Statistical Analysis Plan

A Phase II Randomized, Open-label, Multi-center Study of the Safety and Efficacy of IMCgp100 Compared with Investigator Choice in HLA-A*0201 Positive Patients with Previously Untreated Advanced Uveal Melanoma

PROTOCOL NO.	IMCgp100-202 (V5.0, 31MAR2020)
Sponsor:	**Immunocore Ltd**
CRO	**ClinChoice Inc.**
Version:	3.0
Date:	28 October, 2020

NCT03070392

SIGNATURE PAGE

PREPARED BY:

Author

[redacted]

APPROVED BY:

Approver

[redacted]

Signature Page for VV-TMF-11062 v1.0

Reason for signing: Approved	[redacted]
Reason for signing: Approved	[redacted]
Reason for signing: Approved	[redacted]
Reason for signing: Approved	[redacted]
Reason for signing: Approved	[redacted]
Reason for signing: Approved	[redacted]

Signature Page for VV-TMF-11062 v1.0

DOCUMENT REVISON HISTORY

VERSION	DATE	AUTHOR/ UPDATED BY	COMMENTS
0.1	10MAR2017	[redacted]	Initial draft
0.2	28MAR2017	[redacted]	Comments for V0.1 addressed
1.0	19APRIL2017	[redacted]	Comments for V0.2 addressed
2.0	12JUN2020	[redacted]	Further clarification of endpoints and analysis methods to following the protocol amendment
3.0	27OCT2020	[redacted]	• Remove PK parameter from secondary endpoint (Section 2.1). • Deleted subgroup "Prior surgery for management of oligometastatic disease (Yes/No)" (Section 2.6) • Add post treatment anti-cancer therapy start date imputation rule (Section 3.5). • Add death date imputation rule (Section 3.5). • Change ORR/DCR analysis method from logistic regression to CMH test (Section 4.2.1.2). • Add time to QLQ-C30 sustained deterioration censoring rule. • Add time to onset of each AESI, and time to resolution of each AESI summaries/analysis. • Add additional lab CTCAE shift summaries. • Add vital sign shift from pre-dose to post dose at each dosing schedule.

Contents

ABBREVIATIONS

Abbreviation	Definition of Term
AE	Adverse event
AESI	Adverse events of special interest
BICR	Blinded independent central review
C#D#	Cycle # Day #
CMH	Cochran Mantel-Haenszel
CR	Complete response
CRF	Case report form
CV	Coefficient of variation
DCR	Disease control rate
DOR	Duration of response
ECG	Electrocardiogram
EORTC	European Organization for Research and Treatment of Cancer
HGF	Hepatocyte growth factor
HLA-DR	Human leukocyte antigen DR
HRQoL	Health-related quality of life
IDMC	Independent data monitoring committee
IL-1Rα	Interleukin 1 receptor alpha
IRT	Interactive Response Technology
ITT	Intent to treat
LDH	Lactate dehydrogenase
LLoQ	Lower limit of quantification
MCP-1	Monocyte chemoattractant protein-1
MDSC	Myeloid-derived suppressor cell
NC	Not calculable
NCI CTCAE	National Cancer Institute Common Terminology Criteria for Adverse Events
NQ	Non-quantifiable
ORR	Objective response rate
OS	Overall survival
PD	Progressive disease
PFS	Progression free survival
PID	Percentage intended dose
PK	Pharmacokinetics
PR	Partial response
PRO	Patient reported outcome
RAS	Rash Analysis Set
RDI	Relative dose intensity
RECIST	Response Evaluation Criteria In Solid Tumors
REB	Research ethics board
RP2D	Recommended Phase II dose
RP2D-IE	Recommended Phase II dose intra-patient escalation regimen
SAE	Serious adverse event
SAP	Statistical Analysis Plan

Abbreviation	Definition of Term
SD	Stable disease
TBP	Treatment beyond progression
ULN	Upper limit of normal
ULoQ	Upper limit of quantification
UM	Uveal melanoma

1 INTRODUCTION

The purpose of this statistical analysis plan (SAP) is to describe the procedures and the statistical methods that will be used to analyze and report results for Protocol IMCgp100-202. This SAP is based on version 5.0 (amendment 4.0) of the protocol (dated 31 March 2020).

1.1 STUDY OBJECTIVES AND ENDPOINTS

The study objectives and endpoints are described in the table below:

Table 1-1 Objectives and Related Endpoints

Objective	Endpoint
Primary	
The dual primary objectives are: 1) To compare the OS in all patients randomized to the tebentafusp monotherapy versus all patients randomized to the Investigator's Choice monotherapy 2) To compare the OS in all patients randomized to the tebentafusp monotherapy who develop a rash within the first week of treatment versus all patients randomized to the Investigator's Choice monotherapy Both objectives relate to HLA-A*0201 positive patients with advanced UM with no prior treatment	OS, defined as the time from randomization until death by any cause
Secondary	
To characterize the safety and tolerability of single-agent tebentafusp in the intra-patient dose escalation regimen relative to Investigator's Choice	Safety and tolerability: Incidence and severity of AEs and SAEs; changes in safety laboratory parameters, vital signs, and electrocardiogram (QTcF); dose interruptions, reductions, discontinuations, and dose intensity of all administered agents
To characterize the PK profile of single-agent tebentafusp in the intra-patient dose escalation regimen	Mean serum concentrations over time
To assess the anti-tumor efficacy of tebentafusp versus Investigator's Choice with the parameters of PFS, BOR, DOR, time to response, and DCR using RECIST v1.1	PFS BOR DOR Time to response (TTR) DCR (defined as CR or PR, or SD $\geq$ 24 weeks)
To evaluate the treatment and disease impact to HRQoL in patients treated with tebentafusp versus patients treated with Investigator's Choice. HRQoL will be assessed by the EQ-5D,5L and the EORTC QLQ-C30	EQ-5D,5L and EORTC QLQ-C30 change from Baseline over time and between treatment strategies

Objective	Endpoint
To evaluate the incidence of anti-tebentafusp antibody formation following multiple infusions of tebentafusp in the intra-patient dose escalation regimen	Assessments of anti-tebentafusp antibody formation
Exploratory	
To assess potential predictors of efficacy of tebentafusp	Correlation of the expression of T cell infiltration, expression of gp100, HLA-DR, PD-L1, tumoral lymphocyte activation status, and myeloid-derived suppressor cell infiltration and other immune markers evaluated in tumor biopsies with anti-tumor activity
To assess potential pharmacodynamic changes in peripheral cytokine levels observed with tebentafusp and Investigator's Choice	Changes in serum cytokine, chemokines (eg, CXCL9, CXCL10, HGF, IL-1R nd MCP-1), or other analytes in response to treatment
To assess potential clinical benefit after an initial assessment of progressive disease based on RECIST v1.1	Duration of treatment and response for patients treated beyond RECIST v1.1 PD
To assess time to PFS2 for tebentafusp and Investigator's Choice	PFS2, defined as the time from the date of randomization to the subsequent PD following the initial RECIST v1.1 PD, or death
To assess health- and treatment-related medical resource utilization associated with the advanced UM disease pathway	Hospitalizations, concomitant medication use, medical procedures, and other measures of healthcare utilization

AE = adverse event; BOR = best overall response; C_{max} = maximum observed concentration; C_{min} = minimum observed concentration; CR = complete response; Ctrough = drug concentration at X days after dosing; CXCL# = C-X-C motif chemokine ligand #; DCR = disease control rate; DOR = duration of response; EORTC = European Organization for Research and Treatment of Cancer; EQ-5D,5L = EuroQoL-5 Dimensions – 5- levels of disease severity scale; gp100 = glycoprotein 100; HGF = hepatocyte growth factor; HLA-A*0201 = human leukocyte antigen-A*0201; HLA-DR = human leukocyte antigen-DR isotype; HRQoL= health-related quality of life; IL-1R interleukin 1 receptor alpha; MCP-1 = monocyte chemoattractant protein-1; OS = overall survival; PD-L1 = programmed death-ligand 1; PFS = progression free survival; PFS2 = second disease progression; PK = pharmacokinetic; PR = partial response; QLQ-C30 = Quality of life Questionnaire-Core 30; QTcF = QT interval corrected by Fridericia's formula; RECIST = Response Evaluation Criteria In Solid Tumors; SAE = serious adverse event; SD = stable disease; UM = uveal melanoma.

1.2 SUMMARY OF THE STUDY DESIGN

1.2.1 General Study Design and Plan

This is an open label, randomized, multi-center Phase II study of tebentafusp versus Investigator's choice in adult (> 18 years) HLA-A*0201+ patients with advanced UM previously untreated in the advanced or metastatic setting. Prior adjuvant or neoadjuvant therapy is allowed, provided all prior therapy is administered in the localized, curative setting. Patients are to be randomized 2:1 (tebentafusp: Investigator Choice) with randomization was stratified by lactate dehydrogenase (LDH) status (based on central laboratory assessment performed during the screening period) to receive either tebentafusp administered in the intra-patient escalation dosing regimen or Investigator Choice. One cycle of treatment in this study is defined as 3 weeks (21 day cycles).

Patients will be treated with 1 of the following regimens:

- **Arm 1 (tebentafusp):** All patients randomized to Arm 1 will receive tebentafusp following the intra-patient escalation regimen. On C1D1, patients receive 20 mcg, on C1D8, patients receive 30 mcg, and beginning with C1D15 and thereafter, patients will receive the escalated dose of 68 mcg. Due to the anticipated cytokine and chemokine associated toxicity with tebentafusp, patients will be monitored overnight as an inpatient following the weekly doses at C1D1, C1D8, and C1D15.
- **Arm 2 (Investigator Choice):** All patients randomized to Arm 2 will receive Investigator Choice of one of three options: dacarbazine in the standard dosing regimen in UM of 1000 mg/m^2 given on Day 1 of each 21 day cycle; ipilimumab in the approved dosing regimen for unresectable or metastatic melanoma of 3 mg/kg given on Day 1 of each 21 day cycle for a maximum of 4 doses; or pembrolizumab in the approved dosing regimen of 2 mg/kg given to a maximum of 200mg or 200mg administered intravenously where approved locally on Day 1 of each 21 day cycle. The preferred Investigator Choice agent will be selected prior to randomization. No overnight monitoring is required in Arm 2

The study design is depicted in Figure 1 below.

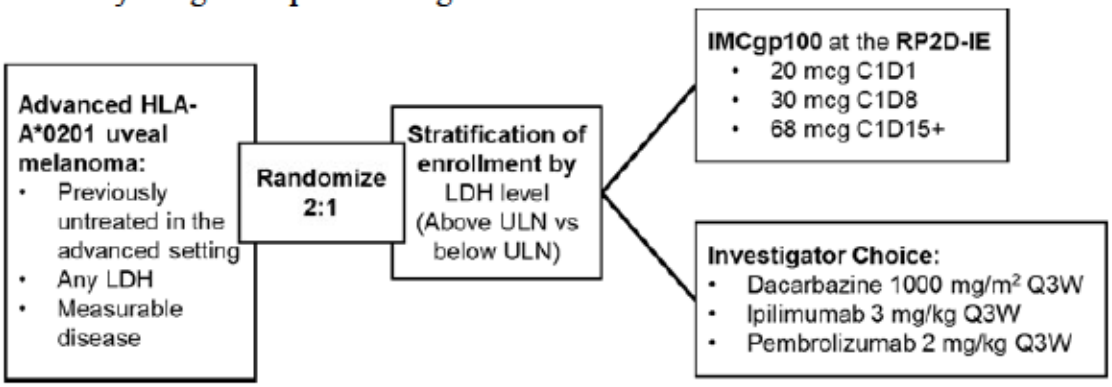


Figure 1. Study Design

LDH = lactate dehydrogenase,; RP2D-IE = recommended Phase II dosing intra-patient escalation regimen; Q3W = dosing repeated every 3 weeks; ULN = upper limit of normal.
Randomized, open label study of tebentafusp versus Investigator Choice in patients with advanced uveal melanoma.

1.2.2 Randomization and Blinding

Patients will be assigned to 1 of 2 randomized treatment arms, Arm 1 (tebentafusp) and Arm 2 (Investigator Choice) in a ratio of 2:1. The treatment assignment to the randomized Arms 1 and 2 will be determined by the Interactive Response Technology (IRT). The randomization numbers will be generated using the following procedure to ensure that treatment assignment is unbiased. A patient randomization list will be produced by the IVRS provider using a validated system that automates the random assignment of patient numbers to randomization treatments, linked to Arms 1 and 2.

