FERRING
PHARMACEUTICALS

Statistical Analysis Plan

Title of trial: A Multicenter, Double-Blind, Randomized, Placebo-Controlled Study Evaluating Three Doses of Subcutaneous Pulsatile GnRH Administered via OmniPod Pump for Ovulation Induction in Female Subjects with Primary Amenorrhea with Hypogonadotropic Hypogonadism
NCT number: NCT01976728
Sponsor trial code: 000070
Date: 07 Dec 2017

STATISTICAL ANALYSIS PLAN

A Multicenter, Double-Blind, Randomized, Placebo-Controlled Study Evaluating Three Doses of Subcutaneous Pulsatile GnRH Administered via OmniPod Pump for Ovulation Induction in Female Subjects with Primary Amenorrhea with Hypogonadotropic Hypogonadism

Trial Code 000070

Investigational Product:	LutrePulse OmniPod
Indication:	Primary Hypothalamic Amenorrhea
Phase:	III
Author:	[redacted]
Date of issue:	December 7, 2017
Version:	Amendment 1

The information in this document is confidential and is proprietary to Ferring Pharmaceuticals A/S or another company within the Ferring Group. It is understood that information in this document shall not be disclosed to any third party, in any form, without prior written consent of an authorised officer of Ferring Pharmaceuticals A/S or another company within the Ferring Group

Change log

Version No.	Effective Date	Reason for the Change / Revision	Supersedes
V1.0	Jul 11, 2017	SAP based on Protocol Amendment 6 dated June 30, 2017.	
Amendment 1	Dec 7, 2017	1. Per protocol amendment 7, the definition for the primary efficacy endpoint of ovulation is changed. Subjects who had a confirmed positive serum β-hCG (i.e., 2 positive results) or presence of a gestational sac documented by transvaginal ultrasound (TVUS) is considered as ovulatory, even if the subject did not have any post-baseline P4 ⩾6 ng/mL. 2. The positive LH surge rate is added as a secondary efficacy endpoint.	V1.0

TABLE OF CONTENTS

1 Introduction

This document describes the data analysis specifications for Ferring Pharmaceuticals protocol LutrePulse OmniPod titled: "A Multicenter, Double-Blind, Randomized, Placebo-Controlled Study Evaluating Three Doses of Subcutaneous Pulsatile GnRH Administered via OmniPod Pump for Ovulation Induction in Female Subjects with Primary Amenorrhea with Hypogonadotropic Hypogonadism". This version of the statistical analysis plan dated December 7, 2017 was prepared in accordance with the protocol LutrePulse OmniPod, Trial ID 000070, Amendment 7, dated December 6, 2017. Other related documents are the electronic case report form (eCRF), data definition fields and data set definitions.

The purpose of this Statistical Analysis Plan (SAP) is to provide greater and more technical details of the pre-planned statistical analyses described in the statistical section of the Clinical Trial Protocol. It aims to answer the protocol objectives in a statistically rigorous fashion using methods identified prior to database lock. Any deviations from these methods must be documented in the final clinical study report. Specifications of tables, figures, and data listings are contained in a separate document (Table shells).

1.1 Definitions/Abbreviations

1.1.1 Definition of Terms

Terms	Definitions
Randomised	Subject randomised to trial treatment
Screened	Subject who enters the screening phase
OmniPod	A disposable delivery pump for drug delivery with a handheld wireless controller (the Manager) for SC drug delivery
OmniPod Manager	A wireless, handheld device that programs the OmniPod with the subject's LutrePulse delivery instructions and wirelessly monitors the OmniPod's operation
Unblinded Coordinator	Assigned by the investigator prior to the start of the study and will be unblinded to the subject's study drug assignment (dose only) and will be responsible for the set-up of the OmniPod Manager, distribution and accountability of the drug.

1.1.2 Abbreviations

Abbreviations	Meaning of abbreviations in document
AE	Adverse Event
ANCOVA	Analysis of Covariance
ATC	Anatomical-Therapeutic-Chemical
GLM	General Linear Model
ITT	Intent-to-Treat
FAS	Full Analysis Set
PPS	Per-Protocol Set
SAS	Statistical Analysis System
NPAR1WAY	Non-Parametric One-Way Analysis of Variance
IMP	Investigational Medicinal Product

Abbreviations	**Meaning of abbreviations in document**
HH	Hypogonadotropic Hypogonadism
LH	Luteinizing Hormone
eCRF	Electronic Case Report Form
FSH	Follicle-Stimulating Hormone
ECG	Electrocardiography
E2	Estradiol
P4	Progesterone
PRL	Prolactin
TSH	Thyroid-Stimulating Hormone
TVUS	Transvaginal Ultrasound
FHM	Fetal Heart Movement
GnRH	Gonadotropin-Releasing Hormone
CT	Computed Tomography
MedDRA	Medical Dictionary for Regulatory Activities
MRI	Magnetic Resonance Imaging
mITT	Modified Intent-to-Treat
β-hCG	Beta human Chorionic Gonadotropin

2 Trial Objectives and Endpoints

2.1 Primary Objective

- To compare the ovulation rate in women with primary amenorrhea with hypogonadotropic hypogonadism (HH) following each of 3 doses (10 µg, 15 µg, and 20 µg per pulse) of subcutaneous pulsatile gonadotropin-releasing hormone (GnRH) administered with the LutrePulse OmniPod Pump versus placebo

2.2 Secondary Objectives

- To evaluate clinical pregnancy rate
- To evaluate biochemical pregnancy rate
- To evaluate positive LH surge rate
- To evaluate ovarian follicular development
- To evaluate luteal phase support with GnRH
- To evaluate pharmacodynamic parameters
- To evaluate the safety of LutrePulse OmniPod SC

2.3 Primary Endpoint

- Ovulation rate, calculated as a proportion of subjects with at least 1 post-baseline progesterone (P4) level ≥ 6 ng/mL as measured by a central laboratory a confirmed positive serum β-hCG (i.e., 2 positive results) or presence of a gestational sac documented by transvaginal ultrasound (TVUS).

2.4 Secondary Endpoints

2.4.1 Progesterone Level

- Proportion of subjects with at least 1 post-baseline progesterone (P4) level ≥ 10 ng/mL

2.4.2 Pregnancy

Clinical pregnancy rate

- Proportion of subjects with presence of gestational sac and fetal heart movement (FHM) on transvaginal ultrasound (TVUS) at approximately 2 to 4 weeks after second positive serum beta human chorionic gonadotropin (β-hCG) test

Biochemical pregnancy rate

- Proportion of subjects with a confirmed positive (i.e., 2 positive results) serum beta human chorionic gonadotropin (β-hCG) test 14 + 4 days post LH surge

2.4.3 LH surge detection

- Proportion of subjects with a positive detection of LH surge on their Clearblue tests

2.4.4 Ovarian Follicular Development

- Number of follicles with a mean diameter ≥ 14 mm
- Number of dominant follicles with a mean diameter ≥ 18 mm

2.4.5 Luteal Phase Support

- Maximum P4 levels from Days 19, 21, 23, 25, and 27
- Mean P4 levels from Days 19, 21, 23, 25, and 27

2.4.6 Pharmacodynamic (PD) Endpoints

- FSH and LH change from baseline in relation to the first GnRH pulse on Day 1 and Day 10
- Mean serum FSH and LH levels on Day 1 and Day 10
- Estradiol (E2) serum levels on Day 1 and Day 10

2.4.7 Safety Endpoints

- Adverse events (type, frequency, causality, intensity and seriousness)
- Clinical chemistry, haematology, urinalysis
- Frequency and severity of ovarian hyperstimulation syndrome (OHSS)

3 Trial design

This multicenter, randomized, double-blind, placebo-controlled study will be performed in women with primary amenorrhea with HH who desire pregnancy. Subjects with primary amenorrhea with HH will be treated with 1 of 3 doses of pulsatile GnRH or placebo, administered SC via the OmniPod pump manufactured by Insulet. The treatment duration with LutrePulse OmniPod SC or placebo can be approximately 5 weeks. It is expected that the response to treatment will occur within 2 to 3 weeks after therapy initiation. When LH surge is detected by urine kit, therapy should be continued for approximately another 2 weeks to maintain corpus luteum (continue until confirmatory serum pregnancy test). Subjects will discontinue treatment either at the onset of menses or negative or positive confirmatory serum pregnancy test. Subjects will be required to return to the study site for pharmacodynamic blood draws, TVUS, serum pregnancy tests, and safety follow-up assessments.

The study will consist of 3 periods: Screening (includes progestin challenge test), Treatment (includes pituitary priming and treatment phase), and Follow-up.

3.1 General Design Considerations

Approximately 60 subjects will be randomly assigned to 1 of 4 treatment arms:

- LutrePulse OmniPod SC 10 µg (Test Product)
- LutrePulse OmniPod SC 15 µg (Test Product)
- LutrePulse OmniPod SC 20 µg (Test Product)
- Placebo (10, 15, or 20 µg /pulse as a fixed “dose”)

Subjects will receive consecutive pulses of GnRH or placebo, each delivered at 90-minute intervals, for up to 37 days of treatment. Each OmniPod will be worn for 3 full days (Days 1-3, 4-6, 7-9, etc. up to Day 37) and will be replaced every 4th day (i.e., Day 4, Day 7, Day 10, etc.). New OmniPod pumps should be applied at approximately the same time as the first pump. Randomized fixed doses of either 10 µg, 15 µg, or 20 µg will be programed into the OmniPod Manager and subcutaneously delivered to the subject. For every 10 µg, 15 µg, or 20 µg dose the subject will receive 10 µL, 15 µL, or 20 µL amount of volume with the same concentration of drug per pulsatile injection. Placebo doses will receive the same amount of volume respectively as the Lutrepulse OmniPod dose without active drug substance.

3.2 Determination of Sample Size

The sample size was determined based on the original fixed parallel group study design. The true ovulation rate for LutrePulse OmniPod SC 20µg, 15µg, 10µg, and placebo was assumed to be 70%, 65%, 60%, and 1%, respectively. Under these assumptions, a sample size of 14 subjects per group (total 56 subjects) will provide at least 80% power to detect differences between LutrePulse OmniPod SC 20µg and placebo, 15µg and placebo, and 10µg and placebo in ovulation rate separately with the stratified one-sided Fisher’s exact tests at an overall significance level of 0.025 with the closed testing procedure (Kong et al., 2005). Approximately 60 subjects will be randomized by assuming up to 4 subjects may be ineligible to the Full Analysis Set.

3.3 Randomization, Blinding, and Replacement of Subjects

All eligible subjects will be randomly assigned at least 3 days before Treatment Day 1, based on the computer-generated randomization list provided by the designated unblinded statistician. Any relevant information pertaining to the generation of the randomization list including the block size will only be communicated from the sponsor to the unblinded statistician.

The first 15 eligible subjects will be randomly assigned to 1 of 3 treatment arms: LutrePulse OmniPod SC 10 µg, LutrePulse OmniPod SC 20 µg, or OmniPod SC placebo in a 1:1:1 ratio. The subsequent 25 eligible subjects will be randomly assigned to LutrePulse OmniPod SC 10 µg, LutrePulse OmniPod SC 15 µg, LutrePulse OmniPod SC 20 µg, or OmniPod SC placebo in a 1:2:1:1 ratio. The rest of the eligible subjects will be randomly assigned to LutrePulse OmniPod SC 10 µg, LutrePulse OmniPod SC 15 µg, LutrePulse OmniPod SC 20 µg, or OmniPod SC placebo in a 1:1:1:1 ratio.

To maintain the blind, the placebo subjects will be randomized to the equivalent dosing setting in the Manager for 10 µg /pulse or 20 µg /pulse for the first 15 eligible subjects and 10 µg /pulse, 15 µg /pulse, or 20 µg /pulse for the subsequent randomization schemes.

Table 1 displays the estimated number of subjects randomized to each treatment arm. The three placebo groups with different dose settings will be pooled as a single placebo group for all data summaries and analyses. The LutrePulse 20 µg will be compared to (Placebo I + Placebo II + Placebo III), LutrePulse 15 µg will be compared to (Placebo I + Placebo II + Placebo III), and LutrePulse 10 µg will be compared to (Placebo I + Placebo II + Placebo III).

The randomization assignment will be available to the OmniPod Manager programmer (Unblinded Coordinator) with respect to dose only (but will remain blinded to assigned study drug, whether active or placebo). The randomization assignment will not be available to any other site personnel involved in the conduct and evaluation of the study until the study database is declared clean and is released to the statistician.

Subjects who discontinue early from the study for any reason will not be replaced.

Table 1 Treatment arms

Treatment Arm	Treatment	Estimated # of Patients Entering Treatment	Formulation and Dose of Study Drug	Frequency	Admin. Route
LutrePulse OmniPod SC 20 µg	LutrePulse 20 µg / pulse	15	20 µg dose in white powder reconstituted in 20 µL saline with 0.9% sodium chloride	90-minute intervals daily	SC pulsatile injection Via OmniPod pump
LutrePulse OmniPod SC 15 µg	LutrePulse 15 µg / pulse	15	15 µg dose in white powder reconstituted in 15 µL saline with 0.9% sodium chloride	90-minute intervals daily	SC pulsatile injection Via OmniPod pump

LutrePulse OmniPod SC 10 μg	LutrePulse 10 μg / pulse	15	10 μg dose in white powder reconstituted in 10 μL saline with 0.9% sodium chloride	90-minute intervals daily	SC pulsatile injection Via OmniPod pump
OmniPod SC Placebo I	Placebo 20 μg / pulse	5	20 μL saline with 0.9% sodium chloride	90-minute intervals daily	SC pulsatile injection Via OmniPod pump
OmniPod SC Placebo II	Placebo 15 μg / pulse	5	15 μL saline with 0.9% sodium chloride	90-minute intervals daily	SC pulsatile injection Via OmniPod pump
OmniPod SC Placebo III	Placebo 10 μg / pulse	5	10 μL saline with 0.9% sodium chloride	90-minute intervals daily	SC pulsatile injection Via OmniPod pump

4 Subject Disposition

The number and percentage of randomized subjects treated with Investigational Medicinal Product (IMP), prematurely discontinued, and completing the study will be summarized.

