STATISTICAL ANALYSIS PLAN
Document: CS0935 Rev C, ADVENT Statistical Analysis Plan
Study Title: The FARAPULSE ADVENT Trial: A Prospective Randomized Pivotal Trial of the FARAPULSE Endocardial Ablation System Compared with Standard of Care Ablation in Patients with Paroxysmal Atrial Fibrillation
Study Device: FARAPULSE Endocardial Ablation System
Sponsor: FARAPULSE, Inc.
Author: [redacted]
Date: 12 Oct 2021

Signature: ______________________________________ Date:_______________

Analysis Summary

Study Title: The FARAPULSE ADVENT Trial: A Prospective Randomized Pivotal Trial of the FARAPULSE Endocardial Ablation System Compared with Standard of Care Ablation in Patients with Paroxysmal Atrial Fibrillation

Primary Objective: The objective of this pivotal study is to provide valid scientific evidence that endocardial ablation using the FARAPULSE Endocardial Ablation System is both safe and effective for treating drug-resistant paroxysmal atrial fibrillation (PAF).

The following endpoint definitions from the ADVENT Protocol are provided for convenience. Any discrepancy with the ADVENT Protocol is inadvertent unless specifically identified as a deliberate change (or clarification) of the protocol's definition.

Primary Effectiveness Endpoint

The primary effectiveness endpoint is Treatment Success, a composite defined in the Protocol as:

1. Acute Procedural Success AND
2. Chronic Success, defined as freedom from:
    - a. At the Index/Rescheduled Index Procedure: The use of a non-randomized treatment modality for pulmonary vein isolation (PVI)
    - b. After the Blanking Period:
        - i. The occurrence of any Detectable AF, AFL or AT (excluding CTI-dependent flutter confirmed by EP study)
        - ii. Any cardioversion for AF, AFL or AT (excluding for CTI-dependent flutter)
        - iii. The use of any Type I or Type III anti-arrhythmic medication for the treatment of AF, AFL or AT
    - c. At any time:
        - i. Re-ablation for AF/AFL/AT (other than for CTI-dependent flutter)
        - ii. Use of amiodarone

Endpoint status will be assessed through the Month 12 Assessment (Day 360 ± 30). The incidence of Treatment Success will be assessed for non-inferiority.

Secondary Effectiveness Endpoint:

The secondary effectiveness endpoint is the same as the primary effectiveness endpoint but will be tested for superiority rather than non-inferiority.

Additional Effectiveness Endpoints

1. Acute Procedural Success
2. Acute Vein Success
3. Chronic Success
4. Chronic Success Allowing Re-ablation
5. Chronic Success Allowing AADs
6. Treatment Success Allowing Re-ablation
7. Treatment Success Allowing AADs
8. Treatment Success with PVI/CTI only
9. Early Recurrence of AF
10. Rate of Re-ablation
11. PVI Durability at Re-ablation
12. CTI Ablation Failure
13. Thermal Group RF Ablation / Cryoablation Effectiveness Comparison
14. AF Symptom Assessment
15. Learning Curve Effectiveness Assessment

Primary Safety Endpoint:

The Primary Safety Endpoint for this study is the Composite Safety Endpoint (CSE) defined in the ADVENT Protocol as the proportion of Safety Subjects with one or more of the following device- or procedure-related serious adverse events (SAEs), as adjudicated by the Clinical Endpoints Committee.

Early onset: Any of the following with an Onset Date within 7 days of the Index / Rescheduled Index procedure:

- Death
- Myocardial infarction
- Persistent phrenic nerve palsy
- Stroke
- Transient ischemic attack
- Peripheral or organ thromboembolism
- Cardiac tamponade / perforation
- Pericarditis
- Pulmonary edema
- Vascular access complications
- Heart block
- Gastric motility/pyloric spasm disorders

Late onset: any time through the 12-month assessment

- Pulmonary vein stenosis
- Atrio-esophageal fistula

Secondary Safety Endpoint:

Aggregate PV Cross-Sectional Area

Additional Safety Endpoints:

1. Severe Ablation Complications (PV stenosis, persistent phrenic nerve palsy or atrio-esophageal fistulae)
2. Nonserious / Serious CSEs
3. Post-Blanking Cardioversions
4. Post-Blanking Arrhythmia Hospitalizations
5. Any Related SAE
6. Any Related Stroke or TIA
7. Categorized PV Dimensional Changes
8. Thermal Group RF Ablation / Cryoablation Safety Assessment
9. Learning Curve Safety Assessment

Patient Population: Males and females aged 18 years and to 75 with symptomatic Paroxysmal Atrial Fibrillation (PAF).

Study Design: This is a prospective, adaptive, multi-center, randomized safety and effectiveness pivotal study comparing the FARAPULSE Pulsed Field Ablation System with standard of care ablation with force-sensing RF catheters and cryoballoon catheters indicated for the treatment of PAF.

Number of Subjects: The sample size is determined adaptively. A minimum of 350 and a maximum of 750 subjects are planned to be accrued into the modified Intent-to-Treat population (primary effectiveness analysis population). To achieve a sample size of 750 mITT subjects, up to 900 subjects may be enrolled (750 mITT + up to 105 Roll-In + up to 45 Randomized Subjects who do not meet the criteria for the mITT population). See Section 4.1.

Contents

Statistical Analysis Plan

1 Introduction

PVI as treatment for drug-refractory paroxysmal atrial fibrillation by means of catheter ablation is recommended by the clinical community. There are two frequently used FDA-approved methods used for such ablation: radiofrequency ablation and cryoballoon technology. FARAPULSE has developed a pulsed field energy device to perform endocardial ablation that may have advantages over currently approved devices.

2 Study Design

This investigation is a prospective, randomized (1:1), adaptive, multi-center IDE pivotal study comparing the FARAPULSE Pulsed Field Ablation System to standard of care RF and cryoballoon ablation in subjects with drug-resistant symptomatic PAF.

This study is designed using Bayesian statistical methods, with non-informative prior distributions for all parameters. The appropriate sample size is determined adaptively via a Goldilocks[15] design, in which the adequacy of the sample size is assessed at several pre-specified enrollment milestones.

Throughout this document, the traditional terminology "Treatment Group" (or Subjects) and "Control Group" (or Subjects) is used. These terms will refer to participants assigned to Pulsed Field or Thermal ablation therapy, respectively.

3 Analysis Objectives

The objective of this pivotal study is to provide valid scientific evidence that endocardial ablation using the FARAPULSE Endocardial Ablation System is both safe and effective for treating drug-resistant paroxysmal atrial fibrillation (PAF).

4 Analysis Populations

The population definitions in this section are copied from Section 3.4 of the protocol. Any discrepancy is inadvertent unless specifically identified as a deliberate change (or clarification) of the protocol's definition.

4.1 Population Definitions

4.1.1 General Pre-Screening

Potential subjects screened for enrollment prior to signing informed consent who fail one (1) or more study eligibility criteria determined by routine clinical assessments or the medical record need not be documented.

4.1.2 COVID-19 Pre-Screening

Screening for COVID-19 status will follow each investigational site's then-current requirements. Patients may proceed to informed consent and full screening unless any of the following conditions exist:

- They fail the investigational site's then-current requirements in relation to COVID-19 status and study participation.
- They currently have confirmed, active COVID-19 disease or a positive test for SARS-CoV-2.

- They have had confirmed COVID-19 disease not clinically resolved at least 3 months prior to the Consent Date.

The results of COVID-19 screening may be used as study data for any consented subject.

4.1.3 Screened Subjects

A Screened Subject is any subject who has not failed pre-screening and who has signed an informed consent. Each Screened Subject will be recorded in the Screening Log with the date of informed consent and the result of the screening process.

Should an AE occur that is associated with a screening procedure or test and with onset prior to randomization, it will be recorded in a Screening AE CRF. Such AEs are not part of the ADVENT Trial, its data or analyses and will not be entered into the study database. A listing and summary of Screened Subjects will be provided along with any screening-associated AEs.

4.1.4 Screen Failure Subjects

A Screen Failure Subject is any subject who has signed an informed consent and who is then excluded due to failure of one or more study eligibility criteria at any time before randomization.

Subjects who become Screen Failures will be recorded in the Screening Log including all eligibility criteria that were the cause of the Screen Failure. Screen Failure Subjects will have a Study Exit CRF completed on the date that Screen Failure is determined. No study data on events with an onset date after the Exit Date may be collected or included as part of the study database or analyses.

4.1.5 Enrolled Subjects

An Enrolled Subject is any subject who has signed the informed consent document AND who has been determined to meet all pre-procedural study eligibility criteria prior to randomization.

4.1.6 Roll-In Subjects

A Roll-In Subject is an Enrolled Subject who is treated with the FARAPULSE Pulsed Field Ablation System under the provisions of Protocol Section 5.6. Roll-In Subjects are not randomized and will be analyzed and reported separately.

4.1.7 Intent-to-Treat Subjects

An Intent-to-Treat (ITT) Subject is an Enrolled Subject who is randomized to either the Pulsed Field Group or the Thermal (Control) Group.

4.1.8 Safety Subjects

A Safety Subject is an ITT Subject who has had a study ablation catheter inserted into the body.

4.1.9 Modified Intent-to-Treat Subjects

A Modified Intent-to-Treat (MITT) Subject is an ITT Subject who receives any energy delivery for PVI with the randomized endocardial ablation catheter at an Index / Rescheduled Index Procedure.

4.1.10 Completed Subjects

A Completed Subject is an ITT Subject who completes the Month 12 Assessment either in-person or remotely. The date on which the Month 12 Assessment is completed is the Completed Subject's Exit Date and a Study Exit CRF will be completed. No study data on events with an onset date after the Exit Date may be collected or included as part of the study database or analyses.