The investigational site must declare, prior to randomization, their choice of control therapy for that patient.

Randomization to 1 of the 2 randomized treatment arms will be stratified by LDH levels based on central lab results. The two strata that will be used are (1) Baseline LDH below or equal to the upper limit of normal (ULN) (≤ 250 U/L) and (2) Baseline LDH above the ULN (>250 U/L). LDH levels utilized for stratification will be assessed centrally during the screening period.

1.2.3 Sample Size and Statistical Power Considerations

1.2.3.1 Intent-to-Treat Analysis

OS is the primary endpoint for this study. Assuming a 2:1 randomization ratio of tebentafusp vs. Investigator Choice, 250 events (deaths) are needed in the randomized trial to provide 89% power to detect a difference of survival distribution that can be characterized by a 0.645 hazard ratio (HR) for OS with a 2-sided significance level of 0.045. Assuming OS is exponentially distributed, this may translate to a median OS of 18.6 months in the tebentafusp treated arm and 12 months in the Investigator Choice arm. The smallest treatment effect that would be statistically significant is an OS HR of 0.75 (e.g. 16 vs 12mo).

Considering a non-uniform recruitment of about 33 months and 10% annual drop-out rate, 369 patients need to be randomized in a 2:1 ratio to the 2 arms in order to observe 250 events after 51 months as follows:

- 246 patients to the tebentafusp arm
- 123 patients to Investigator Choice arm

Three analyses of OS are planned: two formal interim analyses and the final analysis. Details of the interim analyses are described in section 5.1.

In order to randomize 369 patients, (assuming a 10% screen failure rate), 410 patients will need to be enrolled. In order to enroll 410 patients, approximately 900 patients will need to be pre-screened (allowing for a 5% attrition rate and assuming 48% of patients are HLA-A*0201 positive). The prevalence of HLA-A*0201 varies depending on the region, so additional patients may be needed to be pre-screened to enroll 410 patients.

1.2.3.2 Rash Analysis Set Analysis

The study is also powered for the analysis of OS in the Rash Analysis Set (see Section 2.2.2). Assuming 50% of the tebentafusp-treated patients develop a rash within the first week of treatment, there will be an approximate 1:1 ratio between patients in the tebentafusp arm and the Investigator's Choice control arm. One hundred sixty-four (164) events (deaths) are needed to provide 89% power to detect a difference in survival distributions that can be characterized by a 0.531 HR for OS with a 2-sided significance level of 0.005. Assuming OS is exponentially distributed, this may translate to a median OS of 22.6 months in the tebentafusp treated arm and 12 months in the Investigator's Choice arm.

2 STATISTICAL CONSIDERATIONS

2.1 GENERAL CONSIDERATIONS

Data analysis of interim and final data will occur when the required number of death events have been observed.

Continuous variables will be summarized using descriptive statistics (number of observations, mean, standard deviation [SD], median, 25th and 75th percentiles, minimum and maximum).

Confidence intervals (CIs) will be 95% and all tests will be two-sided, unless otherwise specified in the description of the analyses. For binomial variables, the normal approximation methods will be employed unless otherwise specified.

P-values will be rounded to four decimal places. If a p-value is less than 0.0001 it will be reported as "<0.0001." If a p-value is greater than 0.999 it will be reported as ">0.999"

The following rules will be followed for reporting results unless stated otherwise:

- Screen failure patients are those who signed the informed consent but were never randomized into the study for any reason. Screen failure data will only be reported in the table of summary of analyses data.
- A month is operationally defined to be 30.4375 days. Six months is operationally defined to be 182.625 days.
- Data will be presented in data listings by patient number. All summaries will be presented by patient number unless otherwise specified.
- Additional summaries of efficacy and other variables may be produced as a separate report(s) for specific regions, as required by local health authorities.

2.2 DEFINITIONS OF ANALYSIS SETS

2.2.1 Intent to Treat Analysis Set

The Intent to Treat (ITT) set comprises all patients assigned to treatment analyzed by the treatment assignment whether or not the patient received the assigned treatment. All patients randomized in the study will be analyzed in the ITT population. The ITT set will be used for all summaries and analyses of demography, baseline characteristics, disposition, medical history, prior anti-cancer therapy and efficacy data summaries and analyses.

2.2.2 Rash Analysis Set

The rash analysis set (RAS) comprises all patients assigned to tebentafusp who develop a rash within the first week of treatment (i.e. Study Day 1 – 7 and prior to the second dose in case the second dose is received early) and all patients randomized to Investigator's Choice regardless of rash. If the RAS analysis described in Section 4.2.1 crosses the pre-specified stopping boundaries (see Section 5.1) and stopping boundaries for the ITT analysis set at the same planned analysis are not crossed, then this analysis set will also be used for demography, baseline characteristics, efficacy, and safety data

summaries and analyses. The AE preferred terms that define the RAS are detailed in Appendix 1. If the stopping boundaries for the RAS OS analysis are crossed, then summaries of demographics, baseline characteristics, and safety will also be run on this analysis set.

2.2.3 Safety Analysis Set

The Safety Analysis Set (SAF) includes all randomized patients who have received at least 1 full or partial dose of tebentafusp or Investigator Choice. Patients will be classified in this set according to initial treatment received. The safety analysis set will be used for all safety summaries.

2.2.4 PK analysis set

The PK analysis set includes patients in the SAF who have at least one measurable PK concentration and who have the relevant date, time and dosing data for this sample.

Table 2-1 Summary of outcome variables and analysis sets

Outcome variable	Analysis Set
Primary Efficacy Data: OS, PFS, ORR (including BOR and % change in tumour size), DoR, time to response, DCR Exploratory Efficacy Data: tumor response, duration of treatment, and PFS2 based on investigator assessment	ITT and RAS
Demography, Baseline Characteristics, Disposition, Protocol Deviations, Medical History, and Prior and Concomitant Medications. PRO data: EQ-5D, EORTC-QLQ-C30 and Health resource utilisation	ITT (and RAS if applicable)
PK data	PK
Safety Data Incidence of anti-tebentafusp antibody formation	SAF (and RAS if applicable)

2.3 MULTIPLE COMPARISONS/MULTIPLICITY

Two interim analyses will be performed using a three-stage adaptive group sequential design. The primary analyses of overall survival will be based on O'Brien-Fleming boundaries and the Lan-DeMets approximation to O'Brien-Fleming boundaries. Details of the type-1 error adjustment is described in Section 5.

The experiment-wise type I error rate will be controlled at α=0.05 two-sided. Ninety-

percent (90%) of the α will be allocated to the ITT analyses (α_{ITT}=0.045). The other 10% will be allocated to the RAS analyses (α_{RAS}=0.005). See section 5.2 for further details on the testing strategy and methods for controlling the type I error and Appendix 2 for a graphical depiction of the testing strategy.

2.4 EXAMINATION OF SUBGROUPS

Subgroup analyses are exploratory and will be conducted by comparing OS and PFS between treatments in the following groups:

- Ethnicity (Hispanic vs. non-Hispanic)
- Gender (Male versus Female)
- Age at screening (<65 versus ≥65)
- ECOG (0 vs. 1)
- Baseline alkaline phosphatase (≤ ULN vs. > ULN)
- LDH (≤ ULN versus > ULN, where ULN = 250 U/L)
- Prior systemic therapy in the adjuvant or neoadjuvant setting (Yes versus No)
    - Yes category further split by:
        - Chemotherapy
        - Immunotherapy
        - Targeted therapy
- Largest metastatic lesion recorded at baseline (≤3.0 cm vs 3.1-8.0 cm vs ≥8.1 cm). The size of the largest metastatic lesion will be determined based on the sizes of the baseline target lesions recorded on the eCRF which are indicated as metastatic or nodal. If there are any patients with no metastatic/nodal lesions (e.g. primary lesions only) then these patients will be grouped into the ≤3.0cm category.
- Region (North America versus others)
- Pre-choice of chemotherapy (ipilimumab, dacarbazine and pembrolizumab). Sites were asked to record for all patients the intended investigator choice treatment prior to randomization.

For subgroups outlined, the HR (tebentafusp: Investigator's choice) and associated CI will be calculated from a Cox proportional hazards model (ties=Efron) that contains the treatment term, factor and treatment-by-factor interaction term. The treatment effect HRs for each treatment comparison along with their CIs will be obtained for each level of the subgroup from this single model.

To estimate the treatment effect and CI for each level of a covariate an example is given below:

Consider the following model where the covariate is coded as a 0,1 two-level factor

$$y = \mu + \beta_1 trt + \beta_2\, cov + \beta_3\, cov.trt$$

The treatment effect is estimated as β_1 in level of 0 of the covariate and as $\beta_1 + \beta_3$ for level 1 of the covariate with corresponding variances of

$$Var(\beta_1) \text{ and}$$

$$Var(\beta_1) + Var(\beta_3) + 2.Cov(\beta_1, \beta_3) \text{ respectively.}$$

The HRs and 95% CIs will be presented on a forest plot including the HR and 95% CI from the ITT population.

In addition, the significance of the interaction term will be assessed to determine if any are significant at the 2-sided 10% level. This approach will identify the factors that independently alter the treatment effect and prevent identification of multiple correlated interactions.

If there are too few events available for a meaningful analysis of a particular subgroup, the relationship between that subgroup and OS/PFS will not be formally analyzed. In this case, only descriptive summaries will be provided.

The assumption of proportionality will be assessed using the ASSESS statement in PROC PHREG which performs the graphical and numerical methods of Lin, Wei and Ying (Lin, Wei and Yin 1993).

No adjustment to the significance level for testing will be made since the subgroup analysis will be considered exploratory and will be supportive for OS and PFS.

P-value < 0.1 suggest there is an interaction between treatment and the factor.

The focus of these analyses will be on the ITT population, but they will also be conducted among the RAS if the RAS analyses of OS cross the pre-specified stopping boundaries.

3 DEFINITIONS AND CONVENTIONS FOR DATA HANDLING

3.1 BASELINE DEFINITION

In general, for efficacy endpoints and patient reported outcomes (PROs), baseline is defined as the last observed measurement prior to randomization of treatment, including unscheduled visits. However, if an evaluable assessment is only available after randomization but before the first dose of randomized treatment then this assessment will be used as baseline.

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

In all summaries change from baseline variables will be calculated as the post-treatment value minus the value at baseline. The % change from baseline will be calculated as (post-baseline value -baseline value)/baseline value x 100.

3.2 STUDY DAY

For the purpose of efficacy data summary, Study Day 1 is defined as the date of randomization to study treatment. For visits (or events) that occur on or after randomization, study day is defined as (date of visit [event] -date of randomization + 1). For visits (or events) that occur prior to randomization, study day is defined as (date of visit [event] -date of randomization). There is no Study Day 0.

For the purpose of safety data summary, Dose 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). There is no Dose Day 0. For listings (such as for AEs) that include the derivation of "days since last dose," this is defined as (event date - date of last dose). Events that occur on the same day as the last dose of study drug will therefore be described as occurring zero days from the last dose of study drug.

3.3 END OF STUDY

The end of the study is defined as 'the last visit of the last patient undergoing the study'. All patients will have completed follow-up for OS up to the final data cut-off, which will occur when approximately 250 deaths have occurred.

An individual patient may end participation in the study for reasons of death, loss to follow-up, withdrawal of consent or the study end is reached (as described above) or the study is terminated early by the Sponsor.

3.4 ANALYSIS VISIT WINDOWS

During the course of the study visits, tests and/or procedures should occur on schedule whenever possible. A visit window of ± 2 days is allowed for all visits where study drug administration is scheduled. For all other visits, a visit window of ± 7 days is allowed, unless otherwise indicated in the protocol. If the study drug infusions are delayed or otherwise moved from the scheduled day, all study assessments will be moved with the delayed study drug infusions. The only exception to moving study assessments with treatment are the radiological and PRO assessments, which must be performed ± 7 days of the scheduled date of the assessment (unless otherwise indicated in the protocol) taking as reference the date of randomization. The protocol specified radiologic assessments should be performed as scheduled every 12 weeks as indicated in the protocol (reference to randomization) and should not follow delays incurred in the treatment period for the accurate assessment of PFS and duration of response endpoints.

For summaries of vital signs, laboratory data, ECG, and PROs etc., assessments will be assigned to calculated visit windows (using study day). The time windows should be exhaustive so that data recorded at any time point have the potential to be summarized. Inclusion within the visit window should be based on the actual date and not the intended date of the visit. For summaries at a patient level, all values should be included, regardless of whether they appear in a corresponding visit-based summary, when deriving a patient level statistic such as a maximum.

