documented non-progression, or date of randomisation if all post-baseline progression assessments were missed.

Note: In the second sensitivity analysis any patient who meets the conditions for both Case 5 and Case 6 will be censored at the earliest of the two possible censoring dates.

Supportive analyses of PFS based on Investigator assessment will also be carried out using each of the three sets of censoring rules defined above.

Additional sensitivity analyses based on the stratified log-rank test and based on the stratified Cox proportional hazards model will be carried out for PFS (using the censoring scheme as defined for the primary analysis), but where, as described for OS in [Section 7.1.1.3](#), the stratification factors (for primary tumour location and for extent of disease) are derived using information recorded by the investigator at the Screening Visit (rather than using the information that the investigator entered into the IxRS at randomisation). The stratification factors to be used in the stratified log-rank test and in the stratified Cox proportional hazards model will be the same set as identified for the corresponding analysis of OS in [Section 7.1.1.3](#).

In case the proportional hazards assumption is not valid supportive analyses using the RMST method may be conducted for PFS to account for a possible non-proportional hazards effect. Here RMST would be evaluated at: (i) the minimum of the maximum observed event time of each arm (minimax event time); (ii) 18 months; and (iii) 12 months. Treatment groups would then be compared in each case using the SAS RMSTREG procedure (SAS Institute Inc. 2020). In addition to a term for treatment the model will include covariables (as CLASS variables) corresponding to the variables that are identified as stratification factors by the method described in [Section 7.1.1.2](#).

For PFS based on BICR assessment, a sensitivity analysis will be carried out for the ITT population based on [Table 1](#), but where patients with a COVID-19 cause of death and with no prior progression will be censored at their time of death. Additional sensitivity analyses of PFS based on BICR assessment may be needed once the nature and extent of the direct and indirect impact of COVID-19 becomes clearer, and such additional analyses will be described within a future version of this SAP.

7.3 Interim Analyses

7.3.1 Determination of Efficacy at the Interim Analyses

Three interim efficacy analyses are planned in addition to the final analysis:

- The first interim analysis (Interim Analysis 1) will evaluate the ORR primary endpoint. It will be performed 28 weeks after 418 patients in the measurable disease stratum have been randomised
- The second interim analysis (Interim Analysis 2) will evaluate the ORR and OS primary endpoints. It will be the final analysis for ORR and the first interim analysis (for demonstration of efficacy) for OS. It will be performed 28 weeks after 644 patients in the measurable disease stratum have been randomised. It is estimated that approximately 425 deaths will be observed by this time.
- The third interim analysis (Interim Analysis 3) will evaluate the OS primary endpoint for which it will be the second interim analysis (for demonstration of efficacy) and it will take place after 541 deaths have been observed.
- The final analysis will evaluate the OS primary endpoint. It will take place after 637 deaths have been observed, and is expected to occur approximately 48.0 months after the first patient is randomised

PFS, the key secondary endpoint will also be assessed at Interim Analysis 2 and approximately 534 patients are expected to have a PFS event at this time.

The assessment of futility at Interim Analysis 1 is described in [Section 7.3.2](#), and the power re-assessment at Interim Analysis 3 is described in [Section 7.3.3](#).

A summary of the planned analyses for demonstration of efficacy, with timings, and primary endpoints to be evaluated, is given in [Table 2](#).

If ORR crosses its efficacy boundary at Interim Analysis 1, and provided that a further assessment of ORR is not required by regulators, then the driver of timing for Interim Analysis 2 will instead be the occurrence of 425 OS events.

Table 2: Analyses Planned, Primary Endpoints Evaluated, Approximate Timing, and Drivers of Timing

Analysis	Primary Endpoints Assessed for Demonstration of Efficacy	Approximate Time (months)	Driver of Timing
Interim Analysis 1	ORR	~25.5 ^{1,2}	$N_{M28} = 418$ patients
Interim Analysis 2	ORR, OS	~33.1 ^{1,2}	$N_{M28} = 644$ patients (~425 OS ^{1,3} events expected at this time)
Interim Analysis 3	OS	~40.0 ^{1,3}	541 OS events
Final Analysis	OS	~48.0 ^{1,3}	637 OS events

1. Approximate times for all 4 looks and approximate number of OS events at Interim Analysis 2 assume a 30-month duration of enrolment with gradual ramp-up (as described in Section 2.4) over the first 12 months, and a constant rate of enrolment for the next 18 months.
2. Approximate times for the first 2 looks also assume that the percentage of patients with non-measurable disease at baseline is exactly 10%.
3. Approximate times for the last 2 looks and the approximate number of OS events at Interim Analysis 2 further assume that: (i) OS events follow an exponential distribution with a 11.7-month median for the control arm (Arm B) as well as a median of 11.7/0.76 months in Arm A; (ii) The rate of discontinuation of treatment and the rate of discontinuation from the study (for Arm A and for Arm B) are both comparable to the rates seen in the gemcitabine plus cisplatin arm of ABC-02; and (iii) A total of 2% of patients will be lost to follow-up for OS.
Note: PFS, the key secondary endpoint, will also be assessed at Interim Analysis 2.
Abbreviations: N_{M28} =Number of patients randomised in the measurable disease stratum with the opportunity for ≥ 28 weeks of follow-up; ORR=objective response rate; OS=overall survival.

7.3.2 Interim Analysis for Assessment of Futility

At Interim Analysis 1, a futility analysis will take place based on the OS endpoint. A total of approximately 258 OS events are expected to have occurred, which represents approximately 41% of the required number (637) of OS events. The futility boundary is $Z_1 < Z_{FUT}$, where Z_1 is derived from a log-rank test stratified by primary tumour location (gallbladder, intra-hepatic, extra-hepatic/ampullary) and extent of disease (locally advanced, metastatic), where $Z_{FUT} = 0.0$. This boundary can be viewed as corresponding to an effect size favouring the control arm. Kaplan-Meier plots will be produced for OS at Interim Analysis 1 to determine if there is any evidence of delayed onset of efficacy. The value of Z_1 from this stratified log-rank test as well as Kaplan-Meier plots will also be derived for the two additional types of sensitivity analyses that are described in [Section 7.1.1.4](#). These two sensitivity analyses have been included with the aim of at least partially taking account of the COVID-19 pandemic on OS. The IDMC will indicate whether or not they recommend stopping the study for futility. Such a recommendation to stop for futility should be made only if the futility boundary is crossed, there is no indication of delayed onset of OS efficacy, and these findings also apply in the COVID-19-related sensitivity analyses.

The futility boundary is non-binding and so it is ignored in all calculations of Type 1 error. However, for OS power calculations, the impact of the futility boundary is taken into account.

The time of Interim Analysis 1 is based on ORR and the exact number of OS events may differ from 258. If the number of OS events at Interim Analysis 1 is ≤ 248 then Z_{FUT} will be re-calculated as follows:

Let d_1 denote the observed number of OS events at Interim Analysis 1, and similarly let d_2 , d_3 , and d_4 denote the planned number of OS events at looks 2-4 as given in [Table 2](#), *i.e.*, $d_2 = 425$, $d_3 = 541$, and $d_4 = 637$. If Z_i , $i = 1, 2, 3, 4$ denotes the test statistic from a log-rank test conducted on the d_i OS events, then $Z_i \sim N(\Delta[d_i/4]^{0.5}, 1)$, $\rho_{Z_i, Z_j} = [d_i/d_j]^{0.5}$ for $i < j$, and the canonical joint distribution holds ([Jennison & Turnbull, 2000](#)). Multiple simulated datasets with this distribution, for a specified $\Delta = \log(\text{HR})$, can be generated using the SAS SIMNORMAL procedure ([SAS Institute, Inc, 2020](#)).

If the observed number of OS events at Interim Analysis 1 is ≤ 248 then the revised value of Z_{FUT} will be calculated (by consideration of successively smaller values, starting from $Z_{\text{FUT}} = 0$) so as to give a maximal power loss of at most 1% for all of $\text{Med}_1 = 14.0, 14.1, 14.2, 14.3$, or 14.4 , where Med_1 represents a value of the OS median for the NUC-1031 in combination with cisplatin treatment group, and where then $\Delta = \log(\text{Med}_1/11.7)$. The power loss will be equal to the proportion of simulated datasets for which $[Z_1 < Z_{\text{FUT}}]$ and $[Z_i \geq Z_{1-\alpha_i}$ for one or more of $i = 2, 3$, or $4]$, where $Z_{1-\alpha_i}$ are the critical z-statistic values from the overall $\alpha = 0.025$ (one-sided) column of [Table 4](#). Here the specific values of Med_1 have been found (by consideration of many possible values for d_1) to provide a range that contains the value of Δ with largest power loss.

7.3.3 Interim Analysis for Power Re-Assessment

If the study has not crossed the efficacy boundary for OS (at Interim Analysis 2 or at Interim Analysis 3) then a power reassessment will be carried out at Interim Analysis 3 which is scheduled to occur after 541 OS events, and projected to take place 40.0 months after the start of randomisation (see [Table 2](#) in [Section 7.3.1](#)), *i.e.*, projected to occur 10 months after the end of enrolment. This power reassessment will use the CHW method ([Cui et al, 1999](#)), which guarantees that the maximum experimentwise Type 1 error will still be controlled at the required level.

This power reassessment will be carried out by an independent unblinded statistician (who is not a member of the IDMC), and a charter (the Power Reassessment Charter) will describe the full procedure. The IQVIA Biostatistician in support of the IDMC process will provide the value of Z_3 (the z-statistic for OS at Interim Analysis 3) to this independent unblinded statistician. The rule is deterministic and this independent unblinded statistician will provide the new required number of OS events only to NuCana. No information will be provided to NuCana except for the new required number of OS events.

The specific details on the algorithm to be used will be pre-specified only in the Power Reassessment Charter. However, the general form of the algorithm is as follows:

If $Z_A < Z_3 < Z_B$ then the target number of OS events at the final analysis is increased to 701, otherwise the target number of OS events at the final analysis remains at 637.

The interval (Z_A, Z_B) could be viewed as being analogous to a "promising zone", and is chosen on the basis of maximizing overall power with the aim of increasing the required number of events to 701 (which is a 10% increase) when 637 events would be projected to have a high chance of just failing to obtain statistical significance. By definition, the value Z_B will be lower than the value (as given in [Table 4](#) within [Section 7.4](#)) required for statistical significance at Interim Analysis 3.

The Power Reassessment Charter itself (including the values Z_A and Z_B) will be shared with the sponsor, but will not be shared with the IDMC. However, after the power reassessment the sponsor will only be told the new required number of OS events, and so the sponsor will not be able to back-calculate the effect size at Interim Analysis 3 because this new required number of events can only take two possible values (637 or 701).

Let Z_3 denote the z-statistic for the primary analysis of OS calculated at Interim Analysis 3 which is based on an observed d_3 (~541) events. Also let Z_4 denote the corresponding z-statistic at the final analysis if the power re-assessment rule described above determines that the target number of OS events at the final analysis is to remain at 637. In addition, let Z_{4U} denote a z-statistic (based on sufficient statistics) that could be derived if the power re-assessment rule instead determines that the target number of OS events at the final analysis is to be 701, and let d_4^* be the observed number of events (~701) at the final analysis. Under the CHW method (Cui et al, 1999) if there is no increase in the target number of events then Z_4 is still used at the final analysis, but if there is an increase in the target number of events then instead at the final analysis Z_{4CHW} is used where

$$Z_{4CHW} = \sqrt{\frac{d_3}{637}} Z_3 + \sqrt{\frac{637 - d_3}{637}} Z_{34}^*$$

and where

$$Z_{34}^* = Z_{4U} \sqrt{\frac{d_4^*}{d_4^* - d_3}} - Z_3 \sqrt{\frac{d_3}{d_4^* - d_3}}$$

is the incremental z-statistic corresponding to the additional $d_4^* - d_3$ OS events.