4.1.11 Incomplete Subjects

An Incomplete Subject is any ITT Subject who does not become a Completed Subject. Anticipated reasons for failure to become Completed Subjects include withdrawal of consent by a subject, death or incapacity of a subject, lost-to-follow-up (LTFU) status and termination by the Investigator.

A LTFU Subject is an ITT Subject who ceases to participate in study follow-up procedures or who fails to respond to contact attempts. The Investigator will attempt to contact such a subject 2 or more times within 60 days and if there is no return to study participation, the subject will be deemed LTFU. The Investigator will document that a minimum of two attempts were made to contact the subject, including sending a certified letter if current address is known, prior to terminating the subject from the study.

The date on which a subject is determined to be an Incomplete Subject is their Exit Date, and a Study Exit CRF will be completed including the reason for becoming an Incomplete Subject. No study data on events with an onset date after the Exit Date of such subjects may be collected or included as part of the study database or analyses. However, if an Incomplete Subject's exit is due to an AE, the Investigator should, if subject consent is documented, follow the subject until the AE has resolved or is considered stable, but in any case not more than 390 days have passed since the Index / Rescheduled Index Procedure. The data related to that AE may be collected and included under such consent.

4.1.12 Per Protocol Subjects

Per Protocol (PP) Subjects are all Subjects without major protocol deviations that would affect the completeness, accuracy and reliability of the study data. Major protocol deviations[1] that exclude subjects from the PP Population are the following:

- Failure to meet any of the following study entry criteria:
  - Arrhythmia different than PAF (Inclusion Criterion 1)
  - Prior cardiovascular procedures, implants or conditions (Exclusion Criteria 3, 4)
  - Medical conditions (Exclusion Criterion 11)
- Failure to receive the randomized ablation treatment for PVI as specified in Protocol Section 5.2 and Protocol Section 5.3.
- Receipt of a prohibited study drug (amiodarone)
- In addition, the data of Severe COVID-19 Subjects will be censored beginning on the Onset Date of infection.

4.2 Pre-Specification for Analysis

Unless otherwise specified for specific endpoints, the principal analysis population for the effectiveness endpoints is the MITT Population, with treatment groups defined by randomization assignment. Analyses in the PP population will also be conducted as sensitivity analyses.

Unless otherwise specified for specific endpoints, the principal analysis population for the safety endpoints is the Safety Population, with treatment groups defined by randomization assignment. Analyses in the PP population will also be conducted as sensitivity analyses.

[1] Major protocol deviations also include such events as failure of informed consent, confidentiality violations and actions that may affect the subject's rights, safety or well-being. These, while serious, generally do not impact the validity of data collected and so are not part of the per protocol exclusions for outcome assessment.

5 Definition of Study Outcomes

Some study definitions are included here for convenience and are derived from the study protocol. Any discrepancy is inadvertent unless specifically identified as a deliberate change or clarification of the protocol's definition.

5.1 Primary Effectiveness Endpoint (Protocol Section 7.4)

The primary effectiveness endpoint is Treatment Success, which is a composite of the following criteria:

1. Acute Procedural Success AND
2. Chronic Success, defined as freedom from:
    a. At the Index/Rescheduled Index Procedure: The use of a non-randomized treatment modality for PVI
    b. After the Blanking Period:
        i. The occurrence of any Detectable AF, AFL or AT (Section 6.3.10) (excluding CTI-dependent flutter confirmed by EP study)
        ii. Any cardioversion for AF, AFL or AT (excluding for CTI-dependent flutter)
        iii. The use of any Type I or Type III anti-arrhythmic medication for the treatment of AF, AFL or AT
    c. At any time:
        iv. Re-ablation for AF/AFL/AT (other than for CTI-dependent flutter)
        v. Use of amiodarone

Endpoint status will be assessed through 360 ± 30 days.

5.2 Secondary Effectiveness Endpoint (Protocol Section 7.5)

The secondary effectiveness endpoint is the same as the primary effectiveness endpoint but will be tested for superiority rather than non-inferiority. The proportion of Pulsed Field Subjects with Treatment Success will be assessed for superiority to the proportion of Thermal Subjects with Treatment Success.

5.3 Additional Effectiveness Endpoints (Protocol Section 7.6)

1. Acute Procedural Success
2. Acute Vein Success
3. Chronic Success
4. Chronic Success Allowing Re-ablation
5. Chronic Success Allowing AADs
6. Treatment Success Allowing Re-ablation
7. Treatment Success Allowing AADs
8. Treatment Success with PVI/CTI only
9. Early Recurrence of AF
10. Rate of Re-ablation
11. PVI Durability at Re-ablation
12. CTI Ablation Failure
13. Control RF Ablation / Cryoablation Effectiveness Comparison
14. AF Symptom Assessment
15. Learning Curve Effectiveness Assessment

5.4 Primary Safety Endpoint

The Primary Safety Endpoint for this study is the Composite Safety Endpoint (CSE) defined in the Protocol as the proportion of Safety Subjects with one or more of the following device- or procedure-related serious adverse events (SAEs) as adjudicated by the Clinical Events Committee (CEC).

Early onset: within 7 days of the Index / Rescheduled Index procedure

- Death
- Myocardial infarction
- Persistent phrenic nerve palsy
- Stroke
- Transient ischemic attack
- Peripheral or organ thromboembolism
- Cardiac tamponade / perforation
- Pericarditis
- Pulmonary edema
- Vascular access complications
- Heart block
- Gastric motility/pyloric spasm disorders

Late onset: any time through the completion of the 12-month assessment

- Pulmonary vein stenosis
- Atrio-esophageal fistula

5.5 Secondary Safety Endpoints

The secondary safety endpoint is Aggregate PV Cross-Sectional Area (Protocol Section 7.2): The paired comparison of change in the computed aggregate PV cross-sectional area between baseline and 3 months between treatment groups.

5.6 Additional Safety Endpoints

1. Severe Ablation Complications (Protocol Section 7.3.1): The proportion of subjects in each treatment group with one or more of the following CSE components (as defined in Protocol Section 7.1, Table 4) will be compared between treatment groups.
    a. PV stenosis
    b. Persistent phrenic nerve palsy
    c. Atrio-esophageal fistula
2. Nonserious / Serious CSEs (Protocol Section 7.3)
3. Post-Blanking Cardioversions (Protocol Section 7.3)
4. Post-Blanking Arrhythmia Hospitalizations (Protocol Section 7.3)
5. Any Related SAE (Protocol Section 7.3)
6. Any Related Stroke or TIA (Protocol Section 7.3)
7. Categorized PV Dimensional Changes (Protocol Section 7.3)
8. Control RF Ablation / Cryoablation Safety Assessment (Protocol Section 7.3)
9. Learning Curve Safety Assessment (Protocol Section 7.3)

6 Multiplicity

The study will be considered successful only if non-inferiority is established at the final analysis for the primary effectiveness endpoint and for the primary safety endpoint. Two secondary superiority tests are also formally pre-specified: (1) Aggregate PV Cross-Sectional Area, and (2) superiority using the primary effectiveness metric of Treatment Success. These will be tested in the above order, each at the 0.025 significance level (one-sided); the hypothesis for superiority of Treatment Success will not be formally tested unless the endpoint of Aggregate PV Cross-Sectional Area has demonstrated statistical significance. Neither secondary test will be conducted if the primary non-inferiority tests for effectiveness and safety do not establish non-inferiority.

7 Comparability Analyses of the Patient Populations

These analyses are intended to determine the similarity of treatment groups and study sites with respect to important demographic or other variables, either known or suspected to have an influence on the outcome variables. The absence of similarity for any baseline variable will identify that variable as a potential covariate in subsequent safety and effectiveness sensitivity analyses. The data for each baseline variable will be presented descriptively. For quantitative variables like age, the mean, standard deviation (SD), median, minimum, and maximum will be presented. For qualitative variables like gender, the number with the characteristic, the total number evaluated, and the percentage will be presented.

7.1 Comparability between Treatments

The comparison of baseline characteristics across treatment groups will apply the following methods. Continuous variables such as age will be compared with two-sided $\alpha=0.05$ Wilcoxon rank sum test or two-sample t-test, and qualitative variables will be analyzed with χ^2 tests, Fisher's exact test or Fisher-Freeman-Halton test. Any important variable found to significantly differ across treatment groups will be considered as a possible covariate in subsequent multivariate safety or effectiveness analyses.

Comparability will be assessed with frequentist statistical tests, but the primary analysis of study results uses Bayesian statistical methods. Tables of baseline demographics or clinical variables in the clinical study report will be summarized using descriptive statistics. Where appropriate, inferential summaries that are consistent with Bayesian methods (such as 95% Bayesian credible intervals) will be presented. The methods used for these summaries are presented in Section 19.

7.2 Comparability between Study Sites

An analysis of comparability across study sites without regard to treatment assignment will be carried out on the same baseline characteristics by the analysis methods described above. Study site differences do not disallow pooling, and in fact the possibility of site-to-site differences is the prime motivator for stratifying the randomization by site. However, variables including study site and the variable found different may be considered as possible covariates in subsequent sensitivity analyses.

8 Patient Accountability and Missing Data

A summary table will provide the total number of subjects enrolled and completed. The subjects eligible for and compliant with each follow-up assessment will be summarized descriptively. Subjects withdrawn will be tabulated with their reasons for withdrawal.

Every effort will be made to collect all data points in the study. The sponsor plans to minimize the amount of missing data by appropriate management of the prospective clinical trial, proper screening of study subjects, and training of participating investigators, monitors and study coordinators. The sponsor will provide a list of subjects who do not complete the trial along with the best information available on why they each left the trial prematurely.

Analyses that are planned for assessing sensitivity of conclusions to missing data are described below in Section 16.

9 Effectiveness Analyses

All effectiveness analyses will be performed in the modified Intent-to-Treat (mITT) population as the principal analyses. Supportive analyses will be conducted in the Per Protocol (PP) population.