For visit based summaries:

- If there is more than one value per patient within a visit window then the closest to the planned study day value should be summarized, or the earlier in the event the values are equidistant from the planned study day. The visit will be missing if no assessment was reported within the specified visit window around the planned study day.
- To prevent very large tables or plots being produced that contain many cells with meaningless data, summary statistics will be presented where at least 10 patients in either treatment group have data recorded at a particular visit.

3.5 MISSING DATA HANDLING RULES

In general, other than for partial dates, missing data will not be imputed and will be treated as missing.

Imputation rule:

Post treatment anti-cancer therapy start date:

- If only day is missing, then impute as the last day of the month.
- If both the day and month are missing, then impute as Jan 1st of the given year if the year is after the last dosing date. Otherwise impute as min of (the last dosing date + 90, last day of the year).
- If the start date is totally missing, then impute as the last dosing date + 90.

Initial diagnosis date:

- If year is missing, do not impute.
- If only day is missing, impute day as 15th of the month.
- If day and month are missing, impute as July 1st.

Death date:

- Imputed as the last known alive date +1 day if the year/month from partial death date is the same as the last known alive date or if completely missing.
- If the death year/month is later after the last known alive date:
    - If missing day only, impute as the 1st of the month.
    - If missing day and month, impute as the 1st of January

Medical history start date:

- If the year is missing, do not impute.
- If only the day is missing, then impute the day as the last day of the month.
- If the day and month are missing, then impute as December 31st.
- If the resulting imputed start date is after the randomization date, then impute as the day before the randomization date.

Medical history end date:

- If the year is missing, do not impute.
- If only the day is missing and event is NOT ongoing, then impute the day as the last day of the month.
- If the day and month are missing and event is NOT ongoing, then impute as December 31st.
- If the resulting imputed end date is after the end of study date and event is NOT ongoing, then impute as the end of study date

Concomitant medication start date:

- If year is missing (or completely missing), 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 end date:

- If year is missing (or completely missing), 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.
- If year and month are present and day is missing, impute day as last day of the *month.*
- If the resulting imputed end date is after the end of study date, then impute as the end of study date

Missing Dates in Adverse Events

Start dates of adverse events will be imputed as follows:

- Completely missing start date will be imputed as the date of first dose
- Start date missing both month and day will be imputed as:
  - the date of first dose if the year of the start date is the same as the date of first dose.
  - otherwise, Jan 1st of the year of the start date will be used.
- Start date missing day will be imputed as:
  - the date of first dose if the year and month of the start date are the same as the date of first dose.
  - otherwise, the 1st of the month of the start date will be used.

Stop dates of adverse events will be imputed as follows:

- Completely missing stop date will be imputed as the date of last dose plus 90 days.

- Stop date missing both month and day will be imputed as Dec 31st of the year of stop date.
- Stop date missing day will be imputed as the last date of the month of the stop date.

After imputation, the imputed date will be compared against the end of study date, if available. If the planned imputed date is later than the end of study date, then the end of study date will be used as the imputed date instead.

Imputation rules for lab values outside of quantification range

Lab values and concentration data below the lower limit of quantification (LLoQ) that are reported as "<LLoQ" or "≤LLoQ" in the database will be imputed by LLoQ x 0.99 for analysis purposes. Lab values above the upper limit of quantification (ULoQ) that are reported as ">ULoQ" or "≥ULoQ" in the database will be imputed by ULoQ x 1.01 for analysis purposes. The original value (including missing values) will be listed.

For PK concentrations, results that are <LLoQ will be treated as 0 for the calculation of summary statistics. For display purposes, all concentrations <LLOQ (mean or individual results) will be set to ½ the LLOQ. For the calculation of PK parameters, concentrations <LLOQ at predose will be set to 0, all other concentrations will be treated as missing. Any missing PK parameter data will not be imputed.

Missing or incomplete data for the EORTC-QLQ-C30 will be handled (in line with the EORTC-QLC-C30 scoring manual, EORTC Quality of Life Group, 2001) as follows:

- PRO domain scores will be calculated if at least 50% of the items that construct the domain have been answered
- If more than 50% of the items are missing, the domain score will be considered missing
- Missing items for PRO domains will not be imputed at an individual level.

For the EQ-5D, 5L, patients are required to complete all five-levels of the descriptive system in order to generate a self-reported health state. Patients with incomplete data in any of the five dimensions will be assumed to have a missing value for the descriptive system of that visit. If the value for the EQ-5D VAS is missing at any visit, the VAS score will be assumed to be missing for that visit. No imputations will be made for either the EQ-5D, 5L descriptive system or the VAS.

Rounding rules for reported percentages

For percentages ≥10%:

- Values ≥X.5 or above round to X+1.
- Values >X but <X.5 round to X.

For percentages <10%:

- Values ≥X.Y5 or above round to X.Y+0.1.

- Values >X.Y but <X.Y5 round to X.Y.

4 STATISTICAL ANALYSES

4.1 PATIENT INFORMATION

4.1.1 Disposition of Patients

Any patients who signed the informed consent, but were not randomized for any reason will be regarded as screen failures and excluded from any outputs.

One summary will provide the number and percent of patients (among all randomized patients) who are represented in each of the analysis sets described in Section 2.2.

The number of randomized patients enrolled by each region (North America, Europe, and Other), country and site will be summarized by randomized treatment group and overall.

Patient disposition will be listed and summarized.

The number and percent of patients who discontinued study treatment and who discontinued the study will be summarized according to the reason for discontinuation. The pre-chosen investigator's choice of treatment will also be displayed as individual columns in the disposition table.

4.1.2 Protocol Deviations

Protocol deviations will be reviewed by the study team prior to the primary and final analyses and without explicit knowledge of a given patient's randomized treatment assignment.

The number and percentage of patients excluded from the PK analysis or biomarker analysis sets will be presented by treatment group and in total.

All important deviations related to the study inclusion or exclusion criteria and study conduct will also be listed and summarized by randomized treatment group.

If the deviations are serious enough to have the potential to impact the primary analysis, sensitivity analyses may be performed.

The following general categories will be considered important deviations. This list is not exhaustive and additional important deviations may be added prior to database lock:

- Informed consent procedure deviation (e.g., no informed consent signed prior to any screening procedure)
- Eligibility criteria deviation (e.g., any inclusion criteria not met or exclusion criteria met)
- Prohibited medication deviation (e.g., patient received disallowed anticancer treatment while on study treatment)

The categorization of these as important deviations is not automatic and will depend on duration and the perceived effect on efficacy and safety. In addition to the programmatic

determination of the deviations above, monitoring notes or summaries will be reviewed to determine any important post-entry deviations that are not identifiable via programming, and to check that those identified via programming are correctly classified. The final classification will be made prior to the primary and final analyses without explicit knowledge of the treatment group assignment for the patient with the deviation in question.

Treatment misallocation, in terms of errors in treatment dispensing following randomization, in addition to incorrect stratifications, will also be summarized and listed. Treatment misallocation includes patients who receive no treatment whatsoever for a period of time due to errors in dispensing of medication/availability of medication. Note, this is not due to tolerability issues where patients may stop taking drug due to AEs.

Patients who receive the wrong treatment at any time will be included in the safety analysis set and analyzed according to the treatment that they actually received. During the study, decisions on how to handle known treatment allocation errors will be made on an individual basis with written instruction from the study team leader and/or statistician.

4.1.3 Demographics and Baseline Characteristics

Demographic and baseline patient characteristics will be summarized. Age group, (grouped as -<65 and ≥ 65), gender, ethnicity, ECOG performance status, and race will be summarized with counts and percentages. Age, weight, height, and body mass index will be described with standard descriptive statistics.

In addition, the number of patients within each subgroup category will be displayed (see section 2.4 for subgroup categories).

Individual patient listings of all demographic and baseline characteristics will also be generated.

4.1.4 Medical History

Medical History will be coded by MedDRA 19.1 or higher version. The number (percent) of patients reporting a history of any disease related medical condition, as recorded on the CRF, will be summarized by system organ class and preferred term for each treatment group and overall. A patient data listing of medical and surgical history will be provided.

Disease specific history will be summarized by treatment group. The presence or absence of metastases at initial diagnosis, initial stage, anatomic site of primary disease, baseline LDH as determined by central laboratory (below or above ULN), stage entry, administration of neoadjuvant therapy, and administration of adjuvant therapy will be summarized with counts and percentages.

4.1.5 Prior and Concomitant Therapy

All investigator terms for medications recorded on the CRF will be coded using the World Health Organization (WHO) Drug Dictionary (Global B3, March 2019). The number (percent) of patients who took prior and concomitant medications will be summarized and listed by treatment, Anatomical Therapeutic Chemical (ATC) Level 2 Classification and WHO Drug preferred term. Prior medications will be defined as

medications that are taken prior to screening with a stop date prior to the first dose of study treatment. Concomitant medications will be defined as medications that (1) started before the first dose of study drug and are continuing at the time of the first dose of study drug, or (2) started on or after the date of the first dose of study drug (or, started at the time of or after the first dose of study drug) up to 90 days following the last dose.

For the purpose of inclusion in prior and/or concomitant mediation or therapy summaries, incomplete medication or radiotherapy start and stop dates will be reviewed by the medical monitor and imputed as neoadjuvant or adjuvant based on their review.

All anti-cancer therapies will be summarized with counts for the ITT set. They will be summarized separately for prior, concurrent, and post-withdrawal of IP anti-cancer therapies. The number of anti-cancer therapies will be summarized using descriptive statistics. Any anti-cancer therapies will also be listed.

The subsequent therapy data will be summarized in terms of therapy given, duration of subsequent therapies (based on start / stop dates), and investigator reported BOR.

4.2 EFFICACY ANALYSES

4.2.1 Primary Efficacy Analysis

Results of all statistical analyses will be presented using a 95% CI and 2-sided p-value unless stated otherwise.

The primary efficacy endpoint is OS, which is defined as the time between the date of randomization and the date of death from any cause in an individual patient. Patients without documentation of death at the time of the analysis will be censored at the last date of known 'alive' status. OS will be followed continuously while patients are treated on trial and every 3 months in the follow up phase.

Survival "sweeps" will be made in the weeks leading up to the anticipated data cut-off for each planned OS analysis. Data cut-off dates will generally be determined by the dates when the pre-specified number of events for a given analysis have first been confirmed. In order to facilitate planning for IDMC meetings, data cut-off dates may be set prior to the planned number of events provided at least 95% of the planned number of events have occurred. If patients are confirmed to be alive, or if the death date is after a given data cut-off date, then such patients will be censored at the data cut-off date. When applicable, death dates may be found by checking publicly available death registries.

The primary analysis of OS in all randomized patients will be analyzed using a 2-sided log rank test stratified by LDH status for generation of the p-value. The hazard ratio (HR) will be estimated using a Cox-proportional hazards model stratified by LDH using the Efron approach for handling ties (Efron 1977), together with the associated profile likelihood 95% confidence intervals for the HR.

The LDH status in the statistical modelling will be based on the values entered into IRT at randomization, even if it is subsequently discovered that these values were incorrect.

A Kaplan-Meier (KM) plot of OS will be presented by treatment group. Median OS with 95% CIs will be presented. In addition, landmark survival estimates at 1 year with

corresponding 95% CIs will also be presented using Kaplan-Meier methodology. KM survival estimates will also be presented at 6 monthly intervals. The median follow-up time for OS and the corresponding 95% confidence interval (using the method of Brookmeyer and Crowley, 1982 with the log-log transformation) will also be summarized using the reverse KM method. The analysis involves the event and censoring rules to be switched (i.e. the patients with documented disease progression or death become 'censored', and the censored patients are treated as the 'event'). Summaries of the number and percentage of patients who have died, those still in survival follow-up, those lost to follow-up and those who have withdrawn consent will be provided.

OS will also be analysed based on an unstratified log rank test as a supportive analysis. An additional sensitivity analysis will evaluate OS in the Safety Population. For the RAS OS analysis, in order to adjust for the potential immortal-time bias during the first week of treatment that determines eligibility into the RAS for tebentafusp patients, a sensitivity analysis will measure OS starting from Study Day 8 (7 days after the start of infusion) for all patients in the analysis set. Patients in the analysis set who die or who would otherwise be censored prior to Study Day 8 will be excluded from the analysis.

Exploratory subgroup analyses will be conducted to assess the consistency in treatment effect across the different subgroup. See section 2.4 for further details.

