Statistical Analysis Plan J2J-OX-JZLC Version 5

EMBER-3: A Phase 3, Randomized, Open-Label Study of Imlunestrant, Investigator's Choice of Endocrine Therapy, and Imlunestrant Plus Abemaciclib in Patients With Estrogen Receptor Positive, HER2 Negative Locally Advanced or Metastatic Breast Cancer Previously Treated With Endocrine Therapy

NCT04975308

Approval Date: 24-Aug-2023

EMBER-3: A Phase 3, Randomized, Open-Label Study of Imlunestrant, Investigator's Choice of Endocrine Therapy, and Imlunestrant plus Abemaciclib in Patients with Estrogen Receptor Positive, HER2 Negative Locally Advanced or Metastatic Breast Cancer Previously Treated with Endocrine Therapy

Protocol Number: J2J-OX-JZLC

Compound Number: LY3484356

Sponsor Name: Eli Lilly and Company

Legal Registered Address: Indianapolis, IN, USA 46285

Regulatory Agency Identifier Number(s): 145311

Document ID: VV-CLIN-122802

Approved on 24 Aug 2023 GMT

Table of Contents

Version History

This Statistical Analysis Plan (SAP) Version 1 for study J2J-OX-JZLC is based on the protocol dated 15MAR2021. This version has been approved prior to the first patient visit of the study.

This SAP Version 2 for study J2J-OX-JZLC is based on the protocol approved on 01 October 2021. This version has been approved prior to the first patient visit of Arm C.

SAP Version 3 is approved before the first interim analysis.

SAP Version 4 is approved before the first interim analysis.

SAP Version 5 is based on the protocol amendment (d) approved on 31 July, 2023 and is approved before the final PFS analysis, but after the futility interim for PFS comparing Arm A vs Arm B in the ITT population and the futility interim for PFS comparing Arm C vs Arm A in the ITT population. Both futility analyses were conducted by the independent data monitoring committee (DMC). Unblinded interim data were only accessible by the independent DMC and the independent Statistical Analysis Center.

SAP Version History Summary

SAP Version	Approval Date	Change	Rationale
1	07/26/2021	Not applicable	Original version
2	01/04/2022	Added Arm C; PFS and OS analyses for Arm C versus Arm A; sample size determination per protocol amendment(a); other editorial changes	Version 2, updated per protocol amendment (a)
3	03/24/2022	Clarified the timing of the second interim analysis for A versus B and the timing of the first interim for C versus A. No changes on the required number of events.	Version 3, clarification on interim analysis timings
4	See approval date on page 1	Per amendment (b): added a third primary PFS endpoint and a third OS endpoint in the *ESR1*-mutation detected population for Arm A versus Arm B, changed the testing scheme to the graphical approach in Section 4.3.3. Per amendment (c):increased the sample size for arms A,B and C in Section 5; changed the enrollment strategy to close all arms simultaneously in Section 1.2 and 5; increased the event number for PFS between A versus B in ITT (and accordingly for PFS between A versus B in *ESR1*-mutation detected population) in Section 4.3 and updated interim analysis plan in Section 4.9, increased the event number for OS analyses and interim analysis plan accordingly in Section 4.4; added an exploratory endpoint for Arm C versus Arm B in Section 4.5. Other changes are either editorial or for clarifications.	Version 4, updated per protocol amendment (b) and (c)
5	See approval date on page 1	Per amendment (d), changed the initial alpha allocation for PFS between Arm A versus Arm B in ITT and PFS between Arm A versus Arm B in *ESR1*-mutation detected population in Section 4.3.3; removed the efficacy interim analysis for Arm A versus Arm B in the ITT population in Section 4.9.3; updated the interim analysis plan for OS analyses due to the change of initial alpha allocation for PFS endpoints in Section 4.4.1. Added more PFS sensitivity analyses in Section 4.3.4. Other changes are either editorial or for clarifications.	Version 5, updated per protocol amendment (d)

Abbreviations: ITT = intention-to-treat; OS = overall survival; PFS = progression-free survival; SAP = statistical analysis plan.

1. Introduction

Study J2J-OX-JZLC is a Phase 3 global, randomized, open-label confirmatory study of 2 comparisons: imlunestrant (Arm A) versus Investigator's Choice of Endocrine Therapy of either fulvestrant or exemestane (Arm B), as well as imlunestrant plus abemaciclib (Arm C) versus imlunestrant (Arm A), for patients with ER+, HER2- locally advanced or metastatic breast cancer previously treated with an aromatase inhibitor (AI), with or without a CDK4/6 inhibitor. Prior treatment with a CDK4/6 inhibitor is expected if this treatment is approved and reimbursed.

The first version of protocol for the study J2J-OX-JZLC was approved on 15 March 2021, which is the basis of the SAP Version 1. SAP Version 1 was approved on 26 July 2021, which was prior to the first patient visit of the study.

The protocol amendment (a) for the study was approved on 01 October 2021, which is the basis of the SAP Version 2. SAP Version 2 was approved on 04 January 2022, which was prior to the first patient visit of the amendment(a).

SAP Version 3 is approved before the first interim analysis. The purpose of SAP Version 3 is to clarify the timing for the second interim analysis of progression-free survival (PFS) for Arm A versus Arm B and the timing for the first interim analysis of PFS for Arm C versus Arm A. The number of events and boundaries for the interim analysis remain the same.

Protocol amendment (b) [approved on 17 August 2022], and protocol amendment (c) [approved on 11 November 2022] are the basis for the SAP Version 4. SAP Version 4 has been approved before the first interim analysis in this study (i.e., futility interim for Arm A versus Arm B in the ITT population, which is projected to occur in Q1 2023).

Protocol amendment (d) [approved on 31 July 2023] is the basis for the SAP Version 5. In this version, the initial alpha allocation for PFS between Arm A versus Arm B in ITT and PFS between Arm A versus Arm B in *ESR1*-mutation detected population has been updated to 0.005 and 0.02 respectively, due to external oral SERD data showing PFS has been primarily driven by the *ESR1*-mutation detected subgroup. The efficacy interim analysis (originally the second interim for PFS between Arm A and Arm B in the ITT population) has been removed since the timing of this interim is projected to be closer to the timing of the final PFS analysis based on the updated enrollment projection. OS interim analysis plan has been updated due to the change in initial alpha allocation accordingly. SAP Version 5 has been approved after the futility interim analysis for PFS between Arm A versus Arm B in the ITT population and the futility interim analysis for PFS between Arm C versus Arm A in the ITT population), but before the final PFS analysis and before any unblinding action of the Sponsor. Both futility analyses were conducted by the independent DMC. Unblinded interim data were only accessible by the independent DMC and the independent Statistical Analysis Center. There is no intent to declare statistical significance for superior efficacy at either futility interim; therefore, the futility interims have no impact on the statistical significance levels for the final analysis.

1.1. Objectives, Endpoints, and Estimands

Objectives	Endpoints
Primary	
• To compare the PFS of imlunestrant (Arm A) to the standard comparator of Investigator's Choice Endocrine Therapy of either fulvestrant or exemestane (Arm B) in the ITT population • To compare the PFS of Arm A to Arm B in the *ESR1*-mutation detected population • To compare the PFS of imlunestrant plus abemaciclib (Arm C) to imlunestrant (Arm A) in the ITT population	• Investigator-assessed PFS (between Arm A and Arm B) in the ITT population • Investigator-assessed PFS (between Arm A and Arm B) in the *ESR1*-mutation detected population • Investigator-assessed PFS (between Arm C and Arm A) in the ITT population
Secondary	
• To compare OS of Arm A to Arm B in the ITT population • To compare OS of Arm A to Arm B in the ESR1-mutation detected population • To compare OS of Arm C to Arm A in the ITT population • To compare other efficacy objectives of Arm A to Arm B, and Arm C to Arm A	• OS between Arm A and Arm B in the ITT population (*key secondary endpoint*) • OS between Arm A and Arm B in the *ESR1*-mutation detected population (*key secondary endpoint*) • OS between Arm C and Arm A in the ITT population *(key secondary endpoint*) • Investigator-assessed ORR, DoR, and CBR • PFS by blinded Independent Review Committee (BIRC)
• To assess the safety and tolerability of each treatment arm	• Including but not limited to AEs, serious AEs, deaths, and clinical laboratory abnormalities per NCI CTCAE v5.0
• To evaluate the effectiveness of Arm A compared to Arm B and Arm C compared to Arm A based on PROs of pain using the Worst Pain NRS	• Time to sustained worsening of the "worst pain" as measured by Worst Pain NRS
• To assess the PK of imlunestrant (Arm A and Arm C) • To assess the PK of abemaciclib and its metabolites (Arm C)	• Plasma concentrations of imlunestrant and abemaciclib
Exploratory	
• To assess exploratory clinical parameters of Arm A compared to Arm B, and Arm C compared to Arm A	• Time to progressive bone metastases • Time to first SRE (defined as either pathological fracture, spinal cord compression, radiation to the bone, or surgery to the bone) • TTC • CFS • PFS2 (from randomization to disease progression on the next line of treatment or death) • Time to worsening of ECOG PS of ≥2
• To explore other PRO and HRQOL parameters of Arm A compared to Arm B, and Arm C compared to Arm A	• Time to worsening of physical function as measured by the physical function score of EORTC IL 19 • Change from baseline as measured by EORTC QLQ-C30 and EQ-5D-5L • Incidence of AE using the PRO-CTCAE item for diarrhea

Objectives	Endpoints
	• Change from baseline as measured by the PGIS-Cancer Symptoms
• To assess clinical efficacy parameters of Arm C compared to Arm B	• PFS • OS
• To explore potential biomarkers related to the ER pathway and/or the pathogenesis of breast cancer	• Biomarker results and efficacy outcomes

Abbreviations: AE = adverse event; CBR = clinical benefit rate; CFS = chemotherapy-free survival; CTCAE = Common Terminology Criteria for Adverse Events; DoR = duration of response; ECOG = Eastern Cooperative Oncology Group; EORTC IL 19 = European Organization for Research and Treatment of Cancer Item Library 19; EORTC QLQ-C30 = European Organization for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30; EQ-5D-5L = 5-level-EuroQol; ER = estrogen receptor; HRQOL = Health Related Quality of Life; ITT = Intention to treat; NCI = National Cancer Institute; NRS = numeric rating scale; ORR = objective response rate; OS = overall survival; PFS = progression-free survival; PGIS = Patient's Global Impression of Symptoms; PK = pharmacokinetics; PROs = patient-reported outcomes; PS = performance status; SRE = skeletal-related event; TTC = time to chemotherapy.

Primary Estimand

The primary research questions are: What is the difference in PFS time between (1) Arm A versus Arm B in the ITT population, (2) Arm A versus Arm B in the *ESR1*-mutation detected population, (3) Arm C versus Arm A in the ITT population, following progression/relapse on an AI, alone or in combination with a CDK4/6 inhibitor, in participants with advanced/metastatic HR+, HER2- breast cancer.

The estimand for the primary research questions is described by the following attributes:

- Population: adult participants with advanced/metastatic HR+, HER2- breast cancer after progression/relapse on prior treatment with an AI, alone or in combination with a CDK4/6 inhibitor, randomized to study interventions.
    - For PFS between Arm A and Arm B in the ITT population, the analysis population is based on approximately 640 participants concurrently randomized to both arms.
    - For PFS between Arm A and Arm B in the *ESR1*-mutation detected population, the analysis population is based on the *ESR1*-mutation detected subset.
    - For PFS between Arm C and Arm A in the ITT population, the analysis population is based on approximately 440 participants concurrently randomized to both arms.
- Endpoint: investigator-assessed PFS, which is defined as the time from randomization until
    - first occurrence of documented disease progression per RECIST 1.1, or
    - death from any cause in the absence of documented progressive disease
- Treatment condition: the randomized study interventions (Arms A, B and C) will be administered until disease progression, unacceptable toxicity, or another protocol-defined reason for study intervention discontinuation (Protocol Section 7). Further details on

study interventions including interventions, concomitant therapy and dose modification can be found in Protocol Section 6 Study Intervention.