9.1 Primary Effectiveness

The analysis of the primary effectiveness endpoint of Treatment Success defined in Protocol Section 7.4 is a test of non-inferiority of the success rate at 12 months. The null and alternative hypotheses are:

$$H_0: P_T \leq P_C - 0.15 \text{ versus } H_A: P_T > P_C - 0.15$$

where P_T is the Treatment Success rate at 12 months in the Treatment Group (Pulsed Field Group), and P_C is the Treatment Success rate at 12 months in the Control Group (RF and cryoablation).

This trial is designed using Bayesian statistical techniques, and the prior distributions for P_T and P_C in these calculations are Beta (0.5, 0.5), which is the Jeffreys non-informative prior. P_T will be determined to be non-inferior to P_C if it can be established that the posterior probability $\Pr(H_A \mid \text{data}) > \Psi_{\text{eff}}$, where $\Psi_{\text{eff}} = 0.956$ is a pre-specified threshold value that controls the one-sided type I error rate (under simulation) at level 0.05. In the absence of missing data, the posterior distributions for P_T and P_C would be conjugate Beta distributions. However, some missing data are expected. Therefore, Bayesian multiple imputation of 12-month outcomes will be employed in the principal analysis. The imputation model accounts for primary outcome information up through the time of censoring. More detail on this imputation is given in Section 13.1.

9.2 Secondary Effectiveness Endpoint Analysis

The secondary effectiveness endpoint is Treatment Success, tested for superiority rather than non-inferiority. The null and alternative hypotheses are:

$$H_0: P_T \leq P_C \text{ versus } H_A: P_T > P_C$$

where P_T is the Treatment Success rate at 12 months in the Treatment Group, and P_C is the Treatment Success rate at 12 months in the Control Group. Modeling will be identical to the test of non-inferiority, and superiority will be concluded if the posterior probability $\Pr(H_A \mid \text{data}) > \Psi_{\text{eff,sup}}$, where the threshold $\Psi_{\text{eff,sup}} = 0.977$ is a pre-specified threshold value that controls the one-sided type I error rate (under simulation) at level 0.025. In the absence of missing data, the posterior distributions for P_T and P_C would be conjugate Beta distributions. However, some missing data are expected. Therefore, Bayesian multiple imputation of 12-month outcomes will be employed in the principal analysis. The imputation model accounts for primary outcome information up through the time of censoring. More detail on this imputation is given in Section 13.1.

9.3 Additional Effectiveness Endpoint Analyses

9.3.1 Acute Procedural Success

The endpoint is Acute Procedural Success. The proportion of subjects with Acute Procedural Success in each treatment group will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical method will be a Bayesian comparison of proportions as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.2 Acute Vein Success per Vein

This endpoint is Acute Vein Success, assessed on a per vein basis. The proportion of veins in each treatment group will be analyzed descriptively: the number with Acute Vein Successes, the total number of veins evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical method will be a Bayesian comparison of proportions as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.3 Chronic Success

The endpoint is Chronic Success. The proportion of subjects with Chronic Success in each treatment group will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.4 Chronic Success Allowing Re-ablation

The endpoint is Chronic Success Allowing Re-ablation. The proportion of subjects with the endpoint in each treatment group will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.5 Chronic Success Allowing AADs

The endpoint is Chronic Success Allowing AADs. The proportion of subjects with Chronic Success Allowing AADs in each treatment group will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.6 Treatment Success Allowing Re-ablation

The endpoint is Treatment Success Allowing Re-ablation. The proportion of subjects with the endpoint in each treatment group will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.7 Treatment Success Allowing AADs

The endpoint is Treatment Success Allowing AADs. The proportion of subjects with Chronic Success Allowing AADs in each treatment group will be analyzed descriptively: the number with with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.8 Treatment Success with PVI/CTI Only

The endpoint is Treatment Success with PVI/CTI Only, excluding subjects receiving extra PVI/CTI ablation. The proportion of subjects with Treatment Success with PVI/CTI Only in each treatment group will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.9 Early Recurrence of AF

The endpoint is Early Recurrence of AF. The proportion of subjects with Early Recurrence of Atrial Fibrillation ≤ Day 90 in each treatment group will be analyzed descriptively: the number of subjects with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T < P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.10 Re-ablation

The endpoint is Occurrence of Re-ablation. The proportion of subjects with the endpoint through 12 months in each treatment group will be analyzed descriptively: the number of subjects undergoing re-ablation, the number evaluated, the percentage and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T < P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.11 PVI Durability at Re-ablation

The endpoint is PVI Durability at Re-ablation, assessed on a per vein basis and also on a per subject basis. For the per-subject analysis, a subject is a "success" if and only if all originally ablated veins have durable isolation. The proportion in each treatment group will be analyzed descriptively: the number with confirmed PVI Durability, the number evaluated, the percentage and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T > P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.12 CTI Ablation Failure

The endpoint is CTI Ablation Failure. The proportion of CTI-ablated subjects with CTI Ablation Failure in each treatment group will be analyzed descriptively: the number with the endpoint, the number evaluated, the percentage and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_T < P_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

9.3.13 Thermal Group RF Ablation / Cryoablation Effectiveness Assessment

The two control group technologies (radiofrequency ablation and cryoablation) will be evaluated and compared using methods analogous to the primary effectiveness analysis, only comparing RF to Cryo groups rather than Treatment to Control. The endpoint is Treatment Success. The proportion of subjects with Treatment Success in each control subgroup (RF, Cryo) will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $P_{RF} > P_{Cryo}$ will be presented. Exploratory frequentist analyses may also be conducted.

9.3.14 AF Symptom Assessment

The occurrence and number of patient-reported AF Symptoms (such as dizziness, palpitations, rapid heartbeat, dyspnea, fatigue, and syncope) at baseline and 12 months will be reported. For each symptom, the frequency of the symptom at each time point, as well as the proportion of subjects who indicate the presence of the symptom at baseline but not at 12 months, will be summarized with counts, percentages, and 95% Bayesian credible intervals (BCIs) for the proportions within each treatment group as well as the difference in proportions between treatment groups. In addition, the count of reported episodes of each symptom type at each time point, as well as the change in reported counts from baseline to 12 months, will be summarized with means, standards of deviation, and 95% Bayesian credible intervals (BCIs) for the average within each group as well as the difference between groups.

9.3.15 Learning Curve Effectiveness Assessment

This is an analysis of both a per-operator and a per-site learning curve in the Treatment group by enrollment sequence. Patient order will be determined by sequential procedure order by either the operator or the site, determining whether the procedure would be placed in the (1-3) group, (4-6) group, or the (>6) group.

The number of subjects with Acute Procedural Success by experience category, the total number evaluated, the percentage, and the results of the Fisher-Freeman-Halton test for each operator and site will be presented. The analysis will also be conducted using Treatment Success as the endpoint. Supportive analyses of procedural metrics may be performed.

10 Safety Analyses

Unless specifically identified below, all safety analyses will be performed in the Safety Population as the principal analyses. Supportive analyses will be conducted in the Per Protocol (PP) population.

10.1 Primary Safety Analysis

The primary safety endpoint for this study is the Composite Safety Endpoint (CSE) defined in Protocol Section 7.1.

The analysis of the primary safety endpoint is a test of non-inferiority of the event rate at 12 months. The null and alternative hypotheses are:

$$H_0: Q_T \geq Q_C + 0.08 \text{ versus } H_A: Q_T < Q_C + 0.08$$

where Q_T is the CSE proportion in the Treatment Group, and Q_C is the CSE proportion in the Control Group. The differences in the proportion of subjects with one or more CSE events between the Treatment Group and the Control Group will be tested with a Bayesian statistical method that is analogous to that used in the primary effectiveness analysis. The prior distributions for Q_T and Q_C in these calculations are Beta (0.5, 0.5), which is the Jeffreys non-informative prior. Q_T will be determined to be non-inferior to Q_C if it can be established that the posterior probability $\Pr(H_A \mid \text{data}) > \Psi_{saf}$, where $\Psi_{saf} = 0.966$ is a pre-specified threshold value that controls the one-sided type I error rate (under simulation) at level 0.05. In the absence of missing data, the posterior distributions for Q_T and Q_C would be conjugate Beta distributions. However, some missing data are expected. Therefore, Bayesian multiple imputation of 12-month outcomes will be employed in the principal analysis. The imputation model accounts for primary outcome information up through the time of censoring. More detail on this imputation is given in Section 13.2.

10.2 Secondary Safety Endpoint Analyses

10.2.1 PV Dimensional Change: Aggregate PV Cross-Sectional Area

The first secondary safety analysis is an analysis of Aggregate PV Cross-Sectional Area. Aggregate PV Cross-Sectional Area approximates the ability of blood to return to the left atrium.

The cross-sectional area of an individual vein is computed using the formula for area of an ellipse, where the longest and shortest axis measurements of the PV diameter serve as the major and minor axes of the ellipse. "Aggregate PV Cross-Sectional Area" is the within-subject sum of the cross-sectional areas of the following veins:

- Left Superior Pulmonary Vein (LSPV)
- Left Inferior Pulmonary Vein (LIPV)
- Right Superior Pulmonary Vein (RSPV)
- Right Inferior Pulmonary Vein (RIPV)

For subjects with a Left Common Pulmonary Vein (LCPV), the Net PV Cross-sectional Area will be computed using the LCPV area in lieu of the LSPV and LIPV areas. The areas of any Right Middle PVs will be ignored, since they are small, rare, and often not ablated.