4.2.2 Secondary Efficacy Analyses

4.2.2.1 PFS

PFS is defined as the time from randomization to the date of first documented progression (per RECIST v 1.1.) as determined by investigator assessment or death due to any cause, whichever occurs first, regardless of whether the patient withdraws from randomized therapy or receives another anti-cancer therapy prior to progression. Patients who have not progressed or died at the time of the analysis will be censored at the time of the last evaluable tumor assessment. Patients who have progressed or died following two or more missed tumor assessments or NE assessments will be censored at the time of the last evaluable tumor assessment prior to the missed/NE assessments.

Two or more missed visit will be determined if the time from the date of the last non-missing evaluable tumor assessment and date of progression/death is >26 weeks (182 days allowing 1-week window for each visit).

The detail on programmatically determining events or censored times is described below:

Table 4-1

PFS Censoring/Event programming logic				
Step	**Situation**	**Situation -sub**	**Date of Event or Censor**	**Event / Censor**
1	**No baseline** radiological tumor	a). no death reported within 2 scan intervals following the date of randomization	Date of Randomization	Censored

PFS Censoring/Event programming logic				
Step	**Situation**	**Situation -sub**	**Date of Event or Censor**	**Event / Censor**
	assessment available **or** **No post baseline** tumor assessment available	b). death reported within 2 scan intervals following the date of randomization	Date of Death	Event
2	**Tumor progression** (PD)	a). PD documented after 2 scan intervals following previous evaluable radiological tumor assessment	Date of previous evaluable radiological assessment	Censored
		b). PD documented within 2 scan intervals following previous evaluable radiological tumor assessment	Earliest of the dates of the component(s) that triggered the progression (eg if PD based on new lesion only then the new lesion date should be used. If PD based on TL and new lesion, then earliest of TL/NL dates should be used)- see footnote	Event
3	**No tumor PD but Death reported**	a). death reported after 2 scan intervals following last evaluable radiological tumor assessment	Date of previous evaluable radiological assessment	Censored
		b). Death within 2 scan intervals following previous evaluable radiological tumor assessment	Date of Death	Event
4	**No tumor progression and no Death reported**	a). patient lost to follow-up or withdrawal of consent	Date of last evaluable radiological assessment	Censored
		b). patient on study no death reported	Date of last evaluable radiological tumor assessment	Censored

RECIST = Response Evaluation Criteria in Solid Tumors
(1) If target, non-target, and new lesion assessments have different dates within a visit, then the earliest of those dates will be considered as the date of the tumor assessment if the assessment for that visit is progressive disease (PD); otherwise the latest date will be used.
(2) Evaluable radiographical tumor assessment refers to an assessment with overall response of complete response (CR), partial response (PR), stable disease (SD), or progressive disease (PD).

Two scan intervals is defined as 182 days (24 weeks + 2 weeks to allow for the 1-week protocol window per scan visit).

If a patient has a RECIST 1.1 PD/death event > 26 weeks following their last evaluable tumour assessment then the patient will be censored at the time of their last evaluable tumour assessment prior to the progression/death event. If a patient has no evaluable post baseline tumour assessments, then death within 26 weeks of randomisation will be regarded as a PFS event, otherwise the patient will be censored at day 1. The timeframe of 26 week is approximately 2 scan intervals accounting for visit windows.

Note: Each scenario above is mutually exclusive and therefore patients should only be included in one of the above scenarios.

The ITT analysis of PFS will follow the proposed analyses for OS including the unstratified analysis as supportive.

Kaplan-Meier plots of PFS will be presented by treatment arm. Summaries of PFS will be provided, including median PFS for each treatment arm and landmark estimates at 3, 6 and 12, 18, and 24 months (week 13, 26 and 52, 78, and 104) with corresponding confidence intervals. Median duration of follow up for PFS and corresponding 95% CI will also be presented based on the reverse KM method as described above for OS.

If a patient has no evaluable visits or does not have baseline data, they will be censored at 0 days unless they die within two visits of baseline in which case the date of death will be used as the progression date.

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

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

- Date of progression will be determined based on the earliest of the dates of the component that triggered the progression

- When censoring a patient for PFS the patient will be censored at the latest of the dates contributing to a particular overall visit assessment

4.2.2.2 ORR and BOR

ORR is defined as the number of randomized patients with at least one visit response of complete response (CR) or partial response (PR) divided by the number of randomized patients presented as a percentage for each treatment arm in the ITT population. Objective responses do not require confirmation since this is a randomized study. The BOR is defined as the best response designation as determined by Investigator assessment up until progression or last evaluable assessment in the absence of

progression. Any complete or partial responses that occur after a further anti-cancer therapy was received will not be included in the numerator for the ORR calculation by RECIST 1.1 but will be included in the RECIST 1.1 listing and flagged.

The ORR will be compared between treatment arms using stratified Cochran Mantel-Haenszel (CMH) test adjusting for the baseline LDH status. The results of the analysis will be presented in terms of an odds ratio together with its associated 95% CI.

A summary of BOR and ORR will also be presented by treatment group.

Waterfall and/or spider plots of the best percentage change in tumor size will be produced. Tumor shrinkage will be assessed using RECIST1.1 tumor response. The absolute change and percentage change from baseline in sum of tumor size at each assessment will be calculated. Tumor size is the sum of the longest diameters (SLD) or short axis (nodal lesions) of the TLs. The percentage change in SLD at each week for which data are available will be obtained for each patient taking the difference between the SLD at each week and the SLD at baseline divided by the SLD at baseline multiplied by 100 [i.e.(week n-baseline)/baseline x100]. The change from baseline will be obtained for each patient taking the difference between the SLD at each week and the SLD at baseline (i.e.,week n-baseline).

The best percentage change in SLD from baseline or the minimum increase from baseline in the absence of a reduction from baseline based on all post-baseline assessments prior to the visit when progression is detected (RECIST v1.1) or start of subsequent anti-cancer therapy.

If best percentage change cannot be calculated due to missing data, a footnote will be used to describe the missing information. For example

- If a patient has no post-baseline assessment and has died;
- If a patient has new lesions or progression of NTLs
- If a patient has withdrawn due to PD and has no evaluable TL data before or at PD

4.2.2.3 Disease Control Rate

Disease control rate (DCR) is defined as the proportion of patients with a BOR of CR or PR or SD recorded at 24 weeks or later (derived as 23 weeks to allow for the 1 week protocol window) after randomization of study drug and prior to any PD event. The estimated DCR and associated 95% confidence interval will be determined by treatment Arm. DCR will also be compared between the arms using stratified CMH test adjusting for the baseline LDH status, as described for ORR.

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

- Complete Response (CR): At least one overall response of CR documented before progression or start of new anti-cancer therapy.
- Partial Response (PR): At least one overall response of PR documented before progression or start of new anti-cancer therapy, with no responses of CR
- Stable disease (SD): At least 1 overall response of SD documented at least 1 scan interval (12 weeks – 7 day window = 77 days) after the date of randomization and before progression and the start of new anti-cancer therapy, with no CR or PR responses.
- Progressive Disease (PD): Progression documented with no response of CR, PR or SD.
- Not Evaluable (NE): All other cases will be categorized as NE. The reasons for NE will be summarized and the following reasons may be used:
  - No evaluable post baseline tumour assessments and death >182 days from randomization, or no death recorded
  - All post-baseline assessments have an overall response of NE
  - New anti-cancer therapy started before first post-baseline assessment
  - SD recorded prior to day 77 from randomization (i.e prior to first protocol tumour assessment visit at 12 weeks +/- 1 week)

Special cases where BOR is NE due to early SD will be classified as 'SD <12 weeks'.

The disease control rate (DCR) is defined as the proportion of patients with BOR of a documented complete response, partial response, and stable disease (CR + PR + SD), based on Response Evaluation Criteria in Solid Tumors (RECIST) 1.1, by investigator assessment. In order to classify SD as the BOR, the assessment must be made a minimum of 77 days from baseline, where baseline is counted from the date of randomization.

4.2.2.4 Duration of response

DOR (per RECIST 1.1) is calculated only for those with a documented response of CR or PR and will be defined as the time from the date of first documented response until the first date of documented progression or death due to any cause.(i.e., date of PFS event or censoring - date of first response + 1). The end of response should coincide with the date of progression or death from any cause used for the RECIST 1.1 PFS endpoint.

The time of the initial response will be defined as the latest of the dates contributing towards the first visit response of CR or PR. If a patient does not progress following a response, then the corresponding DOR will be censored at the PFS censoring time.

Descriptive data will be provided for the DOR in responding patients, TTR (time to response, excluding responses that occur after PD) including the associated swimmer plot (without any formal comparison of treatment arms or p-value attached). Descriptive statistics for DOR will be based on Kaplan-Meier estimates. Median duration of follow up for DoR and corresponding 95% CI will also be presented based on the reverse KM method as described above for OS.

4.2.3 Sensitivity Analyses

For the primary analysis patients will be included based on the strata recorded at the time of randomization in the IRT system. If the central LDH data indicates errors in

stratification at randomization such that the category chosen at randomization does not reflect the central LDH value, then a sensitivity analysis will be done for OS if there are >10% discrepancies. For this sensitivity analysis patients will be stratified based on the strata determined by the central LDH results, not the strata recorded in IRT.

Sensitivity analyses for the secondary endpoint of PFS are as following:

(a) Evaluation-time bias

In order to assess possible evaluation-time bias, which could occur if scans are not performed at the protocol-scheduled time points, the midpoint between the time of progression and the previous evaluable tumor assessment may be analyzed using a stratified log rank test, as described for the main analysis of PFS. For patients who die in the absence of progression, the date of death will be used to derive the PFS time used in the analysis. Note only tumour progression events will be re-derived using the midpoint approach. If a patient has progression on their first scan, then the midpoint between the PD date and randomisation date will be used. Patients will be considered to have had progression outside of the protocol scheduled timepoint if progression falls outside of the 12 weekly +/1 week interval (e.g. outside of days 77-91, 161-175, 245-259 etc)

(b) Attrition bias

Attrition bias will be assessed by

- repeating the PFS analysis except that the actual PFS event times, rather than the censored times, of patients who progressed or died in the absence of progression immediately following two, or more, non-evaluable/missed tumour assessments, will be included.
- In addition, patients who take subsequent therapy prior to their last evaluable RECIST assessment or progression or death will be censored at their last evaluable assessment prior to taking the subsequent therapy.

A reverse KM curve will also be produced where the censoring indicator for the primary PFS analysis will be reversed such that events will be censored, and censored observations will be treated as events. This method will also be used to estimate the median duration of follow up for PFS (and similarly for OS).

The table below summarize the programming logic for sensitivity analysis:

Table 4-2

Analysis	Situation	Situation –sub	Date of Event or Censor	Event / Censor
1		a) Progression occurred at a scan performed outside of the protocol-scheduled	The midpoint between the time of progression and the previous	event

Analysis	Situation	Situation –sub	Date of Event or Censor	Event / Censor
	Evaluation-time bias	time points (i.e PD occurred outside of Day 77-91(week 12), 161-175 (week 24), 245-259 (week 36) etc)	evaluable tumor assessment	
			Actual PD date, if PD within correct scan window	event
		b) PD on first scan and first scan outside of correct time windows	Midpoint between the PD date and randomization date	event
		c) death report without PD	Date of Death	Event
2	**Attrition bias**	a) PD regardless of missed scan visits	Actual progression time (i.e. ignore the 2 missed visit rule)	Event
		b) Died without PD and no subsequent therapy	Date of Death	Event
		c) Received a subsequent therapy prior to PD/death	Last evaluable tumor assessment prior to start date of first subsequent therapy	censor
		d) No PD/death and received a subsequent therapy	Last evaluable tumor assessment prior to start date of first subsequent therapy	censor

4.2.4 Exploratory Efficacy Analyses

4.2.4.1 Time to second PD (PFS2)

The time to second PD (PFS2) will be defined as the time from randomization until second progression or death due to any cause, whichever occurs first. Second progression is defined as any of the following observed at least 4 weeks after initial PD assessment (but excluding the visit to confirm the initial PD per the CRF visit label) per RECIST v1.1:

1) An additional $\geq$20% increase in tumor burden (sum of diameters of both target and new measurable lesions) accompanied by an absolute increase of $\geq$5 mm
2) Unequivocal PD of non-target lesions
3) New non-measurable lesions

Patients who have not had a second progression or died at the time of the analysis will be censored at the time of the last evaluable tumor assessment. PFS2 will be analysed similarly to the primary analysis of PFS.