A subject who has an increased risk of developing multiple gestation/OHSS (4 or more follicles with a mean diameter ≥ 15 mm by TVUS and E2 levels ≥ 1200 pg/mL) will be immediately discontinued from treatment. In addition, any clinically important adverse physical or symptomatic findings/events and/or abnormal laboratory findings that are above the pre-specified critical values defined in the laboratory manual will be grounds for discontinuation from the study. Subjects who fail to respond to study treatments (no LH surge) and withdraw from study treatments with less than 14 days of study treatment exposure will be considered as an early discontinuation. For subjects with at least 14 days of study treatment exposure, the end of study visit is required for the subject to be considered as a completer.

All post-baseline discontinuations will be summarized by reason for early termination. Subjects screened but not randomized/allocated to treatment will be presented with the reason(s) for screen failure in a data listing. All withdrawals from trial will be summarised with frequency and percentage for each category of withdrawal reason**.**

5 Protocol Deviations

The criteria for protocol deviations considered major with the implication of data exclusions from the per-protocol analysis set will be identified and documented by Ferring clinical team/project management based on a review of data listings prior to database lock.

Major protocol deviations may include major violations of the inclusion/exclusion criteria, missing randomization, errors in treatment assignment, significant noncompliance to the dosing administration, or other serious violations.

Number and percentage of subjects with any major and minor deviations will be tabulated for each category of protocol deviation for descriptive purposes by treatment group.

6 Analysis sets

6.1 Intent-To-Treat Analysis Set

All randomized (as planned) subjects will be included in the ITT analysis set.

6.2 Full-Analysis Set

All randomized subjects who are included in the ITT analysis set and received at least 14 days of study drug for pituitary priming will be included in the FAS.

6.3 Modified Intent-To-Treat Analysis Set

All randomized subjects who received at least one pulsatile injection delivery of study drug will be included in the mITT Population.

6.4 Per Protocol Analysis Set

All subjects who are included in the FAS and did not have exclusionary major protocol deviations will be included in the PPS.

6.5 Safety Analysis Set

All subjects who received at least one pulsatile injection delivery of study drug will be included in the Safety Population. Subjects are analysed according to the actual treatment received.

Table 2 Treatment group variables vs. analysis set

		Treatment Variable Used in Analysis		
Analysis Set	**Analysis Set Population Flag**	**Variable Label**	**Variable Name (Character)**	**Variable Name (Numeric)**
Intent-To-Treat Analysis Set	ITTFL	Planned Treatment	TRTP	TRTPN
Full-Analysis Set	FASFL	Planned Treatment	TRTP	TRTPN
Modified Intent-To-Treat Analysis Set	mITTFL	Planned Treatment	TRTP	TRTPN
Per Protocol Analysis Set	PPROTFL	Planned Treatment	TRTP	TRTPN
Safety Analysis Set	SAFFL	Actual Treatment	TRTA	TRTAN

7 Trial population

7.1 Demographics and Other Baseline Characteristics

A summary of demographics and other baseline characteristics will be presented for the subjects in the FAS by treatment group. No statistical testing for baseline differences will be performed.

Categorical (discrete) data will be summarized using numbers and percentages. The percentages will be based on the total number of subjects with a corresponding assessment. Continuous data will be presented, for example, using the number of subjects (N), mean and standard deviation, median, minimum and maximum. All baseline characteristics will be listed.

7.1.1 Demographics

Descriptive statistics of baseline demographics variables (e.g., age, race and ethnic origin, height, weight and BMI) will be summarized for the subjects in the FAS by treatment group.

7.1.2 Vital Signs at Baseline

Baseline vital signs will be summarised in the FAS by treatment group.

7.1.3 Laboratory Efficacy/Pharmacodynamic Parameters at Baseline

Baseline assessments of safety laboratory tests (haematology, chemistry and urinalysis) and other laboratory tests (urine pregnancy, LH, FSH, E2, P4) will be summarized for the subjects in the FAS by treatment group. The baseline values for all laboratory/pharmacodynamics parameters will be the values obtained at the last assessment prior to the first dose of IMP.

7.2 Medical History

Medical and gynecologic history recorded at screening visit will be coded using the Medical Dictionary for Regulatory Activities (MedDRA) at the latest version available at the time of interim or final analysis, and will be listed and summarized by system organ class (SOC) and preferred term (PT).

7.3 Prior and Concomitant Medication

Prior and concomitant medication will be coded using World Health Organisation Drug (WHODrug) Dictionary version September 1, 2013 and summarised by anatomical-therapeutic-chemical (ATC) classification 1st level (alphabetically), ATC classification 2nd level (in decreasing order of frequency) and treatment group. These medications will be tabulated separately for:

1) Prior medication, i.e. medication taken exclusively prior to treatment (i.e. with stop date before date of first IMP administration);
2) Concomitant medication, i.e. medication taken during the treatment period (i.e. medication that was not stopped before date of first IMP-administration).

If the timing of the dose of a concomitant medication cannot be established in relation to the administration of IMP, it will be considered as concomitant medication. Concomitant medication will be subdivided into ongoing (started before first IMP administration) and treatment-emergent (starting after first IMP administration) concomitant medication.

7.4 Physical Examination

Subjects with abnormalities at any screening, baseline, or post-baseline visit will be listed with all physical examination evaluations.

8 Exposure and Treatment Compliance

8.1.1 Extent of Exposure

The total duration of study treatment exposure defined as

Date of last OmniPod application – Date of first OmniPod application +3

will be summarized descriptively and listed for the Safety Population by treatment group.

8.1.2 Treatment Compliance

Summary (descriptive) statistics will be presented for the number of OmniPods applied by each patient and the treatment compliance during the patient's actual treatment period, defined as the actual days of treatment exposure divided by the patient's total treatment exposure duration. The actual days of treatment exposure is the sum of the durations of individual applications. For each OmniPod applicatiom, the duration is calculated as: date of next application – date of current application, or 3, whichever is smaller. For the last application, it is equal to 3.

.

9 Efficacy

A Bayesian adaptive design is adopted to address the following objectives:

1. Demonstrate superiority of the pooled middle and high doses over placebo
2. Evaluate the dose-response relationship

The adaptive design includes two interim analyses, conducted after approximately 9 and 12 patients per arm enrolling in the study, respectively. Each interim analysis provides opportunities for stopping the trial early due to efficacy success utilizing the group sequential strategy or due to efficacy futility through Bayesian predictive probability of success. Furthermore, it allows the low dose to be dropped following the interim look if it shows a much lower response rate compared with the middle and high doses.

Each interim analysis will result in one of the following outcomes:

1. Continue the trial with all arms
2. Continue the trial without the low dose
3. Declare early success, but continue the trial without placebo
4. Declare early success and stop the trial (both placebo and low dose dropped)
5. Stop the trial for futility

When the primary efficacy analysis is not successful at the interim analysis, and if the ovulation rate in the low-dose arm is sufficiently low, satisfying

$$R_{\text{low}} < (R_{\text{high+middle}}) - 0.2, \qquad (1)$$

then the trial will continue without the low dose, resulting in outcome #2 above. Note that R_{low} and $R_{\text{high+middle}}$ in the inequality (1) above are the observed ovulation rates in the low dose arm and the combined middle- and high-dose arms, respectively. However, if inequality (1) is not satisfied, then the trial will continue with all arms (outcome #1).

Conversely, when success on the primary efficacy analysis is declared at an interim analysis, if the response rate in the low-dose arm meets the condition in inequality (1), then the trial is completed at this point (outcome #4) with dose selection performed using the ED80 rule (see below).

In the case the primary efficacy analysis is successful, but the response rate in the low-dose arm is within 20% from the combined low- and high-dose success rate, the trial is not stopped immediately. Rather, the trial will continue to enroll additional patients into the low-, middle- and high-dose arms to further assess dose responses (outcome #3).

Finally, the trial may terminate (outcome #5) as a result of a low predictive probability of success (<0.05) given the data collected at the interim analysis.

At the end of the study, for a trial that achieves efficacy success, the dose to be selected is the ED80 dose, defined as the lowest dose that retains at least 80% of the largest treatment effect, i.e.

$$(R_{\text{ED80}} - R_{\text{placebo}}) \geq (R_{\text{maximum}} - R_{\text{placebo}}) \times 0.8, \qquad (2)$$

where R_{maximum} is the observed ovulation rate in the most effective dose arm, R_{placebo} is the observed ovulation rate in the placebo arm, and R_{ED80} is the observed ovulation rate in the selected dose arm.

A flow diagram is presented in Figure 1 of Section 15.2.1 to illustrate the decision tree for the two-interim adaptive design. Further details, such as the criteria for efficacy success using the group sequential method, the Bayesian predictive probability of future success to determine futility, the rationales behind the 20% threshold to drop low dose, the ED80 dose-selection rule and dose-response evaluations are described in Section 9.2.1.1. Simulation results to assess operating characteristics of the design are given in Section 15.2.

9.1 General Considerations

Sites from all countries/regions (US, Canada) will be pooled for all analyses. No stratification by site will be made in the analysis.

Unless otherwise noted, data will be summarized in tabular format by treatment group using summary tables and data listings. All study data documented on the eCRFs will be included in the study data listings.

All efficacy analyses will be conducted for the FAS. The three placebo groups with different dose settings will be pooled as a single placebo group for all efficacy summaries and analyses.

Continuous variables will be described with the number of non-missing values, mean, standard deviation (SD), median, and minimum/maximum values. Categorical variables will be described with the number and percentage of subjects with each level. Missing values will not be included in the calculation of percentages. All individual subject data will be listed.
The software used for all summary statistics and statistical analyses will be SAS Version 9.2 or later.

9.2 Primary Endpoint(s)

9.2.1 Primary Variable Analysis

The primary efficacy endpoint is the ovulation rate calculated as the proportion of subjects who had at least one post-baseline P4 level ≥ 6 ng/mL or pregnancy confirmed by at least 2 positive serum β-hCG tests or detection of gestational sac on transvaginal ultrasound (TVUS). The denominator will be based on all subjects in the FAS. Subjects who do not have a confirmed pregnancy, or at least one P4 level ≥ 6 ng/mL due to the P4 levels never reaching this threshold, missing data, early withdrawal, or any other reason will be counted as not having ovulation (treatment failure).

The primary efficacy analysis will compare the pooled ovulation rate of LutrePulse OmniPod SC 20 µg and 15 µg to placebo. Let $P_{high+middle}$ and $P_{placebo}$ be the underlying ovulation rates in the combined middle- and high-dose arms and the placebo arm, respectively. The hypotheses for the one-sided statistical testing are:

$$H_0: P_{high+middle} \leq \ P_{placebo},$$

vs.

$$H_0: P_{high+middle} > \ P_{placebo.}$$

The hypotheses are tested using a stratified one-sided Fisher's exact test (Jung et al., 2014) in the FAS including the randomization scheme as the stratification factor. The SAS PROC FREQ with permutation-based Cochran-Mantel-Haenszel test with exact p-value derived from all possible permutations will be employed for the calculation of p-value for the stratified Fisher's exact test.

Jung has shown that the stratified Fisher's exact test is equivalent to the exact stratified Mantel-Haenszel test when exact p-value is derived from all possible permutations.

9.2.1.1 Bayesian Adaptive Design Rules

This section provides details for the decision rules to be implemented in the Bayesian adaptive study design.

Success criterion using group sequential strategy

To account for the multiplicities introduced by early success interim analyses, Pocock boundaries are used. Thus, the success threshold at each analysis corresponds to a nominal p-value < 0.0131 to maintain an overall one-sided type I error rate of 0.025.

The simulation results in Figure 13 for Scenario 7 in Section 15.2.3, as well as the justification provided in Section 15.2.4, show that the type I error is well controlled under the null hypothesis using the success boundary.

Futility criterion

Given the data collected at each interim time point, we compute the Bayesian predictive probability that the primary analysis will meet the success threshold once all arms have n = 15 subjects. If the predictive probability is less than 0.05, then the trial will be declared as a failure and terminate at this point due to futility.

The Bayesian predictive probability is calculated in the following fashion. First, for each arm d, we determine the number of future subjects required to complete the study, m_d. Then, the number of future responders occurring in arm d, Z_d, is generated from a Beta-Binomial distribution:

$$(Z_d | n_{d,} x_d) \sim BetaBinomial\ (m_d, a + x_d, b + n_d - x_d),$$

where $n_{d,} x_d$ are the numbers of subjects and responders in arm *d* observed at interim analysis, and a non-informative prior Beta distribution with parameters *a* = 0.5 and *b* = 0.5 is used independently for each arm.

A large number of predicted values are sampled from the Beta-Binomial distribution for each arm, resulting in 1000 "complete" datasets. The predictive probability of eventual trial success is then the proportion of these complete datasets for which the primary analysis success threshold is met (p-value < 0.0131).

Dropping the low dose

If the observed response rate for the low dose meets the condition in inequality (2), then low dose will be dropped from additional enrolment, should the trial need to continue with the high dose, middle dose and placebo to assess the primary efficacy endpoint at the next analysis timing.

The remaining sample size that had been planned for the low dose will be reallocated among the remaining arms. Thus the sample size for the placebo, middle, and high doses may either be 9, 12, 15, 16, or 17 subjects, depending on whether early success is declared, and whether low dose is dropped at the first or second interim analysis.