This analysis will be conducted on a per-subject basis, using paired (pre- and post-ablation) continuous measurements of Aggregate PV Cross-Sectional area at baseline and 3 months. Within-subject changes in Aggregate PV Cross-Sectional Area (3-month minus baseline) will be compared between Pulsed Field and Thermal groups via a Bayesian version of a t-test, as described in Section 19.1. The null and alternative hypotheses are:

$$H_0\text{: } \mu_T - \mu_C = 0 \text{ versus } H_A\text{: } \mu_T - \mu_C > 0,$$

where μ_T represents the change (3 months minus baseline) in area in the treatment (Pulsed Field) group, and μ_C represents the corresponding change in area in the control (Thermal) group. (If the pulmonary veins tend to narrow more in the Control group than in the Treatment group, the post-minus-pre differences will be "more negative" in the Control group than in the Treatment group, resulting in $\mu_T - \mu_C > 0$.) The null hypothesis will be rejected if the posterior probability $P(H_A \mid \text{data})$ exceeds a threshold value $\Psi_{PV} = 0.975$, corresponding to a one-sided significance level of 0.025.

Exploratory comparisons (e.g., areas and diameters for each vein and for radio frequency vs cryoablation treatment technologies) will also be conducted, and these may use Bayesian or frequentist statistical methods, but these will be outside of the formal familywise error rate control described in Section 6. When diameters are computed, they will be the geometric mean of the longest and shortest axis measurement.

The population for analysis will be MITT Subjects with pre- and post-procedure PV dimensional data related to one or more protocol-specified ablation lesions. Supplemental analyses will be performed using MITT single-procedure subjects and the PP population.

<u>Type I Error Considerations</u>

This analysis is only conducted once (at the final analysis). Therefore, Type I errors can only arise from the test of the model above, without inflation from the sampling plan.

Since the prior distributions for all parameters are uniform, essentially all information in the model is contained within the likelihood. Therefore, a rule that determines $P(H_A \mid \text{data}) > \Psi$ must induce a "critical value" in the data space that is approximately equal to the critical value defined by a $(1-\Psi)$-level significance test in the setting of a standard frequentist hypothesis test, and the type 1 error rate must therefore closely approximate the quantity $1 - \Psi$. This is consistent with the statement of Gelman et al. that "in various simple one-sided hypothesis tests, conventional p-values may correspond with posterior probabilities, under noninformative prior distributions"[12] and the statement of Albert that "a Bayesian probability of a hypothesis is equal to the p-value for one-sided testing problems when a vague prior distribution is placed on the parameter."[13] It follows that using a posterior probability threshold $\Psi_{PV} = 0.975$ for this test will control the type I error rate at the level 0.025.

10.3 Additional Safety Analyses

10.3.1 Severe Ablation Complications

The proportion of subjects in each treatment group with one or more Severe Ablation Complications will be analyzed descriptively: the number with events, the total number of subjects evaluated, the percentage. and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_T < Q_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

10.3.2 Nonserious / Serious CSEs

The proportion of subjects in each treatment group with any device- or procedure-related AE in the CSE definitions of the protocol (without regard for serious or non-serious categorization) will be analyzed descriptively: the number with events, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_T < Q_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

10.3.3 Post-Blanking Cardioversions

The endpoint is Post-Blanking Cardioversions. The proportion of subjects in each treatment group with one or more Post-Blanking Cardioversions will be analyzed descriptively: the number with events, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_T < Q_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

10.3.4 Post-Blanking Arrhythmia Hospitalizations

The endpoint is Post-Blanking Arrhythmia Hospitalizations. The proportion of subjects in each treatment group with one or more Post-Blanking Arrhythmia Hospitalizations will be analyzed descriptively: : the number with events, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_T < Q_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

10.3.5 Any Related SAE

The endpoint is Any Related SAE. The proportion of subjects in each treatment group with one or more Any Related SAEs will be analyzed descriptively: the number with events, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_T < Q_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

10.3.6 Any Related Stroke or TIA

The endpoint is Any Related Stroke or TIA. The proportion of subjects in each treatment group with one or more Related Stroke or TIA will be analyzed descriptively: the number with events, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_T < Q_C$ will be presented. The statistical methods will be a Bayesian comparison of proportions (with predictions) as described in Section 19. Exploratory frequentist analyses may also be conducted.

10.3.7 Categorized PV Dimensional Changes

The endpoint is Categorized PV Dimensional Changes. The distribution of Categorized PV Dimensional Changes between baseline and 3 months in each treatment group will be categorized by treatment group on a per vein basis by maximal % reduction in PV diameters and computed areas into 4 groups: 0-29.9%, 30-49.9%, 50-69.9%, and ≥ 70%).

The number of veins with the amount of change by category, the total number evaluated, the percentage, and the results of the Fisher-Freeman-Halton test will be presented.

10.3.8 Thermal Group RF Ablation / Cryoablation Safety Assessment

The two control group technologies (radio frequency ablation and cryoablation) are to be evaluated and compared using methods analogous to the primary safety analysis, only comparing RF to Cryo groups rather than Treatment to Control. The endpoint is CSE (using the Primary Safety Endpoint Definition). The proportion of subjects with Treatment Success in each control subgroup (RF, Cryo) will be analyzed descriptively: the number with the endpoint, the total number of subjects evaluated, the percentage, and 95% Bayesian credible intervals (BCIs) for the proportion in each group as well as the difference between groups will be presented. The posterior probability that $Q_{RF} > Q_{Cryo}$ will be presented. Exploratory frequentist analyses may also be conducted.

10.3.9 Learning Curve Safety Assessment

This is an analysis of both a per-operator and a per-site learning curve in the Treatment group by enrollment sequence. Patient order will be determined by sequential procedure order by either the operator or the site, determining whether the procedure would be placed in the (1-3) group, (4-6) group, or the (>6) group.

The number of subjects with one or more CSEs by experience category, total number evaluated, the percentage, and the results of the Fisher-Freeman-Halton test for each operator and site will be presented. Supportive analyses of procedural metrics may be performed.

11 Sample Size Estimation

11.1 Safety Estimation

Although the pre-specified statistical analysis methods are Bayesian, standard frequentist considerations are used to guide sample size selection. The sample size for primary safety computed for a Farrington-Manning test with PASS 13[11] indicates that, for 1:1 randomization, 90% power, alpha=0.05 (one-sided), CSE rate in both Groups of 0.08 and non-inferiority margin of 0.08, the study requires a completed sample size of 214 subjects in each Group, for a total of 428. Assuming a 5% rate of loss-to-follow-up (LTFU), 450 subjects are needed.

There is uncertainty in the CSE rate, and this can have dramatic impact on sample size requirements. If the CSE rate is 0.07 in the Pulsed Field group but 0.08 in the Thermal group, 326 subjects are required to achieve 90% power (344 after compensating for 5% LTFU). If the CSE rate is 0.10 in the Pulsed Field group but 0.08 in the Thermal group, 690 subjects are required to achieve 85% power (727 after compensating for 5% LTFU). Interim analyses can also increase the sample size requirement. This uncertainty in the CSE rate motivates the use of a design wherein the sample size is adaptively determined, with possible sample sizes ranging from 350 to 750.

Overall statistical power for the study depends on power for the primary effectiveness test as well as the primary safety test. Power is more thoroughly estimated by simulation and is described in Section 14.

11.2 Effectiveness Estimation

Although the pre-specified statistical methods are Bayesian, standard frequentist considerations are used to guide sample size selection. The sample size required for the primary effectiveness endpoint obtained by PASS 13[11] for 1:1 randomization, 90% power, alpha=0.05 (one-sided), success rates in each group of 0.65 and non-inferiority margin δ=0.15 is 172 subjects in each arm, for a total of 344 subjects. Similar to the safety objective, there is uncertainty in the success rates, especially in the Pulsed Field group. If the success rate in the Pulsed Field group is 0.625 but 0.65 in the Thermal group, 502 subjects are required to achieve 90% power (558 after compensating for up to 10% missing data). If the success rate in the Pulsed Field group is 0.70 but 0.65 in the Thermal group, only 186 subjects are required to achieve 90% power (207 after compensating for missing data). Interim analysis can also increase the sample size requirement. This uncertainty in the Treatment Success rate motivates the use of a design wherein the sample size is adaptively determined, with possible sample sizes ranging from 200 to 550. This overlaps with the range computed for Safety Estimation, but it is expected that the primary safety objective will be the primary driver of sample size; therefore, the minimum possible sample size for this study is set to 350, and the maximum is set to 750.

Overall statistical power for the study depends on power for the primary effectiveness test as well as the primary safety test. Power is more thoroughly estimated by simulation and is described in Section 14.

Number of Enrolled Subjects:

The primary population for the analysis of effectiveness is the MITT Population. The Safety Population should be similar but will not be smaller. Considering the joint sample size needs of the primary safety and effectiveness objectives, the primary effectiveness analysis population may range from a minimum of 350 to a maximum of 750.

To achieve a sample size of 750 MITT Subjects, up to 900 subjects may be enrolled, consisting of 750 MITT and up to 105 Roll-In and up to 45 Randomized Subjects (6%) who do not meet the criteria for the MITT Population. See Section 4.1.

12 Analysis Plan

This trial includes interim analyses for the purpose of determining sample size and possibly stopping enrollment. Once enrollment has stopped, all subjects will be followed for 12 months, and then the primary analysis will occur. There is no interim test of the primary hypotheses.

One blinded administrative analysis will occur when 100 randomized subjects have completed 7 days of follow-up. The analysis of adverse events, Acute Procedural Failure and Chronic Failure will be aggregated across treatment groups and will not examine between-group differences in any fashion. The aggregated results will be used to assess consistency of early outcomes with the parameter assumptions underlying the design of the ADVENT Trial, in support of clinical trial submissions to regulatory authorities in other countries. This analysis is not part of the formal sequential, adaptive plan.

This trial is designed using Bayesian statistical techniques. The appropriate sample size is determined adaptively via a Goldilocks design[15], with possible sample sizes of N=350, 450, 550, 650, and 750.