4.2.4.2 Duration of treatment and response for patients treated beyond RECIST v1.1 PD

A table will be produced to summarize the number of patients who were treated beyond RECIST 1.1 progression and the duration of treatment beyond RECIST 1.1 progression (mean, median, SD, minimum and maximum as well as categories of <12 weeks, 12-24 weeks and > 24 weeks).

In addition, swimmer plots for patients continuing beyond RECIST 1.1 PD will be produced to show the duration of study treatment indicating time of RECIST 1.1 PD and second progression (if applicable).

4.3 PATIENT REPORTED OUTCOMES (PRO)

PRO data will be assessed using two established patient reported outcome instruments (EORTC QLQ-C30 and EQ-5D, 5L). All items/questionnaires will be scored according to published scoring guidelines or the developer's guidelines. Baseline will be defined as the last non-missing assessment prior to randomization.

The PRO assessments will be measured in all patients at specified time points and changes from baseline assessments will be assessed between the 2 treatment groups of tebentafusp and the investigator's choice. More detailed analyses of PRO to support reimbursement will be defined in a supplemental SAP.

4.3.1 EORTC-QLQ-C30

Description of Instrument:

The EORTC-QLQ-C30 is a multi-dimensional PRO instrument designed for assessing the PRO of cancer patients. It is identified as a recommended PRO instrument in an effectiveness guidance document of recommendations for incorporating PRO outcomes into clinical comparative effectiveness studies in adult oncology (Basch et al., 2012).

The EORTC-QLQ-C30 is a 30-item instrument. Patients record their responses to each of the items using closed-ended response options. Items 1–28 have four response options for patients to report on how they interpret their state of health at a given point in time. Items 29 and 30 have seven response options, ranging from 1 – very poor to 7 – excellent.

The 30 items are summarized into 15 PRO domains composed of:

- five functional domains (Physical, Role, Emotional, Cognitive, Social)
- nine symptom domains (Fatigue, Nausea and Vomiting, Pain, Dyspnea, Insomnia, Appetite Loss, Constipation, Diarrhea, Financial Difficulties)
- one Global Health Status/QoL domain.

Scoring the EORTC-QLQ-C30:

EORTC-QLQ-C30 domains will be scored and labelled in accordance with the QLQ-C30 scoring manual (EORTC Quality of Life Group, 2001). Scoring of each domain is the same in all cases.

- The average of the items that contribute to the domain is estimated; this becomes the raw score.
- Linear transformation is applied to standardize the raw score, so that scores range from 0 to 100.

A higher score represents a higher ("better") level of functioning for functional domains, and a higher ("worse") level of symptoms for symptom domains. A higher score for Global Health Status/QoL (calculated from items q29 and q30) represents a "better" level of global health status.

4.3.2 EQ-5D, 5L

Description of instrument

The EQ-5D, 5L is an instrument to measure self-reported overall health status in patients.

The EQ-5D, 5L descriptive system consists of five health dimensions: mobility, self-care, daily activities, pain, and anxiety. Patients rate each of the dimensions on five levels: no problems, slight problems, moderate problems, severe problems, and extreme problems. The responses from each of the five dimensions are combined into a five-digit number describing the respondent's health state and this can then be translated into a utility value by using published population norms.

The EQ-5D, 5L also uses a visual analog scale (VAS) where patients rate their health on a vertical scale with endpoints labelled 'the best health you can imagine' and 'the worst health you can imagine'. The data from the VAS can be used as a quantitative measure of health as judged by the individual patients.

Scoring the EQ-5D, 5L

The five dimensions and five levels of the EQ-5D-5L descriptive system describe 3,125 unique health states. An EQ-5D health state (represented by a five-digit number) may be converted to a single summary EQ-5D index by applying a formula that attaches weights to each of the levels in each dimension based on valuations from general population samples. The data will be combined with published norms for data summaries of the EQ-5D index value. (Reference: https://euroqol.org/eq-5d-instruments/how-can-eq-5d-be-used/where-is-eq-5d-used/)

The score from the EQ-5D VAS will be summarized as reported by the patient.

4.3.3 Health care resource utilization (HRU)

HRU is defined as any consumption of healthcare resources directly or indirectly related to the treatment of the patient and associated with the UM disease pathway. The case report form collects data on four aspects of HRU, as follows:

- inpatient hospitalizations

- concomitant medication
- medical procedures
- other measures of healthcare utilization

4.3.4 Compliance

The EORTC-QLQ-C30 and EQ-5D, 5L instruments are completed at baseline, at C1D1, at Day 1 of every other cycle through C5D1, then every 4th cycle thereafter beginning with C9D1 until (and including at) end of treatment (during the treatment period). Patients entering the disease progression follow-up period will continue with both EORTC-QLQ-C30 and EQ-5D, 5L assessments at 12-week intervals. During the survival follow-up phase, EQ-5D assessments will be continued to be taken every 3 months to inform post-progression health status.

At each assessment time point, the compliance rate for each of these PRO measures is defined as the proportion of patients who complete each instrument out of the expected number of patients who could complete the instruments (i.e., the number of subjects randomized for the baseline visit and still on treatment for post-baseline visits).

For the EORTC-QLQ-C30 at each assessment time point, domain compliance will be calculated as the total number of domain-compliant subjects divided by the total number of subjects who completed the instrument at that assessment point. Patients who complete at least 50% of QLQ-C30 items will be considered domain compliant (in line with the instrument scoring algorithm).

Compliance rates for each instrument and for each domain of the EORTC-QLQ-C30 will be presented overall and by treatment group at each time point Descriptive Statistics

Descriptive statistics (N, mean, standard deviation, median, interquartile range, minimum, and maximum) for each EORTC-QLQ-C30 domain score will be presented for baseline and at each assessment time point in the treatment and pre-progression study periods, including EOT. In addition to the absolute score, the change from baseline, and the % change from baseline will be calculated at each assessment and summarized in the same way.

Similar analyses will be conducted for the EQ-5D VAS, distinguishing data from the treatment and pre-progression phase and data from the survival follow-up phase of the study.

EQ-5D descriptive profiles will be generated at baseline and at each visit in the treatment, pre-progression and survival follow up phases. Responses to each level of the EQ-5D, 5L dimensions will be used to identify patients with 'no problems' (i.e., level 1) and 'problems' (i.e. levels 2 to 5), therefore changing the profile into frequencies of reported problems. The proportion of patients with 'no problems' at each visit will be summarized overall and by treatment group.

4.3.5 MMRM analysis of longitudinal data

The comparison between the two treatment groups of the change from baseline in PRO will be conducted by using mixed model repeated measures (MMRM) analyses to

estimate the extent of the difference between treatments in terms of change from baseline in each of the EORTC-QLQ-C30 domains. Each domain-specific model will include all cycles for which data were available for at least 10 patients in each treatment group. Each model will include covariates to account for the stratification by including a factor for LDH status (LDH above ULN vs normal LDH).

The MMRM model will have change from baseline in domain score as the outcome variable, with treatment group (binary), time point (categorical), LDH status (binary) and baseline score (continuous) as covariates. All analyses will be conducted using the PROC MIXED procedure in SAS.

From these MMRM models, the predicted mean changes from baseline in each domain score for each treatment group will be estimated (least squares means, LSmeans) to assess whether the impact of treatment on PRO differs between the two treatment groups. The LSmeans estimates of mean change from baseline at each cycle will be summarized for each treatment group at each time point and over time.

LSmean estimates over time will be plotted.

4.3.6 Time to sustained deterioration

Time to sustained deterioration will be summarised for all EORTC-QLQ-C30 domains using data from both the treatment and pre-progression phases of the study. A score will be considered to have deteriorated:

- if it is lower (worse) than the baseline score at any cycle for all EORTC-QLQ-C30 functional domains, and the EORTC-QLQ-C30 Global Health Status/QoL domain
- if it is higher (better) than the baseline score at any cycle for all EORTC-QLQ-C30 symptom domains.

A sustained deterioration will be defined as having at least 2 sequential visit responses of deterioration. Time to sustained deterioration will be the time of the first of the sequential deterioration visits. Patients who don't have 2 sequential deteriorations will be censored at the last QLQ-C30 assessment (randomization date if no baseline or no post-baseline QLQ-C30 assessment). Kaplan-Meier plot of time to sustained deterioration will be presented by treatment arm.

Missing data, due to withdrawals or uncompleted domain scores, will be considered to indicate deterioration at that cycle. For each domain, patients will be uniquely classified into one of the following ordered categories based on whether the patient's score has sustained deterioration from baseline and the earliest cycle at which this occurs (if deteriorated):

- deteriorated at cycle 3
- deteriorated at cycle 5
- …
- deteriorated at EOT
- not deteriorated.

Patients who have not deteriorated by the final on-treatment visit will be classified as not deteriorated. The proportion of patients in each treatment group who have sustained deteriorated at each visit will be summarized by treatment group, along with the proportion of patients who demonstrated sustained deteriorated at any time during the study.

4.3.7 Quantifying HRU

Because events that drive HRU can occur multiple times per patient, HRU will be assessed using rates of use, as well as frequency of use per patient and summarized by treatment group.

Descriptive summaries will be undertaken of inpatient hospitalization, concomitant medication, medical procedures, and other measures of healthcare resource use. Summaries will be presented by treatment group and overall. Per patient data will be listed. Summaries will include calculation of the use of each category of resource use (reported as proportions of patients using the resource) and the number of times the resource is used by each patient (summarized by mean, standard deviation, median, interquartile range, minimum, and maximum).

4.4 SAFETY ANALYSIS

All safety analyses will be performed on the Safety Analysis Set. Safety data presented by treatment group will be summarized on an 'as treated' basis. Safety and tolerability variables include treatment-emergent adverse events (TEAEs), deaths, clinical laboratory parameters, vital signs, 12-lead ECG results, physical examinations and extent of exposure. Study Day 1 for all safety analyses is defined as the date of the first dose of study drug. The overall observation period will be divided into 3 mutually exclusive segments:

- Pre-treatment period: from day of patient's screening informed consent to the day before first dose of study medication
- On-treatment period: from day of first dose of study medication to 90 days after last dose of study medication
- Post-treatment period: starting at Day 91 after last dose of study medication

4.4.1 Adverse Events

The adverse event verbatim descriptions (investigator terms from the CRF) will be classified into medical terminology using the Medical Dictionary for Regulatory Activities (MedDRA) and graded by National Cancer Institute Common Terminology Criteria for Adverse Events (NCI CTCAE) grade. Adverse events will be coded to primary System Organ Class (SOC) and preferred term (PT) using the latest version of MedDRA.

Treatment-emergent adverse events (TEAEs) are defined as AEs that started or worsened in severity from the date of first dose (regardless of time) up until 90 days after the last dose of study drug or until start of subsequent anti-cancer therapy, whichever occurs first. Only anti-cancer therapies with start dates that are greater than the date of last dose of

study drug (i.e. subsequent), are considered. In the case where it is not possible to define an AE as treatment-emergent or not, the AE will be classified by the worst case, i.e. treatment emergent.

Only those AEs that are treatment emergent will be included in summary tables. All AEs and serious adverse events (SAEs), treatment emergent or otherwise, will be presented in patient data listings and include date of onset, date of resolution (if AE is resolved), investigator's assessment of severity and relationship to study drug and period (pre-treatment, on treatment or post-treatment). AEs collected in the pre and post treatment windows will be flagged in the listings. There will be a separate listing of SAEs and AEs leading to death.

A patient data listing of all AEs leading to discontinuation from study treatment will also be provided. Any AEs in this period that occur after a patient has received further therapy for cancer (following discontinuation of study therapy) will be flagged in the data listings.

All TEAEs will be summarized descriptively by the MedDRA system organ class (SOC) and preferred term (PT) count with the incidence (n, number of patients) and incidence rate (%, percentage) for each treatment group. Tables will be produced for:

- All TEAEs
- TEAEs causally related (including 'possibly related' or 'related') to study medication
- TEAEs with maximum CTCAE grade 3 or 4
- TEAEs with maximum CTCAE grade 3 or 4, causally related to study medication
- TEAEs with outcome of death
- TEAEs with outcome of death causally related to study medication
- All SAEs (includes both deaths reportable as SAEs and non-fatal SAEs)
- All SAEs causally related to study medication
- TEAEs leading to permanent discontinuation of study therapy
- TEAEs leading to permanent discontinuation of study therapy, causally related to study therapy

For summaries by severity, TEAEs starting after the first dose of study drug with a missing severity will be classified as 'missing'. If a patient reports the same AE more than once within that SOC/PT, the AE with the worst-case severity will be used in the corresponding severity summaries.