Note: If the final analysis under the CHW procedure is based on Z_{4CHW} then the α value will be derived as described in Section 7.4 and illustrated in Table 4 but still using $d_4 = 637$.

If the required number of OS events remains at 637 then as stated in Table 4 within Section 7.4 the difference in medians at the boundary would be 2.16 months if OS is tested at an overall $\alpha=0.020$ one-sided, and would be 2.06 months if OS is tested at an overall $\alpha=0.025$ one-sided. If the required number of OS events is increased to 701 under the power reassessment procedure then (under the same assumptions as given in footnotes 1 and 2 of Table 4) the difference in medians at the boundary would be 2.05 months if OS is tested at an overall $\alpha=0.020$ one-sided, and would be 1.96 months if OS is tested at an overall $\alpha=0.025$ one-sided, *i.e.*, the difference in medians at the boundary would then only be reduced by approximately 0.1 months (~3 days) if the required number of OS events is increased from 637 to 701.

7.4 Control of Type 1 Error

The Maurer & Bretz method is used to provide strong control of Type 1 error across the two primary endpoints and the key secondary endpoint, as well as across the interim analyses (Maurer & Bretz, 2013). Figure 1 shows the initial one-sided α allocated to each of the two primary endpoints, and the arrows indicate how α is recycled between the endpoints.

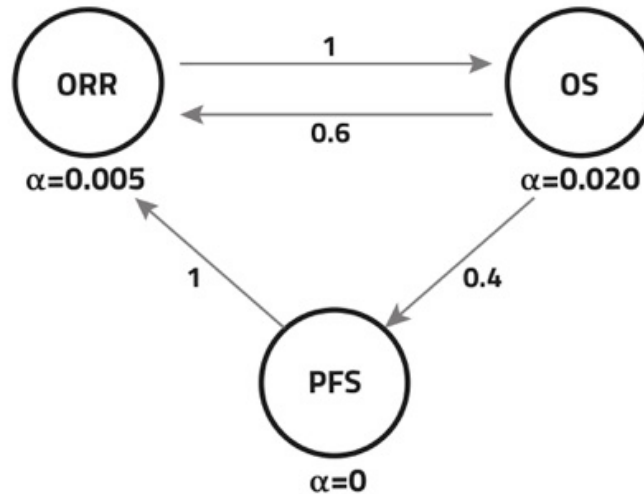


Figure 1: Type 1 error recycling between the two primary endpoints, and the key secondary endpoint

Further details are provided below for each of the primary endpoints and for the key secondary endpoint.

ORR

ORR will be tested using an overall $\alpha=0.005$ one-sided initially. If the null hypothesis for OS is rejected (at Interim Analysis 2, Interim Analysis 3, or at the Final Analysis) then ORR can be tested using an overall $\alpha=0.017$ one-sided. If in addition the null hypothesis for PFS is rejected (at Interim Analysis 2), then ORR can be tested using an overall $\alpha=0.025$ one-sided. The Lan-DeMets O'Brien-Fleming-like α -spending function (Lan & DeMets, 1983) will be used to control the Type 1 error across the two looks for ORR, and Table 3 gives the bounds and boundary properties for ORR testing. The critical values for this α -spending function are derived using the SAS SEQDESIGN procedure with the ERRFUNCBOF design method (SAS Institute, Inc, 2020).

If ORR has not already obtained statistical significance (using the overall $\alpha=0.005$ one-sided) at either Interim Analysis 1 or Interim Analysis 2, then if the null hypothesis for OS is rejected (at Interim Analysis 2, Interim Analysis 3, or at the Final Analysis), but the null hypothesis for PFS is not rejected (at Interim Analysis 2) then the p-value for ORR at Interim Analysis 2 can be compared to the updated alpha value of 0.0160 one-sided (from the $\alpha=0.017$ 1-sided column in Table 3) to determine statistical significance for the ORR primary endpoint. However, if the null hypothesis for OS and the null hypothesis for PFS are both rejected then the p-value for ORR at Interim Analysis 2 can be compared to the updated alpha value 0.0233 one-sided (from the $\alpha=0.025$ 1-sided column in Table 3) to determine statistical significance for the ORR primary endpoint.

The table below, including the required critical values, will be updated using the actual numbers of patients (in the stratum corresponding to measurable disease at baseline who have the opportunity for ≥ 28 weeks of follow-up), if these differ from those given in Table 3.

Table 3: Efficacy boundaries and properties for ORR

Analysis	Value	$\alpha=0.005$ (1-sided)	$\alpha=0.017$ (1-sided)	$\alpha=0.025$ (1-sided)
Interim Analysis 1: 65% $N_{M28} = 418^1$	Critical p-value (1-sided)	0.0005	0.0031	0.0054
	Cumulative α (1-sided)	0.0005	0.0031	0.0054
	Odds Ratio at boundary ²	~ 2.11	~1.88	~ 1.80
	Difference in proportions at boundary ²	~ 14.1%	~11.6%	~ 10.7%
	Cumulative Power ³	29.6%	50.7%	58.3%
Interim Analysis 2: 100% $N_{M28}: 644^1$	Critical p-value (1-sided)	0.0048	0.0160	0.0233
	Cumulative α (1-sided)	0.0050	0.0170	0.0250
	Odds Ratio at boundary ²	~ 1.63	~1.50	~ 1.46
	Difference in proportions at boundary ²	~ 8.6%	~7.1%	~ 6.5%
	Cumulative Power ³	80.0%	90.2%	92.6%
<p>1. N_{M28} denotes the number of patients in the stratum corresponding to measurable disease at baseline who have the opportunity for ≥ 28 weeks of follow-up.</p> <p>2. The approximate odds ratio and approximate difference in proportions at the boundary assume that ORR=19% is observed in the control arm.</p> <p>3. The cumulative power is based on assumed true proportions of 31% vs. 19%.</p>				

OS

OS will be tested using an overall $\alpha=0.020$ one-sided. If the null hypothesis for ORR is rejected (at either Interim Analysis 1 or Interim Analysis 2) then OS can be tested using an overall $\alpha=0.025$ one-sided. The Lan-DeMets O'Brien-Fleming-like α -spending function ([Lan & DeMets, 1983](#)) is used to control the Type 1 error across the three looks for OS, and [Table 4](#) gives the bounds and boundary properties for OS testing. The critical values for this α -spending function are derived using the SAS SEQDESIGN procedure with the ERRFUNCOBF design method ([SAS Institute, Inc, 2020](#)).

[Table 4](#), including the required critical values, will be updated using the actual number of OS events if these differ from the counts given in the table.

Table 4: Efficacy boundaries and properties for OS

Analysis	Value	$\alpha=0.020$ (1-sided)	$\alpha=0.025$ (1-sided)
Interim Analysis 2: 67% N ~ 828 ~ 425 OS events	Z-statistic	2.6198	2.5082
	Critical p-value (1-sided)	0.0044	0.0061
	Cumulative α (1-sided)	0.0044	0.0061
	HR at boundary ¹	~ 0.776	~ 0.784
	Difference in medians at boundary (months) ^{1,2}	~ 3.39	~ 3.22
	Cumulative Power ³	58.3%	62.6%
Interim Analysis 3: 85% N = 828 541 OS events	Z-statistic	2.3164	2.2204
	Critical p-value (1-sided)	0.0103	0.0132
	Cumulative α (1-sided)	0.0116	0.0150
	HR at boundary ¹	~ 0.819	~ 0.826
	Difference in medians at boundary (months) ^{1,2}	~ 2.58	~ 2.46
	Cumulative Power ³	81.4%	83.8%
Final Analysis: 100% N = 828 637 OS events	Z-statistic	2.1366	2.0490
	Critical p-value (1-sided)	0.0163	0.0202
	Cumulative α (1-sided)	0.0200	0.0250
	HR at boundary ¹	~ 0.844	~ 0.850
	Difference in medians at boundary (months) ^{1,2}	~ 2.16	~ 2.06
	Cumulative Power ³	90.9%	92.2%
1. The approximate HR and the approximate difference in medians at the boundary assumes proportional hazards. 2. The approximate difference in medians at the boundary also assumes an 11.7 month observed control arm median. 3. Cumulative power is based on an assumed true HR of 0.76. These power values allow for the possibility that the study could stop for futility at Interim Analysis 1. Abbreviations: HR=hazard ratio; OS=overall survival.			

PFS

If the null hypothesis for ORR (at Interim Analysis 1 or Interim Analysis 2) and the null hypothesis for OS (at Interim Analysis 2, Interim Analysis 3, or at the Final Analysis) are both rejected using the Maurer & Bretz procedure, then PFS can be tested at the $\alpha=0.025$ one-sided significance level. However, if only the null hypothesis for OS is rejected then PFS can be tested at the $\alpha=0.008$ one-sided significance level.

PFS will be analysed at Interim Analysis 2 only, at which point 534 patients are expected to have had a PFS event. This number of patients with a PFS event is based on a control median of 7.6 months, a HR of 0.74, 30-month duration of enrolment with gradual ramp-up over the first 12 months, and a constant rate of enrolment for the next 18 months.

7.5 Additional Secondary Efficacy Endpoints

7.5.1 Duration of Response

DoR as determined by BICR, using RECIST v1.1 criteria, is defined as the time from the first occurrence of a confirmed objective response to the time of disease progression, or death due to any cause, whichever occurs first. The time of progression or death, and censoring, will each be based on the rules used for the primary analysis of PFS as described in [Section 7.2.1.2](#).

DoR will be summarised, for the subgroup of patients in the ITTMD28 population with confirmed response, using the Kaplan-Meier method and will also be displayed graphically. The median event time will be estimated for each treatment group, together with its CI using the [Brookmeyer & Crowley \(1982\)](#) method. Estimates of the 25th and 75th percentiles will also be derived together with their corresponding CIs.

In addition, DoR based on Investigator assessment will be summarised in the same way, and this will be considered a supportive analysis.

For DoR, a sensitivity analysis will be carried out using the rules for the time of disease progression, death, and censoring as used for the primary analysis of PFS described in [Section 7.2.1.2](#), except that patients with a COVID-19 cause of death and with no prior progression will be censored at their time of death. Additional sensitivity analyses of DoR may be needed once the nature and extent of the direct and indirect impact of COVID-19 becomes clearer, and such additional analyses will be described within a future version of this SAP.

7.5.2 18-Month and 12-Month Survival

The proportion of patients alive at 18 months (OS18) will be defined as the Kaplan-Meier estimate of OS at 18 months. For each treatment group, the 95% CI will also be derived using Greenwood's method. A comparison between treatment groups for OS18 will also be carried out using the method described in [Klein et al \(2007\)](#).

The proportion of patients alive at 12 months (OS12) will be analysed in the same way as OS18.

7.5.3 Disease Control Rate

DCR is the proportion of the N_{M28} (defined in [Section 7.1.2.2](#)) patients within the treatment group that have BOR of CR, PR, or SD. The primary analysis of DCR is based on BICR assessment and a supportive analysis will also be conducted based on Investigator assessment.

7.6 Patient-Reported Quality of Life

Patient-reported QoL will be assessed using the EORTC-QLQ-C30 with the QLQ-BIL21 module and the EQ-5D-5L.

The ECOG PS scale (for which the planned analyses are described in [Section 8.4](#)) will be used to assess patients' ability to perform daily living tasks and their range of basic physical ability.