The sample size determination algorithm is as follows. At a pre-defined interim analysis point (e.g., when $N=N_i$ subjects have been accrued into the MITT population; N_i = 350, 450, 550 and 650), four predictive probabilities are calculated:

$PP_{int,saf}$ = Predictive probability that the trial will establish non-inferiority for safety, once the current sample (N_i) is followed to 12 months.
$PP_{max,saf}$ = Predictive probability that the trial will establish non-inferiority for safety, assuming the maximum sample size (750) is attained and followed to 12 months.
$PP_{int,eff}$ = Predictive probability that the trial will establish non-inferiority for effectiveness, once the current sample (N_i) is followed to 12 months.
$PP_{max,eff}$ = Predictive probability that the trial will establish non-inferiority for effectiveness, assuming the maximum sample size (750) is attained and followed to 12 months.

If $PP_{int,saf}$ exceeds a suitably high threshold $W_{saf,i}$, <u>and</u> if $PP_{int,eff}$ exceeds a suitably high threshold $W_{eff,i}$, accrual of subjects will stop because final success looks probable with the current sample. On the other hand, if $PP_{max,saf}$ is less than a suitably low threshold $F_{saf,i}$, <u>or</u> if $PP_{max,eff}$ is less than a suitably low threshold $F_{eff,i}$, subject accrual will stop because final success is unlikely. In either case, follow-up continues to 12 months for all subjects, at which time the final analysis will occur. If neither of these conditions obtains, enrollment will continue to the next larger sample size (subject to the maximum of 750). These analyses are termed "Interim Sample Size Looks."

Table 1 displays the thresholds against which the predictive probabilities above are compared.

Table 1 Posterior Probability Thresholds for Stopping Accrual

	Effectiveness		**Safety**	
Interim Sample Size Look	**Promise Threshold $W_{eff,i}$**	**Futility Threshold $F_{eff,i}$**	**Promise Threshold $W_{saf,i}$**	**Futility Threshold $F_{saf,i}$**
N=350	0.95	0.05	0.95	0.05
N=450	0.90	0.10	0.90	0.10
N=550	0.85	0.10	0.85	0.10
N=650	0.80	0.10	0.80	0.10

Once subject accrual has stopped, all accrued subjects will be followed to 12 months, and then the inferential tests of non-inferiority described in Sections 9.1 and 10.1 will occur. This is the only time that non-inferiority is tested.

13 Imputations and the Predictive Model

This section describes the prediction model that is used to impute the 12-month outcomes for subjects who have not yet reached 12 months. This model computes Bayesian posterior predictive probabilities and serves as the basis for the predictive probability of future trial success that is computed at the interim sample size looks. With a slight change to the specification of a prior distribution, a nearly identical model serves as the "imputer's model" in a Bayesian multiple imputation that is used to account for missing data in the final inferential tests of the primary endpoints.

13.1 Effectiveness

Outcomes at 12 months are binary and are denoted "E" (for having experienced a primary outcome Event, e.g., AF recurrence) or "N" (for not having experienced an event). At any sample size look, many subjects will not have completed 12 months, and the 12-month outcomes for those subjects will be imputed according to statistical modeling, as described below.

At the time of analysis, and separately for each treatment group, subjects will fall into one of the following categories:

A. Final status (E or N) is known. This can happen because the subject has had a known primary event at any time up through the 12-month assessment. It can also happen if the subject is event-free as of the 12-month assessment.

B. Subject is censored without known event prior to the 12-month assessment. Such subjects will have their 12-month outcome imputed to be E with a probability that is subject-specific and is based on how long that subject has been followed and also on the event-free survival experience of other subjects (in the same treatment group), calculated via a Bayesian piecewise exponential survival model, as described below. Imputed outcomes are based on the subject-specific survival probabilities at Follow-up Day 360.

In this survival model, Time 0 days corresponds to the procedure date. The hazard function h(τ) is piecewise constant over the intervals (0, 90 days], (90, 104 days], (104, 150 days], (150–210 days), and (210 – 360 days]. Justification for this partition of the time axis is given later in this section. Specifically, let the survival function S(t) represent the probability that a subject in this category is event-free at time *t*, where

$$S(t) = e^{-\int_0^t h(\tau)d\tau}, \text{ and}$$

$$h(\tau) = \begin{cases} \lambda_1, & \tau \epsilon \ (0, 90 \ days] \\ \lambda_2, & \tau \epsilon \ (90, 104 \ days] \\ \lambda_3, & \tau \epsilon \ (104, 150 \ days] \\ \lambda_4, & \tau \epsilon \ (150, 210 \ days] \\ \lambda_5, & \tau \epsilon \ (210, 360 \ days] \end{cases}$$

Let $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$. We assign each λ_i (i = 1, 2, 3, 4, 5) an independent Gamma prior distribution, and because of the well-known Poisson-Gamma conjugacy, the posterior distribution f(λ_i | data) also follows a Gamma distribution. Samples from the posterior distributions of λ_1, λ_2, λ_3, λ_4, and λ_5 can be drawn directly, without the need for Markov chain Monte Carlo techniques.

For a subject j censored at $R_j < 360$ days, the probability θ_j that this subject has an event by Day 360 can be computed as

$$\theta_j = \text{Prob}[T \leq 360 \mid T > R_j]$$

$$= \text{Prob}[R_j < T \leq 360] \, / \, \text{Prob}[T > R_j]$$

$$= (S(R_j) - S(360)) \, / \, S(R_j)$$

where $S(t) = \exp(-H(t))$ and $H(t)$ is the cumulative hazard function

$$H(t) = \int_0^t h(\tau) d\tau$$

$$= \begin{cases} \lambda_1 t, & t \in [0, 90] \; days \\ 90\lambda_1 + (t - 90)\lambda_2, & t \in (90, 104] \; days \\ 90\lambda_1 + (104 - 90)\lambda_2 + (t - 104)\lambda_3, & t \in (104, 150] \; days \\ 90\lambda_1 + (104 - 90)\lambda_2 + (150 - 104)\lambda_3 + (t - 150)\lambda_4, & t \in (150, 210] \; days \\ 90\lambda_1 + (104 - 90)\lambda_2 + (150 - 104)\lambda_3 + (210 - 150)\lambda_4 \; + & \\ \qquad + (t - 210)\lambda_5 \, , & t \in (210, 360] \; days \end{cases}$$

The distribution of θ_j can thus be computed from the piecewise constant form of $h(\tau)$ to be a function of Λ and R_j. This can be easily accomplished multiple times by first drawing gamma deviates for λ_1, λ_2, λ_3, λ_4, and λ_5 and subsequently transforming these into samples from the subject-specific distribution of θ_j for subject j. The unobserved final binary event status (E, N) for subject j is then imputed multiple times from a Bernoulli distribution with parameter θ_j (once for each sampled value of θ_j).

Prior Distributions in the Predictive Model

Parameters λ_1, λ_2, λ_3, and λ_4 are independently assigned a Gamma(0.5, 0.001) prior distribution. This approximates the Jeffreys non-informative prior but is proper. Because of Poisson-Gamma conjugacy, the posterior distribution $f(\lambda_i \mid \text{data})$ follows a Gamma($0.5 + D_i$, $0.001 + T_i$) distribution, where D_i is the number of events that occurs in interval i and T_i is the total amount of follow-up over all subjects in interval i (in days).

Parameter λ_5 needs special consideration. At the time of the interim sample size looks, it is expected that there will not be enough follow-up experience in the 210-to-360-day window to reliably estimate the λ_5 parameter without some moderate help from an informative prior. Therefore, λ_5 is assigned a Gamma(5, 10000) prior distribution, equivalent to observing 5 events in 10000 patient-days, or 5 events in about 67 subjects followed for the entirety of the 210-to-360-day interval. The influence of this prior can be understood by recognizing that the ratio $S(360)/S(210) = \exp(-\lambda_5 \times (360-210))$, where $\lambda_5 \sim \text{Gamma}(5,10000)$, as shown in **Figure 1**. In other words, S(360 days) is expected to have decreased to roughly 90–95% of the value at S(210 days).

This is consistent with the results of Fire and Ice[3] in **Figure 2**, where it is observed that $1-S(360) \approx 0.35$ and $1-S(210) \approx 0.29$, leading to $S(360)/S(210) = 0.65/0.71 = 0.916$. Other trials[14,16,17,18] have shown similar trends in the 210-to-360-day window.

Figure 1 Prior distribution for survival decrease between 210 and 360 days: S(360)/S(210)

Exp(–λ_5×(360–210)), where λ_5 ~ Gamma(5, 10000)

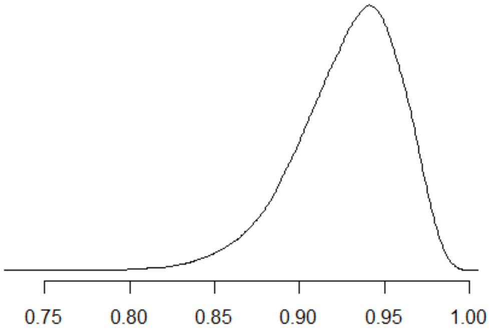
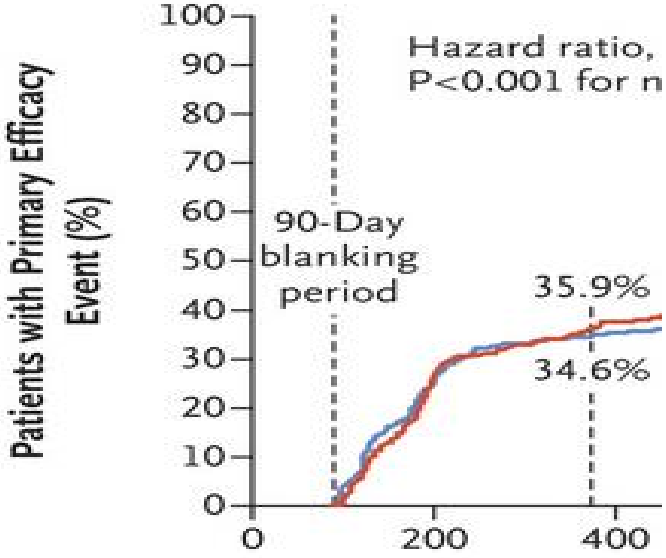


Figure 2 Time to Primary Outcome from Fire and Ice[3]



Rationale for the Time Axis Partition of the Piecewise Exponential Model

The piecewise exponential model has a hazard function that is piecewise constant (or "flat"). With a fine enough partition, a piecewise exponential model can be used to approximate any hazard function[9]. However, to produce stable estimates of hazard parameters, a reasonable number of events must be observed in each "piece" of the partition, so too fine a partition is impractical. The partition (0–90, 90–104, 104–150, 150–210, 210–360 days) is justified primarily by evidence observed in the Fire and Ice trial[3], where the time to primary event is depicted in **Figure 2** (from reference 3), in which it is observed that the shapes of the curves appear to change at approximately 90 days (the end of the blanking period), two weeks later (104 days), and around the 6-month visit window (starting at 180 – 30 = 150 days and ending at 180 + 30 = 210 days). From 210 to 360 days, the curve rises at a reasonably constant exponential rate. A similar assessment of the time-to-event curves from other trials[14,16,17,18] indicates that, although this partition of the time axis may be finer than necessary, it can also appropriately model their respective hazards.