The 20% threshold to drop low dose is chosen based on the ED80 dose-selection rule, as the maximum of the most effective treatment effect can reach 100% (with the success rate in the placebo group expected to be close to 0). It ensures that when the threshold is met, the low dose does not qualify as the ED80 dose. Furthermore, we present in Table 3 below the Bayesian predictive probabilities of selecting low dose as ED80, should it proceed to a later analysis, for a range of borderline interim analysis observations where the 20% threshold is just met. A non-informative prior Beta (0.5, 0.5) is used for all treatment arms and 10000 predictions were generated. The results show that such probabilities are small: <10% in all cases and <5% in all but three cases. Therefore, 20% is a justifiable choice for the threshold to drop low dose early.

Table 3 Bayesian Predicted Probability of Selecting Low Dose as ED80 at a Later Analysis when Low Dose was Dropped Early

N per arm at interim analysis	N per arm at later analysis	Observed n (%) of successes at interim analysis			Diff* (%)	Probability of selecting low dose as ED80 if continue with low dose
		High-dose	Mid-dose	Low-dose		
9	12	6 (67.8)	4 (44.4)	3 (33.3)	-22.2	0.0207
		5 (55.6)	5 (55.6)	3 (33.3)	-22.2	0.0735
		5 (55.6)	4 (44.4)	2 (22.2)	-27.8	0.0059
		5 (55.6)	3 (33.3)	2 (22.2)	-22.2	0.0062
		4 (44.4)	4 (44.4)	2 (22.2)	-22.2	0.0159
	15	6 (67.8)	4 (44.4)	3 (33.3)	-22.2	0.0304
		5 (55.6)	5 (55.6)	3 (33.3)	-22.2	0.0672
		5 (55.6)	4 (44.4)	2 (22.2)	-27.8	0.0136
		5 (55.6)	3 (33.3)	2 (22.2)	-22.2	0.0201
		4 (44.4)	4 (44.4)	2 (22.2)	-22.2	0.0555
12	15	8 (75.0)	6 (50.0)	4 (33.3)	-25.0	0.0015
		7 (58.3)	6 (50.0)	4 (33.3)	-20.9	0.0231
		7 (58.3)	5 (41.7)	3 (25.0)	-25.0	0.0012
		7 (58.3)	4 (33.3)	3 (25.0)	-20.8	0.0014
		6 (50.0)	5 (41.7)	3 (25.0)	-20.8	0.0211
		6 (50.0)	4 (33.3)	2 (16.7)	-25.0	0.0011
		5 (41.7)	5 (41.7)	2 (16.7)	-25.0	0.0059
		5 (41.7)	4 (33.3)	2 (16.7)	-20.8	0.0173

Diff* is the difference in observed success rates between the low-dose group vs. the pooled high- and mid-dose groups

Dose-response evaluation and dose-selection

For trials that meet the primary analysis success threshold, we will subsequently select the ED80 dose, the lowest effective dose satisfying inequality (2). The ED80 dose will be labeled as such in the primary efficacy analysis table.

- As noted earlier, when the primary efficacy success is already declared at an interim analysis, but the success rate in the low-dose arm is not sufficiently low compared to the combined low- and high-dose success rate, the trial will continue to enroll more patients in the three dose arms only. Although no additional patients will be randomized to placebo, the success rate estimated from the enrolled placebo patients will be used in inequality (2) to determine the ED80 for the study.

9.2.1.2 Sensitivity Analyses

The primary comparison between the pooled middle- and high-dose arms vs. the placebo arm for the primary variable will be conducted on the following analysis sets as sensitivity analyses:

- PPS
- mITT
- ITT

9.3 Secondary Endpoint(s)

9.3.1 Progesterone Level

The proportion of subjects with at least 1 post-baseline progesterone (P4) level ≥10 ng/mL will be compared between the pooled middle- and high-dose arms vs. the placebo arm using the same method to be applied to the primary variable analysis in Section 9.2.1.

9.3.2 Pregnancy

Clinical and biochemical pregnancy rates in LutrePulse OmniPod SC 20µg and 15µggroups will be compared between the pooled middle- and high-dose arms vs. the placebo arm using the same method to be applied to the primary variable analysis in Section 9.2.1.

9.3.3 LH surge detection

The positive LH surge rate in LutrePulse OmniPod SC 20µg and 15µg groups will be compared between the pooled middle- and high-dose arms vs. the placebo arm using the same method to be applied to the primary variable analysis in Section 9.2.1.

9.3.4 Ovarian Follicular Development

Number of follicles with a mean diameter ≥ 14 mm and number of dominant follicles with a mean diameter ≥ 18 mm, collected from Days 10, 11, until LH surge or Day 21, in LutrePulse OmniPod SC 20µg, 15µg, and 10µg groups will be compared to those in the placebo group at each time point (analysis visit windows defined in Section 15.4) using Wilcoxon's rank sum test stratified on the randomization scheme (van Elteren, 1960). SAS PROC FREQ with modified ridit score (MODRIDIT option) will be used for the stratified Wilcoxon rank sum test. Summary statistics including median will be presented.

9.3.5 Luteal Phase Support

The maximum P4 level and the mean P4 level from Days 19, 21, 23, 25, and 27 values in LutrePulse OmniPod SC 20µg, 15µg, and 10µg groups will be compared to those in the placebo group using an analysis of covariance (ANCOVA) model including the randomization scheme and treatment group as factors and baseline value as covariate.

Whenever the ANCOVA is performed, the least squares mean difference of the treatment groups, and 95% confidence intervals and p-values for the LutrePulse OmniPod SC 20µg, 15µg, and 10µg groups versus the placebo group based on the fitted linear model will be reported. Type III sum-of-squares for the least-squares means will be used for the statistical comparisons. PROC GLM will be used for ANCOVA in SAS.

The baseline is the value obtained at the last assessment prior to the first dose of IMP.

9.4 Pharmacodynamic Parameters

The mean FSH and LH on Day 10 calculated from all post-dose time points (i.e. 5, 15, 30, 60, and 90 minutes after the first dose), as well as the mean change from Day 10 pre-treatment (prior to first dose) in LutrePulse OmniPod SC 20µg, 15µg, and 10µg groups will be compared to those in the placebo group using an ANCOVA model including the randomization scheme and treatment group as factors and Day 10 pre-treatment value as covariate.

The E2 level on Day 10 (measured only at pre-treatment) as well as the change from baseline (Day 1 pre-treatment) in LutrePulse OmniPod SC 20µg, 15µg, and 10µg groups will be compared to those in the placebo group using an ANCOVA model including the randomization scheme and treatment group as factors and baseline value as covariate.

The change in the maximum post-treatment value of FSH and LH on Day 1 and Day 10 from the pre-treatment value taken on Day 1 and Day 10, respectively, will be presented with descriptive statistics by treatment group.

The FSH and LH levels on Day 1 and Day 10 will be presented graphically by treatment group (i.e. scatter plots over all time points: Baseline, 5, 15, 30, 60, and 90 minutes after the first dose).

10 Safety

10.1 General Considerations

Safety parameters will be evaluated for the Safety Analysis data set.

The three placebo groups with different dose settings will be pooled as a single placebo group for all safety summaries and analyses.

10.2 Adverse Events

An AE is any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have to have a causal relationship with this treatment.

All Adverse Events (AEs) and unanticipated adverse device effects will be listed, documenting all information as on the eCRF. Verbatim terms of AEs will be coded to the MedDRA preferred terms (PT) and related system organ classes (SOC) using the Medical Dictionary for Regulatory Activities (MedDRA). The latest MedDRA version available at the time of interim analysis or final analysis will be applied.

A treatment emergent adverse event (TEAE) will be any adverse event occurring after start of IMP and within 5 days after the last application, or a pre-treatment adverse event or pre-existing medical condition that worsens in intensity after start of IMP and within 5 days after the last application. All adverse events occurring after 5 days from the last application will be categorized as post-treatment adverse events.

Note: Collection of AEs on the eCRF will begin after the informed consent form (ICF) is obtained from the patient. Any AE prior to the signed ICF will be recorded in the medical history.

Written narratives will be issued for all serious AEs (including deaths) and AEs leading to withdrawal.

10.2.1 Partial/Missing Dates for Adverse Events

For any adverse event that has a missing or incomplete start date, a query will be issued to capture the full date. If only a partial date is reported for the start of an AE, a complete date will be estimated using the following algorithm:

- Only the year is reported (i.e., missing the month and day): The date of the first dose of investigation product will be used as the start of the AE.

- The year and month are reported (i.e., missing the day): The first day of the month will be imputed or the date of the first dose, whichever is later.

The month and day are reported (i.e., missing the year): Only situations where the year is self-evident will be allowed.

No imputation will be applied to the incomplete AE end date. The rationale is that one should have a start date for all subject AEs but may not have end dates for AEs that resolve after subject is off study or AEs that are continuing.

10.2.2 Overview of Treatment-Emergent Adverse Events

An AE overview summary table will be prepared including the number of subjects reporting an AE, the percentage of subjects (%) with an AE, and the number of events (E) reported, for the following categories:

- Treatment-emergent adverse events
- Deaths
- Serious adverse events
- Adverse events leading to withdrawal
- Severe adverse events
- Adverse drug reactions

10.2.3 Incidence of Adverse Events

Treatment-emergent adverse events will be summarised in a Table by SOC and PT for MedDRA. The Table will display the total number of subjects reporting an AE, the Treatment-emergent adverse events (TEAEs) will be tabulated by MedDRA SOC and PT by treatment group. The total number of subjects reporting a TEAE, the percentage of subjects (%) with a TEAE, and the number of TEAEs reported will be presented. The order of SOCs presented in tables will be in alphabetic order of SOCs according to MedDRA. Within each SOC, the preferred terms will be shown in decreasing frequency of occurrence.

Summary tables will be prepared for:

- All adverse events
- Adverse events by causality (related/unrelated)
- Adverse events leading to death
- Adverse events by intensity
- Serious adverse events
- Adverse events leading to withdrawal

Note: Patients who have multiple events in the same SOC and/or preferred term will be counted only once at each level of summation (overall, by SOC, and by preferred term) in the tables. For summaries of AEs by severity, only the highest severity of AE will be counted at each level of summation (overall, by SOC, and by preferred term) in the tables. For summaries of related AEs, patients with more than one related AE will be counted only once at each level of summation (overall, by SOC, and by preferred term) in the tables. Treatment-related means 'Reasonable possibility' of relationship to investigational product according to investigator. If a subject has an AE with unknown intensity, then the subject is counted in the category of 'Severe'. If a subject has an AE with unknown relationship to IMP then the subject is counted in the category of 'Reasonable possibility'.

Supporting data listings will be provided for:

- All adverse events sorted by centre and subject no.
- All adverse events sorted by MedDRA Preferred Term
- Serious adverse events
- Adverse events leading to death
- Adverse events leading to withdrawal

10.3 Safety Laboratory Variables

Baseline for all laboratory analyses will be the values obtained at the last assessment prior to the first dose of the investigational medicinal product (IMP). Treatment-emergent laboratory data will include tests completed after the first dose of IMP through the residual time of drug effect. End of trial will include the last post-baseline observation during the trial.

Laboratory variables will be grouped under "Haematology", "Clinical Chemistry" or "Urinalysis"

10.3.1 Summary Statistics

Mean change and mean percentage (%) change from baseline at end of trial will be presented for each laboratory variable. In addition, descriptive statistics, i.e., the number of subjects with data, mean (standard deviation), median, minimum, and maximum values, will be presented for observed values and change from baseline at each time-point for each laboratory variable.

All laboratory parameters, values, units, normal ranges, flags collected (normal/abnormal) in the clinical database will be included in by-patient listings for further medical review.

10.3.2 Laboratory Variable Changes Relative to Normal Range

Changes relative to normal ranges are presented with shift tables with total number of subjects, and number and % of subjects who experienced a shift by treatment period. The following categories for shift tables are defined:

- Low: Values which are below the lower reference range limit;
- Normal: Values which are within the lower and upper reference range;
- High: Values which are above the upper reference range limit.
- Absent: No value for measured variable (for urinalysis only)
- Present: Any value obtained for measured variable (for urinalysis only)

For all haematology and clinical chemistry variables, shift tables will be prepared to compare baseline values to the worst in-treatment value. More specifically, for haematology and clinical chemistry, tables presenting the changes from *Low* or *Normal* to *High* and from *High* or *Normal* to

Low will be provided. For urinalysis variables, shift tables will summarise the number (%) of subjects who had “absent” values at baseline and “present” values during the treatment period.

10.3.3 Markedly Abnormal Changes

A summary table will be prepared that displays for each laboratory variable the number and percentage of subjects in each treatment group with normal baseline values who had at least one pre-specified markedly abnormal value anytime during the treatment period. Pre-specified markedly abnormal criteria for laboratory tests are given in Section 15.5.

10.3.4 Data Listings

Data listings will be prepared by treatment group and centre for all subjects with any abnormal laboratory value at any time-point (including screening, baseline).

10.3.5 Urinalysis

Incidence of changes in urinalysis will be summarised by treatment group. A summary table with number of subjects with change from *Absent* at baseline to *Present* during trial will be summarised.

10.4 Vital Signs and ECGs

10.4.1 Vital Signs

Baseline for all vital signs analyses will be the values obtained at the last assessment prior to the first dose of IMP. Treatment-emergent vital signs data will include tests completed after the first dose of IMP through the time of residual drug effect. End of trial will include the last post-baseline observation during the trial.

10.4.1.1 Summary Statistics

Mean change and mean percentage (%) change from baseline at end of trial will be presented for each vital signs variable. In addition, descriptive statistics, i.e., the number of subjects with data, mean (standard deviation), median, minimum, and maximum values, will be presented for observed values and change from baseline at each time-point for each vital signs variable.