- Intercurrent-event strategies (IES):
    - Post-study intervention discontinuation anticancer therapy started prior to disease progression is handled with the While on Treatment strategy (ie, consider the assessment of endpoint up until the time that post study intervention discontinuation anticancer therapy is taken).
    - Extended time without adequate assessment prior to disease recurrence/progression is handled with the While on Treatment strategy (ie, consider the assessment of endpoint up until the occurrence of extended time without adequate assessment).
- Population-level summary measure:
    - Hazard ratio of PFS in Arm A versus Arm B in the ITT population, in Arm A versus Arm B in the *ESR1*-mutation detected population or in Arm C versus Arm A in the ITT population estimated using a stratified Cox regression model (Cox 1972).
    - P-value of a stratified log-rank test of PFS comparing Arm A versus Arm B in the ITT population, comparing Arm A versus Arm B in the *ESR1*-mutation detected population or comparing Arm C versus Arm A in the ITT population.

Rationale for IES: The interest lies in the treatment effect without the confounding effect of other anticancer therapy or extended time without adequate assessment.

- A post study intervention discontinuation anticancer therapy taken prior to disease progression (or death) will confound the treatment effect in terms of PFS. If the anticancer therapy is taken, future disease progression is confounded by the effect of the new therapy. The participant will be censored and only the time prior to the post study intervention discontinuation anticancer therapy will be considered in analysis.
- Disease progression (or death) observed after an extended time without adequate tumor assessment may have occurred much earlier but is not reported because the scheduled assessment was not done. This inadequate observation may introduce bias to PFS estimates. If extended time without adequate assessment occurs, the participant will be censored and only the time up to the last adequate tumor assessment will be considered in analysis.

1.2. Study Design

In the protocol amendment (a), participants will be randomized 1:1:1 between 3 treatment arms (Arm A: Arm B: Arm C) and will be treated until disease progression or other discontinuation criteria are met (Protocol Section 7).

- Arm A: Imlunestrant 400 mg orally QD on Days 1 to 28 of a 28-day cycle
- Arm B: Investigator's Choice Endocrine Therapy
    - Exemestane 25 mg orally QD on Days 1 to 28 of a 28-day cycle **OR**

    - Fulvestrant 500 mg intramuscularly on Days 1 and 15 of Cycle 1, then on Day 1 of Cycle 2 and beyond
- Arm C: Imlunestrant + Abemaciclib
    - Imlunestrant 400 mg orally QD on Days 1 to 28 of a 28-day cycle
    - Abemaciclib 150 mg orally BID on Days 1 to 28 of a 28-day cycle

Arm C was added to the study (amendment a) after first patient visit for Arms A and B. All arms will be closed at the same time. Randomization of participants will continue in Arms A and B (1:1) until amendment (a) is approved and implemented, at which point participants will be randomized 1:1:1 (A:B:C) until the target enrollment for arms A and B (a total number of approximately 640 participants) is reached.

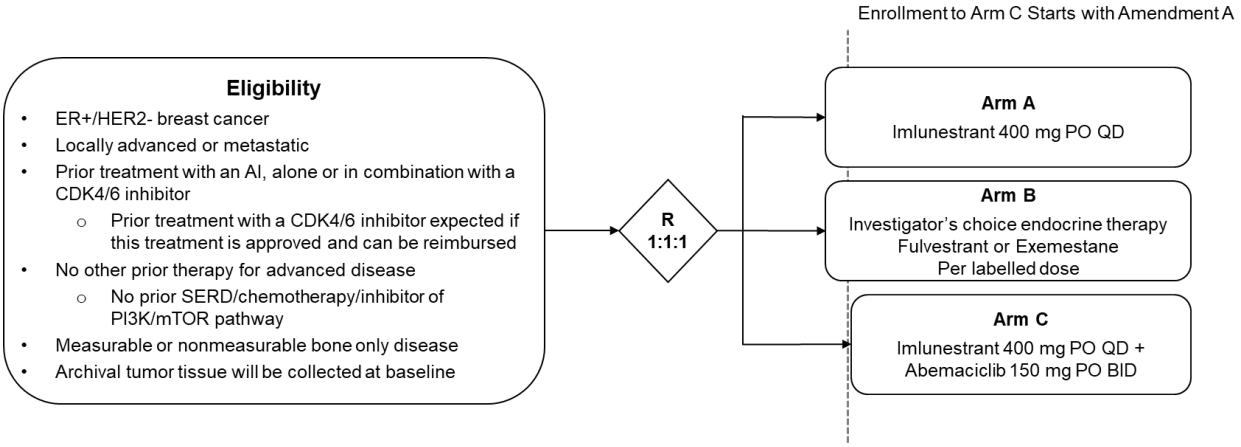
Investigator's Choice Endocrine Therapy (fulvestrant or exemestane) must be selected prior to randomization. Participants will be randomized using the following stratification factors:

- previous treatment with any CDK4/6 inhibitor (yes versus no)
- presence of visceral metastases (yes versus no); visceral includes lung, liver, brain, pleural, and peritoneal involvement, and
- region (East Asia versus North America/Western Europe versus Others).

The primary study objectives are to

- compare the PFS of Arm A to Arm B in the ITT population
- compare the PFS of Arm A to Arm B in the *ESR1*-mutation detected population, and
- compare the PFS of Arm C to Arm A in the ITT population.

The schema of the study is shown below.



Note: *ESR1*-mutation status will be centrally determined in plasma by Guardant 360 ctDNA assay from a blood draw at baseline.
Abbreviations: AI = aromatase inhibitor; CDK = cyclin-dependent kinase; ctDNA = Circulating tumor DNA; ER+ = estrogen receptor positive; FPV = first patient visit; HER2- = human epidermal growth factor receptor 2 negative; mTOR = mammalian target of rapamycin; n = number of participants; PI3K = phosphoinositide 3-kinase; PO = orally; QD = once daily; R = randomization; SERD = selective estrogen receptor degrader.

2. Statistical Hypotheses

The primary objective (1) is to demonstrate superior PFS of Arm A versus Arm B in the ITT population. Thus, letting $S_A(t)$ and $S_B(t)$ denote the PFS survival functions of Arm A and Arm B respectively, the hypotheses to be tested are as follows:

- Null hypothesis H0: $S_A(t) = S_B(t)$ against alternative hypothesis H1: $S_A(t) > S_B(t)$

The primary objective (2) is to demonstrate superior PFS of Arm A versus Arm B in the *ESR1*-mutation detected population. Thus, letting $S_{A,\ ESR1}(t)$ and $S_{B,\ ESR1}(t)$ denote the PFS survival functions of Arm A and Arm B respectively, the hypotheses to be tested are as follows:

- Null hypothesis H0: $S_{A,\ ESR1}(t) = S_{B,\ ESR1}(t)$ against alternative hypothesis H1: $S_{A,\ ESR1}(t) > S_{B,\ ESR1}(t)$

The primary objective (3) is to demonstrate superior PFS of Arm C versus Arm A in the ITT population. Thus, letting $S_C(t)$ and $S_A(t)$ denote the PFS functions of Arm C and Arm A respectively, the hypotheses to be tested are as follows:

- Null hypothesis H0: $S_C(t) = S_A(t)$ against alternative hypothesis H1: $S_C(t) > S_A(t)$

2.1. Multiplicity Adjustment

To adjust for multiplicity and control, the overall type I error rate at 0.025 (1-sided), a graphical approach (Bretz et al. 2009; Mauer and Bretz 2013) will be used to test the 3 PFS primary endpoints and the 3 key secondary overall survival (OS) endpoints. Besides, the Lan-DeMets stopping boundaries (O'Brien-Fleming type) will be used to derive the boundaries for the interim and final analyses for PFS and OS in order to control the Type I error rate. Further details are provided in Section 4.3.3.

3. Analysis Populations

For the purposes of analysis, the following analysis populations are defined:

Population	Description
ITT	All participants randomly assigned to study treatment, regardless of whether they take any doses of study treatment, or if they took the correct treatment. Participants will be analyzed according to the treatment group to which they were assigned. The full ITT population is all participants randomly assigned to study treatment, regardless of concurrent enrollment or not.
Safety	All participants randomly assigned to study treatment and who take at least 1 dose of study treatment. Participants will be analyzed according to the study treatment they actually received.
ORR Evaluable	The subset of participants from the ITT population who have measurable disease per RECIST v1.1 at baseline.
Per-protocol (PP)	All randomized participants (ITT population) who do not have important protocol deviations (IPD) that could potentially affect the efficacy conclusions of the study. See Section 6.2 for details.
Analysis Population for Arm A vs Arm B in the ITT population	All participants randomized to Arm A and Arm B (N = approximately 640)
Analysis Population for Arm A vs Arm B in the *ESR1*-mutation detected population	The subset of participants with *ESR1* mutation detected in the analysis population for Arm A versus Arm B in the ITT population
Analysis Population for Arm C vs Arm A in the ITT population	Participants concurrently randomized to Arm A and Arm C (N = approximately 440)

Abbreviations: ITT = intention-to-treat; N = total number of participants; ORR = objective response rate; OS = overall survival; PFS = progression-free survival; RECIST 1.1 = Response Evaluation Criteria in Solid Tumors, version 1.1.

The ITT population is used to analyze endpoints related to the efficacy objectives, and the safety population is used to analyze the endpoints and assessments related to safety.

A participant listing of analysis population details will be provided. This listing will be presented by treatment arm and will include investigator site, participant identifier, inclusion/exclusion flag for each population, and reason for exclusion from each population. All participants entering the trial will be included in this listing.

4. Statistical Analyses

4.1. General Considerations

Statistical analysis of this study will be the responsibility of Lilly or its designee.

Continuous variables will be summarized using descriptive statistics (that is, number of patients, mean, median, standard deviation, minimum, and maximum). Categorical variables will be summarized by frequency and its corresponding percentage.

PFS and OS endpoints will be tested according to the graphical approach. All other tests of treatment effects will be conducted at a 1-sided significance level of 0.025, unless otherwise stated, and all confidence intervals (CIs) will be given at a 2-sided 95% level.

The assumptions for each statistical method will be evaluated. If there is violation of assumptions, alternative statistical methods may be used.

Any change to the data analysis methods described in the protocol will require an amendment ONLY if it changes a principal feature of the protocol. Any other change to the data analysis methods described in the protocol, and the justification for making the change, will be described in Section 4.10 (if applicable) and the clinical study report (CSR). Additional exploratory analyses of the data will be conducted as deemed appropriate.

Statistical analysis will be performed using SAS software (SAS, version 9.4 or higher).

4.1.1. Definitions

Definitions of efficacy, safety, and patient-reported outcome (PRO) analysis variables are listed in respective sections of the SAP. Other variables are listed below alphabetically:

- Age (years): age at informed consent start date; birth month and day are imputed to be 01 July because only birth year is collected through electronic case report form (eCRF).
- Baseline Measurement: unless otherwise specified, the last non-missing measurement prior to the first dose of study drug.
- Duration: duration is calculated as
    - duration (days): (end date – start date + 1)
    - duration (weeks): (end date – start date + 1)/7
    - duration (months): (end date – start date + 1)/30.4375
      (days in months = (1/12) * average number of days in a year)
    - duration (years): (end date – start date + 1)/365.25
- Duration of disease: (randomization date - diagnosis of cancer date + 1)
- Study Day (safety analyses): study day is calculated as assessment date – first dose date + 1 day if the assessment is done on or after the first dose day. If the assessment is done prior to the first dose day, study day will be calculated as assessment date – first dose date. Date of first dose is defined as Study Day 1.

- Study Day (efficacy analyses): study day is calculated as assessment date – randomization date + 1 day if the assessment is done on or after randomization. If the assessment is done prior to randomization, study day will be calculated as assessment date – randomization date. Date of randomization is defined as Study Day 1, unless otherwise stated.
- Time-to-Event: the event or censoring time (days) is calculated as date of event/censoring – randomization date + 1.
- Analysis Visit: analysis visit (AVISIT) will be derived according to the protocol SOA for the purpose of by-visit analyses if deemed appropriate.

4.1.2. Handling of Dropouts or Missing Data

All analyses and descriptive summaries will be based on the observed data. Unless otherwise specified, missing data will not be imputed or "carried forward." Rules for handling dropouts or missing data are listed by type of analysis alphabetically.

- Adverse event (AE) or concomitant therapy:
    - The missing day of onset of an AE or start date of a concurrent therapy will be set to
        - first day of the month that the event occurred, if the onset yyyy-mm is after the yyyy-mm of first study treatment;
        - the day of the first study treatment, if the onset yyyy-mm is the same as yyyy-mm of the first study treatment; or
        - the date of informed consent, if the onset yyyy-mm is before the yyyy-mm of the first treatment.
    - The missing day of resolution of an AE or end date of a concurrent therapy will be set to
        - the last day of the month of the occurrence. If the patient died in the same month, then set the imputed date as the death date.
    - If the onset date of an AE or start date of a concurrent therapy is missing both the day and month, the onset date will be set to
        - 01 January of the year of onset, if the onset year is after the year of the first study treatment;
        - the date of the first treatment, if the onset year is the same as the year of the first study treatment; or
        - the date of informed consent, if the onset year is before the year of the first treatment.