For summaries by causal relationship to study drug, TEAEs with a missing relationship to study drug will be regarded as 'missing'.

For summaries of TEAEs leading to permanent discontinuation of study drug, these will be identified by using the "Action taken with the study drug" variable collected on the eCRF, where the variable is equal to 'Drug permanently discontinued'. These will be flagged in the listing.

Patients will be counted only once within a SOC and PT, even if the patient experienced more than one TEAE within a specific SOC and PT.

To adjust for potential differences on duration of drug exposure, the exposure adjusted incidence rate (EAIR) per 100 patient years will also be reported. The EAIR is defined as the number of subjects exposed to the drug who experience the event divided by the total exposure time of all subjects who are at risk of the event. For subjects with no reported event, the exposure time is the time from the date of first dose of study drug up until 90 days after the last dose of study drug or until the start of subsequent anti-cancer therapy, whichever occurs first (i.e. up to the end of the safety follow up period/subsequent therapy). For subjects who experience the event, the exposure time is the time from the date of first dose of study drug to the start date of the first event.

TEAEs with an incidence rate of at least 10% or TEAEs with a maximum grade of at least grade 3 with an incidence rate of at least 5% will be summarized as well.

4.4.1.1 AE of Special Interest

Incidence of TEAEs of Special Interest will be presented by TEAE of special interest categories and PT. Based on data from the ongoing and completed clinical trials, the Sponsor considers the following to be tebentafusp adverse events of special interest (AESI):

- Cytokine release syndrome (CRS)
- Rash
- Elevated liver enzymes

Due to evolution in grading and terminology, there may likely be under-reporting of CRS as an AE. Therefore, the incidence and severity of CRS will be based on a medical review utilizing the ASTCT consensus grading for CRS (Lee, et. al, 2019). By incorporating AE, concomitant medication, and vital sign data, a determination will be made as to whether CRS occurred after each patient's dose and, if so, at what grade. Using this approach, the incidence of CRS as an AESI will not match reports of CRS as an AE in AE-based tables.

Rash will be identified using a medically-approved list of PTs provided by the Sponsor. See Appendix 1 for details.

Elevated liver enzymes will be based on the MedDRA SMQ "Drug related hepatic disorders – comprehensive search." Both Narrow and Broad scope terms will be considered.

Incidence of TEAEs for Rash and Elevated Liver Enzyme preferred terms will also be summarized broken down by maximum CTCAE grade (Total, Grade 1, Grade 2, Grade 3, Grade 4, and Grade 5). CRS will be summarized by maximum grade but will not be summarized by preferred terms since the determination of CRS will not be purely based on AE reporting.

Toxicity plots will be produced for AESIs to show the frequency and timing, with respect to when the events occurred while patients were on-treatment.

Summaries for time to onset of each AESI (first AESI will be considered if a subject has multiple records of same AESI) will be provided.

Kaplan-Meier estimates along with Kaplan-Meier curves for the time to resolution of each AESI (first AESI will be considered if a subject has multiple records of same AESI) will be produced. Time to resolution of an AESI is defined as the date of resolution – date of onset + 1 for those subjects with first AESI resolved, and min (date of last dose plus 90 days, death date, data cut-off date) – date of onset + 1 for subjects with first AESI ongoing (i.e. censored subjects).

4.4.1.2 Deaths

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

- Death due to disease under investigation
- AE with outcome of death
- Treatment related AE with outcome of death
- Death due to other reason
- Death reason recorded as unknown by investigator

A corresponding listing will also be produced and will indicate if any AEs leading to death were within the treatment emergent period or >90 days after the date of last dose of study drug, or after a subsequent anti-cancer therapy.

4.4.2 Clinical Laboratory Parameters

For laboratory tests covered by the NCI CTCAE version 4.03 the study team will grade laboratory data accordingly. For laboratory tests covered by NCI CTCAE, a grade 0 will be assigned for all non-missing values not graded as 1 or higher. For laboratory tests where grades are not defined by CTCAE, results will be graded by the low/normal/high classifications based on laboratory normal ranges.

The following summaries will be provided for laboratory data. In general, these summaries will be repeated for haematology, biochemistry and urinalysis:

- Shift from baseline to each protocol scheduled visit and to the worst on-treatment/follow-up value according to NCI CTCAE grading system for quantitative measurements, as well as for urinalysis categorical measurements, will be provided. In addition, a shift table by ≥1 grade, ≥2 grades, and ≥3 grades shift from baseline to the worst during on-treatment/follow-up period will be produced for haematology and biochemistry respectively.
- For measurements that have an NCI CTCAE grading system, 'worst' will be defined as the maximum (i.e. most severe) Grade obtained during treatment; A missing value for Grade due to a missing laboratory value is considered to be the

least severe. **A special note:** For laboratory tests that have CTCAE grades defined for both a lower and higher level of extremity, separate shift from baseline summaries will be presented for each extreme of abnormality, e.g. (i) Lymphocytes Absolute Count – *Low* and (ii) Lymphocytes Absolute Count – *High*.

- For urinalysis categorical measurements, 'worst' will be defined as the maximum (i.e. most severe) result, where the results ranked in order from minimum severity to maximum are: 'Not done', 'NEG', '+', '++', and '+++'.
- Listing of all clinically relevant laboratory data with values flagged to show the corresponding NCI CTCAE grades and the classifications relative to the laboratory normal ranges. The worst grade will be flagged in the listing.

For both haematology and biochemistry, the following plots will be produced:

- Scatter plot of worst on-treatment/follow-up value versus baseline value, with values expressed as multiples of the lower limit normal or upper limit normal. Note that where applicable, parameters will be presented separately for both above and below the normal range, in the cases where 'worst' is defined as both extremes of abnormality.
- Line Plots with error bars will be used to display the mean changes over time for laboratory parameters

To assess potential drug-induced liver injury (DILI) using Hy's Law, the following summaries for liver function tests will be produced:

- Incidence of patients with ALT or AST $\geq$ 3xULN and total bilirubin $\geq$ 2xULN within specified time intervals. The time intervals that will be summarized are the following:
    - ALT/AST and total bilirubin elevation results at any time during on-treatment/follow-up.
    - ALT/AST elevation result within one week (+/- 7 days) of total bilirubin elevation result, during on-treatment/follow-up.
    - ALT/AST elevation result within one day (+/- 1 day) of total bilirubin elevation result, during on-treatment/follow-up.
    - ALT/AST elevation results within one day (+/- 1 day) of total bilirubin elevation results with a duration of $\geq$ 7 days, during on-treatment/follow-up. Duration is calculated as consecutive days (subsequent lab date – previous lab date) where the elevated levels of both (ALT or AST $\geq$ 3xULN) and total bilirubin $\geq$ 2xULN are maintained, without going below the potential Hy's Law criteria.
- A scatter plot of worst on-treatment/follow-up value for ALT versus total bilirubin, with values expressed as multiples of the upper limit normal will be produced. This scatterplot will be repeated for AST versus total bilirubin.

4.4.3 Vital Signs, Physical Examination Findings, and ECG

4.4.3.1 Vital Signs

Vital signs (body temperature, pulse rate, respiratory rate, and blood pressure) must be performed before dosing, and after the treatment administration (at least twice after administration) and as per institutional standards.

The following summaries will be provided for vital signs data:

- Shift from pre-dose to the post-dose at each dosing schedule will be summarized for each vital sign.
- Shift from baseline to the worst on-treatment/follow-up value (for all quantitative measurements) according to the predefined criteria provided in the table below:
    - The 'worst' on-treatment/follow-up value will be defined as the maximum (i.e. most severe) abnormality criterion obtained during treatment. A missing value due to a missing vital signs value is considered to be the least severe.

Variable	Unit	Low	High
SBP	mmHg	< 90 mmHg Post-baseline: Decrease from baseline of: (1) < 15 mmHg (2) ≥ 15 mmHg and < 35 mmHg (3) ≥ 35 mmHg	CTCAE Grade 1: 120 to 139 mmHg CTCAE Grade 2: 140 to 159 mmHg CTCAE Grade 3: ≥ 160 mmHg Post-baseline: Increase from baseline of: (1) < 15 mmHg (2) ≥ 15 mmHg and < 35 mmHg (3) ≥ 35 mmHg
DBP	mmHg	< 60 mmHg	CTCAE Grade 1: 80 to 89 mmHg CTCAE Grade 2: 90 to 99 mmHg CTCAE Grade 3: ≥ 100 mmHg
Respiratory Rate	breaths/min	<10 breaths/min	Baseline: ≥ 20 breaths/min Post-baseline: Do 1 of following: (1) If Low/Normal at baseline: ≥ 20 breaths/min OR change from baseline ≥ 10 breaths/min (2) If High at baseline: ≥ 20 breaths/min

Variable	Unit	Low	High
Pulse rate	bpm	Baseline: ≤ 60 bpm Post-baseline: Do 1 of following: (1) If Normal/High at baseline: ≤ 50 bpm AND change from baseline ≤ -15 bpm (2) If Low at baseline: ≤ 50 bpm	Baseline: ≥ 100 bpm Post-baseline: Do 1 of following: (1) If Low/Normal at baseline: ≥ 100 bpm AND change from baseline ≥ 15 bpm (2) If High at baseline: ≥ 100 bpm
Body Temperature	°C	N/A	CTCAE Grade 1: 38.0 – 39.0 °C CTCAE Grade 2: >39.0 – 40.0 °C CTCAE Grade 3: >40.0 °C for consecutive <=24 hours CTCAE Grade 4: >40.0 °C for consecutive >24 hours
Weight	kg	Baseline: N/A Post-baseline: percentage change from baseline ≤ -10.0 %	Baseline: N/A Post-baseline: percentage change from baseline ≥ 10.0 %

Note: Since Weight at baseline has no Low or High markedly abnormal criteria, the baseline value will be represented in summary tables and listings as "Baseline". For other vital signs variables and weight post-baseline, a value that is not Low nor High will be represented as "Normal".

The following plots will be produced for all assessment time points and the repeated for all assessments after first dose of study drug up to a specific period of time.

- Plot of patient profiles over time for all patients with at least one hypotension or cytokine release syndrome event of CTCAE grade ≥ 3
- Plot of patient profiles over time for all patients with at least one SBP decrease from baseline ≥ 35mmHg

Plots with error bars will be used to display the mean changes over time for vital signs, SBP and DBP Values meeting markedly abnormal criteria will be flagged in the listing.

4.4.3.2 Physical Examinations

Physical examination data will be presented in listing and all the abnormalities will be flagged.

4.4.3.3 ECG

The following ECG parameters will be reported for this study:

- QT, QTcB, QTcF, PR, RR and QRS Intervals
- Overall assessment of ECG (Investigator's judgment):

- Normal
- Abnormal, Not Clinically Significant (ANCS)
- Abnormal, Clinically Significant (ACS)

The following summaries will be provided for ECG data:

- Shift from baseline to the worst on-treatment/follow-up value according to the predefined criteria for abnormality for prolonged QTcB and QTcF (see criteria below).
    - A value below the lowest threshold of concern, i.e. ≤ 450 msec, will be assigned a value of 'Normal'. The 'worst' on-treatment/follow-up value will be defined as the maximum (i.e. most severe) abnormality criterion obtained during treatment. A missing value due to a missing ECG value is considered to be the least severe
- Shift from baseline to worst on-treatment/follow-up value for PR, RR and QRS interval abnormalities.
    - Normal range criteria will be used to define 'worst' as both extremes of abnormality (lowest and highest). A missing value due to a missing ECG value is considered to be the least severe.

Values meeting markedly abnormal criteria will be flagged in the listing.

For rating the 'worst' on-treatment/follow-up value for QTcB and QTcF absolute values, the following categories will be used, in the order of least to most severe (top to bottom):

- Missing result
- ≤ 450 msec (i.e. 'Normal')
- >450-480 msec
- >480-500 msec
- >500 msec

For rating the 'worst' on-treatment/follow-up value for QTcB and QTcF change from baseline values, the following categories will be used, in the order of least to most severe (top to bottom):

- Missing result
- ≤ 30 msec increase from baseline
- 30 to ≤ 60 msec increase from baseline> 60 msec increase from baseline

For rating the 'worst' on-treatment/follow-up value for PR and QRS intervals the values will be graded by the following low/normal/high classifications based on the normal range criteria, in order of least to most severe (top to bottom):

PR Interval:

- Missing result
- <120 msec (i.e. 'Low')
- 120-200 msec (i.e. 'Normal')
- >200 msec (i.e. 'High')

QRS interval:

- Missing result
- <80 msec (i.e. 'Low')
- 80-120 msec (i.e. 'Normal')
- >120 msec (i.e. 'High')

RR interval:

- Missing result
- <600 msec (i.e. 'Low')
- 600-1200 msec (i.e. 'Normal')
- >1200 msec (i.e. 'High')

4.4.4 Pharmacokinetics

Due to limited PK data collection, PK parameters will not be derived for this study. Serum PK concentration over time will be listed, summarized and displayed graphically based on the PK analysis set.