7.6.1 EORTC-QLQ-C30 with the QLQ-BIL21 module

The EORTC QLQ-C30 is a self-administered cancer-specific questionnaire with multi-dimensional scales. It consists of both multi-item scales and single-item measures, including five functional domains, a global QoL domain, three symptom domains, and six single symptom measures. Scoring of the EORTC QLQ-C30 data will be completed following the procedures recommended by the EORTC Study Group on Quality of Life. For each domain or single-item measure, a linear

transformation will be applied to standardise the raw score to range between 0 and 100. QoL data will be analysed to test for clinically meaningful differences between the two treatment groups. Questionnaire compliance rates will be ascertained for each treatment group at each measurement timepoint. Mean baseline scores for each subscale and summary scores will be calculated. Listings will also be provided which where applicable specify the reasons for each patient why the whole EORTC QLQ-C30 questionnaire was not completed at any scheduled visit, and where applicable also provide reasons why data for any of the 30 specific items is missing.

The principal QoL analysis will be based on Kaplan-Meier estimates and log-rank tests on time to definitive QoL deterioration, as measured by the EORTC QLQ-C30 physical function and global QoL scores, in each treatment group. The event of interest is the time to a definitive QoL deterioration, which is defined as the minimum time where two consecutive post-baseline visits have a 5-point or greater deterioration from baseline scores by subtracting baseline scores for each individual from his/her own scores. The 5-point change has been reported to represent “a little” change in QLQ-C30 (Osoba *et al*, 1998, Cocks *et al*, 2012). It has been shown that this is a degree of change that is perceptible to patients (Osoba *et al*, 1998). A minimal change approach is being taken by using a 5-point decrement in physical functioning so that any perceptible change could be captured, but a more definite change may be a 10-point decrement. Indeed, Raman *et al* (2018) have reported asymmetries between improved and worsened states of physical functioning, with worsening requiring roughly a three-fold difference in numeric score compared with the improved state (see Table 5).

Secondary QoL analyses will include analyses of the domains of the QLQ-C30 in which patients are categorised as either improved, stable, or worsened (Raman *et al*, 2018), as described in Table 5.

Table 5: QLQ-C30 domain minimal clinically important differences (MCIDs)

Domain	Improved	Worsened
Physical functioning	≥5	<-15
Role functioning	≥12	<24
Emotional functioning	≥8	<18
Cognitive functioning	≥5	<14
Social functioning	≥8	<20
Fatigue	<-13	≥19
Pain	<9	≥21
Nausea & vomiting	≥-4	≥9
Appetite	<9	≥19

The domain specific minimal clinically important difference (MCID) changes will be used to classify patients as improved, stable, or worsened. As an example, using Table 5, patients with ≥5-point change in Physical Functioning will be classified as improved. If their score is between +5 and -15, then they will be classified as stable. If their score has <-15 change, then they will be classified as worsening.

The Pain domain of the QLQ-C30 will also be summarised relative to changes in analgesic use. Patients' use of analgesics will be categorised into 2-categories (Increased or Stable) relative to their baseline use. Change in actual pain medication will be defined as (Increased/Stable) where a change (Increase) is defined as a doubling of the dose of the same medication OR a switch to a more potent medication.

The EORTC QLQ-BIL21 is a validated disease-specific module for measuring QoL in patients with cholangiocarcinoma and cancer of the gallbladder that is designed to be used with EORTC QLQ-C30. The QLQ-BIL21 is self-administered and consists of 21 questions: 3 single-item assessments relating to treatment side-effects, difficulties with drainage bags/tubes, and concerns regarding weight loss, in addition to 18 items grouped into 5 scales: eating symptoms (4 items), jaundice symptoms (3 items), tiredness (3 items), pain symptoms (4 items), and anxiety symptoms (4 items). Using an anchor-based approach, Kaupp-Roberts *et al* (2016) have reported meaningful changes in each of the domains of the QLQ-BIL21, as shown in Table 6.

Table 6: QLQ-BIL21 domain minimal clinically important differences (MCIDs)

QLQ-BIL21 Domain	Difference
Eating	18
Jaundice	7
Tiredness	22
Pain	15
Anxiety	14
Treatment side effects	16
Drains	18
Weight loss	1

Analysis of QLQ-BIL21 will be carried out in a similar manner as QLQ-C30, using the domain specific MCIDs shown in Table 6. Listings will also be provided which where applicable specify the reasons for each patient why the whole EORTC QLQ-BIL21 questionnaire was not completed at any scheduled visit, and where applicable also provide reasons why data for any of the 21 specific items is missing.

Perceived weight loss is an important indicator of patient well-being and is captured by the QLQ-BIL21. Actual weight patient change from baseline will also be summarised with weight loss defined as a negative change from baseline of 5% or greater.

7.6.2 EuroQoL Health questionnaire instrument (EQ-5D-5L)

The EQ-5D-5L questionnaire will also be administered as a part of this study to assess health-related QoL. EQ-5D is an international, validated, standardised, generic questionnaire for describing and valuing health-related QoL (Rabin & de Charro, 2001). EQ-5D was developed by the EuroQol group in order to provide a simple, generic measure of health for clinical and economic appraisal. This instrument generates a preference-based health-state utility score (EQ-5D utility index) and an overall health state score based on a visual analogue scale (EQ-5D VAS).

EQ-5D is designed for self-completion by respondents and is ideally suited for use in clinics and face to face interviews. It is cognitively undemanding, taking only a few minutes to complete. Instructions to respondents are included in the questionnaire. The most recent version of the EQ-5D is the EQ-5D-5L, which was developed to improve the instrument's sensitivity and to reduce ceiling effects. The number of dimensions (mobility, self-care, usual activities, pain/discomfort, anxiety/depression) has not changed; however, the new version includes 5 levels of severity in each of the existing dimensions in place of three ([EuroQol Group, 2015](#)).

7.7 Health Economics

Health economics will be assessed through collection of core health resource use information, using eCRFs to capture inpatient procedure codes, total number of days in hospital and outpatient visits. Data collected on concomitant medication will also be used in this economic analysis. For the economic modelling, costs will be imputed on the basis of representative country unit costs at the point of analysis using standard fee schedules. Health outcomes will be quantified using QALYs. Quality adjustments will be based on patients' responses to the EQ-5D-5L health status measure which will be administered at baseline, before each cycle of therapy, and each point of follow up as part of the QoL questionnaire. Finally, a cost-utility analysis will be conducted by creating incremental cost-utility ratios for each of the treatment groups.

7.8 Pharmacokinetics

PK analyses are outside the scope of this SAP. A separate PK analysis plan and report will be provided.

7.9 Efficacy Analyses by Subgroups

For OS numbers of events by treatment group in the ITT population, together with HRs (derived from an unstratified Cox proportional hazards model with a single term for treatment within the model) will be provided separately for each of the following subgroups:

- Primary tumour site: gallbladder, intra-hepatic, extra-hepatic, ampullary
- Stage of Disease: Metastatic disease, Locally advanced disease
- ECOG PS (at baseline): 0, 1
- Region: Asia, Non-Asia (with non-Asia also subdivided and provided separately for North America/Western Europe/Australasia combined, and for Central/Eastern Europe/Rest of the World combined)
- Gender: Male, Female
- Age (at baseline): <65, ≥65
- Measurable disease at baseline: Yes, No (based on BICR)

The study is powered on the basis of the overall treatment effect and is not designed to have high power for separate subgroup analyses of the primary endpoints. However, for OS 95% and 80% two-sided CIs (derived from an unstratified Cox proportional hazards model with a single term for treatment within the model) will be provided separately for each of the above subgroups.

For ORR analyses in terms of estimates (of ORR, as well as counts for CR and for PR) by treatment group in the ITTMD28 population, together with odds ratios and difference in proportions of subjects with ORR, each calculated from raw counts, will be given for each of the following subgroups:

- Primary tumour site: gallbladder, intra-hepatic, extra-hepatic, ampullary
- Stage of Disease: Metastatic disease, Locally advanced disease

The odds ratio for ORR will have 95% and 80% two-sided CIs derived using the Baptista-Pike mid-p method (Fagerland *et al*, 2015). For the difference in proportions of patients with ORR 95% and 80% two-sided CIs will be derived using the Newcombe method without continuity correction (Newcombe, 1998, method 10).

Power calculations for subgroup analyses of ORR are provided in Appendix 4, for overall subgroup sizes of 20, 30, 40, or 50 across multiple sets of possible ORR values. Based on these calculations a rationale is provided there for the proposal that 20 patients in total be viewed as sufficient for an adequate assessment of ORR within a subgroup.

8 SAFETY SUMMARIES

All safety analyses will be performed using all patients in the Safety population. No formal statistical comparison between the two treatments groups is planned.

Only results with assessment dates through the end of the safety observation period (see Section 4.5) will be included in any safety analyses, unless otherwise specified. For AE listing-type tables and for the AE data listing, those events within the safety observation period which started or increased in severity after initiating NPACT will be flagged. Similarly, for other safety listing-type tables or data listings, values recorded within the safety observation period after initiating NPACT will be flagged.

An IDMC will monitor safety during the study on a regular basis. The committee will operate independently from the Sponsor and the clinical Investigators. Refer to Section 12.16 of the Study Protocol and the separate IDMC charter for further details.

8.1 Adverse Events, Serious Adverse Events, and Deaths

Adverse event terms recorded on the eCRF will be mapped to preferred terms and system organ class (SOC) using MedDRA version 20.0 (or higher). The severity of AEs will be measured by CTCAE version 5.0 (or higher) guidelines, and for those AEs not included in CTCAE the severity will be categorised as described in Section 12.4 of the Study Protocol.

SAEs will be defined as described in Section 12.1 of the Study Protocol. The relationship of an AE to study medication is categorised as described in Section 12.3 of the Protocol.

Events (including patient deaths) exempt from being reported as AEs or SAEs follow the rules as described in Protocol Section 12.10.

A TEAE is defined as any event with an onset date on or after initiation of study medication (and up to 30 days after the last dose of study medication), or which are present at the time of starting study medication, but which subsequently increase in severity.

For the purpose of calculating treatment emergence and inclusion in summary tables, incomplete onset dates will be imputed as detailed in Section 4.4.

An overall summary of AEs will be provided with the number and percentage of patients who experienced the following types of events during the safety observation period (unless otherwise noted) in each treatment group:

- Patients with any TEAE
- Patients with a Definitely or Probably Related TEAE
- Patients with a Serious TEAE
- Patients with a Serious Definitely or Probably Related TEAE
- Patients with a Worst-Grade 3 or 4 TEAE
- Patients with a Worst-Grade 3 or 4 Definitely or Probably Related TEAE
- Patients with a Worst-Grade 4 TEAE
- Patients with a Worst-Grade 4 Definitely or Probably Related TEAE
- Patients with a Grade 5 TEAE
- Patients with a Grade 5 Definitely or Probably Related TEAE
- Patients with a TEAE leading to Treatment Discontinuation of both components (cisplatin and NUC-1031 [Arm A], cisplatin and gemcitabine [Arm B]) of study medication
- Patients with a TEAE leading to Treatment Discontinuation of the cisplatin component of study medication
- Patients with a TEAE leading to Dose Reduction of either component (cisplatin or NUC-1031 [Arm A], cisplatin or gemcitabine [Arm B]) of study medication
- Patients with a TEAE leading to Dose Reduction of the cisplatin component of study medication
- Patients with a TEAE leading to Dose Reduction of the NUC-1031 (Arm A) or gemcitabine (Arm B) component of study medication
- Patients with a TEAE leading to Interruption of either component (cisplatin or NUC-1031 [Arm A], cisplatin or gemcitabine [Arm B]) of study medication

The following additional summaries in terms of incidences and percentages will be provided:

- TEAEs by SOC and preferred term (all CTCAE Grades)
- TEAEs by SOC and preferred term, separately by worst severity (CTCAE Grade 1/2, 3, 4, or 5)
- TEAEs by SOC and preferred term, separately by strongest relationship
- TEAEs CTCAE Grades 3 or higher by SOC and preferred term, separately by strongest relationship
- Deaths due to all causes by primary cause of death (Progressive disease, SAE, COVID-19, Drug-induced liver injury (DILI), and Unknown).
- TEAEs of CTCAE Grade 5 (deaths) by SOC and preferred term
- Serious TEAEs by SOC and preferred term
- Serious TEAEs by SOC and preferred term, separately by worst severity (CTCAE Grade 1/2, 3, 4, or 5)
- Serious TEAEs by SOC and preferred term, separately by strongest relationship
- TEAEs leading to Treatment Discontinuation by SOC and preferred term
- TEAEs leading to Dose Reduction by SOC and preferred term

- TEAEs leading to Interruption of Study Medication by SOC and preferred term

Listing-type tables will also be provided for SAEs, TEAEs leading to study medication discontinuation (of cisplatin alone or of both components of study medication), TEAEs leading to reduction in dose of study medication (reduction for both components of study medication, reduction for cisplatin alone, or reduction for NUC-1031 [Arm A] or gemcitabine [Arm B] alone), TEAEs leading to interruption of either component of study medication, TEAEs leading to death, and deaths due to all causes.