  - If the resolution date of an AE or end date of a concurrent therapy is missing both the day and month, the date will be set to:
    - 31 December of the year of occurrence. If the patient died in the same year, then set the imputed date as the death date.
  - If the date is completely missing, then no imputation will be done and the event will be considered as treatment emergent with unknown onset date, unless the end date rules out the possibility.

- Diagnosis date, the following conventions will be used for imputing partial dates:
  - If only the day of the month is missing, the 15th of the month will be used to replace the missing day.
  - If both the day and the month are missing, "Jul 1" will be used to replace the missing information.

- General rule for imputing other dates (excluding the dates used in the efficacy analyses):
  - If only the day is missing, then assign Day 15 of the month, or the date of death if the patient died prior to 15th of the same month to the day.
  - If month is missing, then the date will be set to July 1 of the year, or the date of death if the patient died prior to July 1 of the same year.

  However, in all cases, after imputation, check if the imputed date is logically consistent with other relevant date variable(s) and make appropriate correction if necessary.

- Time-to-event analysis: all censored data will be accounted for using appropriate statistical methods. This information will be copied directly from the protocol. Additional general considerations may be added.

4.2. Participant Dispositions

A detailed description of patient disposition will be provided, including a summary of the number and percentage of participants enrolled (i.e., randomized) in the study, and treated as well as number and percentage of participants completing the study, defined as the participants who are evaluable for the primary endpoint, or discontinuing the study (overall and by reason for discontinuation). Reasons for the screen failures will also be summarized.

4.3. Primary Endpoint Analysis

4.3.1. Definition of Endpoint

The primary endpoint is investigator-assessed PFS. PFS is defined as the time from randomization to the date of first documented progression of disease or death from any cause in the absence of disease progression using Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1 criteria (Eisenhauer et al. 2009). Participants known to be alive and without disease progression will be censored according to the censoring scheme detailed in Section 4.3.2.

4.3.2. Main Analytical Approach

The Kaplan-Meier method (Kaplan and Meier 1958) will be used to estimate the PFS curves. Median PFS and PFS rates at various time points with 95% CIs will be estimated for each arm. The comparison of PFS curves between treatment arms will be conducted by a stratified log-rank test as the primary analysis, stratified by the randomization strata. The treatment effect will be estimated by hazard ratio with its corresponding 95% CIs using the stratified Cox proportional hazard model (Cox 1972) with treatment as the only covariate, stratified by the randomization strata. A detailed PFS event/censoring scheme is provided in the table below.

PFS Censoring Scheme

Situation	Event/Censor	Date of Event or Censor
Tumor progression or death	Event	Earliest date of PD or death
No tumor progression and no death	Censored	Date of last adequate tumor assessment, per RECIST 1.1 criteria, or date of randomization (whichever is later)
	Unless	
No baseline radiologic tumor assessment available	Censored	Date of randomization
No adequate postbaseline tumor assessment available and death reported after 2 scan intervals following randomization	Censored	Date of randomization
New systemic anticancer therapy prior to tumor progression or death	Censored	Date of last adequate tumor assessment, per RECIST 1.1 criteria, prior to start of new therapy or date of randomization (whichever is later)
Tumor progression or death documented immediately after 2 or more missing scan intervals following last adequate tumor assessment or randomization (whichever is later)	Censored	Date of last adequate tumor assessment prior to 2 or more missing scans, per RECIST 1.1 criteria, or date of randomization (whichever is later)

Abbreviations: PD = progressive disease; RECIST 1.1 = Response Evaluation Criteria in Solid Tumors, version 1.1.

[a] Symptomatic deterioration (that is, symptomatic progression that is not radiologically confirmed per RECIST 1.1 criteria) will not be considered as tumor progression.

[b] Adequate tumor assessment per RECIST 1.1 criteria refers to an assessment with 1 of the following responses: CR, PR, SD, or PD.

[c] The 2-scan interval is counted from the date of last adequate tumor assessment to the date of next 2 scheduled tumor assessments plus 8 days (adjusted by tumor assessment window).

[d] If there are multiple dates associated with 1 assessment, the assessment date will be set to the first date when the overall response is PD and the last date otherwise.

For the primary objective of PFS for Arm A versus Arm B in the ITT population, one interim PFS analysis will be conducted when approximately 192 (40% information fraction) investigator-assessed events have been observed in Arm A and Arm B, and the final PFS analysis will be conducted when approximately 480 investigator-assessed events have been observed in the same population. The interim analysis will allow the trial to stop early due to futility.

For the primary objective of PFS for Arm A versus Arm B in the *ESR1*-mutation detected population, the final analysis will be conducted when approximately 192 investigator-assessed events have been observed in the *ESR1*-mutation detected subset. The comparison of PFS curves

between treatment arms will be conducted by a stratified log-rank test as the primary analysis, stratified by the randomization strata, excluding region. Region is excluded in the stratified analysis for the subset to reduce the number of strata from 12 to 4. The treatment effect will be estimated by hazard ratio with its corresponding 95% CIs using the stratified Cox proportional hazard model with treatment as the only covariate. Similarly, region will be excluded in the stratified analysis.

For the primary objective of PFS for Arm C versus Arm A in the ITT population, 1 interim PFS analysis will be conducted when approximately 100 events have been observed among the approximately 440 participants concurrently randomized to Arm A and Arm C. This interim analysis will allow the trial (Arm A versus Arm C comparison) to stop early due to futility. The final analysis will be conducted when approximately 248 events have been observed among the participants concurrently randomized to both arms. The PFS between Arm C and Arm A will only be tested hierarchically based on the graphical approach.

See further details about the interim analyses in Section 4.9.3.

4.3.3. Graphical Approach

To adjust for multiplicity and control the overall family-wise type I error rate at 0.025 (1-sided), the graphical approach (Bretz et al. 2009; Mauer and Bretz 2013) will be used to test the 3 PFS primary hypotheses and the 3 OS key secondary hypotheses. The corresponding hypotheses are as follows.

Primary hypotheses:

H_1: PFS between Arm A and Arm B in the ITT population

H_2: PFS between Arm A and Arm B in the *ESR1*-mutation detected population

H_3: PFS between Arm C and Arm A in the ITT population

Key secondary hypotheses:

H_4: OS between Arm A and Arm B in the ITT population

H_5: OS between Arm A and Arm B in the *ESR1*-mutation detected population

H_6: OS between Arm C and Arm A in the ITT population

Initially, the overall 1-sided significance level of $\alpha = 0.025$ will be split between H_1 and H_2, with H_1 tested at the 1-sided significance level of $\alpha = 0.005$, and H_2 tested at the 1-sided significance level of $\alpha = 0.02$. No significance level ($\alpha = 0$) is initially assigned to H_3, H_4, H_5 and H_6. Figure 4.1 represents the graph with initially allocated significance levels at each node, and the associated weights for each directed edge from these respective nodes.

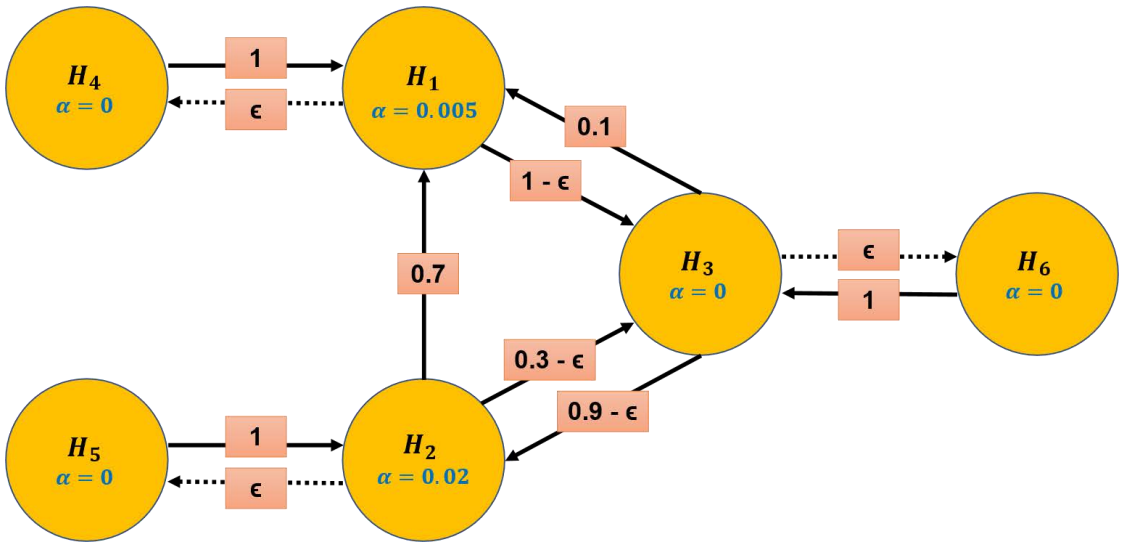


Figure 4.1. Initial graphical representation of testing scheme. Small edge weights are represented by dotted lines. ϵ denotes an infinitesimally small number ($\epsilon = 10^{-4}$).

The initial α-split between H_1 and H_2, and the corresponding edge weights for the various directed edges of the initial graph in Figure 4.1 were chosen based on an optimization algorithm. Various operating trial characteristics and metrics for the primary and key secondary objectives (i.e., marginal power, conditional power, etc.) were evaluated using extensive simulation studies. These metrics were calculated for a grid of varying assumptions: possible values of edge weights in the graph ranging from 0 to 1, independent/dependent correlation structures between the test statistics for the different hypotheses, prevalence of *ESR1*-mutation in ITT population, etc. Optimum values of edge weights for the graph in Figure 4.1 were selected based on the assessed trial metrics mentioned above from all possible combinations of the different parameters. Dotted lines in Figure 4.1 from H_1 to H_4, H_2 to H_5, and H_3 to H_6 denoted by ϵ represent edges with infinitesimally small weights. ϵ is chosen to be 10^{-4} in this setup.

The testing procedure in the graphical approach is carried out by testing each hypothesis at its local significance level. If a hypothesis can be rejected at a specific stage, i.e., the corresponding primary or key secondary objective is positive and demonstrates superiority, its significance level is reallocated to 1 of the other non-rejected hypotheses and the edge weights of the graph are updated based on a pre-specified algorithm (Bretz et al. 2009). Alpha levels will be recalculated if deemed appropriate. Note that for vertices with edge weight ϵ, no significance level is essentially passed when the corresponding hypothesis is rejected. These infinitesimal small ϵ-edge weights are updated into non-infinitesimal positive values only if no other outgoing edges except ϵ-edges remain. Description of this update step for reallocation of the significance levels along with details on ϵ-calculus can be found in Bretz et al. (2009) and Bretz et al. (2011). The testing step for each of the remaining non-rejected hypotheses can be carried out with the updated local significance levels obtained from the previous update step. This can lead to additional hypotheses being rejected in 2 possible ways: (i) a hypothesis which was not tested before because no local significance level was available previously before the update step can now be tested, or (ii) a hypothesis that we failed to reject earlier can now be tested at a higher local significance level. Local significance levels are further reallocated after the testing step, and this procedure is repeated until no more hypotheses can be rejected.

Note that the proposed graphical testing scheme ensures that H_3 will only be tested if at least 1 of H_1 or H_2 is rejected, and the key secondary OS hypotheses (H_4, H_5 and H_6) will be tested only if the corresponding primary PFS hypotheses are rejected. Figure 4.2a, Figure 4.2b, and Figure 4.2c represent the updated graphs from Figure 4.1 after H_1 is rejected, H_2 is rejected, or if both H_1 and H_2 are rejected at their original testing levels. Note that the local significance levels for the non-rejected hypotheses and the edge weights are updated in each case.