10.4.1.2 Markedly Abnormal Changes

Summary tables will be prepared displaying the number and percentage of subjects with normal baselines who had one or more pre-specified markedly abnormal treatment-emergent values, according to the definition in Section 15.5.

10.4.1.3 Data Listings

Data listings will be prepared by treatment group and centre for all subjects with any abnormal vital signs value at any time-point (including screening, baseline).

10.4.2 ECGs

12-lead ECG will be collected for subjects who report chest pain or are symptomatic of chest pain any time during the study (unscheduled visit).

Treatment-emergent ECG data will include evaluations completed after the first dose of IMP through the time of residual drug effect.

Data listings will be prepared by treatment group and centre for all subjects with 12-lead ECG value at any time-points during the study.

10.5 Other Safety Variables

10.5.1 Pregnancy Test

Pregnancy Test results will be presented in by-patient listings for further medical review. Any positive test results for pregnancy reported during study will be noted in the clinical study report.

10.5.2 Serological Test

Serological test results for hepatitis B, hepatitis C, and HIV I and II will be presented in by-patient listings.

10.5.3 Ovarian Hyperstimulation Syndrome (OHSS)

The frequency and severity of ovarian hyperstimulation syndrome will be presented in by-patient listings and summary tables as for AEs.

11 Interim analyses

Two interim analyses are planned after approximately 9 and 12 patients per arm complete the study. The details of the interim analysis strategy are described in Section 9.2.

An independent data monitoring committee consisting of two unblinded clinicians and one unblinded statistician will be set up to review the interim analysis study results and make recommendations for the direction of the study, in accordance with the adaptive design decision-making rules specified in Section 9.2.1. In addition, independent statistician and programmer will be identified to generate the data analysis output. Individuals involved in the first and second interim analysis of the study will not be involved in the conduct of the study afterwards; individual patient identification will not be released to anyone who is directly involved in the conduct of the study.

The interim analysis process, the measures used to protect the blind and the integrity of the study, the communication plan, and the confidentiality agreement will be described in a separate document.

12 Deviations from protocol analysis

13 References

[1] International Federation of Pharmaceutical Manufacturers and Associations. Medical Dictionary for Regulatory Activities (MedDRA). Version 10.1. Reston, Virginia, USA; 2008.

[2] WHO Collaborating Center for International Drug Monitoring. WHO Drug Dictionary. July 2008 edition. Uppsala, Sweden; 2008.

[3] SAS Institute Inc. SAS Version 9.1. Cary, NC, USA; 2002-2003.

[4] Kong, L, Koch, G, Liu, T, Wang, H. Performance of some multiple testing procedures to compare three doses of a test drug and placebo. Pharmaceutical Statistics. 2005;4,25-35.

[5] Jung, S.H. Stratified Fisher's exact test and its sample size calculation. Biometrical J. 2014;56. doi: 10.1002/bimj.201300048.

[6] van Elteren, P. H. (1960). "On the combination of independent two-sample tests of Wilcoxon," Bulletin of the International Statistical Institute, 37, 351-361.

14 Tables, Listings and Figures

The document with tables, figures and listings (TLF) shells will be presented in a separate document.

15 Appendix

15.1 Schedule of Events

Procedure	Screening (Day -60 to Day -3)	Treatment Day 1	Treatment Day 10	Follicle ≥ 18 mm and LH surge	Days 19, 21, 23, 25, and 27	14 (+4 days) post LH surge and 3 days later	2 to 4 weeks from 2nd positive serum ß-hCG test	End-of-Study Visit[a]
Informed consent	X							
Inclusion/exclusion	X							
Medical/gynecologic history	X							
Physical examination	X							X
Gynecological exam	X							X
Vital signs	X							X
Safety laboratory tests	X							X
Hepatitis B, C, & HIV screen	X							
Hormones: PRL, TSH	X							
TVUS	X		X –	X[b]			X[c]	
Semen analysis	X							
Progestin challenge test[d]	X							
Serum pregnancy test	X					X		
Concomitant medications	X	X	X	X	X	X	X	X
Urine pregnancy test		X						X[e]
Blood samples for E_2	X	X	X	X				
Hormones: LH, FSH	X	X[f]	X[f]					
Blood samples for P_4	X	X			X			
Randomization		X[g]						
OmniPod application[h]		X						
Clearblue testing[i]				X				
Begin sex intercourse[j]				X				
Adverse events		X	X	X	X	X	X	X
Follow up phone call[k]								X

a. The end-of study visit will be performed for all subjects who discontinue early or the day of 2nd negative serum pregnancy test or documented fetal heart movement for subjects with clinical pregnancy.

b. Follicular size monitoring

c. If fetal heartbeat is not documented, the subject may return in 7 days for another TVUS.

d. Subjects with uterine bleeding within 5 days of the last dose of medroxyprogesterone acetate will be considered screen failures.

e. A urine pregnancy test should be performed at the End-of-Study visit for those who discontinue the study early and have no positive pregnancy results during the study.

f. On Day 1 (pre-treatment baseline) and Day 10 (post-treatment) LH and FSH samples will be collected at 0 hour (first sample taken prior to dosing) and 5, 15, 30, 60, and 90 minutes after the first dose is administered.
g. Randomization will occur 3 days before Treatment Day 1.
h. The first OmniPod applied will remain on the subject until Day 4, when the second OmniPod will be applied at approximately the same time that the first was applied; each OmniPod will be worn for 3 full days.
i. When follicles ≥ 14 mm are detected on TVUS, Clearblue tests are dispensed to the subject to be administered at home each day at approximately the same time between 8 AM and 11 AM until a positive test result occurs.
j. Intercourse to begin at confirmed Clearblue LH surge
k. For subjects who achieve biochemical pregnancy – collect ongoing pregnancy rate and birth data

15.2 More on the Bayesian Adaptive Design

15.2.1 Design Flow Diagram

Figure 1 illustrates the possible decisions that may be made at each interim analysis. There are 15 possible pathways to reach a trial decision, annotated by the circled numbers on the plot.

For example, the paths numbered 1, 3, 4, 6, 8, 9, 10, 12, and 14 represent paths ending in success, and are shaded by green. Of those, paths numbered 1, 3, 4, 6, 8, and 9 declare early success at an interim, with paths 8 and 9 continuing enrolment to the final analysis to assess dose response.

Paths numbered 2, 5, and 7 result in early futility at an interim analysis and paths 11, 13, and 15 fail at the final analysis. All of these are shaded in pink.

The boxes shaded blue indicate trials for which the low dose is dropped at an interim analysis, but enrolment continues on the placebo, middle, and high dose arms. Orange boxes indicate that enrolment continues with all arms.

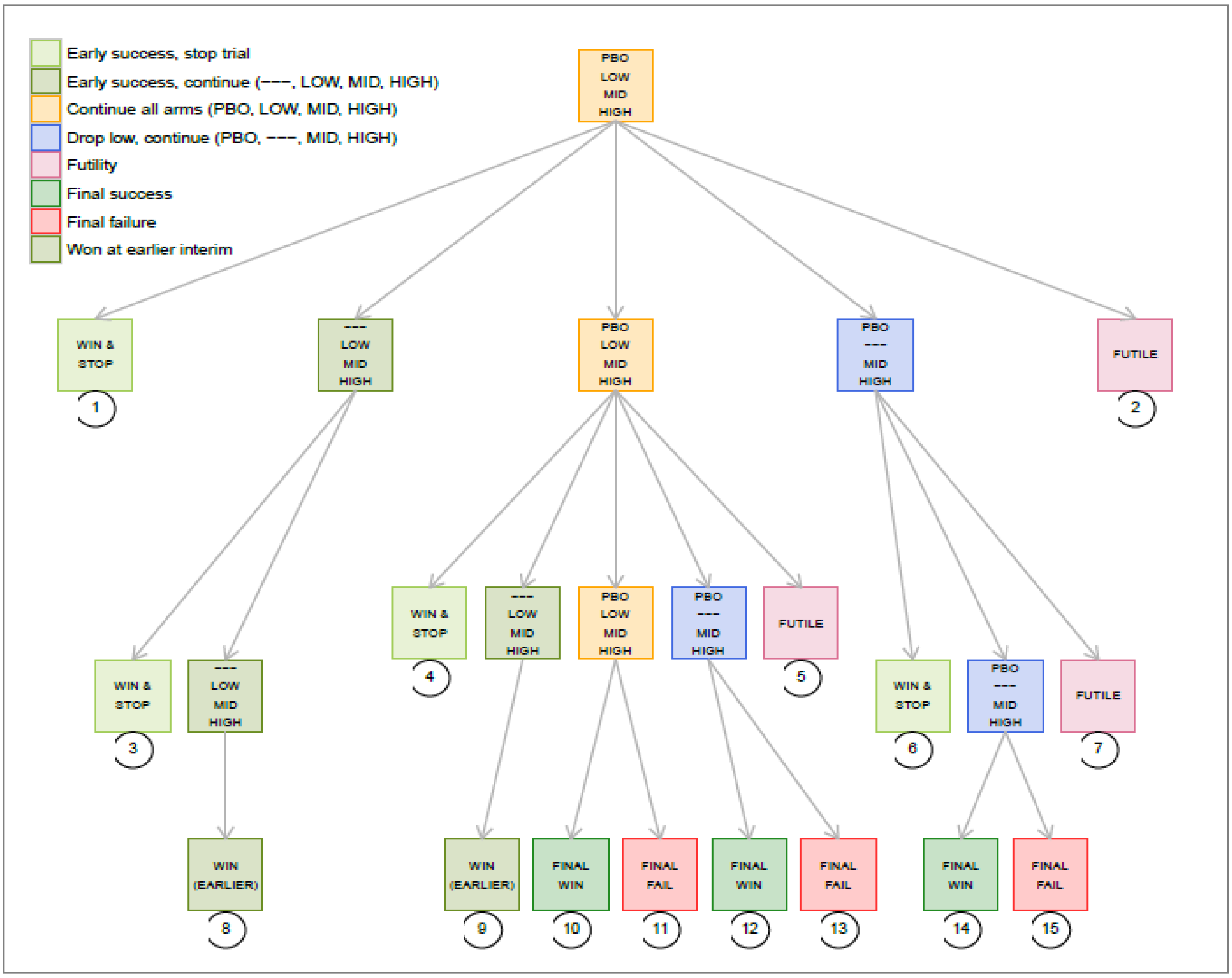


Figure 1 Flowchart of possible trial outcomes

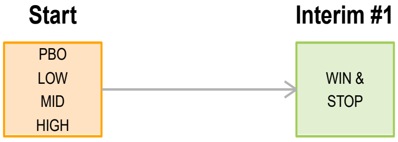
15.2.2 Example Trials

Example Trial 1

A schematic of the first example trial is shown in Figure 2. The trial pathway is indicated by the colored boxes, with the observed data shown below.

The first interim occurs after 9 subjects per arm have outcomes. In this example, there were 0 responders on the placebo arm and 10 responders pooled across the middle and high doses. The p-value for the pooled arms versus placebo is 0.0052, which crosses the early success threshold. Furthermore, there was only 1 responder on the low dose, so that the difference in observed response rates is 0.667 (0.778 – 0.111). Since this difference is larger than 0.20, the low dose is determined not to be the best arm and is dropped from the study.

Since both the placebo and the low dose have been dropped, the trial stops accrual. The high dose has the highest response rate, and is also the ED80 (shaded in blue in the data table) since the middle dose does not achieve at least 80% of the effect seen on the high dose.



	n	y	y/n
placebo	9	0	0
low	9	1	0.111
mid	9	3	0.333
high	9	7	0.778
pooled	18	10	0.556

p-value: 0.0052

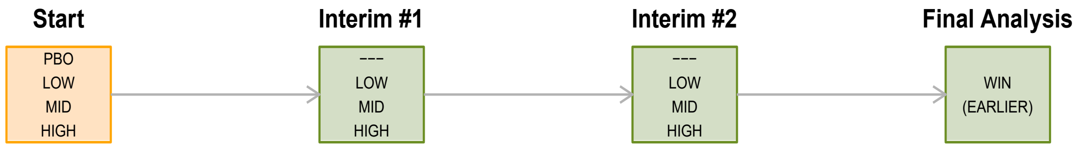
Figure 2 Simulation results from Example Trial 1

Example Trial 2

The second example trial is shown in Figure 3. At the first interim, this trial is similar to the first, except on the low dose. In this trial, early success is declared, but since the low dose response rate is equal to the pooled response rate, the low dose will not be dropped. Thus, since superiority to control has been demonstrated, the placebo arm is dropped, but accrual continues on the investigational arms.

The second interim occurs when each remaining arm has $n = 12$ subjects. Since no additional subjects have been recruited to the placebo arm, this row is greyed out in the data table. Again, the low dose performs similarly to the pooled middle and high dose arms. Thus, the trial continues to the final analysis with the low, middle, and high dose arms.