Note: The eCRF does not include a category of COVID-19 as the primary cause of death and Section 7.1.1.4 describes the process by which cases where 'COVID-19' is the primary cause of death will be identified. The eCRF also did not include a category of DILI as the primary cause of death and so here also the identification of such cases is carried out in a two-stage process: (i) searching the primary cause of death on the death details page of the eCRF for the text term 'LIVER' or for text terms starting with 'HEPAT'; and then (ii) carrying out a manual review to identify those cases from (i) where DILI is the primary cause of death. Analyses will then use DILI or COVID-19 (as applicable) as the primary cause of death in place of the the reason recorded on the eCRF. For the analyses of safety data (but not for the sensitivity analyses of OS described in section 7.1.1.4) if any patients have both DILI and COVID-19 identified in the way described above as the primary cause of death then DILI will take precedence over COVID-19.

8.1.1 Treatment-Emergent COVID-19 Adverse Events, Including their Impact on the Study

The overall incidence of patients with a confirmed case of treatment-emergent (TE) COVID-19 will be derived and a corresponding listing will be provided. In addition, the overall incidence of patients with a confirmed case of TE COVID-19 will be presented by severity. The overall incidence of patients discontinuing study treatment due to TE COVID-19 will also be derived and a corresponding listing will be provided. In addition, a listing from the death details page of the eCRF will be provided for patients with COVID-19 as the primary cause of death.

Note: The procedure for identifying patients with COVID-19 as the primary cause of death is described in Section 7.1.1.4. For identifying COVID-19 AEs the following two-stage procedure will be carried out: (i) cases that are potentially COVID-19 AEs are found by first searching AE preferred terms and high level terms for the text 'COVID', 'VIRUS', 'CORONA', or '19'; and then (ii) all cases found by the search in (i) will be manually reviewed to identify cases where 'COVID-19' is an AE.

Listings will also be provided for:

- (i) Patients delaying, interrupting, or discontinuing treatment due to the patient's confirmed or suspected COVID-19 infection; and
- (ii) Patients taking a concomitant medication for a COVID-19 infection.

Additional analyses of the impact of COVID-19 on the study may ultimately be needed once the nature and extent of the direct or indirect impact of COVID-19 becomes clearer, and such additional analyses will then be described within a future version of this SAP.

8.2 Laboratory Parameters

All clinical laboratory testing will be performed at local sites. Laboratory tests may be performed either on the day of a treatment visit, or during the 3 days prior to a treatment visit.

Samples will be analysed for the following parameters:

- Haematology: white blood cell (WBC) count, differential WBC count, red blood cell count, haemoglobin, haematocrit and platelets
- Coagulation parameters: prothrombin time/international normalised ratio and activated partial thromboplastin time
- Chemistry: sodium, potassium, magnesium, urea or blood urea nitrogen, creatinine, glucose, phosphate, total protein, albumin, adjusted calcium, total bilirubin, bicarbonate, chloride, uric acid, alkaline phosphatase, aspartate aminotransferase (AST), alanine aminotransferase (ALT) and lactate dehydrogenase

The investigator will identify and record clinically significant laboratory test results. The investigator will also record laboratory test results as an AE if it meets any of the four criteria as specified in Section 12.12 of the Protocol. Estimated glomerular filtration rate will also be derived at each applicable visit using the Cockcroft-Gault formula ([Cockcroft & Gault, 1976](#)).

All laboratory data parameters and visits will include flags for values above or below laboratory reference ranges. Toxicity grades will be assigned programmatically by applying the National Cancer Institute (NCI) CTCAE version 5.0 (or higher) guidelines. Only results with assessment dates through the end of the safety observation period will be included in summary tables.

Laboratory summaries will be presented in SI units. Continuous laboratory test results will be summarised by treatment group using descriptive statistics for actual values and for changes from baseline by scheduled visit. Box-Whiskers plots may also be presented at each scheduled visit (with visits shown on x-axis) for some laboratory parameters. For continuous laboratory test results which are below or above the quantification level, the imputed values as described in [Section 4.4](#) will be used for deriving the grade and then summarised.

Laboratory test results that are identified by the investigator as being clinically significant will be summarised in terms of incidence and in listing form.

Incidence tables will be provided for each laboratory parameter reported as a TEAE in which patient counts (and percentages) are given for each CTCAE grade separately (based on worst grade), for Grades 3-4 combined (based on worst grade), and for Grades 1-4 combined. For laboratory parameters which have CTCAE grades for both low and high, the summaries will be produced separately and labelled by the name of the condition, *e.g.*, sodium will be summarised as hyponatraemia and hypernatremia.

Shift tables from baseline to each scheduled study visit will be produced for each laboratory parameter using individual NCI CTCAE grades. Corresponding shift tables will also be produced from baseline to worst value recorded at a post-baseline visit.

Shift tables will also be provided based on reference ranges.

Additional summaries will also be provided for ALT, AST, alkaline phosphatase, total bilirubin, neutrophils, haemoglobin, and platelets.

Cases of potential drug-induced liver injury are defined by the stopping criteria defined in the FDA DILI guidance document (FDA Guidance for Industry: Drug-induced liver injury: premarketing clinical evaluation, 2009) if they meet any of the following four criteria:

1. Treatment-emergent ALT or AST $>8 \times$ ULN
2. Treatment-emergent ALT or AST $>5 \times$ ULN for more than 2 weeks
3. Treatment-emergent ALT or AST $>3 \times$ ULN with total bilirubin $>2 \times$ ULN
4. Treatment-emergent ALT or AST $>3 \times$ ULN and International Normalised Ratio (INR) >1.5

Patients, who meet any of the criteria above and any of the following criteria might potentially meet criteria for Hy's law:

- No initial finding of cholestasis (elevated serum ALP)
- No other reason can be found to explain the increases in ALT/AST and bilirubin or INR, such as viral hepatitis A, B, or C; pre-existing or acute liver disease; biliary duct obstruction or another drug capable of causing the observed injury

For any possible Hy's law cases, patient profiles for all liver function tests will also be produced.

For each of criteria 1-4, counts will be provided for patients meeting the criterion based on ALT but not AST, based on AST but not ALT, based on both ALT and AST, or based on either AST or ALT. Counts will also be provided for patients meeting any of criteria 1-4.

Laboratory parameters will also be summarised in listing form within which the CTCAE grade will be indicated, and flags will be included for those values above or below the reference range.

8.3 Vital Signs

The following vital signs will be summarised:

- Systolic blood pressure (mmHg)
- Diastolic blood pressure (mmHg)
- Pulse rate (bpm)
- Weight (kg)

The investigator will identify and record clinically significant vital sign results. The investigator will also record vital sign results as an AE if it meets any of the four criteria as specified in Section 12.13 of the Protocol.

Vital sign results that are identified by the investigator as being clinically significant will be summarised in terms of incidence and in listing form.

The vital signs test results will also be summarised by treatment group using descriptive statistics for actual values and for changes from baseline by scheduled visit.

8.4 ECOG Performance Status

For the purpose of evaluating safety, ECOG will be summarised in terms of shift from baseline for, at minimum, the worst value recorded after the initiation of study treatment.

Frequencies of ECOG worsening of $\geq +1$ and $+2$ change from baseline to worst value after first dose will also be summarised.

8.5 Electrocardiogram

Standard 12-lead ECG measurements will be obtained at screening and prior to dosing on Day 1 of all cycles. In addition, ECG measurements will be taken within 30 minutes post-infusion on C1D1, C1D8, C2D1, and C2D8. ECG readings should be performed in triplicate. The QTc interval will be calculated using the Fredericia formula and averaged.

Only results with assessment dates through the end of the safety observation period (see [Section 4.5](#)) will be considered for summaries. The following categorical summaries based on the patients' worst value will be presented:

- number of patients with triplicate average QTc >450 msec (for male patients only)
- number of patients with triplicate average QTc >480 msec (separately by gender and for both genders combined)
- number of patients with triplicate average QTc >500 msec (separately by gender and for both genders combined)
- number of patients with increase in triplicate average QTc from baseline of >60 msec (separately by gender and for both genders combined)
- number of patients with increase in triplicate average QTc from baseline of >30 msec (separately by gender and for both genders combined)

Separate shift tables will also be produced for male patients for QTc comparing the baseline value to the patient's worst value using each of the 450, 480, and 500 msec cut-points. Similar shift tables will be produced for female patients using the 480 and 500 msec cut-points.

Shift tables will be produced from baseline to the patient's worst value based on Investigator's overall assessment (normal, abnormal not clinically significant, abnormal clinically significant).

QTc (using the Fredericia formula), QT interval, PR interval, RR interval, QRS interval, and heart rate will also be summarised (on the basis of triplicate averages) in terms of absolute values and change from baseline values.

ECG findings and ECG values will also be provided in listings.

The analyses to be carried out on the data from the robust QT sub-study will be summarised in a separate SAP.

9 REFERENCES

- Agresti A. *Categorical Data Analysis*, 3rd edition, 2013, John Wiley, Hoboken; pages 69–70.
- Bogaerts J, Ford R, Sargent D, *et al.* Individual patient data analysis to assess modifications to the RECIST criteria. *Eur J Cancer* 2009; 45:248–260.
- Brookmeyer R & Crowley J. Confidence interval for the median survival time. *Biometrics* 1982; 38:29-41.
- Clopper CJ & Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; 26:404–413.
- Cockroft DW & Gault MH. Prediction of creatinine clearance from serum creatinine. *Nephron* 1976; 16:31–41
- Cocks K, King MT, Velikova G, *et al.* Evidence-based guidelines for interpreting change scores for the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. *Eur J Cancer* 2012; 48:1713–1721.
- Cui L, Hung HMJ, and Wang S. Modification of sample size in group-sequential clinical trials. *Biometrics* 1999; 55:853–857.
- Eisenhauer EA, Therasse P, Bogaerts J, *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* 2009; 45:228–247.
- European Medicines Agency (EMA): Appendix 1 to the Guideline on the Evaluation of Anti-cancer Medicinal Products in Man. 2012.
- EuroQol Group. Available at <https://euroqol.org/eq-5dinstruments/eq-5d-5l-anout>. Accessed 25 February 2019.
- Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. *Statistical Methods in Medical Research* 2015; 24:224–254.
- FDA Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. December 2018.
- FDA Guidance for Industry: Clinical Trial Endpoints for the Approval of Non-Small Cell Lung Cancer Drugs and Biologics. April 2015.
- FDA Guidance for Industry: Drug-Induced Liver Injury: Premarketing Clinical Evaluation. July 2009.
- ICH Harmonised Tripartite Guideline E9. Statistical principles for clinical trials. 5 February 1998.
- Jennison C & Turnbull BW. *Group- sequential methods with applications to clinical trials*, 2000, Chapman & Hall/CRC, Boca Raton.
- Kaupp-Roberts SD, Yadegarfar G, Friend E, *et al.* Validation of the EORTC QLQ-BIL21 questionnaire for measuring quality of life in patients with cholangiocarcinoma and cancer of the gallbladder. *Br J Cancer* 2016; 115:1032–1038.
- Klein JP, Logan B, Harhoff M, *et al.* Analyzing survival curves at a fixed point in time. *Statist Med* 2007; 26:4505–4519.