Plasma concentration values below the lower limit of quantification (<LLOQ) will be handled as follows:

- If, at a given time point, the minimum observed value is <LLOQ the plasma concentrations is <LLOQ, the geometric mean, and geometric mean ± SD will not be calculated. Values LLOQ will be set to zero (0) for the calculation of arithmetic mean, standard deviation, median, and CV.
- If more than 50%, but not all, of the concentrations are NQ, the geometric mean, CV, geometric mean ± SD, arithmetic mean and SD will be reported as not calculable (NC).

- If all the concentrations are <LLOQ, the geometric mean and arithmetic mean will be reported as <LLOQ and the CV, geometric mean ± SD and SD as NC.
- The number of values above LLOQ will be reported for each time point along with the total number of collected values. If data are available for fewer than 3 patients, no summary statistics other than minimum, maximum and n will be presented.

Mean and individual plasma concentrations will be displayed graphically on linear and semilogarithmic scales. For display purposes, mean and individual concentrations which are <LLOQ (100 pg/mL) will be displayed as ½ the LLOQ (50 pg/mL) graphically.

4.4.5 Tolerability

Tolerability of study treatment will be assessed by summarizing the number of treatment dose interruptions and dose reductions. Reasons for dose interruptions and dose reductions will be listed by patient and summarized.

4.4.6 Extent of Exposure

The following data will be summarised to describe exposure to gp100 and Investigator choice

A cycle is defined as 21 days = 3 weeks, and there is 1 dose administration visit every week.

- Number of cycles started = total number of cycles of study drug received (including partial cycles and reduced doses)
- Number of cycles completed = total number of complete cycles of study drug received without any interruption (but including reduced doses)
- Duration of treatment (days) = (date of treatment discontinuation – date of first study drug administration + 1). Note the date of treatment discontinuation is the date recorded on the "End of Treatment" CRF. Patients for whom a treatment discontinuation date has not been entered on the EOT CRF (e.g. still ongoing treatment at the time of data cut off for reporting or died prior to making a decision to discontinue treatment) then the earliest of (last dose date + 6 days) or the date of death will be used for deriving duration of treatment.
- Duration of interruption (days) = (start date of next study drug administration following interruption - date of visit in which interruption started). Note that this is calculated for each interruption a given patient has and is only calculated where study drug administration restarts following interruption.
-
- Relative dose intensity (%) = (Total Actual Dose received / Total Planned dose) × 100

4.5 IMMUNOGENICITY

ADA (Anti-drug Antibody) analysis will be used to evaluate immunogenicity responses

for tebentafusp.

The following ADA data will be provided by a third party vendor:

- ADA result
  - Positive: Final result in sample results spreadsheet (SRS) is Positive
  - Negative: Final result in SRS is Negative and PK value is < drug tolerance limit of assay (200 ng/mL, which is equivalent to 200,000 pg/mL)
  - Inconclusive: Final result in SRS is Negative and PK value is ≥ drug tolerance limit of assay (200 ng/mL, which is equivalent to 200,000 pg/mL) or unknown
  - Unevaluable: Sample was unable to be analyzed (insufficient volume, wrong matrix, etc.)
- Titer value
- Neutralizing ADA (NAb) result
- Neutralizing ADA titer value

The following table gives the variables that will be derived for ADA results and titers. These variables will be additionally derived for Neutralizing ADA (NAb) results and titers in a similar fashion.

Variable	Definition
Baseline ADA result & Baseline ADA titer	Closest sample prior to dosing
Pre-existing ADA	Subject with Positive baseline ADA result (without a boost in titer in response to study drug administration). See below for definition of Treatment-boosted (≥4-fold) ADA.
ADA *prevalence* at Baseline	The number of subjects with a Positive ADA result at baseline as a percentage of the total number of subjects tested at baseline for ADA.
ADA Evaluable Subset	All subjects who received at least one dose of study drug and have at least one ADA assessment post-baseline. This subset of the SAF will be used for determining ADA incidence.
Max titer	Highest titer value post-baseline.
Peak (or max) fold increase in titer	Ratio of max post-baseline titer to baseline titer (calculated only for subjects with a Positive ADA result at baseline).
Treatment-induced ADA	Subject in the ADA Evaluable Subset who has a positive ADA sample post-baseline with a Negative ADA result at baseline.
Treatment-induced ADA *incidence*	Number of treatment-induced ADA subjects / number of subjects in ADA Evaluable Subset with a Negative ADA result at baseline.

Variable	Definition
Treatment-boosted (≥4-fold) ADA	Subject in the ADA Evaluable Subset who has a Positive ADA sample at baseline and a Positive ADA sample post-baseline with a titer that has a peak (or max) fold increase in titer ≥4 compared to baseline.
Treatment-boosted (≥4-fold) ADA *incidence*	Number of treatment-boosted (≥4-fold) ADA subjects / number of subjects in ADA Evaluable Subset with a Positive ADA result at baseline.
ADA Incidence (ADA Positive Subjects)	All subjects with a treatment-induced or treatment-boosted ADA response (see definitions above) or subjects with positive post-baseline ADA sample but do not have a baseline ADA sample in the ADA Evaluable Subset.
ADA Negative Subjects	All subjects without a treatment-induced nor treatment-boosted ADA response in the ADA Evaluable Subset (can include subjects classified as Pre-existing ADA).
ADA Status	Three categories: a) Unevaluable: Patient has no post-baseline ADA samples. b) Positive: See "ADA Incidence (ADA Positive Subjects)" definition above. c) Negative: See "ADA Negative Subjects" definition above (can include subjects classified as Pre-existing ADA).
ADA Onset (as applicable)	For subjects with a treatment-induced ADA response: number of days from first dose of study drug to the first instance of Positive ADA. Therefore, ADA Onset = (date of first instance of Positive ADA – date of first dose of study drug + 1).
ADA Duration (as applicable)	For subjects with a treatment-induced ADA response: number of days from the first instance of Positive ADA to last instance of Positive ADA for a subject, such that a subsequent Negative ADA follows the last instance of Positive ADA. Therefore, ADA Duration = (date of last Positive ADA* – date of first instance of Positive ADA + 1). *The last Positive has to exist such that a Negative ADA follows the last instance of Positive ADA. If the date of last Positive ADA result is the *final* ADA assessment, then ADA duration will be calculated above, but in this case, the duration will be concatenated with '+' to imply that the ADA duration is *at least* the calculated number of days.
Transient ADA response	ADA Positive subject (post-baseline) with at least one subsequent Negative result, after the last Positive result and the ADA Duration is < 20 weeks (i.e. < 140 days).
Persistent ADA response	Subject with either (i) an ADA Duration ≥ 20 weeks (i.e. ≥ 140 days), regardless of whether intervening sample results are Positive or Negative; or, (ii) if last sample is ADA Positive (i.e. where ADA

Variable	Definition
	Duration is concatenated with '+'as described above).

Missing ADA values will be reported as is in data listings, as 'No Result' or 'No Sample'.

4.5.1 ANALYSIS OF ADA

The ADA analysis comprises two parts; the ADA status summary at a patient-level and the study ADA summary. The safety analysis set will be used for all ADA summaries and listings.

Details are summarised below:

4.5.1.1 Patient ADA Status

For each patient, the final ADA status (that takes into account the cumulative ADA sample results for that patient) will be summarized. This will include:

- ADA status (Unevaluable, Positive, Negative)
- ADA Characterisation (Treatment-induced, Treatment-boosted, or Pre-existing ADA)
- Time to ADA Onset for treatment-induced ADA patients
- ADA Duration category (Transient ADA response or Persistent ADA response)
- Max titer for treatment-induced ADA patients
- Peak fold increase in titer for treatment-boosted ADA patients
- Neutralizing ADA (NAb) activity
- Neutralizing ADA (NAb) max titer

4.5.1.2 Study ADA Summary

At a study level, the ADA data summaries will include:

- The number (%) of patients who are ADA-positive at baseline ('ADA prevalence at Baseline').
- The number (%) of evaluable patients ('ADA Evaluable Subset').
- The number (%) of patients who are ADA-positive at follow-up ['ADA Incidence (ADA Positive Subjects)'].
- The number (%) patients with a treatment-induced ADA (from baseline negative) – see Treatment-induced ADA incidence definition above.
- The number (%) patients with a treatment-boosted (≥4-fold) (from baseline positive) - see Treatment-boosted (≥4-fold) ADA incidence definition above.
- The max titer from patients with treatment-induced ADA. Descriptive statistics including the median, IQR of the max titer will also be shown.
- The peak fold increase in titer among patients with treatment-boosted (≥4-fold)

ADA. Descriptive statistics including the median, IQR of peak titer fold increases will also be shown.

- A graphical representation of time to ADA onset and ADA duration for patients with treatment-induced ADA. Descriptive statistics (median, minimum and maximum) may also be summarised for time to ADA onset and ADA duration.

Two versions of data summaries for ADA will be presented, for the regular ADA results and a repeat for the Neutralizing ADA (NAb) results.

For all patients, all Sample ADA values (including regular ADA results and NAb results) collected at baseline and post-baseline will be listed.

4.6 OTHER EXPLORATORY ANALYSIS

4.6.1 Biomarkers

The following are exploratory, biomarker-related objectives of the study:

- Correlation of the expression of T cell infiltration, expression of gp100, human leukocyte antigen DR (HLA-DR), PD-L1, tumoral lymphocyte activation status, and myeloid-derived suppressor cell (MDSC) infiltration and other immune markers evaluated in tumor biopsies with anti-tumor activity
- Changes in serum cytokine, chemokines (eg, CXCL9, CXCL10, hepatocyte growth factor [HGF], interleukin 1 receptor alpha [IL-1Rα], and monocyte chemoattractant protein-1 [MCP-1]), or other analytes in response to treatment

Data from other tebentafusp trials is still emerging to help shape the direction of the exploratory biomarker work. Once a clearer understanding of the key biomarkers of interest is known, a more detailed analysis plan may be developed to document the biomarker exploration work. This may be added into the SAP prior to the primary analysis or may be documented in a separate biomarker analysis plan.

5 INTERIM ANALYSIS AND DATA MONITORING COMMITTEE (DMC)

5.1 INTERIM ANALYSIS

Two interim analyses will be performed using a 3-stage group sequential design (**Error! Reference source not found.**). Separate analyses will be carried out for each of the two primary objectives, but the analyses will occur at the same time and the timing will be driven by the number of events in the ITT population. The first interim analysis will be based on approximately 60% of the events (150 events) and the second interim analysis will be based on approximately 80% of the events (200 events). Analyses of OS will be based on O'Brien-Fleming boundaries (O'Brien and Fleming, 1979). The Lan-DeMets approach (Lan and DeMets, 1983) that approximates the O'Brien-Fleming spending function will be used to adjust for situations where the actual number of events up to the data cut-off date for a given interim analysis does not match the planned number.

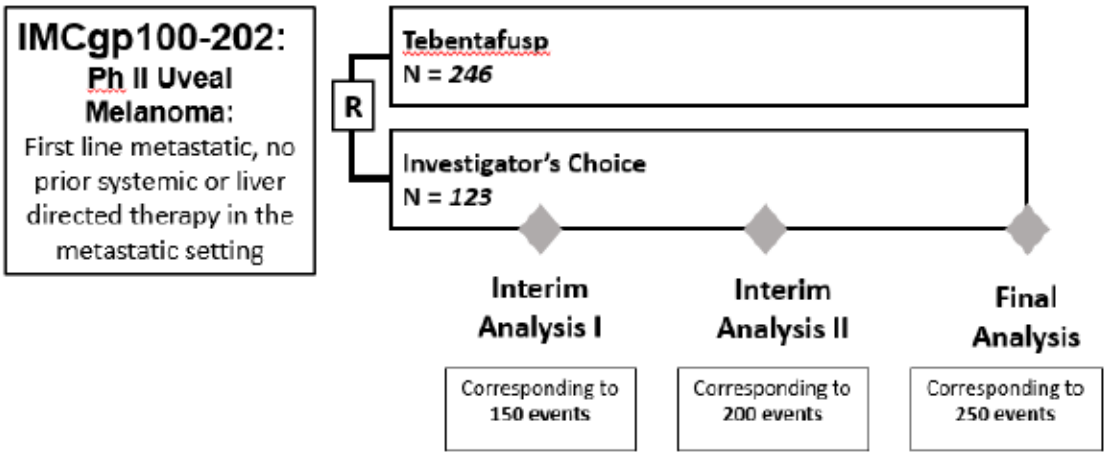


Figure 5.1. Study Design and Interim Analyses for the ITT Population

As described in Section 1.2.3, α=0.045 will be devoted to the ITT analyses and α=0.005 will be devoted to analyses in the RAS, thereby controlling the overall experiment-wise error rate at 0.05. The significance level for the interim OS analyses will be calculated by specifying the information fraction for each analysis. The information fraction is calculated as the actual number of OS events up to the data cut-off date divided by the total required number of events for the final analysis. For example, for an interim analysis conducted after 150 death events, the information fraction would be entered as 0.60 (150/250 events). This would result in a stopping boundary on the p-value scale of 0.006 (2-sided) for the first interim analysis. The first interim analysis of OS will occur after the end of accrual.