At the final analysis, the high dose has the highest observed response rate with 11 responders out of 15 subjects (73.3%). The low dose, with 9 responders (60.0%) is selected as the ED80.



	n	y	y/n
placebo	9	0	0
low	9	5	0.556
mid	9	3	0.333
high	9	7	0.778
pooled	18	10	0.556

	n	y	y/n
placebo	9	0	0
low	12	8	0.667
mid	12	5	0.417
high	12	10	0.833
pooled	24	15	0.625

	n	y	y/n
placebo	9	0	0
low	15	9	0.6
mid	15	8	0.533
high	15	11	0.733
pooled	30	19	0.633

p−value: 0.0052

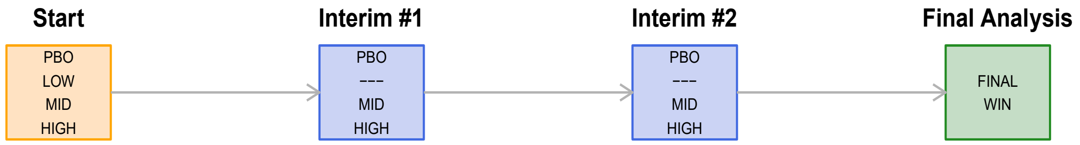
Figure 3 Simulation results from Example Trial 2

Example Trial 3

In the third example (Figure 4), the pooled response rate at the first interim is 44.4% (8 total responders, with 3 on the middle dose and 5 on the high dose) versus 0% on placebo. The p-value is 0.0197 which does not meet the early success threshold. The predictive probability of eventual trial success is 86.1%, so the trial does not stop for futility. The low dose response rate is only 22.2%, which is more than 20% lower than the pooled response rate. Thus, the low dose is dropped, and accrual continues on the placebo, middle, and high dose arms.

The second interim occurs when the remaining arms each have 12 subjects. Since the last interim, there is one new responder on the high dose, and no new responders on the placebo or middle dose arms. The p-value for the pooled versus placebo is 0.0139, which is not sufficient to declare early success. The predictive probability is still above the futility threshold, so the trial continues to the final analysis.

Since the low dose was dropped at the first interim with only 9 subjects, the remaining 6 slots that would have been allocated to the low dose have been redistributed to the remaining arms. Thus, each arm at the final analysis has $n = 17$ subjects rather than $n = 15$. The p-value for the Fisher Exact test is 0.0011, so the trial meets success at the final analysis. The high dose is selected as the ED80.



	n	y	y/n
placebo	9	0	0
low	9	2	0.222
mid	9	3	0.333
high	9	5	0.556
pooled	18	8	0.444

Predictive probability: 0.861
p−value: 0.0197

	n	y	y/n
placebo	12	0	0
low	9	2	0.222
mid	12	3	0.25
high	12	6	0.5
pooled	24	9	0.375

Predictive probability: 0.85
p−value: 0.0139

	n	y	y/n
placebo	17	0	0
low	9	2	0.222
mid	17	4	0.235
high	17	10	0.588
pooled	34	14	0.412

p−value: 0.0011

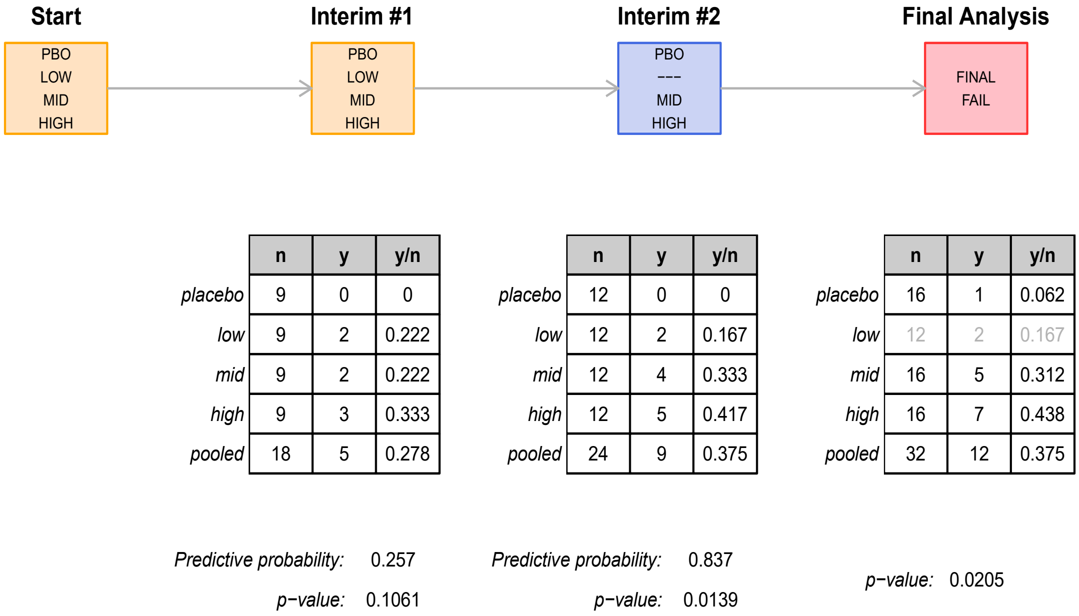
Figure 4 Simulation results from Example Trial 3

Example Trial 4

At the first interim of example trial 4 (Figure 5), the response rate is a modest 20-30% on the investigational arms. No responses were observed on the placebo. The trial does not declare early success, but has high enough predictive probability to continue. Additionally, the low dose is not sufficiently lower than the pooled arms to be dropped. Thus, the trial continues enrolling subjects to all arms.

At the second interim, there are $n = 12$ subjects allocated to each arm. No additional responses were observed on the low dose, which had only 2 responders (16.7%). The response rate for the pooled middle and high doses is now 41.7%. Neither the early success nor futility criteria were met. However, the low dose response rate is no longer within 20% of the pooled, so it will be dropped. The 3 remaining slots for the low dose are redistributed so that each remaining arm will receive one additional subject.

The final analysis occurs with $n = 16$ subjects each on the placebo, middle, and high dose arms. The p-value for the primary analysis is 0.0205, resulting in trial failure.



	n	y	y/n
placebo	9	0	0
low	9	2	0.222
mid	9	2	0.222
high	9	3	0.333
pooled	18	5	0.278

Predictive probability: 0.257
p−value: 0.1061

	n	y	y/n
placebo	12	0	0
low	12	2	0.167
mid	12	4	0.333
high	12	5	0.417
pooled	24	9	0.375

Predictive probability: 0.837
p−value: 0.0139

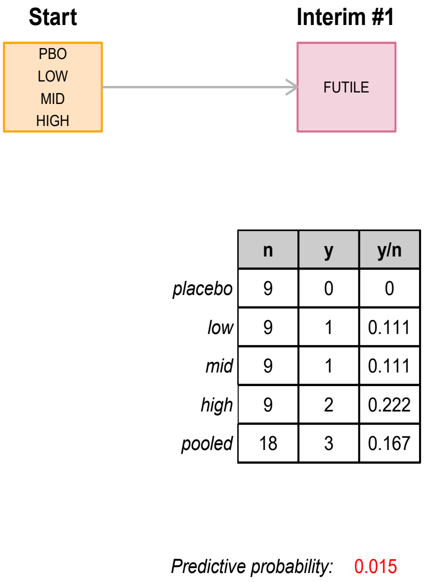
	n	y	y/n
placebo	16	1	0.062
low	12	2	0.167
mid	16	5	0.312
high	16	7	0.438
pooled	32	12	0.375

p−value: 0.0205

Figure 5 Simulation results from Example Trial 4

Example Trial 5

Figure 6 shows the fifth example trial. In this example, only 3 responses combined were observed on the middle and high dose arms, for a rate of 16.7%. The predictive probability of eventual trial success is 0.015, resulting in early futility.



	n	y	y/n
placebo	9	0	0
low	9	1	0.111
mid	9	1	0.111
high	9	2	0.222
pooled	18	3	0.167

Predictive probability: 0.015

Figure 6 Simulation results from Example Trial 5

15.2.3 Simulation Studies

The operating characteristics of this trial were determined through trial simulation. We hypothesized several scenarios for the underlying performance of the control and investigational arms, and simulated the entire trial multiple times under each scenario. In each virtual trial, the interim analyses were conducted according to the pre-specified rules, and results were tracked for each trial, including whether the trial was successful, whether the low dose was dropped, the sample size per arm, etc. This section describes the algorithm and the parameters that were used to simulate subject-level data for the virtual trials. These assumptions are used only in the generation of virtual subjects for the simulations, and are not part of the analysis of subject data (virtual or real).

We simulate individual subject responses from a Bernoulli distribution where the true response rates per arm are given in the following table for different scenarios. We assume that the placebo rate is 1% in all scenarios. We consider a null scenario in which all arms are equal to placebo, and scenarios that vary the treatment effect. In some scenarios, the low dose is similar to the middle and high doses while in other scenarios, it is considerably lower.

Table 4 Response rate profiles

	Placebo	Low	Middle	High
3	0.01	0.01	0.01	0.01
4	0.01	0.10	0.20	0.40
5	0.01	0.10	0.40	0.40
6	0.01	0.10	0.40	0.50
7	0.01	0.20	0.60	0.60
8	0.01	0.50	0.65	0.75
9	0.01	0.60	0.65	0.70

Operating Characteristics

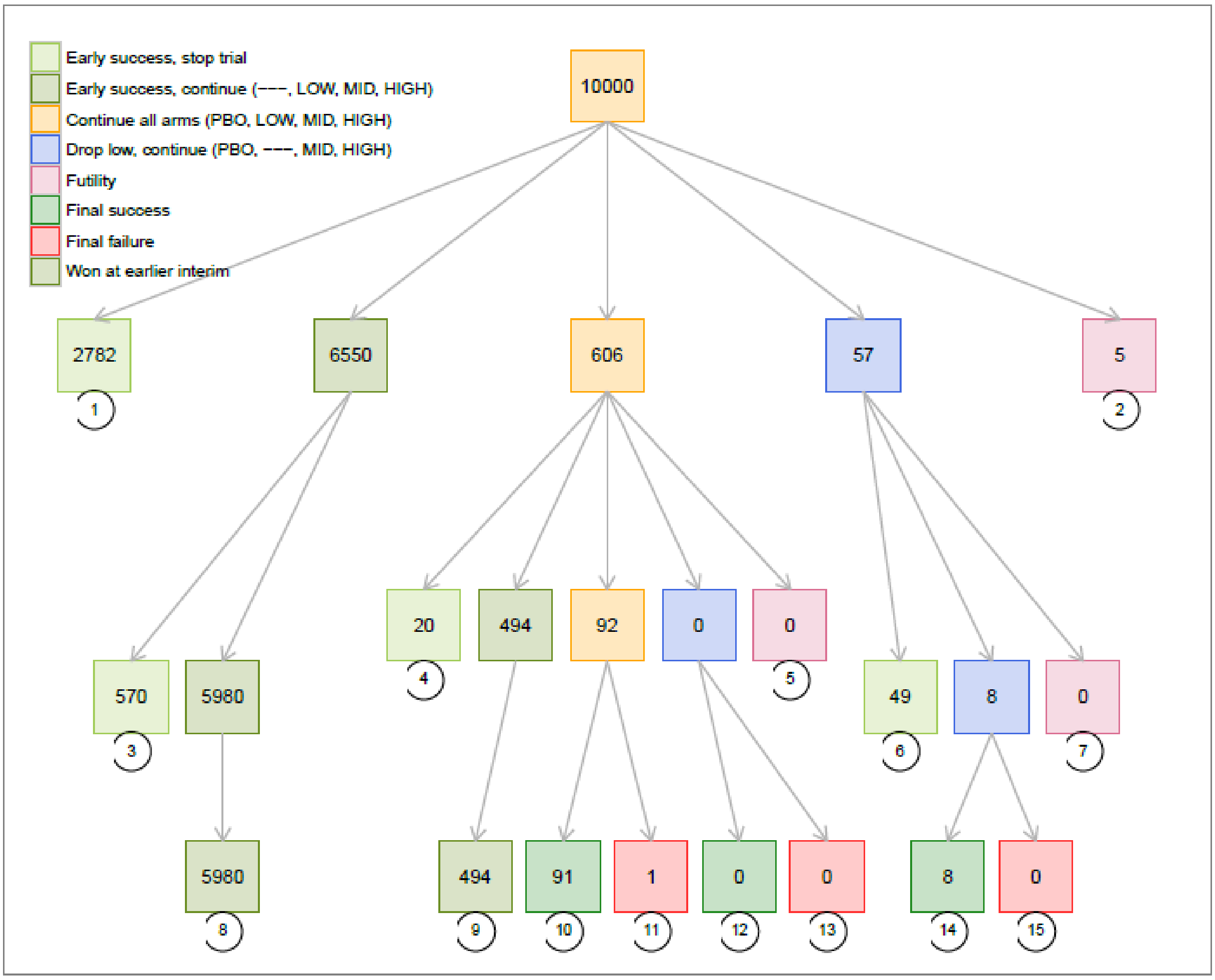
Operating characteristics of the design are presented graphically so that all possible trial pathways may be represented. Additional summaries presented in tabular form include:

- The proportion of trials with the following outcomes:
- Success (either at an interim or final analysis). For non-null scenarios, this number is the power of the trial
- Early success (all trials that met the early success criterion, regardless of whether enrolment continued without the placebo arm in order to address dose response)
- Early futility
- The average sample size per arm and overall
- The proportion of trials for which each arm was selected as the ED80 (note: no arm is selected for unsuccessful trials)

For all operating characteristics in this section, the results are based on 10000 virtual trials per scenario.

Scenario 1: 0.01, 0.60, 0.65, 0.70

In this scenario, all doses are highly effective. The power of the trial exceeds 99%, with most trials declaring early success at an interim. Of the 10000 simulated trials, 9332 trials (93.3%) declared early success at the first interim. Of those, 2782 trials also dropped the low dose at the same time (ending the trial), and 6550 dropped the placebo and continued with only the investigational arms. Of those trials, 570 trials dropped the low dose at the second interim, while the remaining 5980 trials continued to the final analysis with all investigational arms. The average sample size reflects the tendency of trials to declare early success, with the placebo arm having, on average, 9.2 subjects (smaller than the investigational arms). In truth, the low dose is quite effective, so it is rarely dropped, and it is selected as the ED80 in 45.6% of trials.