- Lan KKG & DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; 70:659–663.
- Lipkovich I, Ratitch B, O’Kelly M. Sensitivity to Censored-at-Random Assumption in the Analysis of Time-to-Event Endpoints. *Pharmaceutical Statistics* 2016; 15:216-229.
- Maurer W & Bretz F. Multiple testing in group sequential trials using graphical approaches. *Stat Biopharm Res* 2013; 5:311–320.
- Newcombe RG. Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine* 1998; 17:873–890
- Okusaka T, Nakachi K, Fukutomi A, *et al.* Gemcitabine alone or in combination with cisplatin in patients with biliary tract cancer: a comparative multicentre study in Japan. *Br J Cancer* 2010; 103:469–474.
- Osoba D, Rodrigues G, Myles J, *et al.* Interpreting the significance of health-related quality of life scores. *J Clin Oncol* 1998; 16:139–144.
- Rabin R & de Charro F. EQ-5D: a measure of health status from the QuroQol Group. *Ann Med* 2001; 33:337–43.
- Raman S, Ding K, Chow E, *et al.* Minimal clinically important differences in the EORTC QLQ-C30 and brief pain inventory in patients undergoing re-irradiation for painful bone metastases. *Quality of Life Research* 2018; 27:1089–1098.
- Robins JM, Breslow N, and Greenland S. Estimators of the Mantel-Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 1986; 42: 311-323.
- SAS Institute Inc. 2020. SAS/STAT® 15.2 User's Guide.
- Sato, T. On the variance estimator of the Mantel-Haenszel risk difference. *Biometrics* 1989; 45: 1323-1324.
- Valle JW, Wasan H, Johnson P, *et al.* Gemcitabine alone or in combination with cisplatin in patients with advanced or metastatic cholangiocarcinomas or other biliary tract tumours: a multicentre randomised phase II study - The UK ABC-01 Study. *Br J Cancer* 2009; 101:621–27.
- Valle JW, Wasan H, Palmer DH, *et al.* Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *N Engl J Med* 2010; 362:1273–1281.
- Valle JW, Wasan H, Lopes A, *et al.* Cediranib or placebo in combination with cisplatin and gemcitabine chemotherapy for patients with advanced biliary tract cancer (ABC-03): a randomised phase 2 trial. *Lancet Oncol* 2015; 16:967–78.

Appendix 1: Derivation of the Objective Response Rate Assumed for the Control Arm

The control arm of the NuTide:121 study is gemcitabine at a dose of 1000 mg/m² in combination with cisplatin at a dose of 25 mg/m², each given on Days 1 and 8 of 21-day cycles.

An extensive Medline search was undertaken to identify randomised studies in patients with previously untreated advanced BTC which included an arm for gemcitabine in combination with cisplatin administered on the schedule to be used in NuTide:121.

A total of three such randomised studies in patients with advanced BTC using the reference regimen were identified, which are briefly summarised below and which form the basis for derivation of the value assumed for ORR in the control arm as used within the power and sample size justification section of this SAP ([Section 2.4](#)).

ABC-02

[Valle *et al*, 2010](#) reported on the results from the ABC-02 study in which patients with previously untreated advanced BTC were randomised to gemcitabine in combination with cisplatin or to single-agent gemcitabine. The earlier [Valle *et al*, 2009](#) publication reported on the ABC-01 study (with the same arms as in ABC-02), but that study will not be considered further here because it was expanded into the ABC-02 study, and the ABC-02 study publication includes all patients already reported in ABC-01.

BT-22

[Okusaka *et al*, 2010](#) reported on the BT-22 study in patients with previously untreated advanced BTC within which gemcitabine in combination with cisplatin was compared to single-agent gemcitabine.

ABC-03

[Valle *et al*, 2015](#) reported on the ABC-03 study in which patients with previously untreated advanced BTC were randomised to gemcitabine in combination with cisplatin or gemcitabine in combination with cisplatin plus cediranib.

In the NuTide:121 study, the primary analysis population for ORR will be the ITTMD population, which consists of all patients randomised with measurable disease at baseline (as assessed by BICR). The denominator for the primary ORR analysis will consist of all patients in the ITTMD population. The numerator for the primary ORR analysis in NuTide:121 will consist of all patients with PR or CR with confirmation at least 28 days later, and will be based on BICR assessment.

As regards the denominator, only the ABC-03 study derived this in the way that will be used in the NuTide:121 study. Furthermore, ABC-02 and BT-22 did not report the numbers of patients that had measurable disease at baseline.

Also, for the numerator part of the calculation, none of these three reference studies required confirmation of response (CR or PR) and results in each case were based on Investigator assessment and not BICR. Thus, the three reference studies reported unconfirmed responses based on Investigator assessment.

Therefore, to derive an estimate of ORR for the control arm to use in sample size calculations for the NuTide:121 study, it has been necessary to start with the count data as reported in these three reference publications.

Note: ABC-03 and BT-22 only enrolled patients with ECOG PS of 0 or 1, whereas ABC-02 enrolled patients with ECOG PS of 0, 1 or 2. However, Professor Valle (*personal communication, 2018*) has provided ABC-02 results related to ORR calculation for the subset of patients with ECOG PS of 0 or 1. The count data for these three studies are summarised below in [Appendix 1 Table 1](#).

Appendix 1 Table 1: Gemcitabine in combination with cisplatin – reported counts from ABC-02, BT-22 and ABC-03 studies

Number of Patients	ABC-02 (PS 0, 1 or 2)	ABC-02 (PS 0 or 1 only)	BT-22	ABC-03
Randomised	204	177	42	62
Randomised and treated	200	NS	41	60
Measurable disease at baseline	NS	31	NS	54
Evaluable at baseline and Evaluable post-baseline	161	146	37	50
CR	1	1	0	0
PR	41	38	8	10
SD	89	84	20	25
PD	30	23	9	15
NE post-baseline	NS	NS	NS	4

Abbreviations: CR=complete response; NE=not evaluable; NS=not specified; PD=progressive disease; PR=partial response; PS=performance status; SD=stable disease.

The calculation of ORR for use in the control arm within the sample size section of the SAP is carried out in the following stages in turn:

- Estimation of the number of patients in ABC-02 and BT-22 that had measurable disease at baseline
- Derivation of pooled and meta-analysis-based estimates of ORR in all randomised patients with measurable disease at baseline, but still based on Investigator assessment and without the requirement for confirmation of response
- Adjustment of the estimate to allow for the requirement for confirmation of response
- Further adjustment of the estimate to allow for assessment needing to be BICR-based (rather than based on Investigator assessment)

Estimation of the number of patients in ABC-02 and BT-22 that had measurable disease at baseline

[Appendix 1 Table 1](#) above shows that 8/62 patients in the gemcitabine/cisplatin arm from ABC-03 had non-measurable disease at baseline, and the [Valle *et al*, 2015](#) paper also reported that the gemcitabine/cisplatin/cediranib arm from that same study had 3/62 with non-measurable disease at baseline. Therefore, at baseline 11/124 (8.9%) patients had non-measurable disease at baseline in ABC-03.

The number of patients with non-measurable disease at baseline is not provided in either ABC-02 or BT-22, but this can be estimated as 8.9% of the patients randomised in the gemcitabine/cisplatin arm, *i.e.*, for ABC-02 (PS 0/1 only) we then obtain 15.25 ($=177 \times 0.089$) with non-measurable disease at baseline, and for BT-22 we obtain 3.74 ($=42 \times 0.089$) with non-measurable disease at baseline. This then gives revised counts as in [Appendix 1 Table 2](#) below:

Appendix 1 Table 2: Gemcitabine in combination with cisplatin – counts from ABC-02, BT-22 and ABC-03 studies with estimates for number of patients with measurable disease at baseline in ABC-02 and BT-22

Number of Patients	ABC-02 (PS 0 or 1)	BT-22	ABC-03
Randomised	177	42	62
Non-measurable disease at baseline	15.75 [#]	3.74 [#]	8
Measurable disease at baseline	146 + 15.25 [#]	37 + 1.26 [#]	50 + 4
CR	1	0	0
PR	38	8	10
SD	84	20	25
PD	23	9	15
NE	15.25 [#]	1.26 [#]	4
ORR based on <u>Investigator Assessment with No Requirement for Confirmation</u> , but now based on all randomised patients with measurable disease at baseline	24.19% (39/161.25)	20.91% (8/38.26)	18.50% (10/54)
[#] =imputed values based on an assumed 8.9% being non-measurable at Baseline. Abbreviations: CR=complete response; NE=not evaluable (post-baseline); ORR=objective response rate; PD=progressive disease; PR=partial response; PS=performance status; SD=stable disease.			

Pooled and Meta-Analysis-based Estimates of ORR in all Randomised Patients with Measurable Disease at Baseline, but based on Investigator Assessment Without the Requirement for Confirmation of Response

A pooled estimate of ORR of 22.48% (57/253.51) is then obtained, based on the counts in [Appendix 1 Table 2](#) above, for ORR based on Investigator assessment without the requirement for confirmation. A formal fixed-effects meta-analysis of the above estimates was also conducted and gave an overall estimate of 22.29%, with $p=0.351$ for Cochran's Q-statistic.

This 22.29% represents the best estimate of ORR based on Investigator assessment without a requirement for confirmation, *i.e.*, it is now based on the denominator that will be used in NuTide:121, but it does not yet use the appropriate numerator, *i.e.*, it does not yet allow for the requirement for confirmation of response and it is still based on Investigator assessments (rather than based on BICR assessments).

Adjustment to Allow for the Requirement of Confirmation of Response

[Bogaerts et al \(2009\)](#) assessed the use of RECIST v1.1 criteria in 16 previous studies in solid tumours with a total of 6,512 patients and considered initially responding (CR or PR) patients who had a follow-up evaluation (or a progression) within eight weeks. They found that 1,858 out of 2,207 initial responders (84.2%) had confirmed responses.

With application of this multiplier of 0.842 to allow for the requirement for confirmation, the unconfirmed ORR estimate of 22.29% in randomised patients with measurable disease at baseline would be decreased to approximately 18.77%.

Further Adjustment to Allow for BICR Rather than Investigator Assessment of Response

There are no published randomised studies in patients with BTC which report results for ORR based on BICR assessment. However, five recent papers reporting on large randomised studies in particular solid tumours were identified for which results were given for ORR assessed by BICR and ORR based on Investigator assessment, with confirmation required in each case. These studies (which were in renal cell carcinoma or in non-small cell lung cancer) included a total of 2469 randomised patients. After simple pooling of the BICR-based ORR results from these studies and simple pooling of the corresponding ORR results based on Investigator assessment, a ratio of 1.034 was derived for BICR-based ORR divided by ORR based on Investigator assessment.

If this multiplier is applied to the 18.77% estimate, then we obtain 19.41% ($= 1.034 \times 18.77\%$). Then, after rounding to the nearest integer percentage we obtain 19% as an estimate for ORR based on all randomised patients with measurable disease at baseline, allowing for the requirement for confirmation, and adjusting for the need for BICR.

This 19% value is then used for sample size calculation related to ORR within [Section 2.4](#) of the SAP.

Appendix 2: Response Evaluation Criteria in Solid Tumours (RECIST) version 1.1

This appendix is identical to Appendix 4 from the Study Protocol.