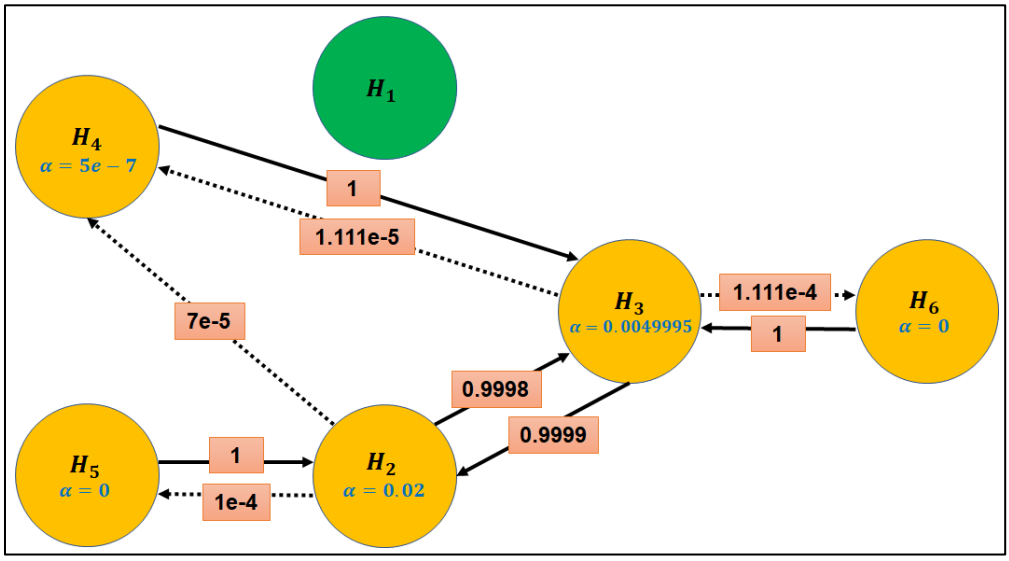


Figure 4.2a. Graphical representation of testing scheme when H1 is rejected for illustration. Small edge weights are represented by dotted lines. Edge weights are rounded off to 4 decimal places.

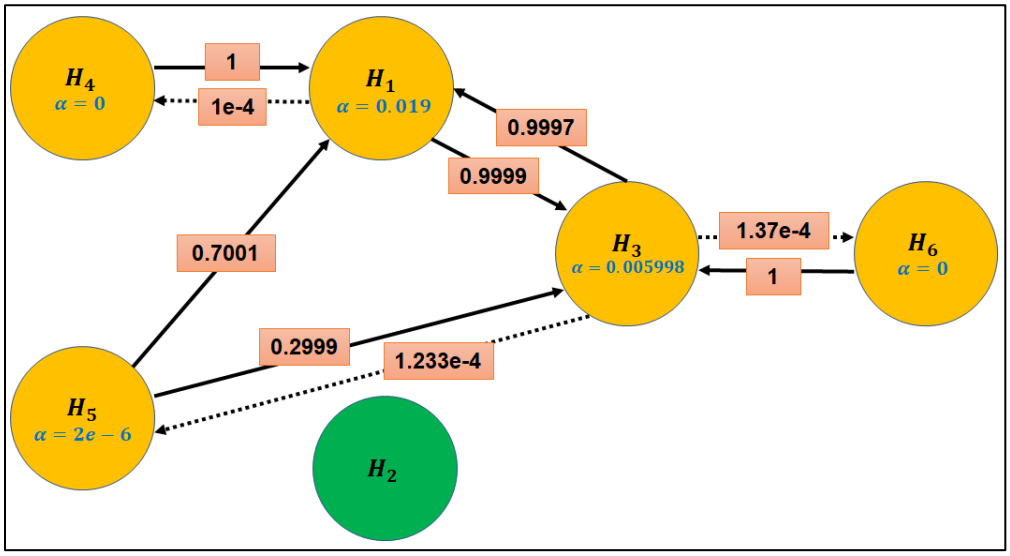


Figure 4.2b. Graphical representation of testing scheme when H2 is rejected for illustration. Small edge weights are represented by dotted lines. Edge weights are rounded off to 4 decimal places.

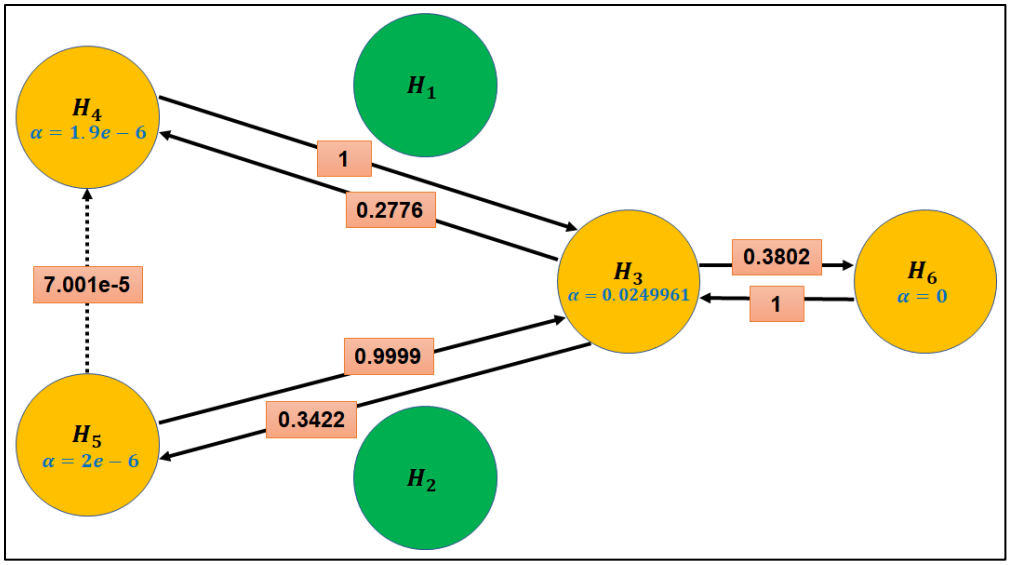


Figure 4.2c. Graphical representation of testing scheme when both H1 and H2 are rejected for illustration. Small edge weights are represented by dotted lines. Edge weights are rounded off to 4 decimal places.

Taking H_1 as an example, H_1 is initially tested based on the allocated 1-sided local significance level of $\alpha = 0.005$ from the initial α-split. If H_1 cannot be rejected at this significance level, it can be re-tested based on a higher significance level of $\alpha = 0.019$ if H_2 is rejected based on the initially allocated 1-sided local significance of $\alpha = 0.02$ (Figure 4.2b). If H_1 cannot still be rejected, it can be re-tested again at an even higher 1-sided local significance level of $\alpha = 0.02499645$ if both H_2 and H_3 are rejected (Figure 4.4). Similarly, if H_2 cannot be rejected based on the allocated initial 1-sided local significance level of $\alpha = 0.02$, it can be re-tested again at a higher 1-sided local significance level of $\alpha = 0.0249989$ if both H_1 and H_3 are rejected (Figure 4.3). As H_1, H_2 and H_3 are all event-driven, their final analyses may occur at different times (e.g., the hypotheses can be tested in the following order: first H_1, then H_2 and finally H_3). If re-testing of any hypotheses needs to occur at a later timepoint, the p-value associated with the original test statistic (observed at the prespecified final analysis for the corresponding hypothesis) will be compared against the updated local significance level. The analysis based on longer follow-up (after the final analysis) will not serve as basis for inferential testing.

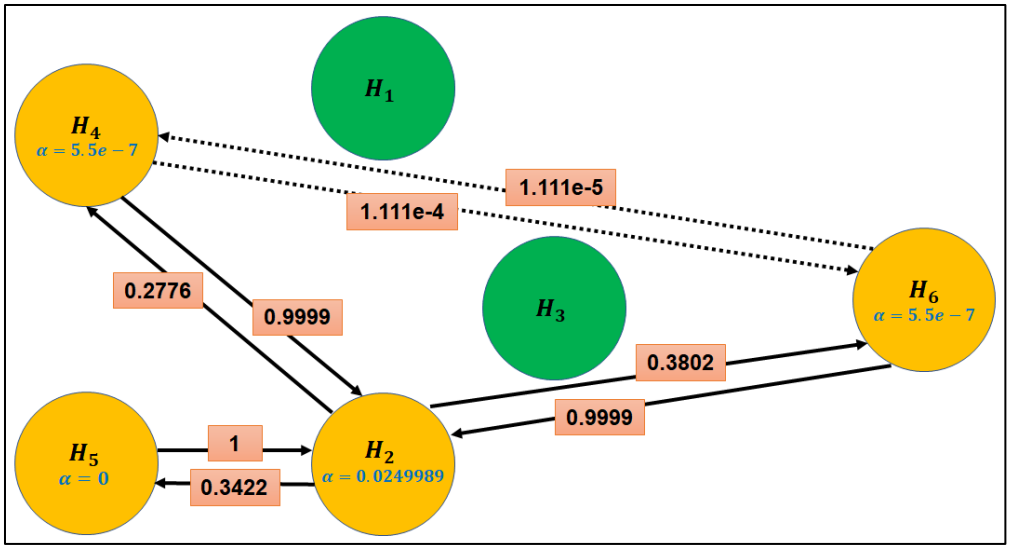


Figure 4.3. **Graphical representation of testing scheme when both H_1 and H_3 are rejected for illustration. Small edge weights are represented by dotted lines. Edge weights are rounded off to 4 decimal places.**

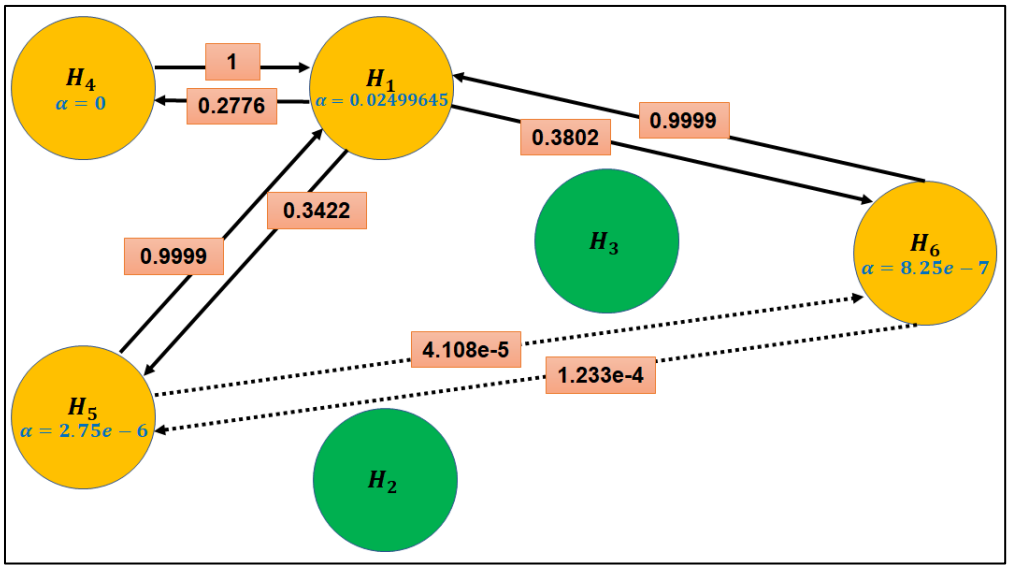


Figure 4.4. **Graphical representation of testing scheme when both H_2 and H_3 are rejected for illustration. Small edge weights are represented by dotted lines. Edge weights are rounded off to 4 decimal places.**

4.3.4. Sensitivity Analyses

Multiple sensitivity analyses for the primary PFS analysis will be conducted as defined below:

- Using different rules for censoring (details provided in the table below)
- Using an unstratified log-rank test and unstratified Cox model
- Using stratification factors based on the case report form (CRF) data if available

- Using a multivariate Cox regression model constructed by selecting variables among all the potential variables such as the variables used in the subgroup analyses, using stepwise selection method, with an entry p-value of 0.05 and an exit p-value of 0.1. The treatment factor will be kept out of the model throughout the covariate selection process and only added to the final model.
- Including randomization scheme as another stratification factor in the stratified log-rank test and stratified Cox model for comparing Arm A versus Arm B, as randomization scheme (2-arm 1:1 randomization versus 3-arm 1:1:1 randomization) has been used to randomize patients in the randomization stage. That is to say, the randomization scheme for "Arm A versus Arm B" is A:B (1:1) & A:B:C (1:1:1).
- For PFS comparing Arm A versus Arm B in the *ESR1*-mutation detected population, in addition to the preplanned analyses above, PFS analysis in the *ESR1*-mutation not detected population will also be conducted.
- For PFS comparing Arm C versus Arm A in the ITT population, using the full ITT population (i.e., all participants randomized to both arms will be used).