The stopping boundaries are described in the table 5-1 below:

Table 5-1 Stopping Boundaries for Given Alpha Levels and Information Fractions (IF)

Analysis Set (allocated alpha)	Analysis Number	OS Events (IF)	Lower Z-Boundary	Upper Z-boundary	Nominal Alpha (2-sided)	Cumlative Alpha
RAS (α_{RAS}=0.005)	First Interim	99 (60%)	-3.732	3.732	0.0002	<0.001
	Second Interim	131 (80%)	-3.198	3.198	0.0014	0.001
	Final	164 (100%)	-2.838	2.838	0.0045	0.005
ITT (α_{ITT}=0.045)	First Interim	150 (60%)	-2.724	2.724	0.006	0.006
	Second Interim	200 (80%)	-2.336	2.336	0.019	0.021
	Final	250	-2.073	2.073	0.038	0.045

		(100%)				
ITT (α_{ITT}=0.05)	First Interim	150 (60%)	-2.669	2.669	0.008	0.008
	Second Interim	200 (80%)	-2.289	2.289	0.022	0.024
	Final	250 (100%)	-2.031	2.031	0.042	0.050

For the first interim analysis, the RAS analysis will be conducted first. If the stopping boundaries are crossed, then the full alpha (α=0.005) will be added to the ITT analyses, resulting in a total α_{ITT}=0.05. If, at a particular interim analysis, stopping boundaries are crossed for the RAS analysis but not the ITT analysis, then the ITT alpha allocation for future analyses will be adjusted as described by Maurer and Bretz (Maurer and Bretz, 2013), Stopping boundaries for the ITT analyses will therefore be adjusted accordingly. Secondary endpoints such as PFS and BOR will not be formally tested in the RAS (see Section 5.2 and Appendix 2).

5.2 Multiple Testing Strategy

In order to provide strong control of the type I error rate, the primary endpoint of OS and key secondary endpoints, PFS and BOR, will be tested in this sequential order for the ITT analyses. If any previous analysis in the sequence is not statistically significant, the alpha cannot be transferred to subsequent analyses. Since neither PFS nor BOR will be tested on an interim basis, both endpoints will be tested at an overall alpha level of 0.045 (two-sided) or 0.05 if the alpha from the RAS analysis is transferred to the ITT analyses (see Section 5.1 and Appendix 2).

The final analysis of PFS will be based on 274 progression and death events. Assuming a median PFS of 5 months in the tebentafusp arm and 3.3 months in the Investigator's Choice arm, an HR of 0.66, and the same accrual and drop-out assumptions described earlier for OS, the analysis of PFS is expected to have 90% power to demonstrate a difference in the survival distributions between the two arms. Due to the higher hazard of progression events relative to death events, the required number of PFS events is expected to occur prior to the first interim analysis of OS. If the required number of events have not occurred at the time of the first OS interim analysis, then the PFS analysis will occur at a later time when 274 progression events have occurred.

The analysis of BOR will occur after all 369 randomized patients have been followed for approximately 9 months (ie, three planned assessments) for response. It will only be formally tested for statistical significance if null hypotheses for OS and PFS in the ITT population are both rejected.

5.3 Data Monitoring Committee

An IDMC will be established to provide oversight of safety and efficacy considerations in the current protocol (IMCgp100-202). The IDMC will act in an advisory capacity and

make recommendations regarding steps to ensure both patient safety and the ethical integrity of the trial. The voting members of the committee are external to Immunocore and will not otherwise be involved with the trial. The IDMC will include three clinicians experienced in oncology/melanoma and one statistician. Specific details regarding IDMC responsibilities, governance, and documentation will be described in a separate charter that is reviewed and approved by the IDMC members. Immunocore has primary responsibility for design and conduct of the study.

The IDMC will recommend if the trial should continue in accordance with the protocol. The IDMC will also monitor trial safety data 3 times annually, or at a frequency described in the IDMC charter and efficacy data at the interim analysis (stage 2), to evaluate the overall benefit-risk profile to ensure the ongoing protection of the patients enrolled in the study.

6 SUMMARY OF CHANGES IN THE PLANNED ANALYSES

Changes in planned analyses from one version of the protocol to the next are documented in protocol amendments. In this SAP, there are no additional changes in the planned analyses from the version of the protocol referenced in Section 1.0 to report.

7 REFERENCES

Basch E, Abernethy AP, Mullins CD. Recommendations for Incorporating Patient-Reported Outcomes Into Clinical Comparative Effectiveness. Research in Adult Oncology. J Clin Oncol 2012; 30:4249-4255.

Lan KK and DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika. 1983; 70(3):659-63.

O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35(3):549–56.

Efron, B., "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association*, 1977; 72: 557–565.

EORTC Quality of Life Group. (2001) EORTC QLQ-C30 Scoring Manual. EORTC Quality of Life Group 2001, http://groups.eortc.be/qol/manuals

FDA Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims, 2009.

Lin, D., Wei, L. J., and Ying, Z. "Checking the Cox Model with Cumulative Sums of Martingale-Based Residuals," Biometrika, 80, 557–572, 1993.

Maurer W, F Bretz, Multiple Testing in Group Sequential Trials Using Graphical Approach, Statistics in Biopharmaceutical research, 2013

8 APPENDIX

APPENDIX 1: PROCESS FOR CATEGORIZING SKIN TOXICITIES ASSOCIATED WITH TEBENTAFUSP, INCLUDING RASH

Since gp100 is expressed on melanocytes on the skin, tebentafusp was expected to induce a skin rash and indeed this was observed. These AE may present in different ways and be reported by investigators using different AE preferred terms (PTs). Therefore, for the purpose of detailed safety evaluations, Immunocore designed a process to determine a composite list of rash:

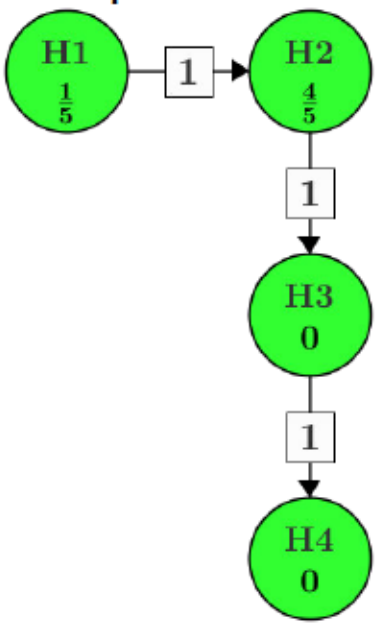
1) All unique MedDRA PTs under the System Organ Class (SOC) of "Skin and subcutaneous tissue disorders" from study IMCgp100-102 were listed and reviewed by the study's medical monitor.
2) PTs were then grouped into only one of several skin toxicity composite terms based this medical review:
    a. Rash
    b. Pruritis
    c. Pigment change
    d. Erythema
    e. Edema
    f. Dry skin
    g. Other changes
3) Due to their suspected relationship to the tebentafusp mechanism of action, the following PTs from other SOCs were also added to these categories:
    a. Eye pruritus (to Pruritis)
    b. Eyelash hypopigmentation (to Pigment change)
    c. Periorbital oedema (to Edema)
4) The process was repeated for the first-in-human study IMCgp100-01, which included some new PTs in the Skin and subcutaneous tissue disorders SOC that were not reported in study IMCgp100-102 and a few PTs from other SOCs (Eyelash discolouration, Erythema of the eyelid, Skin abrasion, Eyelid eodema). All of these PTs were added to the relevant composite lists.
5) The composite lists were reviewed by oncologists who were high volume enrollers on the tebentafusp clinical trials. Based on their input, only a few PTs were adjusted by from one composite list to another.
6) The resulting list of rash PTs is: Blister, Dermatitis acneiform, Dermatitis allergic, Dermatitis bullous, Dermatitis exfoliative, Drug eruption, Eczema, Palmar-plantar erythrodysaesthesia syndrome, Papule, Psoriasis, Rash, Rash erythematous, Rash generalised, Rash macular, Rash maculo-papular, Rash papular, Rash pruritic, Rash vesicular, Skin abrasion, Skin exfoliation, Urticaria
7) Prior to any snapshot for a formal efficacy analysis in Study IMCgp100-202 ("the 202 study") the following steps will occur:
    a. A list of terms in the "Skin and subcutaneous tissue disorders" SOC from the 202 study that have not already occurred in studies IMCgp100-01 or IMCgp100-102

will be reviewed in a blinded manner. Clinicians who review the list will not have knowledge of the treatment group the term occurred in or the timing of the event relative to first dose.

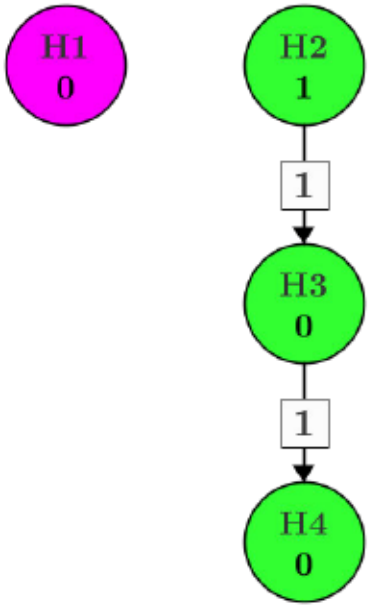
b. Terms that are agreed to be a part of the rash phenomenon will be added to the list above in Step #6 and used to determine the Rash Analysis Set (along with the study day of the event).

APPENDIX 2 MULTIPLE TESTING ALPHA ALLOCATION: GRAPHICAL APPROACH

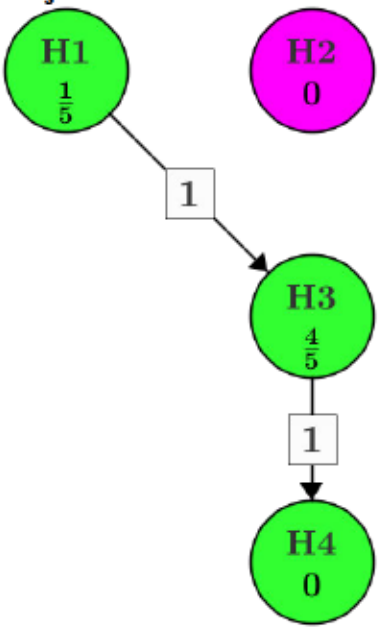
Initial Graph:



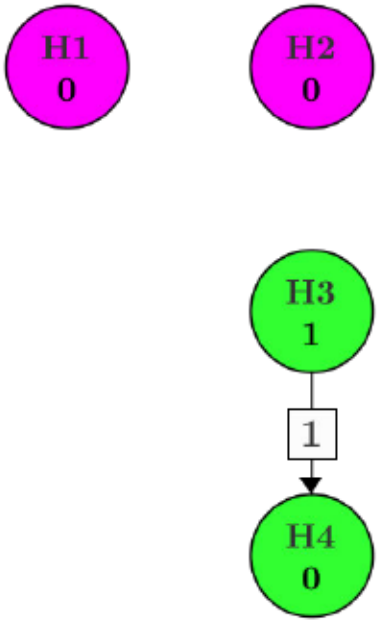
H1 Rejected:



H2 Rejected:



H1 & H2 Rejected:



H1 = OS in the Rash Analysis Set (RAS), H2 = OS in ITT, H3 = PFS in ITT, H4 = BOR in ITT