The following paragraphs provide a detailed description of the implementation of RECIST criteria (v1.1; [Eisenhauer et al, 2009](#)), with slight modifications and additional clarifying text to highlight what is required for this study.

MEASURABILITY OF TUMOUR AT BASELINE DEFINITIONS

At baseline, tumour lesions/lymph nodes will be categorised as measurable or non-measurable as follows.

a. Measurable Tumour Lesions

Tumour Lesions. Tumour lesions must be accurately measured in at least one dimension (longest diameter in the plane of measurement is to be recorded) with a minimum size of:

- 10 mm by CT or MRI scan (CT/MRI scan slice thickness/interval no greater than 5 mm)
- 10 mm calliper measurement by clinical examination (lesions that cannot be accurately measured with callipers should be recorded as non-measurable)

Malignant Lymph Nodes. To be considered pathologically enlarged and measurable, a lymph node must be ≥ 15 mm in the short axis when assessed by CT scan (CT scan slice thickness recommended to be no greater than 5 mm). At baseline and in follow-up, only the short axis will be measured and followed. See also notes below on “Baseline Documentation of Target and Non-Target Lesions” for information on lymph node measurement.

b. Non-Measurable Tumour Lesions

Non-measurable tumour lesions encompass small lesions (longest diameter < 10 mm or pathological lymph nodes with ≥ 10 to < 15 mm short axis), as well as truly non-measurable lesions. Lesions considered truly non-measurable include leptomeningeal disease, ascites, pleural or pericardial effusion, inflammatory breast disease, lymphangitic involvement of skin or lung, peritoneal spread, and abdominal masses/abdominal organomegaly identified by physical examination that is not measurable by reproducible imaging techniques.

c. Special Considerations Regarding Lesion Measurability

Bone lesions, cystic lesions, and lesions previously treated with local therapy require particular comment, as outlined below.

Bone lesions:

- Bone scan, PET scan, or plain films are not considered adequate imaging techniques to measure bone lesions. However, these techniques can be used to confirm the presence or disappearance of bone lesions.
- Lytic bone lesions or mixed lytic-blastic lesions, with identifiable soft tissue components, that can be evaluated by cross-sectional imaging techniques such as CT or MRI can be considered measurable lesions if the soft tissue component meets the definition of measurability described above.
- Blastic bone lesions are non-measurable.

Cystic lesions:

- Lesions that meet the criteria for radiographically defined simple cysts should not be considered malignant lesions (neither measurable nor non-measurable) since they are, by definition, simple cysts.
- Cystic lesions thought to represent cystic metastases can be considered measurable lesions if they meet the definition of measurability described above. However, if non-cystic lesions are present in the same patient, these are preferred for selection as target lesions.

Lesions with prior local treatment:

- Tumour lesions situated in a previously irradiated area or in an area subjected to other loco-regional therapy are usually not considered measurable unless there has been demonstrated progression in the lesion.

TARGET LESIONS: SPECIFICATIONS BY METHODS OF MEASUREMENTS

a. Measurement of Lesions

All measurements should be recorded in metric notation, using callipers if clinically assessed. All baseline evaluations should be performed as close as possible to the treatment start and never more than 14 days before the beginning of the treatment.

b. Method of Assessment

The same method of assessment and the same technique should be used to characterize each identified and reported lesion at baseline and during study. Imaging-based evaluation (CT scan, MRI, or PET-CT) should always be used in this study.

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

Chest X-Ray. Chest X-ray should not be used in assessment of lesion size or for detecting new lesions. If, however, a new lesion is detected by X-ray then it should be confirmed by CT scan, MRI, or PET-CT.

CT, MRI, PET-CT. CT is the best currently available and reproducible method to measure lesions selected for response assessment. This guideline has defined measurability of lesions on CT scan on the basis of the assumption that CT slice thickness is 5 mm or less. When CT scans have slice thickness greater than 5 mm, the minimum size for a measurable lesion should be twice the slice thickness. MRI and PET-CT are also acceptable. PET scans are not considered adequate to measure lesions, PET-CT scans may be used providing that the measures are obtained from the CT scan and the CT scan is of identical diagnostic quality to a diagnostic CT (with IV and oral contrast).

If prior to enrolment it is known that a patient is unable to undergo CT scans with IV contrast because of allergy or renal insufficiency, the decision as to whether a non-contrast CT scan or MRI scan (without IV contrast) will be used to evaluate the patient at baseline and during the study should be guided by the tumour type under investigation and the anatomic location of the disease.

For patients who develop contraindications to contrast after a baseline contrast CT scan is done, the decision as to whether non-contrast CT or MRI (enhanced or non-enhanced) scan will be performed should also be based on the tumour type and the anatomic location of the disease and should be optimised to allow for comparison with the prior studies if possible.

Each case should be discussed with the radiologist to determine if substitution of these other approaches is possible and, if not, the patient should be considered not evaluable from that point forward. Care must be taken in measurement of target lesions on a different modality and interpretation of non-target disease or new lesions since the same lesion may appear to have a different size using a new modality.

Ultrasound. Ultrasound is not useful in assessment of lesion size and should not be used as a method of measurement.

Endoscopy, Laparoscopy, Tumour Markers, Cytology, Histology. The utilisation of these techniques for objective tumour evaluation cannot generally be advised.

TUMOUR RESPONSE EVALUATION

ASSESSMENT OF OVERALL TUMOUR BURDEN AND MEASURABLE DISEASE

To assess objective response or future progression, it is necessary to estimate the overall tumour burden at baseline and to use this as a comparator for subsequent measurements. Measurable disease is defined by the presence of at least one measurable lesion, as detailed above.

BASELINE DOCUMENTATION OF TARGET AND NON-TARGET LESIONS

When more than one measurable lesion is present at baseline, all lesions up to a maximum of five lesions total (and a maximum of two lesions per organ) representative of all involved organs should be identified as target lesions and will be recorded and measured at baseline. This means in instances where patients have only one or two organ sites involved, a maximum of two lesions (one site) and four lesions (two sites), respectively, will be recorded. Other lesions (albeit measurable) in those organs will be recorded as non-target lesions (even if the size is >10 mm by CT, MRI, or PET-CT scan).

Target lesions should be selected on the basis of their size (lesions with the longest diameter) and be representative of all involved organs but, additionally, should lend themselves to reproducible repeated measurements. 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 that can be measured reproducibly should be selected.

Lymph nodes merit special mention since they are normal anatomical structures that may be visible by imaging even if not involved by tumour. As noted above, pathological nodes that are defined as measurable and may be identified as target lesions must meet the criterion of a short axis of ≥ 15 mm by CT scan. Only the short axis of these nodes will contribute to the baseline sum. The short axis of the node is the diameter normally used by radiologists to judge if a node is involved by solid tumour. Nodal size is normally reported as two dimensions in the plane in which the image is obtained (for CT scan, this is almost always the axial plane; for MRI scan the plane of acquisition may be axial, sagittal, or coronal). The smaller of these measures is the short axis. For example, an abdominal node that is reported as being 20 mm x 30 mm has a short axis of 20 mm and qualifies as a malignant, measurable node. In this example, 20 mm should be recorded as the node measurement. All other pathological nodes (those with short axis ≥ 10 mm but < 15 mm) should be

considered non-target lesions. Nodes that have a short axis <10 mm are considered non-pathological and should not be recorded or followed.

Lesions irradiated within 3 weeks prior to C1D1 may not be counted as target lesions.

A sum of the diameters (longest for non-nodal lesions, short axis for nodal lesions) for all target lesions will be calculated and reported as the baseline sum of diameters. If lymph nodes are to be included in the sum, then, as noted above, only the short axis is added into the sum. The baseline sum of diameters will be used as a reference to further characterize any objective tumour regression in the measurable dimension of the disease.

All other lesions (or sites of disease), including pathological lymph nodes, should be identified as non-target lesions and should also be recorded at baseline. Measurements are not required and these lesions should be followed as “present,” “absent,” or in rare cases “unequivocal progression.”

In addition, it is possible to record multiple non-target lesions involving the same organ as a single item on the eCRF (*e.g.*, “multiple enlarged pelvic lymph nodes” or “multiple liver metastases”).

RESPONSE CRITERIA

a. Evaluation of Target Lesions

This section provides the definitions of the criteria used to determine objective tumour response for target lesions.

- Complete response (CR): disappearance of all target lesions
 - Any pathological lymph nodes (whether target or non-target) must have reduction in short axis to <10 mm
- Partial response (PR): at least a 30% decrease in the sum of diameters of target lesions, taking as reference the baseline sum of diameters
- Progressive disease (PD): at least a 20% increase in the sum of diameters of target lesions, taking as reference the smallest sum on study (nadir), including baseline
 - In addition to the relative increase of 20%, the sum must also demonstrate an absolute increase of at least 5 mm
 - The appearance of one or more new lesions is also considered progression.
- Stable disease (SD): neither sufficient shrinkage (compared to baseline) to qualify for PR nor sufficient increase (taking as reference the smallest sum of diameters while on study) to qualify for PD

b. Special Notes on the Assessment of Target Lesions

Lymph Nodes. Lymph nodes identified as target lesions should always have the actual short axis measurement recorded (measured in the same anatomical plane as the baseline examination), even if the nodes regress to <10 mm on study. This means that when lymph nodes are included as target lesions, the sum of lesions may not be zero even if CR criteria are met since a normal lymph node is defined as having a short axis <10 mm.

Target Lesions That Become Too Small to Measure. While on study, all lesions (nodal and non-nodal) recorded at baseline should have their actual measurements recorded at each subsequent evaluation, even when very small (*e.g.*, 2 mm). However, sometimes lesions or lymph nodes that

are recorded as target lesions at baseline become so faint on CT scan that the radiologist may not feel comfortable assigning an exact measure and may report them as being too small to measure.

When this occurs, it is important that a value be recorded on the eCRF as follows:

- If it is the opinion of the radiologist that the lesion has likely disappeared, the measurement should be recorded as 0 mm.
- If the lesion is believed to be present and is faintly seen but too small to measure, a default value of 5 mm should be assigned and BML (below measurable limit) should be ticked. (Note: It is less likely that this rule will be used for lymph nodes since they usually have a definable size when normal and are frequently surrounded by fat such as in the retroperitoneum; however, if a lymph node is believed to be present and is faintly seen but too small to measure, a default value of 5 mm should be assigned in this circumstance as well and BML should also be ticked).

To reiterate, however, if the radiologist is able to provide an actual measure, that should be recorded, even if it is below 5 mm, and, in that case, BML should not be ticked.

Lesions That Split or Coalesce on Treatment. When non-nodal lesions fragment, the longest diameters of the fragmented portions should be added together to calculate the target lesion sum. Similarly, as lesions coalesce, a plane between them may be maintained that would aid in obtaining maximal diameter measurements of each individual lesion. If the lesions have truly coalesced such that they are no longer separable, the vector of the longest diameter in this instance should be the maximal longest diameter for the coalesced lesion.

c. Evaluation of Non-Target Lesions

This section provides the definitions of the criteria used to determine the tumour response for the group of non-target lesions. Although some non-target lesions may actually be measurable, they need not be measured and, instead, should be assessed only qualitatively at the timepoints specified in the protocol.

- CR: disappearance of all non-target lesions
 - All lymph nodes must be non-pathological in size (<10 mm short axis)
- Non-CR/Non-PD: persistence of one or more non-target lesion(s)
- PD: unequivocal progression of existing non-target lesions
 - The appearance of one or more new lesions is also considered progression

d. Special Notes on Assessment of Progression of Non-Target Disease

When the Patient Also Has Measurable Disease. In this setting, to achieve unequivocal progression on the basis of the non-target disease, there must be an overall level of substantial worsening in non-target disease in a magnitude that, even in the presence of SD or PR in target disease, the overall tumour burden has increased sufficiently to merit discontinuation of therapy. A modest increase in the size of one or more non-target lesions is usually not sufficient to qualify for unequivocal progression status. The designation of overall progression solely on the basis of change in non-target disease in the face of SD or PR of target disease will therefore be extremely rare.