PFS Censoring Scheme for Sensitivity Analyses

Definition	Situation	Event/ Censor	Date of Event or Censor
SA1: Ignoring new anticancer therapy prior to tumor progression or death	New anticancer therapy started before treatment discontinuation and 1. Tumor progression or death after the start date of the new therapy 2. No tumor progression and no death	1. Event 2. Censored	1. Earliest date of PD or death 2. Date of last adequate tumor assessment, per RECIST 1.1 criteria, or date of randomization (whichever is later)
SA2: Ignoring all censoring rules defined in the PFS Censoring Scheme table	1. No baseline radiologic tumor assessment available 2. Else: a. Tumor progression or death; b. No tumor progression and no death	1. Censored 2. a. Event; b. Censored	1. Date of randomization 2. a. Earliest date of PD or death; b. Date of last adequate tumor assessment, per RECIST 1.1 criteria, or date of randomization (whichever is later)
SA3: Ignoring absence of adequate postbaseline tumor assessment	No adequate postbaseline tumor assessment available and death reported after 2 scan intervals following randomization	Event	Death
SA4: Ignoring missing tumor assessments	PD or death documented after 2 or more missing scan intervals following last adequate tumor assessment or randomization (whichever is later)	Event	Earliest date of PD or death

Abbreviations: PD = progressive disease; PFS = progression-free survival; SA = sensitivity analysis.

Other sensitivity analyses for PFS may be conducted if deemed appropriate. The PFS analyses may also be conducted in the PP population if deemed appropriate.

4.4. Secondary Endpoints Analysis

4.4.1. Key Secondary Endpoint

4.4.1.1. Definition of Endpoint(s)

Overall survival is defined as the time from randomization until death from any cause. If the participant is alive or lost to follow-up at the time of analysis, OS data will be censored on the last date the participant is known to be alive.

4.4.1.2. Main Analytical Approach

OS curves, median OS, and OS rates at 1 year, 2 years, and 3 years with 95% CI for each treatment arm will be estimated using the Kaplan-Meier method. OS will be compared between treatment arms using a log-rank test stratified by the same factors as PFS, as the primary analysis for OS. The corresponding hazard ratio between treatment arms will be estimated using the stratified Cox regression model. The inferential analysis of OS will be based on the stratified analyses.

The analysis population for each OS comparison is the same as PFS. The OS endpoints will be hierarchically tested according to the graphical approach described in Section 4.3.3.

In the following subsection, we will present the boundary tables for H_4, H_5 and H_6 for OS interim analyses assuming that H_1, H_2 and H_3 are all rejected, intended to be illustrative only. The graph and the allocation of α for different OS endpoints will depend on the actual trial outcomes observed. The graph shown in Figure 4.5 displays the alpha level assigned to each endpoint under this scenario.

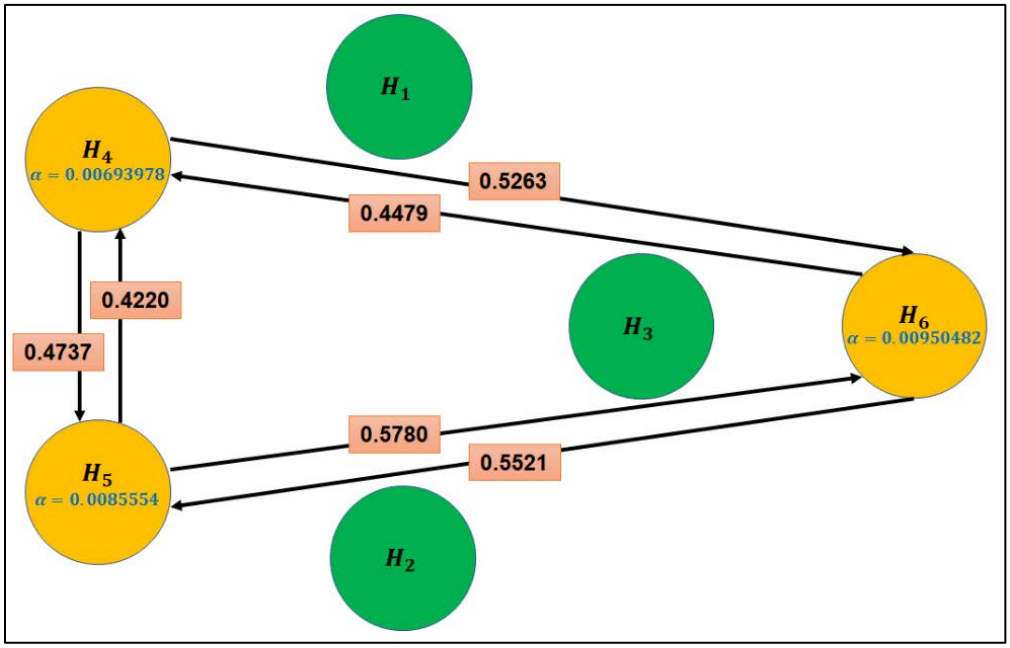


Figure 4.5. Graphical representation of testing scheme after H1, H2 and H3 are rejected.

4.4.1.3. Interim Analyses for OS Endpoints

For OS between Arm A and Arm B in the ITT population, the following analyses are planned:

OS look 1	At the time of final analysis of PFS between Arm A vs Arm B in the ITT population
OS look 2	Approximately 255 OS events
OS look 3	Approximately 330 OS events
Final OS	Approximately 390 OS events (estimated to be 3 years after the final analysis of PFS)

Abbreviations: ITT = intention-to-treat; OS = overall survival; PFS = progression-free survival.

The first OS analysis will be conducted at the time of final analysis of PFS between Arm A vs Arm B in the ITT population if PFS is significant. . According to the graphical approach, if PFS (H_1) is not significant after the final analysis, OS (H_4) will not be statistically evaluated. The Lan-DeMets spending function (O'Brien-Fleming type) will be used to determine the boundaries for the interim and final analyses of OS, specifically

$$\alpha^*(t_k) = 2\left(1 - \Phi\left(\Phi^{-1}(1-\alpha/2)/\sqrt{t_k}\right)\right),$$

where t_k is the information fraction at time k, Φ is the standard normal cumulative distribution function, and Φ^{-1} is the standard normal quantile function. The p-value boundary at each analysis is summarized in Table 4.1 assuming OS for Arm A versus Arm B in the ITT population will be tested at the alpha level of 0.00693978, which is the alpha level assuming all PFS hypotheses H_1, H_2 and H_3 are rejected. The actual boundaries will be updated based on observed number of events and the actual alpha level that is passed to H_4 from previous tests, using software (e.g., EAST version 6.5).

Table 4.1. Stopping Boundaries for Each OS Analysis Between Arm A and Arm B in the ITT Population for Illustration

Analysis	Critical P-value Boundary	Critical HR Boundary	Cumulative Type I Error Rate
OS look 1	1.4E-05	0.504	1.4E-05
OS look 2	0.0008	0.675	0.0008
OS look 3	0.0031	0.740	0.0033
Final OS	0.0059	0.775	0.0069

Abbreviations: ITT = intention-to-treat; OS = overall survival.

For OS between Arm A and Arm B in the *ESR1*-mutation detected population (H_5), the following analyses are planned assuming that the final analysis for Arm A versus Arm B in the ITT population and the final analysis for Arm A versus Arm B in the *ESR1*-mutation detected population occur at the same time.

OS look 1	At the time of final analysis of PFS between Arm A vs Arm B in the ITT population
OS look 2	At the time of OS look 2 in ITT
OS look 3	At the time of OS look 3 in ITT
Final OS	Approximately 155 OS events (estimated to be 3 years after the final analysis of PFS)

Abbreviations: ITT = intention-to-treat; OS = overall survival; PFS = progression-free survival.

Table 4.2 summarizes the boundaries based on the alpha level of 0.0085554 assuming all PFS hypotheses H_1, H_2 and H_3 are rejected. The actual boundaries will be updated based on the actual alpha level assigned to H_5. The Lan-DeMets spending function (O'Brien-Fleming type) will be used to determine the boundaries for the interim and final analyses of OS between Arm A and Arm B in the *ESR1*-mutation Detected Population. The actual boundaries will be updated based on observed number of events and the actual alpha level assigned to H_5.

Table 4.2. Stopping Boundaries for Each OS Analysis Between Arm A and Arm B in the ESR1-mutation Detected Population for Illustration

Analysis	Number of Events[a]	Critical P-value Boundary	Critical HR Boundary	Cumulative Type I Error Rate
OS look 1	65	4.9E-05	0.380	4.9E-05
OS look 2	105	0.0014	0. 557	0.0014
OS look 3	135	0.0044	0.637	0.0048
Final OS	155	0.0070	0.674	0.0086

Abbreviation: OS = overall survival.

[a] Number of Events for OS look 1 to OS look 3 were estimated based on the OS assumptions.

Similarly, multiple looks are planned for the OS endpoint between Arm C and Arm A in the ITT population (H_6). The Lan-DeMets spending function (O'Brien-Fleming type) will be used to determine the boundaries for the interim and final analyses of OS between Arm C and Arm A. Table 4.3 summarizes the boundaries based on the alpha level of 0.00950482 assuming all PFS hypotheses H_1, H_2 and H_3 are rejected. The actual boundaries will be updated based on observed number of events and the actual alpha level assigned to H_6.

Table 4.3. Stopping Boundaries for Each OS Analysis Between Arm C and Arm A for Illustration

Analysis	Number of Events[a]	Critical P-value Boundary	Critical HR Boundary	Cumulative Type I Error Rate
OS look 1 (at final PFS for Arm C vs Arm A in the ITT population)	65	2E-07	0.285	2E-07
OS look 2 (at OS look 2 for Arm A vs Arm B in the ITT population)	135	0.0003	0.555	0.0003
OS look 3 (at OS look 3 for Arm A vs Arm B in the ITT population)	185	0.0020	0.655	0.0021
OS look 4 (at final OS for Arm A vs Arm B in the ITT population)	225	0.0046	0.707	0.0053
Final OS	260	0.0078	0.741	0.0095

Abbreviations: ITT = intent to treat; OS = overall survival.

[a] Number of Events for OS look 1 to OS look 4 were estimated based on the OS assumptions.

The final OS analysis will be conducted when approximately 260 OS events have been observed in the analysis population for Arm C versus Arm A, which is estimated to be 4 years after the final PFS analysis of Arm C versus Arm A.

Overall survival analyses may be collapsed if they are expected to occur within a similar timeframe (e.g., within approximately 2 months).

4.4.1.4. Sensitivity Analyses

Multiple sensitivity analyses for OS will be conducted as defined below:

- Using an unstratified log-rank test and an unstratified Cox model
- Using stratification factors based on the CRF data if available
- Using a multivariate Cox regression model
- Including randomization scheme as another stratification factor in the stratified log-rank test and stratified Cox model for comparing Arm A versus Arm B (2-arm 1:1 randomization vs 3-arm 1:1:1 randomization).
- For OS comparing Arm A versus Arm B in the *ESR1*-mutation detected population, in addition to the preplanned analyses above, OS analysis in the *ESR1*-mutation not detected population will also be conducted.
- For OS comparing Arm C versus Arm A in the ITT population, using the full ITT population (i.e., all participants randomized to both arms will be used).

Other sensitivity analyses for OS may be conducted if deemed appropriate.

4.4.2. Supportive Secondary Endpoints

All supportive secondary endpoint analyses will be conducted for each comparison (Arm A versus Arm B, and Arm C versus Arm A) per the prespecified analysis population for each comparison.

Objective response rate (ORR) is defined as the proportion of participants who achieve a confirmed best overall response of CR or PR. The ORR with 95% CI will be summarized for each treatment arm and compared between treatment arms using the Cochran-Mantel-Haenszel test adjusting for the randomization strata. The analysis of ORR will be conducted in the ORR evaluable population. The Mantel-Haenszel estimate of the common risk difference of ORR with 95% CI between 2 arms will also be estimated.

Clinical benefit rate (CBR) is defined as the number of participants who achieve a best overall response of CR, PR, or SD ≥ 24 weeks divided by the total number of participants randomized to the corresponding treatment arm. The CBR with 95% CI will be summarized for each treatment arm and compared between treatment arms using the Cochran-Mantel-Haenszel test adjusting for the randomization strata. The Mantel-Haenszel estimate of the common risk difference of CBR with 95% CI between 2 arms will also be estimated.

Duration of response (DoR) is defined as the time from the date measurement criteria for CR or PR (whichever is first recorded) are first met until the first date that disease is recurrent or objective progression is observed, per RECIST 1.1 criteria, or the date of death from any cause

in the absence of objectively determined disease progression or recurrence. The DoR will be censored according to the same scheme as the main scheme for PFS. Median DoR with 95% CI and curves for each treatment arm will be estimated using the Kaplan-Meier method. The analysis of DoR will be based on the participants who achieve an objective response (CR or PR).