Censoring Procedures

The model in part B, above, is based on censoring or event times up through post-blanking day Day 360. Subjects without a known event will be considered censored at the later of their last Event Monitor or last Holter Monitor. Since the 12-month follow-up window extends to Follow-up Day 390, all events and censoring times > 360 days will be assigned a time of 360 Days when fitting the piecewise exponential models.

Predictive Probabilities and Posterior Inference

At every analysis, the imputations described above are executed M=5000 times, resulting in M completed datasets, filled with only E or N final outcome values. With the Beta(0.5, 0.5) prior distributions specified in Section 9.1, each of the M completed datasets implies a conjugate Beta posterior distribution for both P_T and P_C. From these, the posterior probability $\Pr(P_T > P_C - 0.15 \mid$ m^{th} completed dataset$)$ is computed. This is used in the following ways:

- At each sample size look: The proportion of these M completed datasets that result in $\Pr(P_T > P_C - 0.15 \mid m^{th}$ completed dataset$) > \Psi_{eff}$ is the predictive probability of future success for the effectiveness objective. As described in Section 12, this is used to determine whether to curtail enrollment.

- At the final analysis: For the final analysis, a nearly identical model serves as the "imputer's model" in a Bayesian multiple imputation that is used to account for any missing data in the inferential tests of the primary endpoint. The critical difference in the two models is that, when used for multiple imputation, all parameters of the piecewise exponential model (including λ_5) are assigned non-informative Gamma(0.5, 0.001) prior distributions. Averaging the values $\Pr(P_T > P_C - 0.15 \mid m^{th}$ completed dataset$)$ over the set of M completed datasets integrates over the posterior predictive distribution of imputed values and results in a final (combined) posterior probability $\Pr(P_T > P_C - 0.15 \mid$ data$)$ that is then compared to the threshold Ψ_{eff} to determine whether non-inferiority has been statistically established. This is Bayesian multiple imputation.

  A test for superiority (secondary objective, conducted at the final analysis) is similar but requires computing $\Pr(P_T > P_C + 0 \mid m^{th}$ completed dataset$)$. Averaging these over the set of M completed datasets integrates over the posterior predictive distribution of imputed values and results in a final (combined) posterior probability $\Pr(P_T > P_C \mid$ data$)$ that is compared to the threshold $\Psi_{eff, sup}$ to determine whether superiority has been statistically established. This is Bayesian multiple imputation.

13.2 Safety

Statistical modeling for safety is similar to that used for effectiveness, but the piecewise exponential model is simpler. Outcomes at 12 months are binary and are denoted "E" (for having experienced a primary outcome Event) or "N" (for not having experienced an event). At any sample size look, many subjects will not have completed 12 months, and the 12-month outcomes for those subjects will be imputed according to statistical modeling, as described below.

At the time of analysis, and separately for each treatment group, subjects will fall into one of the following categories:

A. Final status (E or N) is known. This can happen because the subject has had a known primary event at any time up through the 12-month assessment. This can also happen if the subject is event-free as of the 12-month assessment.

B. Subject is censored without known event prior to the 12-month assessment. Such subjects will have their 12-month outcome imputed to be E with a probability that is subject-specific and is based on how long that subject has been followed and also on the event-free survival experience of other subjects (in the same treatment group), calculated via a Bayesian piecewise exponential survival model, as described below. Imputed outcomes are based on the subject-specific survival probabilities at Follow-up Day 360.

In this survival model, Time 0 days corresponds to the procedure date. The hazard function h(τ) is piecewise constant over the intervals (0, 7 days] and (7 – 360 days]. Specifically, let the survival function S(t) represent the probability that a subject in this category is event-free at time *t*, where

$$S(t) = e^{-\int_0^t h(\tau)d\tau}, \text{ and}$$

$$h(\tau) = \begin{cases} \lambda_1, & \tau \in [0, 7\ days] \\ \lambda_2, & \tau \in (7, 360\ days] \end{cases}$$

Let $\Lambda = \{\lambda_1, \lambda_2\}$. We assign λ_1 and λ_2 a Gamma prior distribution, and because of the well-known Poisson-Gamma conjugacy, the posterior distribution f(λ_i | data) also follows a Gamma distribution. Samples from the posterior distributions of λ_1, and λ_2 can be drawn directly, without the need for Markov chain Monte Carlo techniques.

For a subject *j* censored at $R_j < 360$ days, the probability θ_j that this subject has an event by Day 360 can be computed as

$$\theta_j = \text{Prob}[T \leq 360 \mid T > R_j]$$

$$= \text{Prob}[R_j < T \leq 360] \ / \ \text{Prob}[T > R_j]$$

$$= (\ S(R_j) - S(360)\ ) \ / \ S(R_j)$$

where S(t) = exp(-*H(t)*) and *H(t)* is the cumulative hazard function

$$H(t) = \int_0^t h(\tau)d\tau = \begin{cases} \lambda_1 t, & t \in [0, 7]\ days \\ 7\lambda_1 + (t-7)\lambda_2, & t \in (7, 360]\ days \end{cases}$$

The distribution of θ_j can thus be easily computed to be a function of Λ and R_j. This can be easily and quickly accomplished multiple times by first drawing samples from the gamma posterior distribution of λ_1 and λ_2 and subsequently transforming these into samples from the subject-specific distribution of θ_j for subject *j*. The unobserved final binary event status (E, N) for subject *j* is then imputed multiple times from a Bernoulli distribution with parameter θ_j (once for each sampled value of θ_j).

<u>Prior Distributions in the Predictive Model</u>

Parameters λ_1 and λ_2 are independently assigned a Gamma(0.5, 0.001) prior distribution. This approximates the Jeffreys non-informative prior but is proper. Because of the well-known

Poisson-Gamma conjugacy, the posterior distribution f(λ_i | data) follows a Gamma(0.5 + D_i, 0.001 + T_i) distribution, where D_i is the number of events that occurs in interval *i* and T_i is the total amount of follow-up over all subjects in interval *i* (in days).

Rationale for the Time Axis Partition of the Piecewise Exponential Model

The partition of the time axis is motivated by the nature of the endpoint; by definition, most of the component events that make up the composite safety event (CSE) occur within 7 days. Late onset components can happen between 7 days and 360 days, but these are expected to be rare.

Censoring Procedures

The model in part B, above, is based on censoring or event times up through Follow-up Day 360. Subjects without a known event will be considered censored at the last scheduled or unscheduled case report form (CRF) date. When fitting the piecewise exponential models, all events and censoring times > 360 days will be assigned a time of 360 Days.

Predictive Probabilities and Posterior Inference

At every analysis, the imputations described above are executed M=5000 times, resulting in M completed datasets, filled with only E or N final outcome values. With the Beta(0.5, 0.5) prior distributions specified in Section 10.1, each of the M completed datasets implies a conjugate Beta posterior distribution for both Q_T and Q_C. From these, the posterior probability Pr($Q_T < Q_C + 0.08$ | m^{th} completed dataset) is computed. This is used in the following ways:

- At each sample size look: The proportion of these M completed datasets that results in Pr($Q_T < Q_C + 0.08$ | m^{th} completed dataset) > Ψ_{saf} is the predictive probability of future success for the safety objective. As described in Section 12, this is used to determine whether to curtail enrollment.

- At the final analysis: For the final analysis, an identical model serves as the "imputer's model" in a Bayesian multiple imputation that is used to account for any missing data in the inferential tests of the primary endpoint. Averaging the values Pr($Q_T < Q_C + 0.08$ | m^{th} completed dataset) over the set of M completed datasets integrates over the posterior predictive distribution of imputed values and results in a final (combined) posterior probability Pr($Q_T < Q_C + 0.08$ | data) that is compared to the threshold Ψ_{saf} to determine whether non-inferiority has been statistically established. This is Bayesian multiple imputation.

14 Operating Characteristics of the Design

The design described above has been subjected to extensive simulation in order to evaluate its anticipated performance in a variety of scenarios. In the simulation, power is determined as the proportion of simulated trials that result in a simultaneous declaration of non-inferiority for both effectiveness and safety at the final analysis. All results are based on 10,000 simulated trials (per scenario studied), with predictive probabilities and multiple imputation implemented using M=5000 observations from the relevant posterior predictive probability distributions.