Early success, stop trial
Early success, continue (---, LOW, MID, HIGH)
Continue all arms (PBO, LOW, MID, HIGH)
Drop low, continue (PBO, ---, MID, HIGH)
Futility
Final success
Final failure
Won at earlier interim
10000
2782
1
6550
606
57
5
2
20
4
494
92
0
0
5
570
3
5980
49
6
8
0
7
5980
8
494
9
91
10
1
11
0
12
0
13
8
14
0
15

Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
0.999	0.990	0.000

Mean Sample Size

Placebo	Low	Mid	High	Total
9.2	13.1	13.1	13.1	48.6

Dose Selection

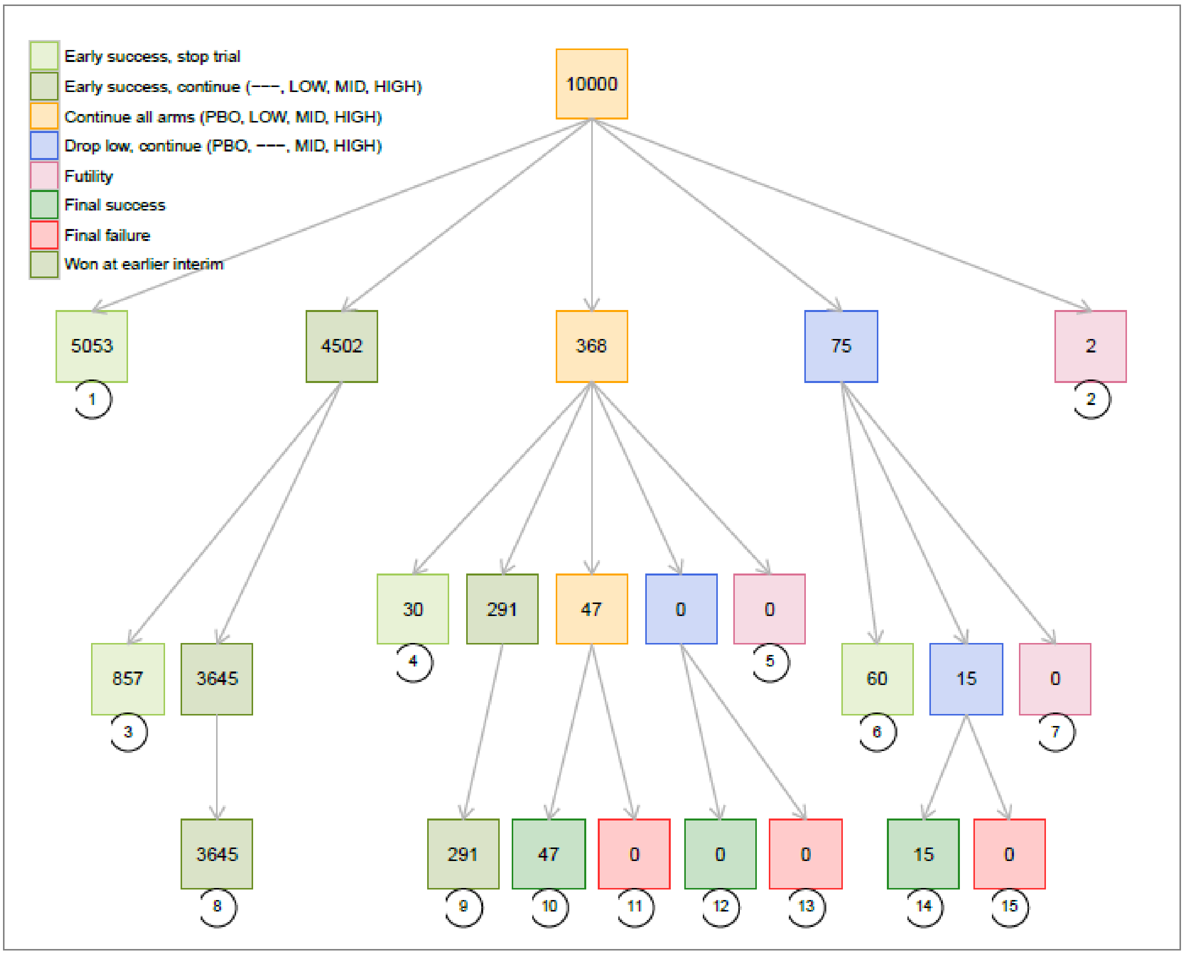
None	Low	Mid	High
0.001	0.456	0.356	0.188

Figure 7 Operating Characteristics for Scenario 1: 0.01, 0.60, 0.65, 0.70

Scenario 2: 0.01, 0.50, 0.65, 0.75

In this scenario, there is a larger range between the low dose and high dose response rates. The low dose response rate is exactly 20% lower than the combined middle and high doses.

This scenario also has quite high power, and is more likely to stop the trial early due to declaring early success and dropping the low dose. Roughly half of the trials (50.5%) ended at the first interim, with an additional 45% declaring early success at the first interim, but continuing the trial (without the placebo arm). The average sample size tends to be smaller in this scenario, since more trials end early. The middle dose is most likely to be selected as the ED80.


Early success, stop trial
Early success, continue (---, LOW, MID, HIGH)
Continue all arms (PBO, LOW, MID, HIGH)
Drop low, continue (PBO, ---, MID, HIGH)
Futility
Final success
Final failure
Won at earlier interim
10000
5053
1
4502
368
75
2
2
30
4
291
47
0
0
5
857
3
3645
60
6
15
0
7
3645
8
291
9
47
10
0
11
0
12
0
13
15
14
0
15

Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
1.000	0.994	0.000

Mean Sample Size

Placebo	Low	Mid	High	Total
9.2	11.7	11.7	11.7	44.2

Dose Selection

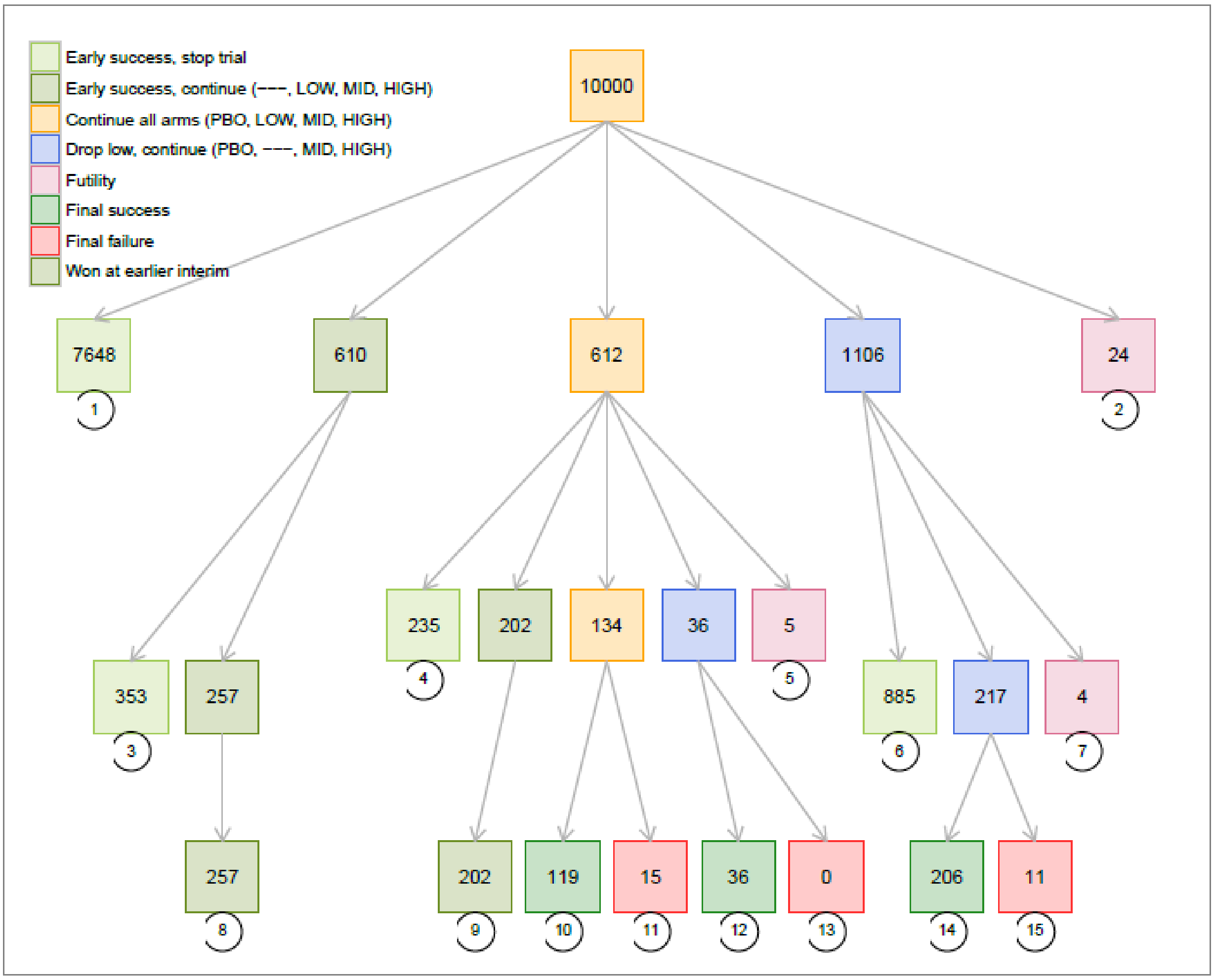
None	Low	Mid	High
0.000	0.182	0.482	0.335

Figure 8 Operating Characteristics for Scenario 2: 0.01, 0.50, 0.65, 0.75

Scenario 3: 0.01, 0.20, 0.60, 0.60

In this scenario, the middle and high doses are equivalent, each having a 60% response rate (less effective than previous scenarios, but still highly effective relative to placebo). The low dose response rate is 40% lower than the combined middle and high doses, so we expect it to be dropped in a large percentage of trials.

In fact, since the pooled middle and high doses are effective and the low dose is inferior to pooled, most trials are able to declare early success and stop the trial early: 76.5% of trials declare success and stop at the first interim, and an additional 14.7% (353+235+885 trials) ended with success at the second interim. There were 459 trials that declared early success at an interim, but continued enrolling the low dose to the final analysis. The middle dose was selected as the ED80 in 72% of trials.


Early success, stop trial
Early success, continue (---, LOW, MID, HIGH)
Continue all arms (PBO, LOW, MID, HIGH)
Drop low, continue (PBO, ---, MID, HIGH)
Futility
Final success
Final failure
Won at earlier interim
10000
7648
1
610
612
1106
24
2
353
3
257
235
4
202
134
36
5
5
885
6
217
4
7
257
8
202
9
119
10
15
11
36
12
0
13
206
14
11
15

Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
0.994	0.958	0.003

Mean Sample Size

Placebo	Low	Mid	High	Total
9.7	9.5	10	10	39.2

Dose Selection

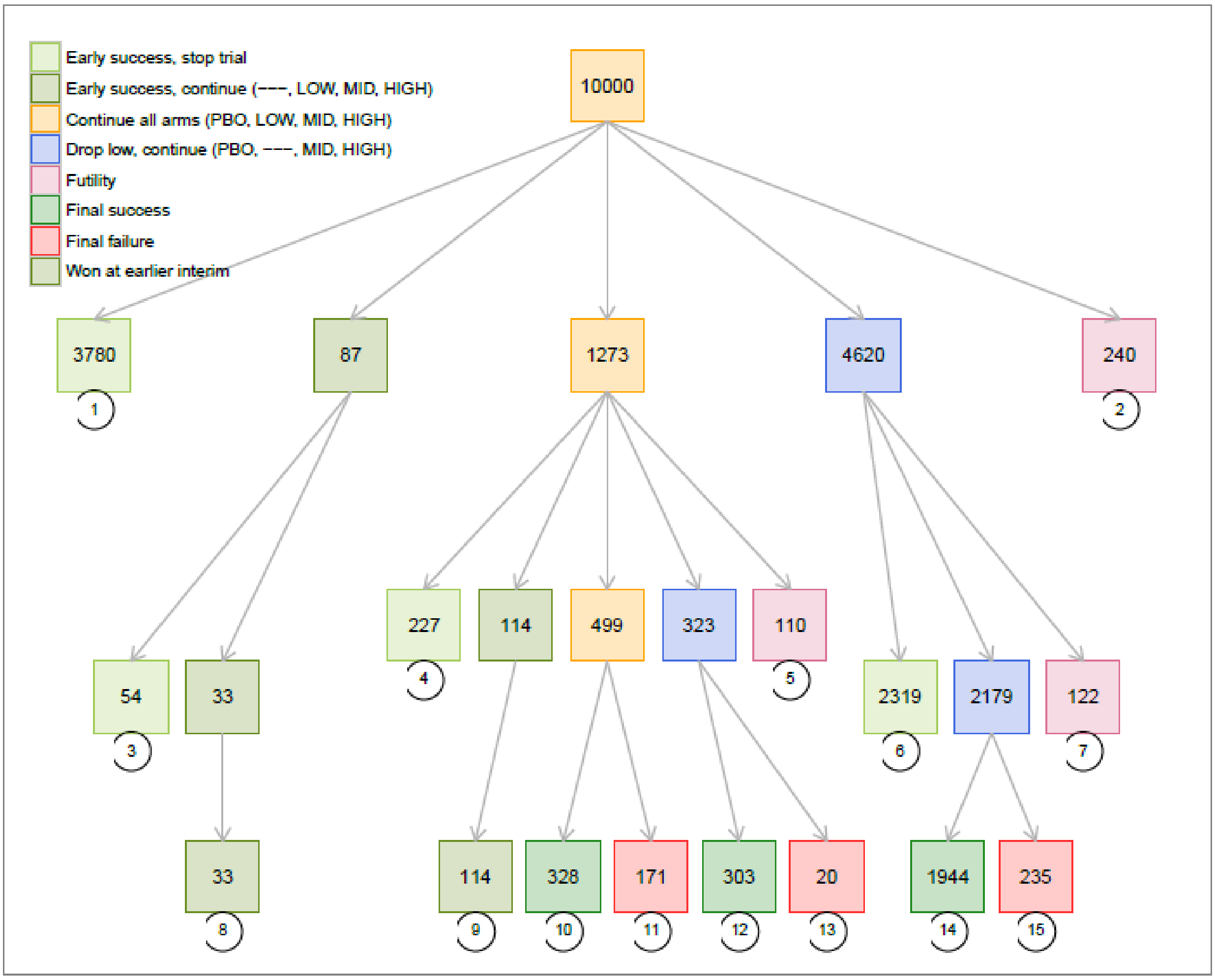
None	Low	Mid	High
0.006	0.003	0.720	0.271

Figure 9 Operating Characteristics for Scenario 3: 0.01, 0.20, 0.60, 0.60

Scenario 4: 0.01, 0.10, 0.40, 0.50

In this scenario, the pooled response rate for the middle and high doses is 45%. This smaller effect is reflected in lower power relative to the previous scenarios.