Progression-free survival by blinded independent review committee (BIRC) is defined the same way as the primary endpoint of PFS. For BIRC analysis, scans will be collected and reviewed in all randomized participants based on RECIST version 1.1. PFS as assessed by BIRC intends to evaluate the reliability of the treatment effect based on the investigator-assessed PFS. PFS as assessed by BIRC will be analyzed using the same methods as the investigator-assessed PFS. PFS as assessed by BIRC is not intended to provide an alternative means of definitive analysis, but it may be useful to evaluate bias in local assessments. Discordance rates (i.e., differences in assessment of progression between investigator and BIRC) will be summarized for each arm (Amit, et al. 2011). Specifically, differential discordance will be described using early discrepancy rate and late discrepancy rate differences.

BIRC-Assessed Versus Investigator-Assessed Disease Progression

	BIRC	
Investigator	PD	No PD
Investigator PD	a=a1+a2+a3	b
No PD	c	d

Abbreviations: PD = progressive disease; BIRC = blinded independent review committee.

a1: number of agreements on timing and occurrence of PD.

a2: number of times investigators declare PD later than BIRC.

a3: number of times investigators declare PD earlier than BIRC.

The early discrepancy rate (EDR) quantifies the frequency with which the investigator assessment declares progression early relative to BIRC within each arm and is defined as:

$$EDR = (b+a3)/(a+b)$$

The late discrepancy rate (LDR) quantifies the frequency with which the investigator assessment declares progression later than BIRC within each arm and is defined as:

$$LDR = (c+a2)/(b+c+a2+a3)$$

The EDR and LDR will be summarized for each treatment arm and the differential discordance around each measure can be defined as the rate on the experimental arm minus the rate on the control arm. A negative differential discordance for the EDR and/or positive differential discordance for the LDR are suggestive of a bias in the investigator-assessed PFS favoring the experimental arm.

The ORR, CBR, and DoR by BIRC will also be summarized for each arm.

Adverse Events

The Medical Dictionary for Regulatory Activities (MedDRA) Preferred Term (PT) derived from the verbatim term will be used when reporting AEs by MedDRA terms. The MedDRA Lower Level Term will be used in the treatment-emergent computation. Severity grades will be assigned by the investigator using National Cancer Institute - Common Terminology Criteria for Adverse Events (NCI-CTCAE) version 5.0.

Preexisting conditions are defined as AEs that either are ongoing at informed consent and or end on or after informed consent. Pre-existing conditions will be included in the listing of AE so that the history of AEs can be traced.

Treatment-emergent adverse events (TEAEs) are events that first occurred or worsened in severity after baseline. Treatment-emergent adverse events will be summarized by System Organ Class (SOC) and by decreasing frequency of PT within SOC.

Adverse event analyses will include summaries of the following:

- Overview of adverse events
- TEAEs, including severity (any grade and grade ≥3) and possible relationship to study drug
- Cumulative TEAE incidence at 3 months, 6 months, and 1 year, will be provided in accordance with the draft EMA guideline (EMA 2017) on evaluation of anticancer medical products in humans if deemed appropriate. In case where the time on therapy is longer, additional time points up to approximately 5 years may also be considered.
- Serious adverse events, including possible relationship to study drug
- Adverse events leading to dose adjustments/omissions
- Discontinuations from study treatment due to AEs or death
- Time to onset for selected TEAEs
- **Adverse events of special interest** (AESI): Categories of AESI may be modified as the understanding of the safety of imlunestrant increases. The final list of categories will be maintained at both compound and study level and reported in the CSR.
- **Consolidated AEs** are composite AE terms consisting of synonymous PTs to allow meaningful interpretation of the AE data. The final list of consolidated AE categories and PTs will be maintained at both compound and study level and reported in the CSR.

Deaths

- Deaths (all deaths and deaths within 30 days of treatment discontinuation) and their primary cause (study disease progression, AE, other)
- Adverse events leading to death

Laboratory Abnormalities

The severity of laboratory results will be classified according to NCI-CTCAE. The laboratory toxicity by worst NCI-CTCAE grade and shifts in toxicity grading from baseline to the worst post-baseline grade will be summarized. Abnormal laboratory parameters will be listed.

Shift to low/high tables will include the number and percentage of patients within each baseline category (baseline value is low, normal, high, or missing) versus each postbaseline category (worst value is low, normal, high, or missing) by treatment arm.

The analyses of adverse events, deaths, and laboratory abnormalities will be conducted in the safety population.

Time to sustained worsening of worst pain (as measured by the Worst Pain NRS) is defined as the time from randomization to the first increase (≥2 points) in the weekly average of the worst pain score with confirmation in the next consecutive week. See Section 4.6 for further details.

Pharmacokinetic/Pharmacodynamic Analyses

PK parameters for imlunestrant in plasma (for example, clearance, volume of distribution) and inter-individual PK variability will be computed using nonlinear mixed-effect modeling implemented in NONMEM. Covariate effects, such as age, weight, sex, and creatinine clearance, on the PK parameters of imlunestrant in plasma will also be investigated.

Biomarker data collected in this study may be used in a population PK/pharmacodynamic model.

4.5. Exploratory Endpoints Analysis

All the exploratory analyses will be conducted for each comparison (Arm A versus Arm B and Arm C versus Arm A) per the prespecified analysis population for each comparison if the number of events is sufficient, unless otherwise stated.

Time to progressive bone metastases is defined as the time from randomization to the date of earliest development of new bone metastases or unequivocal progression of current bone lesions. Participants not known to have progressive bone metastases will be censored at the date of last documented tumor assessment. Time to progressive bone metastases will be summarized for each treatment arm using the Kaplan-Meier method and will be compared between 2 arms using the log-rank test.

Time to first skeletal-related event (SRE) is defined as the time from randomization to the first SRE defined as either pathological fracture, spinal cord compression, radiation to the bone, or surgery to the bone. Participants not known to have an SRE will be censored at the date of last documented assessment. Time to first SRE will be summarized for each treatment arm using the Kaplan-Meier method and will be compared between 2 arms using the log-rank test.

Time to chemotherapy (TTC) is defined as the time from randomization to the initiation of first post-discontinuation chemotherapy. Participants who die prior to the initiation of chemotherapy will be censored at the date of death. TTC will be summarized for each treatment arm using the Kaplan-Meier method and will be compared between 2 arms using the log-rank test.

Chemotherapy-free survival (CFS) is defined as the time from randomization to the initiation of first post-discontinuation chemotherapy or death, whichever is earlier. Participants not known to have initiated chemotherapy will be censored at the last documented assessment. CFS will be summarized for each treatment arm using the Kaplan-Meier method and will be compared between 2 arms using the log-rank test.

Progression-free survival 2 is defined as the time from randomization to objective disease progression on the next line of treatment or death, whichever is earlier. Participants alive and not known to have a second objective progressive disease will be censored on the latest date known to be alive and without a second objective disease progression. PFS2 will be summarized for each treatment arm using the Kaplan-Meier method and will be compared between 2 arms using

the log-rank test. Sensitivity analysis for PFS2 will be performed if deemed appropriate. For example, instead of objective disease progression, we may conduct analysis including clinical disease progression.

Time to worsening of ECOG PS of ≥2 is defined as the time from randomization to the date when ECOG PS score of ≥2 was observed for the first time. Participants not known to have such worsening will be censored at the last time when no worsening of ≥2 is observed. Time to worsening of ECOG PS of ≥2 will be summarized for each treatment arm using the Kaplan-Meier method and will be compared between 2 arms using the log-rank test.

PFS and OS will be compared between Arm C and Arm B using the stratified log-rank test. The treatment effect will be estimated by hazard ratio with its corresponding 95% CIs using the stratified Cox proportional hazard model. The analysis population for Arm C versus Arm B is all concurrently randomized participants between 2 arms. Subgroup analyses (e.g. based on stratification factors) will be conducted if deemed appropriate. Other efficacy endpoints may also be compared between Arm C and Arm B if deemed appropriate.

4.6. Patient-Reported Outcomes Analyses

The patient reported outcomes (PROs) will be used to compare changes in cancer-related symptoms, physical function, adverse effect of diarrhea, and other health-related quality of life (HRQoL) outcomes between treatment arms and generate health utility data. The analyses for PROs will be conducted in the ITT population. The following PRO instruments are utilized:

- Worst pain NRS
- EORTC-QLQ-30
- EORTC IL19: Physical Function
- EQ-5D-5L
- PGIS (Patient's Global Impression of Severity)-Cancer Symptoms
- mBPI-SF
- PRO-CTCAE Items for Diarrhea and Injection Site Pain and Swelling

Time to sustained worsening of worst pain (based on the Worst Pain NRS) is defined as the time from randomization to the first increase (≥2 points) in the weekly average of the worst pain score with confirmation in the next consecutive week. Participants not known to have sustained worsening will be censored at the last documented assessment. Time to sustained worsening of worst pain will be summarized for each arm by the Kaplan-Meier method and will be compared between the arms using the stratified log-rank test. Additional cutoff values of increase (e.g., ≥3 points, ≥4 points) may be explored as sensitivity analyses if deemed appropriate. Sensitivity analysis will be done by including death as part of the event definition. Additional sensitivity analyses for time to worsening without confirmation in the next consecutive week may be performed if deemed appropriate. A standard analgesic categorization algorithm such as WHO analgesic ladder will be used to analyze the data, along with the worst pain, as an additional analysis. This will be detailed in the PRO SAP.

Time to worsening of physical function (based on either EORTC-QLQ-30 or EORTC IL19) is defined as the time from randomization to the first ≥10-point decrease from baseline with confirmation at the next cycle. Participants not known to have worsening will be censored at the last documented assessment. Sensitivity analysis will be done by including death as part of the

event definition. Time to worsening of physical function will be summarized for each arm by the Kaplan-Meier method, and the stratified log-rank test will be used to compare between the 2 arms.

For EORTC-QLQ-30, EORTC IL 19, EQ-5D-5L, mBPI-sf, PGIS, a summary of change from baseline will be provided. For each participant with data from baseline and at least 1 post-baseline visit, change from baseline at each time point, and the maximum change from baseline score will be calculated for each scale of each instrument if appropriate.

PRO-CTCAE items for diarrhea and injection site pain and swelling will be summarized by visit and overall for each treatment arm.

Compliance will be assessed for each instrument. The compliance rate will be calculated at baseline, per week (if appropriate) and per cycle. The number of participants with expected assessments at each post-baseline visit is the number of participants who have received the study drugs in a previous visit. For the Worst Pain NRS, participants will be considered compliant if they have completed ≥50% of the daily assessments during each period (e.g., week, cycle). A 7-day average score regardless of dosing will be calculated for each consecutive week starting from cycle 1 day 2, if the participant is compliant for the period (completed 4 days out of the 7-day period). A 7-day average score based on any available data may also be calculated as sensitivity analysis even if the compliance rate is lower than 50%. For other instruments, the number of missing and incomplete questionnaires and/or assessments by visit will be summarized for each instrument and treatment arm.

The cycle 1 day 1 (pre-dose) visit will be considered as baseline for all PRO analyses. The analysis for change from baseline will be based on the participants who have baseline and at least 1 post-baseline data.

Further details about PRO analyses will be described in a separate PRO SAP.

4.7. (Other) Safety Analyses

The other safety analyses will be conducted in the safety population.

4.7.1. Extent of Exposure

The number of cycles received, dose omissions, dose reductions, dose delays, and dose intensity will be summarized for all treated patients by treatment arm/study drug as deemed appropriate. The derivations for each study drug are provided below.

Imlunestrant/Exemestane/Abemaciclib

- duration of therapy (weeks) = (date of last dose - date of first dose + 1) /7
- cumulative dose (mg) = sum of all doses actually received
- dose intensity (mg/week) = (cumulative dose) ÷ (duration of therapy)
- planned dose intensity (mg/week) = assigned daily dose (mg) * 7
- relative dose intensity (%) = (dose intensity / planned dose intensity)*100

Fulvestrant

- duration of therapy (weeks) = (date of last cycle Day 1 - date of first dose + 28) /7, if the last dosing date is in cycle 1, then duration of therapy (weeks)= (date of last dose in cycle 1-date of first dose+14) / 7
    - if the patient died or was lost to follow-up within the date of last dose +length of interval (14 days or 28 days), then duration of therapy = (date of death or date of last contact – date of first dose)/7
- cumulative dose (mg) = sum of all doses
- dose intensity (mg/week) = (cumulative dose level) ÷ (duration of therapy)
- planned dose intensity (mg/week) = planned dose per infusion (mg) / 4
- relative dose intensity (%) = (dose intensity / planned dose intensity) * 100

Duration of therapy for patients with certain characteristics may be summarized by treatment arm.