Table *2* displays the power and sample size behavior of the design in the specific case that $P_T = P_C = 0.65$ and $Q_T = Q_C = 0.08$. Of the 10,000 trials in this simulation, 1900 simulated trials stopped enrolling at the

N=350 analysis because the interim results were promising (i.e., the predictive probabilities of future success for both effectiveness and safety exceeded their respective thresholds); 40 trials stopped enrolling at the N=350 analysis for futility (i.e., at least one of the predictive probabilities of future success, for efficacy or for safety, was less than its respective threshold). Of the 8060 trials that did not stop at the N=350 analysis, 3523 stopped enrolling for promising trend at the N=450 analysis, while 69 stopped enrolling for futility. The table shows other possibilities, including the fact that 813 of the 10,000 trials enrolled the full maximum of N=750. The final column ("SSavg") indicates the expected sample size to be 504 subjects, calculated as the weighted average of the various sample sizes, weighted by their frequency of occurrence. Of all 10,000 trials, 9635 ended in a simultaneous determination of non-inferiority for both effectiveness and safety, so the estimated power for this scenario is 0.9635.

Table 2 Operating Characteristics for Target Scenario, assuming PC = 0.650, QC = 0.08.

			Probability of Stopping Enrollment at …					
P_T	Q_T	Power	N=350	N=450	N=550	N=650	N=750	SSavg
0.650	0.080	0.9635	0.1900	0.3523	0.2402	0.1189	0.0813	504
			0.0040	0.0069	0.0032	0.0032		

In the columns labeled N=350, N=450, N=550, and N=650, the upper entry for each scenario is the estimated probability that the trial will stop enrolling at that size for promising trends; the lower entry for each scenario is the estimated probability that the trial will stop enrolling at that size for futility. All estimates are based on 10,000 simulated trials.

Table 3 displays similar operating characteristics as the values P_T and Q_T assume different values—some more optimistic than those in **Table *2***, and some less so. Power is seen to range from 0.7891 in the first row (most pessimistic scenario), with the largest average sample size, to 0.9926 in the last row (most optimistic), with the smallest average sample size. Overall, the design displays good power for scenarios of interest, and the sample size determination algorithm responds appropriately as it learns the strength of the accruing evidence.

Table 3 Operating Characteristics for Various Scenarios, assuming P_C = 0.650, Q_C = 0.08.

Scenario			Probability of Stopping Enrollment at					
P_T	Q_T	Power	N=350	N=450	N=550	N=650	N=750	SSavg
0.625	0.100	0.7891	0.0648	0.1917	0.2109	0.1729	0.2454	578
			0.0240	0.0398	0.0261	0.0244		
0.650	0.100	0.8270	0.0937	0.2380	0.2342	0.1528	0.1781	552
			0.0247	0.0352	0.0209	0.0224		
0.700	0.100	0.8405	0.1778	0.2941	0.1870	0.1144	0.1280	516
			0.0247	0.0285	0.0237	0.0218		
0.625	0.080	0.9302	0.1246	0.2810	0.2529	0.1566	0.1567	542
			0.0059	0.0095	0.0074	0.0054		
0.650	0.080	0.9635	0.1900	0.3523	0.2402	0.1189	0.0813	504
			0.0040	0.0069	0.0032	0.0032		
0.700	0.080	0.9767	0.3352	0.3916	0.1685	0.0615	0.0322	456
			0.0031	0.0038	0.0020	0.0021		
0.625	0.070	0.9417	0.1594	0.3052	0.2323	0.1403	0.1425	529
			0.0042	0.0074	0.0047	0.0040		
0.650	0.070	0.9820	0.2391	0.3620	0.2241	0.1061	0.0604	488
			0.0022	0.0032	0.0016	0.0013		
0.700	0.070	0.9926	0.4339	0.3855	0.1295	0.0374	0.0108	430
			0.0005	0.0012	0.0007	0.0005		

In the columns labeled N=350, N=450, N=550, and N=650, the upper entry for each scenario is the estimated probability that the trial will stop enrolling at that size for promising trends; the lower entry for each scenario is the estimated probability that the trial will stop enrolling at that size for futility. All estimates are based on 10,000 simulated trials (per scenario).

The Type I error rate of the design is controlled (according to the principle of intersection-union testing) by ensuring that the type I error rate of each co-primary objective is at the level 0.05, regardless of whether the other co-primary objective concludes non-inferiority. For effectiveness, this is estimated as the proportion of simulated trials that result in a (false) determination of non-inferiority for effectiveness when $P_C = 0.65$ and $P_T = P_C - 0.15$. For safety, this is estimated as the proportion of simulated trials that result in a (false) determination of non-inferiority for safety when $Q_C = 0.08$ and $Q_T = Q_C + 0.08$. These simulations were conducted using 10,000 simulated trials per scenario, with predictive probabilities and multiple imputation implemented using M=5000 observations from the relevant posterior predictive probability distributions.

Table *4* displays the type I error and sample size behavior of the design. The first scenario is for the specific case that $P_T = P_C - 0.15$ (and assuming that the safety sample size assessments would always indicate stopping enrollment for promise and the final safety analysis would always end by demonstrating non-inferiority). Here it is seen that the observed type I error rate for effectiveness is 0.0487, and the sample size algorithm tends to curtail enrollment early for futility (1731 of 10,000 trials curtailed enrollment for futility at N=350, 2409 of 10,000 trials curtailed enrollment for futility at N=450, etc.). The second scenario is for the specific case $Q_T = Q_C + 0.08$ (and assuming that the effectiveness sample size assessments would always indicate stopping enrollment for promise and the final effectiveness analysis would always end by demonstrating non-inferiority). Here the observed type I error rate for safety is 0.0528, and once again the sample size algorithm tends to curtail enrollment early for futility (4394 of 10,000 trials curtailed for futility at N=350, 2554 curtailed enrollment for futility at N=450, etc.). In both cases, the type I error rate is within the range of what is to be expected in a simulation of n=10,000 iterations, calculated as a 95% probability interval $0.05 \pm 1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043 = (0.0457, 0.0543)$. Therefore, the type I error rate of this design is considered controlled at the level 0.05.

Table 4 Operating Characteristics for Null Scenarios, assuming $P_C = 0.650$, $Q_C = 0.08$

Null Scenario			Probability of Stopping Enrollment at					
P_T	Q_T	Power	N=350	N=450	N=550	N=650	N=750	SSavg
0.500	N/A	0.0487	0.0077	0.0150	0.0149	0.0146	0.2445	551
			0.1731	0.2409	0.1662	0.1231		
N/A	0.160	0.0528	0.0088	0.0127	0.0120	0.0119	0.0735	457
			0.4394	0.2554	0.1119	0.0744		

In the columns labeled N=350, N=450, N=550, and N=650, the upper entry for each scenario is the estimated probability that the trial will stop enrolling at that size for promising trends; the lower entry for each scenario is the estimated probability that the trial will stop enrolling at that size for futility. All estimates are based on 10,000 simulated trials (per scenario).

15 Simulation Details and Sensitivity

In order to generate data for the study of operating characteristics, several assumptions about the data mechanism must be made. These assumptions are intrinsic to the generation of artificial data for the simulation exercise but are not part of the analysis that is conducted on any data (real or simulated). This section describes these assumptions and also the sensitivity of the design to many of these assumptions.

15.1 Independence of Safety Events and Effectiveness Events

Event incidence and timing for Primary Effectiveness Events and Primary Safety Events are assumed to be independent processes and are generated independently of each other. However, as described below, the missingness mechanism and follow-up censoring mechanism applied to those data are related, since LTFU applies at the level of the subject rather than the data point.

15.2 Event Times for Effectiveness Events

Simulated effectiveness data are generated via a time-to-event mechanism using a piecewise exponential model. This mechanism assigns Time 0 to be the day of index procedure and uses a piecewise constant hazard for the generation of event times. For the purpose of assessing robustness to the form of this hazard function, three different functions are considered. Operating characteristics resulting from these three variants are shown in Section 15.8.

Method ***Eff_1***

The default piecewise-constant hazard for generation of simulated data is

$$h(\tau) = \begin{cases} \lambda_1, & \tau \epsilon\ (0, 90\ days] \\ \lambda_2, & \tau \epsilon\ (90, 104\ days] \\ \lambda_3, & \tau \epsilon\ (104, 150\ days] \\ \lambda_4, & \tau \epsilon\ (150, 210\ days] \\ \lambda_5, & \tau \epsilon\ (210, 360\ days] \end{cases}$$

The partition of the time axis is motivated by results from the Fire and Ice trial[3] (see **Figure 2** on page 24) and aligns with the partition used in the analysis model of Section 13.1.

Let the vector $\Lambda = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\}$. The values of Λ for data generation are $\lambda_1 = 0.00011167$, $\lambda_2 = 0.002197976$, $\lambda_3 = 0.003163208$, $\lambda_4 = 0.002839089$, and $\lambda_5 = 0.000494053$. These values are thought to reasonably represent what will occur in this study in that they induce an event rate of 0.35 at 360 days, which is evidenced by the fact that

$$S(360\text{ days}) = \exp(-1\times\{\lambda_1(90–0) + \lambda_2(104–90) + \lambda_3(150-104) + \lambda_4(210–150) + \lambda_5(360–210)\})$$

$$= 0.65$$

and also S(90 days) = 0.99, S(104 days) = 0.96, S(150 days) = 0.83, S(210 days) = 0.70. (These values approximate the complements of the survival rates observed in **Figure 2**.)

In order to generate data with varying values of S(360), we scale the vector Λ by a constant k, noting that

$$S(360\text{ days}) = \exp(-k\times\{\lambda_1(90–0) + \lambda_2(104–90) + \lambda_3(150-104) + \lambda_4(210–150) + \\ + \lambda_5(360–210)\})$$

and for any desired value of S(360), such as 0.70, 0.625, 0.60, or 0.50, we can algebraically solve for k and subsequently generate time-to-event data according to a piecewise exponential distribution with this new vector of parameters $k\Lambda$. In this manner, time-to-event data are generated for every desired P_T and P_C value in the tables of operating characteristics. Notably, this data generation mechanism assumes proportionality of hazards between the treatment groups, but the predictive model in the analysis (described in Section 13.1) models the treatment groups separately and does not impose this assumption.