When the Patient Has Only Non-Measurable Disease. The same general concepts apply here as noted above; however, in this instance, there is no measurable disease assessment to factor into the interpretation of an increase in non-measurable disease burden. Because worsening in non-target disease cannot be easily quantified (by definition: if all lesions are truly non-measurable), a useful test that can be applied when assessing patients for unequivocal progression is to consider if the increase in overall disease burden based on the change in non-measurable disease is comparable in magnitude to the increase that would be required to declare PD for measurable disease; that is, an increase in tumour burden representing an additional 73% increase in volume (which is equivalent to a 20% increase in diameter in a measurable lesion). Examples include an increase in a pleural effusion from “trace” to “large” or an increase in lymphangitic disease from localised to widespread or may be described in protocols as “sufficient to require a change in therapy.” If unequivocal progression is seen, the patient should be considered to have had overall PD at that point. Although it would be ideal to have objective criteria to apply to non-measurable disease, the very nature of that disease makes it impossible to do so; therefore, the increase must be substantial.

e. New Lesions

The appearance of new malignant lesions denotes disease progression; therefore, some comments on detection of new lesions are important. There are no specific criteria for the identification of new radiographic lesions; however, the finding of a new lesion should be unequivocal, that is, not attributable to differences in scanning technique, change in imaging modality, or findings thought to represent something other than tumour (for example, some “new” bone lesions may be simply healing or flare of pre-existing lesions). This is particularly important when the patient’s baseline lesions show PR or CR. For example, necrosis of a liver lesion may be reported on a CT scan report as a “new” cystic lesion, which it is not.

A lesion identified during the study in an anatomical location that was not scanned at baseline is considered a new lesion and will indicate disease progression.

If a new lesion is equivocal, for example because of its small size, continued therapy and follow-up evaluation will clarify if it represents truly new disease. If repeat scans confirm there is definitely a new lesion, then progression should be declared using the date of the initial scan.

EVALUATION OF RESPONSE

a. Timepoint Response (Overall Response)

It is assumed that at each protocol-specified timepoint, a response assessment occurs. [Appendix 2 Table 1](#) provides a summary of the overall response status calculation at each timepoint for patients who have measurable disease at baseline.

When patients have non-measurable (therefore non-target) disease only, [Appendix 2 Table 2](#) is to be used.

Appendix 2 Table 1: Timepoint response: patients with target lesions (with or without non-target lesions)

Target Lesions	Non-Target Lesions	New Lesions	Overall Response
CR	CR	No	CR
CR	Non-CR/non-PD	No	PR
CR	Not evaluated	No	PR
PR	Non-PD or not all evaluated	No	PR
SD	Non-PD or not all evaluated	No	SD
Not all evaluated	Non-PD	No	NE
PD	Any	Yes or no	PD
Any	PD	Yes or no	PD
Any	Any	Yes	PD

Abbreviations: CR=complete response; NE=not evaluable; PD=progressive disease; PR=partial response; SD=stable disease.

Appendix 2 Table 2: Timepoint response: patients with non-target lesions only

Non-Target Lesions	New Lesions	Overall Response
CR	No	CR
Non-CR/non-PD	No	Non-CR/non-PD ^a
Not all evaluated	No	NE
Unequivocal PD	Yes or no	PD
Any	Yes	PD

^a“Non-CR/non-PD” is preferred over “stable disease” for non-target disease since stable disease is increasingly used as an endpoint for assessment of efficacy in some studies; thus, assigning “stable disease” when no lesions can be measured is not advised.

Abbreviations: CR = complete response; NE = not evaluable; PD = progressive disease.

b. Missing Assessments and Not-Evaluable Designation

When no imaging/measurement is done at all at a particular timepoint, the patient is not evaluable at that timepoint. If only a subset of lesion measurements are made at an assessment, usually the case is also considered not evaluable at that timepoint, unless a convincing argument can be made that the contribution of the individual missing lesion(s) would not change the assigned timepoint response. This would be most likely to happen in the case of PD. For example, if a patient had a baseline sum of 50 mm with three measured lesions and, during the study, only two lesions were assessed, but those gave a sum of 80 mm; the patient will have achieved PD status, regardless of the contribution of the missing lesion.

If one or more target lesions were not assessed either because the scan was not done or the scan could not be assessed because of poor image quality or obstructed view, the response for target lesions should be “unable to assess” since the patient is not evaluable. Similarly, if one or more non-target lesions are not assessed, the response for non-target lesions should be “unable to assess” except where there is clear progression. Overall response would be “unable to assess” if either the target response or the non-target response is “unable to assess,” except where this is clear evidence of progression as this equates with the case being not evaluable at that timepoint.

Appendix 2 Table 3: Best overall response allowing for the requirement of confirmation for PR and CR

Overall Response at First Timepoint	Overall Response at Subsequent Timepoint	Best Overall Response
CR	CR	CR
CR	PR	SD, PD, or PR ^a
CR	SD	SD, provided minimum duration for SD was met; otherwise, PD
CR	PD	SD, provided minimum duration for SD was met; otherwise, PD
CR	NE	SD, provided minimum duration for SD was met; otherwise, NE
PR	CR	PR
PR	PR	PR
PR	SD	SD
PR	PD	SD, provided minimum duration for SD was met; otherwise, PD
PR	NE	SD, provided minimum duration for SD was met; otherwise, NE
NE	NE	NE

^a If a CR is truly met at the first timepoint, any disease seen at a subsequent timepoint, even disease meeting PR criteria relative to baseline, qualifies as PD at that point (since disease must have reappeared after CR). Best response would depend on whether the minimum duration for SD was met. However, sometimes CR may be claimed when subsequent scans suggest small lesions were likely still present and in fact the patient had PR, not CR, at the first timepoint. Under these circumstances, the original CR should be changed to PR and the best response is PR.

Note: For a best overall response of SD the timepoint on which this is based must be at least 8 weeks after C1D1. Abbreviations: CR=complete response; NE=not evaluable; PD=progressive disease; PR=partial response; SD=stable disease.

c. Special Notes on Response Assessment

When nodal disease is included in the sum of target lesions and the nodes decrease to “normal” size (<10 mm), they may still have a measurement reported on scans. This measurement should be recorded even though the nodes are normal in order not to overstate progression should it be based on increase in size of the nodes. As noted earlier, this means that patients with CR may not have a total sum of “zero” on the eCRF.

Patients with a global deterioration of health status requiring discontinuation of treatment without objective evidence of disease progression at that time should be reported as “symptomatic deterioration.” Every effort should be made to document objective progression after discontinuation of treatment. Symptomatic deterioration is not a descriptor of an objective response; it is a reason for stopping study therapy. The objective response status of such patients is to be determined by evaluation of target and non-target disease as shown in Appendix 2 Table 1-Table 3.

For equivocal findings of progression (*e.g.*, very small and uncertain new lesions; cystic changes or necrosis in existing lesions), treatment may continue until the next scheduled assessment. If at the next scheduled assessment progression is confirmed, the date of progression should be the earlier date when progression was suspected.

If a patient undergoes an excisional biopsy or other appropriate approach (*e.g.*, multiple passes with large core needle) of a new lesion or an existing solitary progressive lesion that following serial sectioning and pathological examination reveals no evidence of malignancy (*e.g.*, inflammatory cells, fibrosis, *etc*), then the new lesion or solitary progressive lesion will not constitute disease progression.

The present study is one in which patients with advanced disease are eligible (*i.e.*, primary disease still or partially present), the primary tumour should also be captured as a target or non-target lesion, as appropriate. This is to avoid an incorrect assessment of CR if the primary tumour is still present but not evaluated as a target or non-target lesion.

Appendix 3: Determination of Best Overall Response in Patients having Intermediate Timepoint Responses of NE

[Appendix 2 Table 3](#) describes the rules for determination of BOR allowing for the requirement of confirmation for PR and CR. However, as noted in section 4.4.4 of [Eisenhauer et al \(2009\)](#), when confirmation of response is required it is necessary to define how missing data/assessments will be addressed in determination of response.

Appendix 3 here provides rules for defining BOR when patients have a timepoint response of PR or CR, followed by one or more intermediate visits with timepoint responses of NE, and then followed by a timepoint response of PR or CR (at least 28 days after the previous PR or CR).

The first case of this type is

- 1) CR timepoint response, followed by one or more NE timepoints responses, followed by CR timepoint response - BOR = CR

For Case 1 above where BOR is classified as CR, this also requires that the timepoint responses with NE either had no imaging assessment in all cases, or if one or more assessments with NE did have imaging undertaken then all of the following were satisfied at each such timepoint:

- (i) any subset of target and non-target lesions that were assessed are found to have disappeared;
- (ii) any lymph nodes assessed have short axis <10 mm; and
- (iii) no new lesions were identified

The rule above for Case 1 patients would take precedence over the BOR classification given in [Appendix 2 Table 3](#), e.g., if the first CR timepoint response was immediately preceded by a PR timepoint response (giving a sequence of PR-CR-NE-CR) then [Appendix 2 Table 3](#) would have classified such a patient as having BOR = PR, but the rule above re-classifies such a patient as having BOR = CR.

The next cases considered here are:

- 2) PR timepoint response, followed by one or more NE timepoints responses, followed by PR timepoint response - BOR = PR
- 3) PR timepoint response, followed by one or more NE timepoints responses, followed by CR timepoint response - BOR = PR

For Cases 2 and 3 where BOR is classified as PR, this also requires that the timepoint responses with NE either had no imaging assessment in all cases, or if one or more assessments with NE did have imaging undertaken then all of the following were satisfied at each such timepoint:

- (iv) the subset of target lesions that were assessed have a sum of diameters such that if all of the non-lymph node target lesions not assessed had a zero value, and the lymph node target lesions not assessed had a 10 cm value, then it would still be possible to demonstrate at least a 30% decrease relative to baseline;
- (v) any non-target lesions are non-PD; and
- (vi) no new lesions were identified

For Case 3 if the CR timepoint response is followed by a further CR then the BOR classification given in [Appendix 2 Table 3](#) takes precedence over the rule given above, e.g., a sequence of PR-NE-CR-CR would be classified as BOR = PR above, but [Appendix 2 Table 3](#) would have classified such a patient as having BOR = CR and the CR takes precedence over PR in calculation of BOR.

The remaining case considered here is:

- 4) CR timepoint response, followed by one or more NE timepoints responses, followed by PR timepoint response - BOR would use the rules for CR followed by PR as given in [Appendix 2 Table 3](#) to classify as PR, SD, or PD, but the additional conditions (iv)-(vi) above would need to be satisfied for assessments with NE in order for the BOR to be classified as PR

Appendix 4: Power Calculations for Subgroup Analyses of ORR

Section 7.9 indicates that confidence intervals will be provided for ORR by primary tumour site (gallbladder, intra-hepatic, extra-hepatic, ampullary) and by stage of disease (metastatic disease, locally advanced disease). Estimates of ORR (as well as counts for CR and for PR) by treatment group, together with odds ratios and difference in proportions of subjects with ORR will be given for these subgroups. In addition, the odds ratio for ORR will have 95% and 80% two-sided CIs derived using the Baptista-Pike mid-p method (Fagerland *et al*, 2015). For the difference in proportions of patients with ORR, 95% and 80% two-sided CIs will also be derived using the Newcombe method without continuity correction (Newcombe, 1998, method 10).