4.7.2. Additional Safety Assessments

Electrocardiograms

Electrocardiogram (ECG) will be summarized by visit and by treatment arm. A summary of change from baseline (by visit) and the corresponding AEs will also be provided.

Vital Signs

All vital signs (e.g., temperature, blood pressure, pulse rate, height, weight, heart rate) will be summarized by visit and by treatment arm. Treatment emergent abnormal changes in vital signs will also be summarized by treatment arm.

4.8. Other Analyses

4.8.1. Participant Characteristics

Demographics and baseline disease characteristics will be summarized by treatment arm.

Disease characteristics will include the following:

- Initial pathological diagnosis
- Disease stage (Stage IIA, Stage IIB, etc.)
- Histopathological diagnosis grade (G1, G2, etc.)
- Baseline Eastern Cooperative Oncology Group (ECOG) performance status (PS)
- Site of disease (liver, lung, etc.)
- Number of sites involved (1, 2, or 3+)
- Measurable disease at baseline (yes versus no)
- Endocrine resistance (primary versus secondary)

- Progesterone receptor status (positive versus negative)
- Nature of disease (visceral metastases, bone only metastases, or other)
- HER2 status (low versus negative)
- Prior CDK4/6 inhibitor (yes versus no)
    - If yes, setting (adjuvant versus metastatic)
    - If yes and received prior CDK4/6 inhibitor in the metastatic setting, length of the prior CDK4/6 inhibitor (≥ 12 months versus <12 months)
- Type of prior CDK4/6 inhibitor (e.g., palbociclib, ribociclib, abemaciclib, dalpiciclib)
- Prior therapy history
- ESR1 mutation status

Nature of disease, number of sites involved will be derived from the location codes of the targe and non-target lesions. All patients with at least 1 lesion on the baseline target lesion form will be considered as having measurable disease at baseline. Primary endocrine resistance is defined as relapse during the first 2 years of adjuvant endocrine therapy (ET), or progressive disease within the first 6 months of first-line ET for advanced/metastatic breast cancer. Secondary endocrine resistance is defined as relapse while on adjuvant ET but after the first 2 years or relapse within 12 months of completing adjuvant ET, or progressive disease 6 months after initiating ET for advanced/metastatic breast cancer.

4.8.2. Historical Illnesses/Preexisting Conditions

Historical illnesses and preexisting conditions (using MedDRA Preferred Terms) will be summarized by treatment arm.

4.8.3. Prior Therapy

Prior radiotherapy, surgery, and systemic therapy will be summarized by treatment arm. Prior radiotherapy and surgery will be categorized by reason for regimen. Prior systemic therapies will be categorized by type of regimen (endocrine therapy, chemotherapy, etc.) and reason for regimen (neoadjuvant, adjuvant, locally advanced, or metastatic). Frequency of each specific therapy will be tabulated within each type of therapy and per reason for regimen.

4.8.4. Concomitant Therapy

A summary of preferred names of concomitant medications by treatment arm by decreasing frequency will be reported.

4.8.5. Post-Study Treatment Therapy

The numbers and percentages of participants receiving poststudy anticancer therapies will be provided by type of therapy (surgery, radiotherapy, or systemic therapy), and by drug class and/or name, overall and by line of therapy.

4.8.6. Treatment Compliance

Treatment compliance for imlunestrant, exemestane, or abemaciclib will be assessed through pill counts at each cycle. Compliance will be calculated as the ratio of total dose taken to the total assigned dose (minus any dose adjustments and doses omitted/withheld for medical or logistical reasons). A patient will be considered noncompliant if he or she takes <80% or >125% of the planned doses. Compliance information for exemestane will be collected via number of doses at each cycle.

Compliance for fulvestrant is assured as the drug will be administered at the investigator site.

4.8.7. Duration of Stable Disease

Duration of stable disease is defined as the time from randomization to the date of objective progression of disease or death from any cause in the absence of disease progression. Participants will be censored using the same scheme as the main scheme for PFS.

4.8.8. Change in Tumor Size

Percent change in tumor size, defined as (post-baseline sum of target lesion measurements-baseline sum of target lesion measurements)/baseline sum of measurements, will be summarized by cycle and by treatment arm. Mean percent change in tumor size will also be plotted by treatment arm. Repeated measures analysis may also be performed. Best percent change in tumor size is defined as the maximum post-baseline reduction (minimum percent change). The best percent change from baseline (per investigator assessment) will be presented per patient in a waterfall plot. The analysis of change in tumor size will be conducted in the ORR evaluable population.

4.8.9. Time to Response

Time to response (TTR) is defined as the time from the date of randomization to the date measurement criteria for CR or PR (whichever is first recorded) are first met. TTR will be summarized for each treatment arm. The analysis of TTR will be based on the participants who achieve an objective response (CR or PR).

4.8.10. Follow-up Time

Follow-up time is defined as the time from the date of randomization until death from any cause or last date the patient is known to be alive and under follow-up. Median follow-up time will be estimated using Kaplan-Meier estimation of potential follow-up (“reverse Kaplan-Meier”) (Schemper and Smith 1996). The inverse of the censoring rules for the OS will be used (i.e., considering all censoring times for OS as event times (times when the patient is known to be still alive and under follow-up) and censoring patients who had OS events at the date of death.

4.8.11. Medical Resource Utilization

Frequency counts of hospitalizations, emergency room visits, radiation, surgery, transfusion, and analgesic use will be summarized descriptively for each arm.

Duration of hospital stays and average number of emergency room visits will be reported by treatment arm.

4.8.12. Subgroup Analyses

Subgroup analyses of PFS and OS will be performed for potential prognostic subgroup variables, including but not limited to

- All baseline stratification factors
- *ESR1* mutation status (mutation detected versus mutation not detected), for the analyses comparing Arm A versus Arm B and Arm C versus Arm A in the ITT population
- Measurable disease at baseline (yes versus no)
- Bone only disease at baseline (yes versus no)
- Age (<65 years versus ≥65 years)
- Region (North America, Europe, Asia, and Other)
- Race (Caucasian, Asian, and Other)
- Progesterone receptor status (positive versus negative)
- Baseline ECOG PS (0 versus 1)
- Number of sites involved (1 versus 2 versus 3+).
- Intended use of endocrine therapy in the control arm (fulvestrant versus exemestane)
- Endocrine resistance (primary versus secondary)
- Prior CDK4/6 inhibitor (yes versus no)
    - If yes, setting (adjuvant versus metastatic)
    - If yes and received prior CDK4/6 inhibitor in the metastatic setting, length of the prior CDK4/6 inhibitor (≥ 12 months versus <12 months)
- Type of prior CDK4/6 inhibitor (palbociclib versus ribociclib versus abemaciclib versus dalpiciclib)
- HER2 status (low [IHC 1+ and IHC 2+] versus negative [IHC=0])
- Prior therapy history, for example, subgroup variables derived per inclusion criteria 3a-c in the protocol if deemed appropriate
    - patient progressing while on or within 12 months of adjuvant treatment (see 3a),
    - patient progressing more than 12 months after completion of (neo)adjuvant treatment with subsequent progression on or after only 1 line of therapy (see 3b),
    - patient presenting metastatic disease de novo with subsequent progression on or after only 1 line of therapy (see 3c)

If a level of a factor consists of fewer than 5% of total number of events, analysis within that level may be omitted. The level of "missing/unknown" may be dropped if deemed appropriate. Other subgroup analyses may be performed as deemed appropriate. A forest plot will be used to present the results graphically. P-values for the interaction between the treatment and subgroup variables will be reported.

Subgroup analyses may also be performed for safety analyses (such as for summary of TEAE), including but not limited to age group, gender, race, region, menopausal status and baseline *ESR1* mutation status.

Subgroup analyses for Japan and China regulatory submissions will be described in a separate SAP addendum.

4.8.13. Biomarker Analyses

Biomarkers related to treatment, mechanism of action, and/or cancer will be measured and analyzed. The association of biomarker and clinical outcome will be assessed via single-marker and/or multi-marker analysis. Baseline *ESR1* mutation status (mutation detected versus mutation not detected) will be summarized. Baseline is defined as pre-treatment sample collected on the first date of dose (C1D1) when available. If the C1D1 pre-treatment sample fails or is not provided or not sufficient plasma is provided, the screening sample will be used as baseline.

4.8.14. Important Protocol Deviations

Important protocol deviations that potentially compromise the data integrity and participants' safety will be summarized. These deviations will include deviations that can be identified programmatically and those which can only be identified by the clinical research associates during monitoring. Important protocol deviations are described in another document within the study Trial Master File.

4.9. Interim Analyses

4.9.1. Data Monitoring Committee

Interim analyses for safety and efficacy will be conducted under the guidance of an independent data monitoring committee (DMC). The DMC will consist of at least 3 members, including 2 clinicians and 1 statistician. The DMC will communicate any recommendations based on interim analysis to the Sponsor senior management designee (SMD). If necessary, the SMD may form an internal review committee (IRC) to review and act upon the recommendations of the DMC. Details will be provided in a separate DMC charter.

4.9.2. Safety Interim Analyses

The DMC will monitor the overall safety of the study. An early safety analysis will be performed after approximately 100 participants have been randomized and had the opportunity to be treated for 1 cycle. The DMC will meet and review data approximately every 6 months thereafter. At the recommendation of the DMC, the frequency of safety interim analyses may be modified.

At each interim analysis, the DMC may recommend the trial continue without modifications, continue with specific modifications, or be stopped for safety concerns. There will be no prespecified rules for stopping the trial due to safety concerns. The DMC members will review unblinded safety data at each interim analysis. If a significant safety signal is identified, the DMC may recommend a protocol amendment, termination of enrollment, and/or termination of study treatment. The recommendations of the DMC will be communicated to the Sponsor SMD.

In the event that safety monitoring uncovers an issue that needs to be addressed by unblinding at the treatment group level, members of the DMC can conduct additional analyses of the safety data. Additionally, unblinding of a limited number of the Sponsor representatives external to the study team may be required for evaluation of selected SAEs for determination of regulatory reporting.

4.9.3. Efficacy Interim Analyses

One interim analysis is planned for the primary endpoint of PFS between Arm A and Arm B, when approximately 192 of the 480 events (40% information fractions, respectively) have been observed in the analysis population as defined in Section 3. The purpose of this interim is to allow the trial to stop early due to futility. The beta-spending function is used to control the type II error rate, which is determined by the gamma family. The DMC should recommend stopping the trial (for the comparison between Arm A versus Arm B) for futility if the hazard ratio is above 1.128. The boundary will be updated based on the actual number of observed events. There is no intent to declare statistical significance for superior efficacy at this interim; therefore, there is no impact on the statistical significance levels for the final analysis.

Table 4.4. Stopping Boundaries for Each Analysis Between Arm A and Arm B in the ITT Population Based on the Initial Allocation of Alpha for Illustration

Analysis	Number of Events	Information Fraction	Critical P-value Boundary	Critical Hazard Ratio Boundary	Cumulative Type I Error Rate	Boundary Crossing Probabilities[a]		
						Hazard Ratio = 0.74	Hazard Ratio = 1	Hazard Ratio = 1.25
Interim 1	192	40%	NA	1.128	NA	0.002	0.203	0.757
Final	480	100%	0.005 (efficacy)	0.790	0.005	0.759	0.005	0

Abbreviations: ITT = intention-to-treat; NA = not applicable.

[a] Boundary crossing probabilities under different hazard ratio assumptions.

For PFS between Arm A and Arm B in the *ESR1*-mutation detected population, if approximately 192 events have been observed in this subset at the time of final PFS analysis in the ITT population, the final PFS analysis for the *ESR1*-mutation detected population will be conducted at the 1-sided alpha level of 0.02 (assuming the initial alpha level based on the graphical approach). If the target number of events in this subset has not been reached at this time, 1 interim analysis for efficacy may be conducted in the subset at this time if deemed appropriate. The Lan-DeMets spending function (O'Brien-Fleming type) will be used to determine the boundaries at the interim and final analyses for PFS in the *ESR1*-mutation detected population. In this situation, DMC will review the final analysis in the ITT population and be instructed to recommend to the SMD that the results for the comparison between Arm A and Arm B be released to the Sponsor if PFS in the ITT population is statistically significant based on the initial alpha. If the number of events in the subset is close enough to the target number of events (e.g., at least 180 events have been observed assuming a target number of 192 events) then the final analysis for PFS in the subset may be performed at this time.