The trial stopped for futility 4.7% of the time (240 trials at the first interim and 232 trials at the second interim). Since the low dose is much lower than the middle and high doses, and since the efficacy of the middle and high doses is lower than previous scenarios, we see more trials here that drop the low dose and continue the trial with only the placebo, middle dose, and high dose. There were 4620 trials that dropped the low dose at the first interim. Of those, 2319 trials met early success at the next interim (thus ending the trial), while 122 of the trials stopped for futility at the next interim and 2179 continued to the final analysis. Thus, for these 2179 trials, the final sample size for the low dose was 9 subjects, while the middle and high doses each had a final sample size of 17 subjects (each arm receiving two extra subjects because of the low dose dropping at the first interim). Additionally, there were 323 trials that continued all arms at the first interim, then dropped the low dose at the second interim. In these trials, the final sample size for the low dose was 12 subjects while the middle and high doses each received one extra subject for a total of 16 subjects at the final analysis.



Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
0.910	0.653	0.047

Mean Sample Size

Placebo	Low	Mid	High	Total
12.1	9.6	12.2	12.2	46.2

Dose Selection

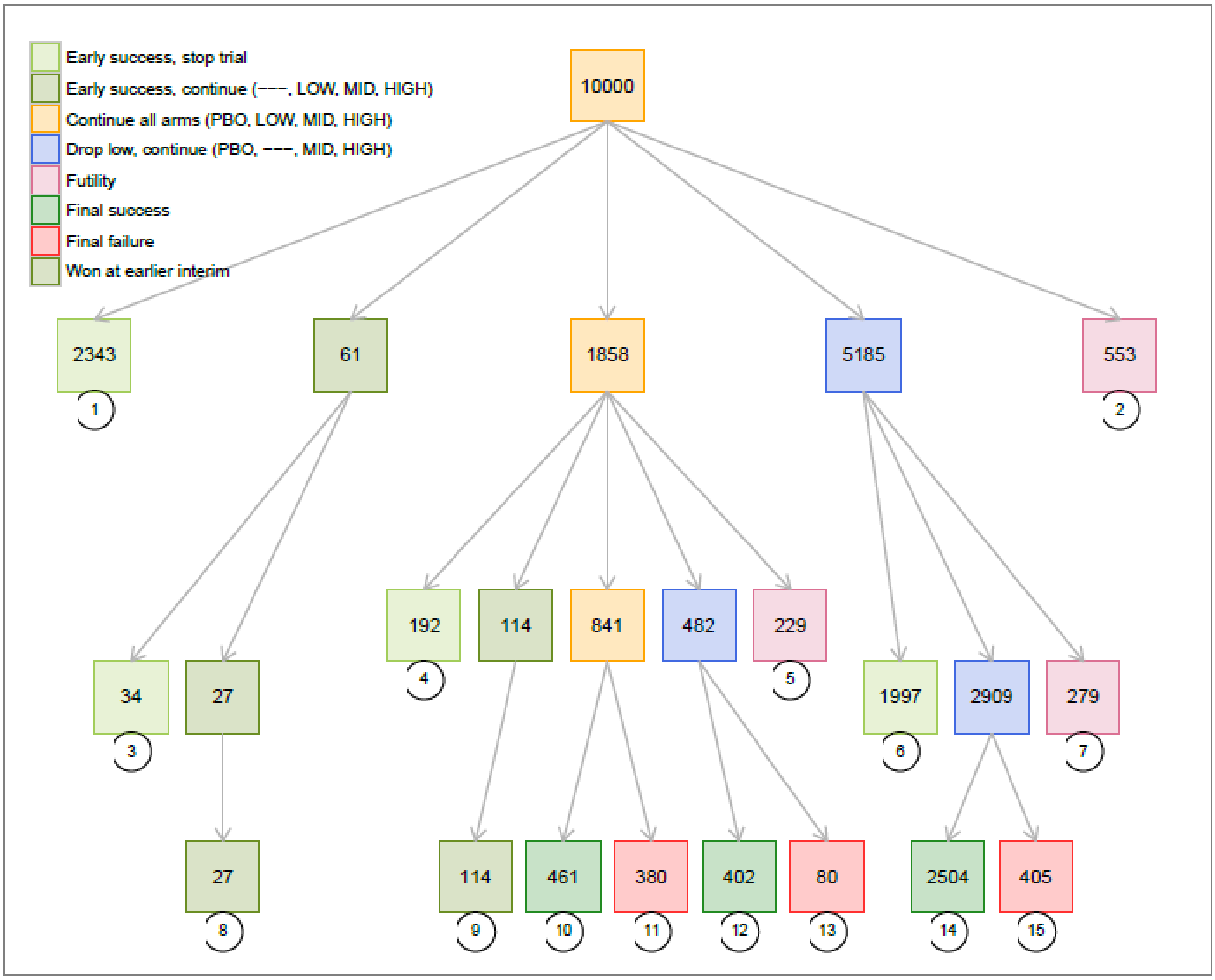
None	Low	Mid	High
0.090	0.001	0.451	0.458

Figure 10 Operating Characteristics for Scenario 4: 0.01, 0.10, 0.40, 0.50

Scenario 5: 0.01, 0.10, 0.40, 0.40

In this scenario, the middle and high doses are equivalent with a 40% response rate. The low dose is 30% lower than the middle and high doses.

Roughly half of the simulated trials (5185 trials, or 51.9%) dropped the low dose at the first interim, but continued enrolling the other arms. An additional 482 trials (4.8%) dropped the low dose at the second interim. With the high probability of dropping the low dose, the average sample size for the low dose is 9.9 while the middle and high doses tend to have larger sample sizes, since they receive re-allocated subjects from the dropped low dose. The power for this scenario is 80.7%. The probability of early futility is 10.6% (5.5% at the first interim and 5.1% at the second interim).



Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
0.807	0.471	0.106

Mean Sample Size

Placebo	Low	Mid	High	Total
13	9.9	13.1	13.1	49

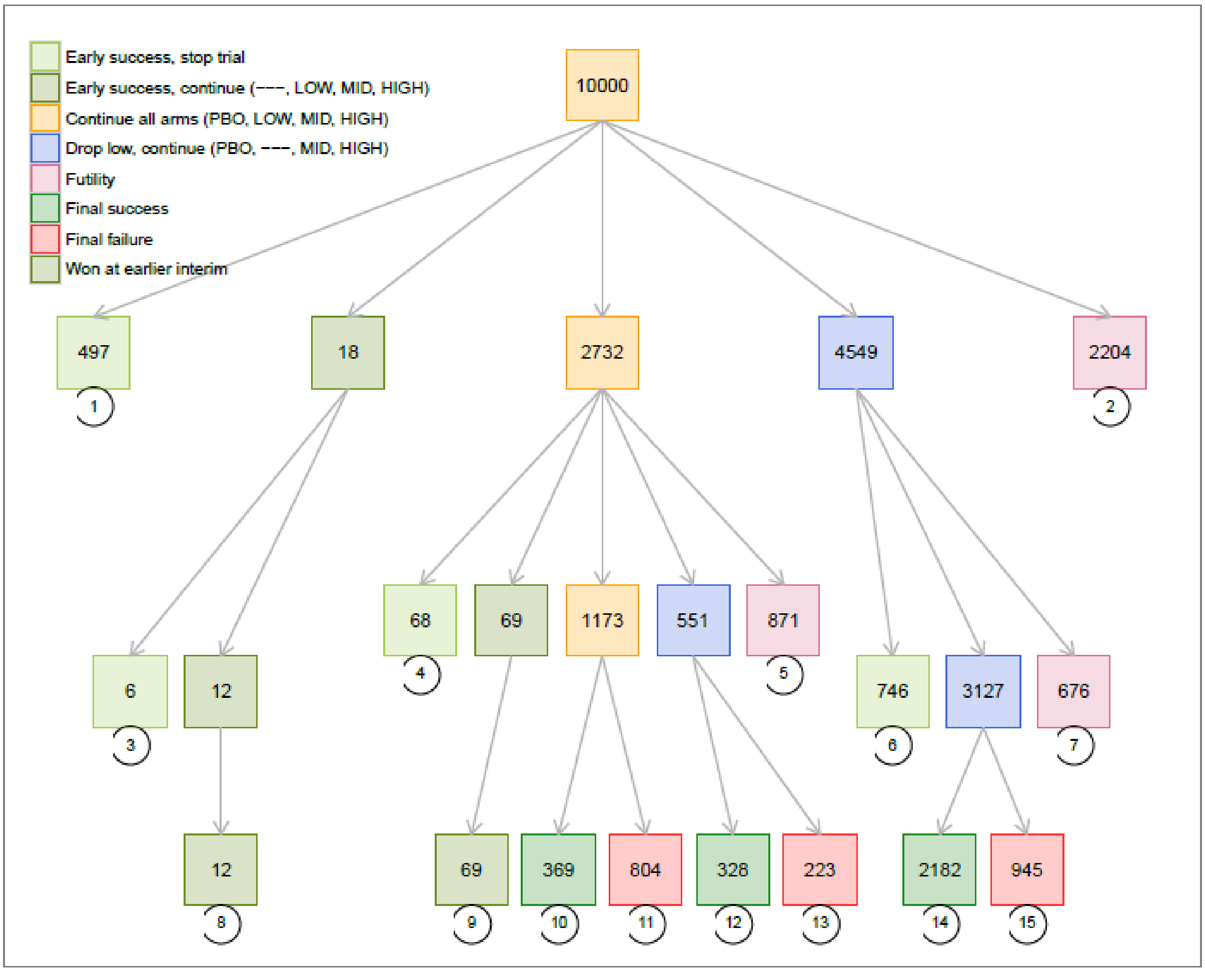
Dose Selection

None	Low	Mid	High
0.193	0.001	0.551	0.255

Figure 11 Operating Characteristics for Scenario 5: 0.01, 0.10, 0.40, 0.40

Scenario 6: 0.01, 0.10, 0.20, 0.40

This scenario is similar to the last, except that the middle dose response rate has been lowered to 20%. Given the lower effect on the middle dose, the power of the trial drops to 42.8%. A larger percentage of trials (37.5%) stop for early futility.


Early success, stop trial
Early success, continue (---, LOW, MID, HIGH)
Continue all arms (PBO, LOW, MID, HIGH)
Drop low, continue (PBO, ---, MID, HIGH)
Futility
Final success
Final failure
Won at earlier interim
10000
497
1
18
2732
4549
2204
2
68
4
69
1173
551
871
5
6
3
12
746
6
3127
676
7
12
8
69
9
369
10
804
11
328
12
223
13
2182
14
945
15

Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
0.428	0.140	0.375

Mean Sample Size

Placebo	Low	Mid	High	Total
13.3	10.2	13.3	13.3	50.2

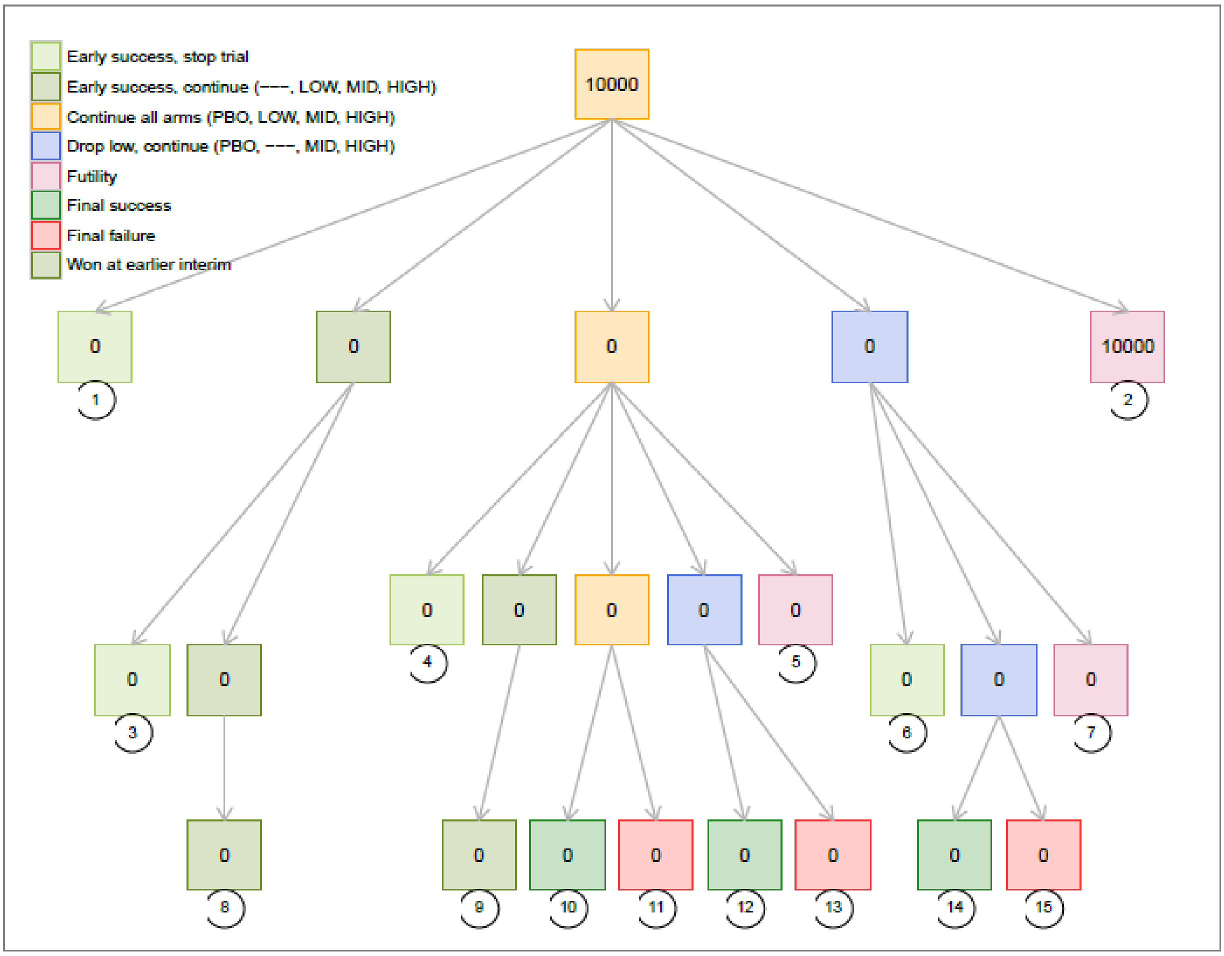
Dose Selection

None	Low	Mid	High
0.572	0.001	0.095	0.332

Figure 12 Operating Characteristics for Scenario 6: 0.01, 0.10, 0.20, 0.40

Scenario 7: 0.01, 0.01, 0.01, 0.01

This is a null scenario in which all arms are equivalent to placebo. In the simulated trials, 100% of trials stopped for futility, and in fact, stopped at the first interim analysis.


Early success, stop trial
Early success, continue (---, LOW, MID, HIGH)
Continue all arms (PBO, LOW, MID, HIGH)
Drop low, continue (PBO, ---, MID, HIGH)
Futility
Final success
Final failure
Won at earlier interim
10000
0
1
0
0
0
10000
2
0
3
0
0
4
0
0
0
0
5
0
6
0
0
7
0
8
0
9
0
10
0
11
0
12
0
13
0
14
0
15

Trial Outcomes

Pr(Success)	Pr(Early Win)	Pr(Early Futility)
0	0	1

Mean Sample Size

Placebo	Low	Mid	High	Total
9	9	9	9	36

Dose Selection

None	Low	Mid	High
1	0	0	0

Figure 13 Operating Characteristics for Scenario 7: 0.01, 0.01, 0.01, 0.01

15.2.4 Control of Type I Error

To understand the type I error, we begin by examining the type I error rate of a fixed trial. That is, we consider a trial with $n = 15$ subjects on the placebo arm and $n = 30$ subjects on the pooled middle and high doses. There are no interims for early stopping, and no adjustments to sample size for dropping the low dose arm.

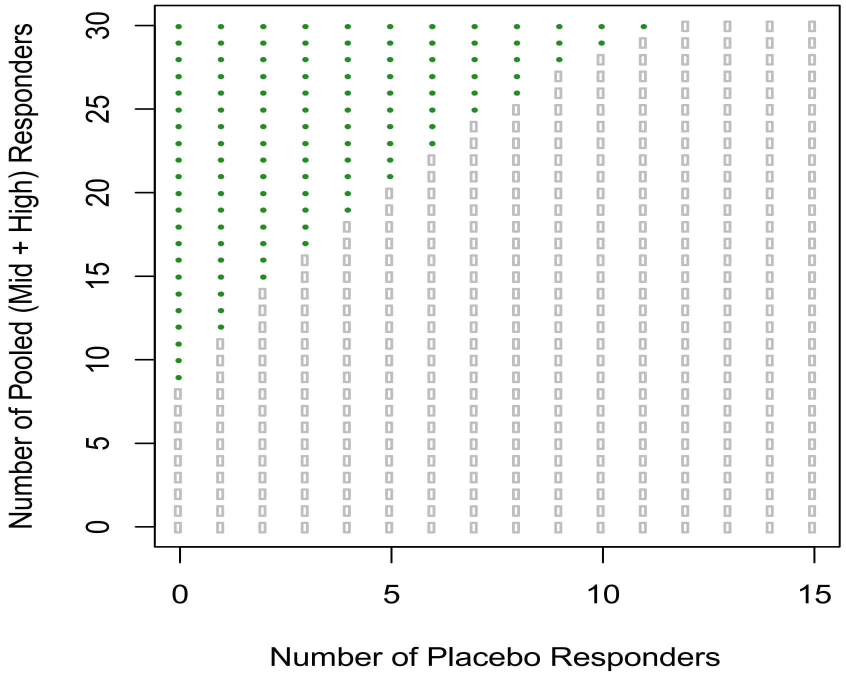
There are 496 possible combinations of outcomes: there may be between 0 and 15 responders on the placebo arm (16 possible outcomes) and between 0 and 30 responders on the pooled investigational arms (31 possible outcomes). We compute the probability of each combination using Binomial distributions. Then we analyze each combination with Fisher's Exact test. The type I error rate is then simply the sum over all successful combinations (p-value < 0.025, since there are no interims in this example), weighted by the probability of each of those combinations.

This computation yields a type I error rate of 1.018×10^{-11}. This is an exact calculation, not simulation-based.

Fisher's Exact Test is known to be conservative, so it is not surprising that the type I error rate is below the nominal 0.025 for the fixed trial. To further understand the small type I error rate, we examine the types of outcomes that would lead to a successful trial. Figure 7 illustrates all possible outcomes for the pooled arms (y-axis) versus all possible outcomes for the placebo (x-axis). Thus each dot on the graph represents a possible combination. The dots shaded in green indicate the combinations leading to success. Out of the 496 possible outcomes, only 118 result in a successful trial. Furthermore, these combinations are unlikely to occur under the null hypothesis of a 1% responder rate in all arms. For example, the probability of observing 9 or more responders out of 30 subjects, assuming a true responder rate of 1% is 1.184×10^{-11}.

Given the extremely low type I error rate for the fixed design, combined with the Pocock boundaries for the adaptive design, the addition of early success interims does not inflate the type I error above the nominal 0.025. In fact, in simulations of 100000 trials under the null hypothesis, no instances of trial success were observed.

In the execution of the trial, there may be slight deviations from the assumptions used in this report. For example, the sample size may not be exactly equal to the planned 9 subjects in each arm at the first interim analysis (e.g. due to blocking). Such small deviations do not pose a threat for type I error inflation since the type I error rate is so far below the nominal value.



(Green = success, Grey = failure)

Figure 14 All possible response outcomes for a fixed trial

15.2.5 Computational Details

The simulations were run using custom code written in the R software package. The stopping boundaries for the group sequential design were obtained from the gsDesign package. R was also used to summarize the simulation output and to create graphics and tables for this report.

15.3 Additional Missing Data Handling Procedures

Unless otherwise noted in previous sections, in general, data will be analysed as received from the clinical database. Hence, missing values including those caused by dropouts will not be imputed.

Missing or partially missing date of the following situations will be imputed for purposes of analysis. It is anticipated that all study-related visit and procedure dates entered into the clinical study database will be complete (i.e. day, month, year are all recorded) and accurate. Any missing or partially missing date of this type will be queried before statistical analyses are performed. If the day, month, and/or year are still unknown, then the dates will be imputed as follows for purposes of analysis:

- If the day of the visit or procedure date is missing, then take the previous visit and add the number of days to the next visit according to the visit schedule.
- If the month of the visit or procedure date is missing then take the previous visit and add the number of months to the next visit according to the visit schedule.
- If the year of the visit or procedure date is missing, then the year will be queried. If the year is unknown, no imputation of year will take place.

Only collected dates will be noted in the patient data listings.

15.4 Analysis Visit Windows

For the analysis of follicular development by visit from study day 10 to study day 21, visit windows are allowed for certain visits per the schedule of assessments. Subject data will not be excluded from analyses due to the subject's failure to comply with the visit schedule.

The following rules as shown in Table 5 will be applied to assign analysis visit to study visits. The analysis visit is determined by the number of days between the date of first study treatment application and the corresponding visit dates for TVUS (for monitoring follicular development).

In the case of multiple assessments at a specific visit, the observation which is closest to the target date will be used. If the observations have the same distance to the target date, the latest one will be used. If baseline assessments are not performed, the last non-missing measurement taken before the first dose will be used as baseline.

The visit windows described in the following table will be applied to the analyses of efficacy endpoints.

Table 5 Analysis visit windows

Study Visit	Target Day	Analysis Visit Window
4*	10	2≤Day≤10
5*	12	11≤Day≤13
6*	15	14≤Day≤16
7*	18	17≤Day≤19
8*	21	20≤Day≤22

Target day is calculated as the study day relative to the first day of IMP administration: date of visit – date of first IMP administration + 1.

15.5 Markedly Abnormal Laboratory Safety Values and Vital Signs

When applying the markedly abnormal criteria defined below, the low/high thresholds will be compared to the laboratory normal reference low/high limits. If the normal reference low limit is smaller than the low threshold, then the normal reference low limit would be used; On the other hand, if the normal reference high limit is larger than the high threshold, then the normal reference high limit would be used.

Table 6: Markedly abnormal Criteria for Laboratory Tests

Variable	Units	Markedly abnormal Criteria	
		Low	**High**
Haematology			
Haemoglobin	g/L	≤ 115	Not applicable
Haematocrit	Ratio	≤ 0.37	≥ 0.56
Total WBC	10^9/L	≤ 2.8	≥ 16.0
Eosinophils	%	Not applicable	≥ 10
Neutrophils	%	≤ 15	≥ 90
Lymphocytes	%	≤ 10	≥ 80
Monocytes	%	Not applicable	≥ 20
Basophils	%	Not applicable	≥ 5
Bands	%	Not applicable	≥ 20
Platelets	10^9/L	≤ 75	≥ 700
Total RBC	10^{12}/L	≤ 3.5	Not applicable
Clinical Chemistry			
AST	IU/L	Not applicable	> 3xULN
ALT	IU/L	Not applicable	> 3xULN
Alkaline phosphatase	IU/L	Not applicable	> 3xULN and 25% increase from baseline
GGT	IU/L	Not applicable	> 3xULN
LDH	IU/L	Not applicable	> 3xULN
Total bilirubin	µmol/L	Not applicable	≥ 1.5xULN
Urea nitrogen	mmol/L	Not applicable	≥ 10.7
Creatinine	µmol/L	Not applicable	≥ 177
Total protein	g/L	≤ 45	≥ 90
Albumin	g/L	≤ 25	≥ 65
Sodium	mmol/L	≤ 130	≥ 155
Potassium	mmol/L	≤ 3.0	≥ 5.8
Chloride	mmol/L	≤ 90	≥ 115
Phosphorus	mmol/L	≤ 0.5	≥ 1.9
Calcium	mmol/L	≤1.8	≥ 3.9
Uric acid	mmol/L	Not applicable	≥ 0.62
Glucose	mmol/L	≤ 2.8	≥ 10
Total cholesterol	mmol/L	Not applicable	≥ 8.0
Urinalysis: Quantitative			

pH	None	≤ 4	Not applicable
Specific gravity	None	≤ 1.005	Not applicable
Urinalysis: Dipstick Chemistries			
Glucose	0 - 4+	Not applicable	Increase of 2 or more units from baseline
Casts		Not applicable	Increase of 2 or more units from baseline
Protein	0 - 4+	Not applicable	Increase of 2 or more units from baseline
Ketones	0 - 4+	Not applicable	Increase of 2 or more units from baseline
Blood (Hgb)	0 - 4+	Not applicable	Increase of 2 or more units from baseline
Urinalysis: Microscopic Variables			
RBC	no./hpf	Not applicable	≥ 10
WBC	no./hpf	Not applicable	≥ 20
Casts	no./hpf	Not applicable	Neg at baseline to positive on-treatment
Bacteria	no./hpf	Not applicable	Neg at baseline to positive on-treatment
Cells	no./hpf	Not applicable	Neg at baseline to positive on-treatment
Crystals	no./hpf	Not applicable	Neg at baseline to positive on-treatment

Table 7: Markedly abnormal Criteria for Vital Signs*

Variable	**Criterion Value**	**Change from Baseline**
Systolic blood pressure	≥ 180 mmHg ≤ 90 mmHg	Increase of ≥ 20 mmHg Decrease of ≥ 20 mmHg
Diastolic blood pressure	≥ 105 mmHg ≤ 50 mmHg	Increase of ≥ 15 mmHg Decrease of ≥ 15 mmHg
Pulse rate	≥ 120 bpm ≤ 50 bpm	Increase of ≥ 15 bpm Decrease of ≥ 15 bpm
Body weight	None	Increase of ≥ 7% Decrease of ≥ 7%
Body temperature	≥ 38.3° C	Increase to ≥ 39.4°C

* To be identified as markedly abnormal, a treatment value must meet the criterion value and also the specified change from baseline.