This appendix provides details on power and other related calculations for subgroup sizes of 20, 30, 40, and 50 (corresponding to 10, 15, 20, and 25 per arm). Appendix 4 Table 1 below identifies the 9 cases in which power calculations have been derived, where θ_1 represents the true ORR in Arm A and θ_2 represents the true ORR in Arm B.

Appendix 4 Table 1: Sets of Parameters values for which Power Calculations have been Derived

Case	θ_1	θ_2	$\theta_1 - \theta_2$	Odds Ratio
1	0.28	0.17	0.11	1.90
2	0.34	0.17	0.17	2.52
3	0.39	0.17	0.22	3.12
4	0.31	0.19	0.12	1.92
5	0.37	0.19	0.18	2.50
6	0.42	0.19	0.23	3.09
7	0.34	0.21	0.13	1.94
8	0.40	0.21	0.19	2.51
9	0.45	0.21	0.24	3.08

Case 4 represents the parameter values assumed for the overall sample size calculations where the corresponding odds ratio is 1.92 and the corresponding difference in proportions is 0.12. Cases 1 and 7 correspond to approximately the same odds ratio as case 4 but assume slightly lower (Case 1) and slightly higher (Case 7) values for the true ORR in Arm B. Only small differences in the assumed ORR for Arm B are considered because the value for θ_2 was derived (as described in Appendix 1) from past data on approximately 250 patients.

For Arm A, in addition to values of θ_1 which correspond to an odds ratio of approximately 1.92 (Cases 1, 4, 7), values which correspond to odds ratios of approximately 2.5 (Cases 2, 5, 8) and approximately 3.1 (Cases 3, 6, 9) are considered. As described in Protocol Section 1.4.3, in the ABC-08 study 7/14 (50%) of patients from the ITT population had a response (BOR of CR or PR). Confirmation of response was not required in ABC-08, but if (as described in Appendix 1 of the SAP) the 0.842 multiplier obtained from Bogaerts *et al* (2009) was applied then an approximation to a confirmed response rate would be 42.1%. If a response rate in Arm A was 42.1% compared to a response rate of 19% in Arm B, then this would give an odds ratio of 3.1. Therefore, in Appendix 4 Table 1 above an odds ratio of 3.1 has been considered together with a value that is midway between 1.9 and 3.1.

Three further cases have been considered where $\Delta = \theta_1 - \theta_2 = 0$ (and so where the odds ratio = 1), i.e., with common assumed true ORRs of 0.17 (Case 10), 0.19 (Case 11), and 0.21 (Case 12).

Calculations in [Appendix 4 Table 2](#) and [Table 3](#) below have all been derived based on exact binomial probabilities. For subgroups with relatively small sample sizes any confidence intervals will necessarily be relatively wide. [Appendix 4 Table 2](#) provides tail probabilities for the observed odds ratio using cut-points of 1.5, 1.75, and 1.9, where the latter is approximately equal to the odds ratio assumed in the overall sample size calculation.

Appendix 4 Table 2: Probabilities that the Observed Odds Ratio is Greater than Particular Cut-Points, With Subgroup Sample Sizes in the Range 20-50

n per arm	Total n	Case	P (ORhat > 1.5)	P (ORhat > 1.75)	P (ORhat > 1.9)
10	20	1	62.3%	51.7%	51.7%
10	20	2	72.8%	61.6%	61.6%
10	20	3	79.8%	69.3%	69.3%
10	20	4	63.5%	51.5%	51.5%
10	20	5	73.3%	61.4%	61.4%
10	20	6	79.9%	69.3%	69.3%
10	20	7	64.3%	51.6%	51.6%
10	20	8	73.5%	61.7%	61.7%
10	20	9	79.8%	69.6%	69.6%
10	20	10 ($\Delta=0$)	37.9%	32.3%	32.3%
10	20	11 ($\Delta=0$)	38.4%	31.3%	31.3%
10	20	12 ($\Delta=0$)	38.7%	30.1%	30.1%
15	30	1	60.3%	54.6%	52.8%
15	30	2	71.9%	67.7%	65.0%
15	30	3	80.1%	76.8%	73.8%
15	30	4	60.4%	55.5%	52.8%
15	30	5	72.1%	68.1%	64.6%
15	30	6	80.4%	76.9%	73.3%
15	30	7	60.9%	56.5%	52.8%
15	30	8	72.7%	68.5%	64.4%
15	30	9	81.0%	76.9%	73.1%
15	30	10 ($\Delta=0$)	35.6%	29.1%	28.8%
15	30	11 ($\Delta=0$)	34.4%	28.2%	27.6%
15	30	12 ($\Delta=0$)	33.1%	27.5%	26.5%
20	40	1	62.8%	54.4%	49.8%
20	40	2	76.8%	68.6%	65.1%
20	40	3	85.7%	78.6%	76.3%
20	40	4	64.2%	54.8%	50.9%
20	40	5	77.8%	68.9%	66.2%
20	40	6	86.3%	79.0%	77.0%

n per arm	Total n	Case	P (ORhat > 1.5)	P (ORhat > 1.75)	P (ORhat > 1.9)
20	40	7	65.8%	55.5%	52.4%
20	40	8	78.9%	69.7%	67.3%
20	40	9	86.9%	79.8%	77.7%
20	40	10 ($\Delta=0$)	32.8%	26.4%	23.3%
20	40	11 ($\Delta=0$)	31.6%	25.3%	21.7%
20	40	12 ($\Delta=0$)	30.8%	24.2%	20.7%
25	50	1	64.2%	56.4%	49.8%
25	50	2	78.4%	72.2%	66.1%
25	50	3	86.9%	82.5%	77.7%
25	50	4	64.6%	57.1%	50.3%
25	50	5	78.4%	72.5%	66.6%
25	50	6	87.0%	82.4%	78.2%
25	50	7	65.1%	57.9%	51.2%
25	50	8	78.7%	72.7%	67.6%
25	50	9	87.3%	82.5%	79.2%
25	50	10 ($\Delta=0$)	30.8%	24.7%	20.5%
25	50	11 ($\Delta=0$)	29.8%	23.6%	19.2%
25	50	12 ($\Delta=0$)	28.9%	22.6%	18.2%

It can be seen that with 10 patients per arm in Cases 3, 6, and 9 (which each have an assumed true odds ratio of 3.1) an observed odds ratio greater than 1.9 is obtained 69%-70% of the time but when the true odds ratio is 1.0 (Cases 10-12) an observed odds ratio greater than 1.9 is obtained approximately 30%-32% of the time. With increasing sample size these probabilities of obtaining an observed odds ratio greater than 1.9 for Cases 3, 6, and 9 compared with Cases 10-12 become: (i) 73%-74% vs. 27%-29% for 15 patients per arm; (ii) 76%-78% vs. 21%-23% for 20 patients per arm; and (iii) 78%-79% vs. 18%-21% for 25 patients per arm.

Even though confidence intervals will necessarily be wide for subgroups with relatively small sample sizes, in [Table 3](#) the following power values have been derived in relation to 80% CIs:

- The probability that the lower limit of the 80% Newcombe CI for the difference in proportions is greater than zero; and
- The probability that the lower limit of the 80% Gart CI for the odds ratio is greater than 1.0.

As indicated in [Section 7.9](#), CIs for the subgroup analyses of ORR for this study will be derived using the Baptista-Pike mid-p method for the odds ratio and by the Newcombe method (without continuity correction) for the difference in proportions. For the power calculations given below the Newcombe method has been used for the difference in proportions, but the Gart method

(Agresti, 2013) which is a modified Wald CI whereby 0.5 is added to each of the four cell counts, has been used for simplicity for the odds ratio.

Note: The Gart method was found by Fagerland *et al*, 2015 to be somewhat conservative (although less conservative than for unmodified Wald based CIs) for small sample sizes. For Cases 10-12 (which each have a true assumed odds ratio of 1.0) the Type 1 Error ranged from 6.5% to 8.7% for the 4 sample sizes considered, *i.e.*, lower than the nominal 10% expected for 80% two-sided CIs. Therefore, it would be expected that the power values given in Table 3 would also be lower than the power from the less conservative Baptista-Pike mid-p method that will actually be used in the subgroup analyses for the odds ratio.

It can be seen from Table 3 that for Cases 3, 6, and 9 the power from the Newcombe 80% two-sided CI is: (i) 46%-49% with 10 patients per arm; (ii) 55%-56% with 15 patients per arm; (iii) 63%-67% with 20 patients per arm; and (iv) 68%-70% for 25 patients per arm. Corresponding power values from the Gart 80% CI for the odds ratio were 1%-5% lower than the values obtained from the Newcombe method (for difference in proportions), but as noted above the Gart method is conservative and so will in general underestimate the power that would be obtained from the Baptista-Pike mid-p method based CIs for odds ratios that will be used in the actual subgroup analyses.

In summary, with 10 patients per arm (20 patients in total within the subgroup) when the true odds ratio for ORR is approximately 3.1 (which was derived on the basis of data from ABC-08) the power from Newcombe 80% two-sided CIs is 46%-49%, the probability that the observed odds ratio will be greater than 1.9 (the value assumed in the studies overall power calculation for ORR) is 69%-70%, and the probability that the observed odds ratio will be greater than 1.5 is 80%. Therefore, it is proposed that 20 patients in total be viewed as sufficient for an adequate assessment of ORR within a subgroup.

Appendix 4 Table 3: Power Values for the Newcombe CI for the Difference in Proportions, and for the Gart CI for the Odds Ratio, with Subgroup Sample Sizes in the Range 20-50

n per arm	Total n	Case	P (Newc. 80% CI > 0)	P (Gart 80% CI > 1)
10	20	1	26.2%	21.8%
10	20	2	36.7%	32.7%
10	20	3	46.2%	42.3%
10	20	4	27.5%	23.9%
10	20	5	38.1%	34.5%
10	20	6	47.7%	43.6%
10	20	7	29.0%	25.7%
10	20	8	39.8%	35.8%
10	20	9	49.3%	44.5%
15	30	1	31.9%	26.0%
15	30	2	44.7%	38.6%
15	30	3	55.3%	50.0%
15	30	4	32.7%	27.2%
15	30	5	45.1%	39.9%
15	30	6	55.6%	51.4%
15	30	7	33.4%	28.6%
15	30	8	45.6%	41.5%
15	30	9	56.1%	53.1%
20	40	1	32.7%	30.3%
20	40	2	49.5%	46.8%
20	40	3	63.3%	60.1%
20	40	4	35.1%	32.7%
20	40	5	51.8%	48.4%
20	40	6	65.1%	60.9%
20	40	7	37.7%	34.6%
20	40	8	54.1%	49.5%
20	40	9	66.8%	61.5%
25	50	1	37.1%	33.4%
25	50	2	54.7%	51.0%
25	50	3	68.2%	65.4%
25	50	4	38.5%	35.1%
25	50	5	55.6%	52.6%
25	50	6	68.9%	66.9%
25	50	7	39.9%	36.9%
25	50	8	56.7%	54.6%
25	50	9	70.0%	68.8%

Appendix 5: Dates that a COVID-19 Vaccine Was First Approved in those Countries Which Randomized Patients in the NuTide:121 Trial

<u>Country</u>	<u>Date that a COVID-19 Vaccine was first approved¹</u>
Australia	25-Jan-21
Canada	23-Dec-20
Czech Republic	21-Dec-20
France	21-Dec-20
Germany	21-Dec-20
Hungary	21-Dec-20
Italy	21-Dec-20
Russia	11Aug-20
South Korea	10-Feb-21
Spain	21-Dec-20
Taiwan	6-May-21
Turkey	14-Jan-21
Ukraine	22-Feb-21
United Kingdom	2-Dec-20
United States	11-Dec-20

¹ Date corresponds to when a COVID-19 was first approved in the country even if this was just under an Emergency Use Authorisation (or equivalent).