Table 4.5 summarizes the boundaries for PFS in the *ESR1*-mutation detected population assuming 1 interim look will be performed, with the look 1 at approximately 180 events. The boundaries will be updated based on the actual number of observed events and the actual 1-sided alpha level that is used for this endpoint per the graphical approach.

Table 4.5. Stopping Boundaries for Each Analysis Between Arm A and Arm B in the ESR1-mutation Detected Population Based on the Initial Allocation of Alpha for Illustration (Assuming 1 Interim Look)

Analysis	Number of Events[a]	Information Fraction	Critical P-value Boundary	Critical Hazard Ratio Boundary	Cumulative Type I Error Rate
Look 1 (at final in ITT)	180	94%	0.016 (efficacy)	0.727	0.016
Final	192	100%	0.016 (efficacy)	0.733	0.020

Abbreviation: ITT = intention-to-treat.

[a] Number of events at Look 1 was estimated based on enrollment and PFS assumptions and assuming that the sample size of *ESR1*-mutation detected subset is approximately 240 patients.

One interim analysis is planned for the primary endpoint of PFS between Arm C and Arm A after approximately 100 of the 248 events (40% information fraction) have been observed in the analysis population for this comparison as defined in Section 3. The purpose of the interim is to allow the trial comparison for Arm C versus Arm A to stop early (stop the enrollment to Arm C if not completed) due to futility. The beta-spending function is determined by the gamma family. The DMC should recommend stopping the trial (for Arm C versus Arm A) for futility if the hazard ratio is above 1.126. The boundary will be updated based on the actual number of observed events. There is no intent to declare statistical significance for superior efficacy at this interim; therefore, there is no impact on the statistical significance levels for the final analysis. The study will continue after the interim analysis if the futility interim is passed.

Table 4.6. Stopping Boundaries for Each Analysis between Arm C and Arm A in the ITT Population Based on the 1-sided Alpha Level of 0.025 for Illustration

Analysis	Number of Events	Information Fraction	Critical P-value Boundary	Critical Hazard Ratio Boundary	Cumulative Type I Error Rate	Boundary Crossing Probabilities[a]		
						Hazard Ratio = 0.7	Hazard Ratio = 1	Hazard Ratio = 1.25
Interim	100	40%	NA	1.126	NA	0.009	0.279	0.705
Final	248	100%	0.025 (efficacy)	0.779	0.025	0.794	0.027	0

Abbreviations: ITT = intention-to-treat; NA = not applicable.

[a] Boundary crossing probabilities under different hazard ratio assumptions.

Only the DMC is authorized to evaluate unblinded safety and efficacy interim analyses. Study sites will receive information about interim results only if they need to know for the safety of their participants.

An interim analysis may be collapsed with any other analyses including a final analysis if they are expected to occur within a similar timeframe (e.g., within approximately 2 months).

Additionally, the final analysis of PFS between Arm A and Arm B in the ITT population may be collapsed with the final analysis in the *ESR1*-mutation detected population or the final analysis between Arm C and Arm A in the ITT population if they are expected to occur within a similar timeframe. At the time of the final evaluation of PFS between Arm A and Arm B in the ITT population or in the *ESR1*-mutation detected population, whichever is later (if they are not collapsed), the Sponsor will only have access to Arm A and Arm B data and will be blinded to

Arm C if the final analysis between Arm A and Arm B is before the final analysis between Arm C and Arm A.

Unblinding details are specified in a separate blinding and unblinding plan document.

5. Sample Size Determination

Participants will be randomly assigned to Arm A, Arm B or Arm C in a 1:1:1 ratio until the target enrollment for arms A and B (a total number of approximately 640 participants) is reached.

Though Arm C was added to the study (amendment a) after first patient visit for Arms A and B, all arms will be closed at the same time. Randomization of patients will continue in Arms A and B (1:1) until amendment (a) is approved and implemented, at which point patients will be randomized 1:1:1 (A:B:C) until the target enrollment is met.

To adjust for multiplicity and control the overall type I error rate at 0.025 (1-sided), the graphical approach (Bretz et al. 2009; Mauer and Bretz 2013) will be used to test the 3 PFS hypotheses and the 3 OS hypotheses. Initially, the overall 1-sided alpha level of 0.025 will be split between PFS for Arm A versus Arm B in the ITT population (H_1) and PFS for Arm A versus Arm B in the *ESR1*-mutation detected population (H_2), with H_1 tested at the 1-sided alpha level of 0.005 and H_2 tested at the 1-sided alpha level of 0.02. Zero alpha is initially assigned to the PFS endpoint for Arm C versus Arm A in the ITT population (H_3) and OS endpoints. Details are described in Section 4.3.3.

As the overall alpha is initially split between H_1 and H_2, the study will be considered positive (for Arm A versus Arm B) if either PFS in the ITT population (H_1) or PFS in the *ESR1*-mutation detected population (H_2) is statistically significant. The power and associated sample sizes for H_1 and H_2 are based on the initial allocation of alpha. The analysis population for the first primary hypothesis (PFS for Arm A versus Arm B in the ITT population) is all participants randomized to Arm A and Arm B. The primary analysis of PFS for Arm A versus Arm B in the ITT population will be performed after approximately 480 investigator-assessed events have been observed (that is, a 25% censoring rate) in Arm A and Arm B. Assuming a PFS hazard ratio of 0.74, a total of 480 events yields at least 76% power to detect superiority of imlunestrant Arm A over Arm B with the 1-sided log-rank test at the significance level of 0.005. If H1 can be tested at the full alpha level of 0.025 after recycling per the graphical approach (Figure 4.4), the same number of events can yield at least 91% power with the 1-sided log-rank test. The median PFS of Arm B is assumed to be 4.3 months, and the hazard ratio of 0.74 amounts to an approximate 1.5-month improvement in median PFS under the assumption of exponential survival distribution. The assumed median PFS of Arm B is estimated based on an unpublished meta-analysis of historical controls.

The analysis population for the second primary hypothesis (PFS for Arm A versus Arm B in the *ESR1*-mutation detected population) is the *ESR1*-mutation detected subset in Arm A and Arm B. The primary analysis of PFS will be performed after approximately 192 investigator-assessed events have been observed in the *ESR1*-mutation detected subset. Assuming a PFS hazard ratio of 0.57, a total of 192 events yields approximately 97% power to detect superiority of Arm A over Arm B with the 1-sided log-rank test at the significance level of 0.02. Assuming the median PFS of Arm B in the subset is 3.6 months, the hazard ratio of 0.57 amounts to an approximate 2.7-month improvement in median PFS under the assumption of exponential survival distribution.

The power and sample size for H_3 is based on the 1-sided alpha level of 0.025 assuming that both H_1 and H_2 are rejected. The analysis population of the third primary hypothesis (PFS for Arm C

versus Arm A in the ITT population) is the participants concurrently randomized to Arm A and Arm C. The primary analysis of PFS will be performed after approximately 248 investigator-assessed events have been observed in this analysis population. Assuming a PFS hazard ratio of 0.7, a total of 248 events yields at least 80% power to detect superiority of Arm C over Arm A with the 1-sided log-rank test at the significance level of 0.025. If the median PFS of Arm A is assumed to be 5.8 months (assuming the target HR for Arm A versus Arm B in the ITT population is met), the hazard ratio of 0.7 then amounts to an approximate 2.5-month improvement in median PFS under the assumption of exponential survival distribution. Per the enrollment assumptions below, it is estimated that 220 participants will be enrolled in Arm C. Thus approximately 440 participants will be concurrently randomized to Arm A and Arm C.

The following enrollment assumptions are considered to estimate the planned total sample size for this study.

1. Participants will be enrolled at a rate of 5, 10, 15, 20, 25, 35, 35, 35, 35, 35, 40, 40, 40, 40/month for the first 14 months, and at a rate of 45/month for the reminder of the enrollment period.
2. Arm C will enter the study 6 months after the first patient visit.
3. Once arm C enters, participants will be randomly assigned to 1 of the 3 arms at a gradually increasing rate until the peak, specifically, at a rate of 15, 15, 18, 18, 36, 36, 36, 36/month for 8 months, and 45/month thereafter.
4. After Arm C is added to the trial, participants may continue to be randomized to arms A and B (at the sites where Arm C is not fully implemented yet), at a monthly rate which is the difference between the rates in assumption 1 and those in assumption 3 in the corresponding months, until amendment a (Arm C) is implemented at all sites.
5. All arms will be closed at the same time.

Under these assumptions, it is estimated that a total number of 860 participants will be enrolled to this study. Specifically, approximately 320 participants will be enrolled in Arms A and B respectively, and approximately 220 participants in Arm C.

The sample size calculation is conducted in EAST version 6.5.

6. Supporting Documentation

6.1. Appendix 1: Clinical Trial Registry Analyses

Additional analyses will be performed for the purpose of fulfilling the Clinical Trial Registry (CTR) requirements.

Analyses provided for the CTR requirements include the following:

- Summary of adverse events, provided as a dataset which will be converted to an XML file. Both Serious Adverse Events and 'Other' Adverse Events are summarized: by treatment group, by MedDRA preferred term.
- An adverse event is considered 'Serious' whether or not it is a treatment emergent adverse event (TEAE).
- An adverse event is considered in the 'Other' category if it is both a TEAE and is not serious. For each Serious AE and 'Other' AE, for each term and treatment group, the following are provided:
    - the number of participants at risk of an event
    - the number of participants who experienced each event term
    - the number of events experienced.
- Consistent with www.ClinicalTrials.gov requirements, 'Other' AEs that occur in fewer than 5% of participants/subjects in every treatment group may not be included if a 5% threshold is chosen (5% is the minimum threshold).
- AE reporting is consistent with other document disclosures for example, the CSR, manuscripts, and so forth.

In addition, the following rules apply in order to meet the requirement for participant flow and accurately represent study completion.

Study Discontinuation Reason	Completed	Not Completed
Participants who had an event (progressive disease or death)	X	
Participants who were off the treatment and were alive at study conclusion	X	
Lost to follow-up*		X
Withdrew consent to study participant (participant or physician)*		X
On study treatment at study conclusion		X

*Include participants only if not meeting the definition for "Completed".

6.2. Appendix 2: Per Protocol Set Definition

Per Protocol analysis set is defined according to important protocol deviations (IPD) defined in the Trial Issue Management Plan (TIMP). Important protocol deviations (IPD) are a subset of protocol deviations that may significantly impact the completeness, accuracy and/or reliability of key study data or that may significantly affect a subject's rights, safety, or well-being. Based on IPD, we define the Per Protocol (PP) Set as a subset of subjects in the ITT population who do

not have an IPD (e.g. clinically important and potentially impact efficacy evaluations) as listed below:

- Received incorrect study drug
- Did not receive study treatment
- Non-compliant patients
- Took prohibited anti-cancer therapy while on study treatment
- Inclusion #1, #2, #3, #5, #6
- Exclusion #15, #16, #17, #19, #20, #29

The PP set will be used to perform sensitivity analysis for the primary efficacy endpoints if deemed appropriate.

7. References

Amit O, Mannino F, Stone AM, et al. Blinded independent central review of progression in cancer clinical trials: results from a meta-analysis. *Eur J Cancer*. 2011;47(12):1772-1778. https://doi.org/10.1016/j.ejca.2011.02.013

Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Stat Med*. 2009;(4):586-604. https://doi.org/10.1002/sim.3495

Bretz F, Posch M, Glimm E, et al. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Bioml J*. 2011;53(6):894-913. https://doi.org/10.1002/bimj.201000239

Cox DR. Regression models and life-tables. *J Royal Stat Soc Ser B*. 1972;34(2):187-220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247. https://doi.org/10.1016/j.ejca.2008.10.026

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481. https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452

Maurer W, Bretz F. Multiple testing in group sequential trials using graphical approaches. *Stat Biopharm Res*. 2013;5(4):311-320. https://doi.org/10.1080/19466315.2013.807748

Schemper M, Smith TL. A note on quantifying follow-up time in studies of failure time. *Control Clin Trials*. 1996;17(4):343-346. https://doi.org/10.1016/0197-2456(96)00075-x

Signature Page for VV-CLIN-122802 v1.0

Approval	PPD 24-Aug-2023 17:19:12 GMT+0000

Signature Page for VV-CLIN-122802 v1.0