The operating characteristics shown in Section 14 were computed using data generated by Method Eff_1.

***Method Eff*$_2$**

This alternative piecewise-constant hazard for generation of simulated data is

$$h(\tau) = \begin{cases} \lambda_1, & \tau \epsilon \ (0, 90\ days] \\ \lambda_2, & \tau \epsilon \ (90, 100\ days] \\ \lambda_3, & \tau \epsilon \ (100, 200\ days] \\ \lambda_4, & \tau \epsilon \ (200, 300\ days] \\ \lambda_5, & \tau \epsilon \ (300, 360\ days] \end{cases}$$

Unlike the partition of Method *Eff*$_1$, the cutpoints of this partition do not align with the cutpoints of the partition used in data analysis (Section 13.1).

The values of Λ for data generation are λ_1 = 0.000055695, λ_2 = 0.010034797, λ_3 = 0.001690763, λ_4 = 0.000680535, and λ_5 = 0.000573122. These values induce S(t) values as follows: S(90 days) = 0.995, S(100 days) = 0.90, S(200 days) = 0.76, S(300 days) = 0.71, and S(360 days) = 0.686. This choice of values is motivated by results from the Toccastar[14] trial. As in Method *Eff*$_1$, the vector Λ is scaled by a factor *k* to produce any desired value for S(360 days).

***Method Eff*$_3$**

This alternative piecewise-constant hazard for generation of simulated data is

$$h(\tau) = \begin{cases} \lambda_1, & \tau \epsilon \ (0, 90\ days] \\ \lambda_2, & \tau \epsilon \ (90, 120\ days] \\ \lambda_3, & \tau \epsilon \ (120, 360\ days] \end{cases}$$

Importantly, and unlike the partition of Method *Eff*$_1$, the cutpoints of this partition and the number of intervals in this partition do not align with the partition used in data analysis (Section 13.1). The values of Λ for data generation are λ_1 = 0.000569925, λ_2 = 0.007879626, and λ_3 = 0.000596254. These values induce S(t) values as follows: S(90 days) = 0.95, S(120 days) = 0.75, and S(360 days) = 0.65. This scenario reflects a higher incidence of events during the blanking period.

As in Method *Eff*$_1$, the vector Λ is scaled by a factor *k* to produce any desired value for S(360 days).

15.3 Event Times for Primary Safety Events

Simulated safety data are generated via a time-to-event mechanism using a piecewise exponential model. This mechanism assigns Time 0 to be the day of index procedure and uses a piecewise-constant hazard for the generation of event times. For the purpose of assessing robustness to the form of this hazard function, two different functions are considered. Operating characteristics resulting these two variants are shown in Section 15.8.

***Method Saf*$_1$**

This mechanism assigns Time 0 to be the day of index procedure, and the default scenario uses a piecewise-constant hazard defined as

$$h(\tau) = \begin{cases} \lambda_1, & \tau \in [0, 7\ days] \\ \lambda_2, & \tau \in (7, 360\ days] \end{cases}$$

The partition of the time axis is motivated by the nature of the endpoint; by definition, most of the component events that make up the composite safety event (CSE) occur within 7 days. Late onset components can happen between 7 days and 360 days, but these are expected to be rare.

The values for λ_1 and λ_2 for data generation are $\lambda_1 = 0.011137363$ and $\lambda_2 = 1.53540\times10^{-5}$. Unadjusted, these induce values of S(7 days) = 0.925 and S(360 days) = 0.920, corresponding to CSE event rates of 7.5% at 7 days and 8.0% at 360 days. In order to generate data with varying values of S(360), we scale the vector Λ by a constant k, noting that

$$\text{S(360 days)} = \exp(-k\times\{\ \lambda_1(7–0) + \lambda_2(360–7)\})$$

and for any desired value of S(360), we can algebraically solve for k and then generate time-to-event data according to a piecewise exponential distribution with this new vector of parameters $k\Lambda$. In this manner, time-to-event data are generated for every desired Q_T and Q_C value in the tables of operating characteristics. Notably, this data generation mechanism assumes proportionality of hazards between the treatment groups, but the predictive model in the analysis (described in Section 13.2) does not impose this assumption.

The operating characteristics shown in Section 14 were computed using data generated by Method Saf_1.

Method Saf_2

This alternative piecewise-constant hazard for generation of simulated data is

$$h(\tau) = \begin{cases} \lambda_1, & \tau \epsilon (0, 7\ days] \\ \lambda_2, & \tau \epsilon (7, 180\ days] \\ \lambda_3, & \tau \epsilon (180, 360\ days] \end{cases}$$

This partition includes one more cutpoint than is in Method Saf_1. The values of Λ for data generation are $\lambda_1 = 0.007327613$, $\lambda_2 = 0.000122991$, and $\lambda_3 = 6.00610\times10^{-5}$. These values induce S(t) values as follows: S(7 days) = 0.95, S(180 days) = 0.93, and S(360 days) = 0.92, corresponding to CSE event rates of 5%, 7%, and 8%, respectively.

As in Method Saf_1, the vector Λ is scaled by a factor k to produce any desired value for S(360 days).

15.4 Censoring Mechanism

The censoring mechanism applied to these simulated event times is induced by the accrual rates, the timing of analyses, and by a random dropout pattern. For every simulated subject, a study start time is assigned according to an assumed accrual pattern (described below), and when an analysis occurs, each simulated subject's event time is censored at that time (if an event has not previously occurred for that subject). In addition, a designated fraction of simulated subjects is randomly chosen to be lost to follow-up at random times, and these subjects are censored at the minimum of the LTFU time and the analysis time (assuming they have not yet had an event). For the subsequent analyses, including the final analysis, censoring times are simply re-computed based on the new time of analysis, and some simulated subjects who did not have an event at one analysis may have an event at a subsequent analysis.

15.5 Lost to Follow-up / Data Missingness

For the safety assessment, simulations were conducted under the assumption that 5% of subjects were lost to follow-up (LTFU) over the 1-year study period. For effectiveness assessment, it was assumed that 7.5% (the same 5% that are lost for safety, plus 2.5% more) are assumed lost, reflecting the possibility that some subjects may be unassessable for effectiveness but still be assessable for safety. In each mechanism, those subjects assumed to be LTFU were assumed to be lost uniformly in time over the 1-year period. Subsequently, subjects were assigned event times. If the event time occurred after the LTFU time, that subject was considered censored at the LTFU time, whereas if the event time occurred before the LTFU time, that subject had an observed event at the event time.

15.6 Control Arm Rates

The operating characteristics in Section 14 assumed that the effectiveness event-free rate in the control group is $P_C = 0.65$, while the safety CSE rate in the control group is $Q_C = 0.08$. Other scenarios, where $P_C \in \{0.60, 0.65, 0.70\}$ and $Q_C \in \{0.06, 0.08, 0.10\}$, in all combinations, were also studied. Results of simulating these scenarios are shown in Section 15.8.

15.7 Rate of Accrual

The speed at which subjects are recruited into the trial is another important assumption embedded within the simulations to determine operating characteristics. The expected rate of accrual (into the primary analysis population) is 33 subjects/month, after an initial ramp-up period for the first 7 months of 2, 5, 10, 15, 20, 25, and 30 subjects/month. The operating characteristics of Section 14 were generated using this assumption.

If enrollment is faster or slower than this assumed rate, the amount of information available at the interim analysis is affected, and thus the operating characteristics could also be affected. Two alternate accrual patterns were also studied:

- Slower accrual assumes 25 subjects/month (steady state) after an initial ramp-up period for the first 7 months of 2, 5, 10, 10, 15, 15, and 20 subjects/month.
- Faster accrual assumes 50 subjects/month (steady state) after an initial ramp-up period for the first 6 months of 5, 10, 20, 25, 30, and 40 subjects/month.

Operating characteristics that result as the rate of accrual varies are shown in Section 15.8.

15.8 Sensitivity of Operating Characteristics

15.8.1 Operating Characteristics when Varying the Rate of Accrual

Table *5* displays operating characteristics when accrual follows the slower pattern (25/month after ramp-up), while **Table *6*** displays operating characteristics when accrual follows the faster pattern (50/month after ramp-up). This is analogous to

Table 3 in Section 14, which followed the expected pattern (33/month after ramp-up). As with **Table 3**, each row of these tables is based on 10,000 simulated clinical trials, with P_C = 0.650 Q_C = 0.080. The hazard functions used for data generation in all three tables were Method Eff_1 (for effectiveness) and Method Saf_1 (for safety).

Comparing **Table 5**,

Table 3, and **Table *6***, it is evident that the power does not meaningfully differ as the enrollment rates change, but the average sample size is smallest with the slow enrollment rate, since the design tends to stop at smaller sizes when there is more follow-up information available at the sample size determination analyses.

Table 5 Operating Characteristics for Various Scenarios, Slow Enrollment Rate (P_T = 0.650, Q_C = 0.08)

Scenario			Probability of Stopping Enrollment at …					
P_T	Q_T	Power	N=350	N=450	N=550	N=650	N=750	SSavg
0.625	0.100	0.7890	0.1007	0.2113	0.1996	0.1549	0.2145	560
			0.0295	0.0396	0.0248	0.0251		
0.650	0.100	0.8254	0.1570	0.2496	0.2026	0.1303	0.1547	532
			0.0247	0.0335	0.0226	0.0250		
0.700	0.100	0.8398	0.2528	0.2622	0.1630	0.1033	0.1197	502
			0.0214	0.0316	0.0224	0.0236		
0.625	0.080	0.9288	0.1839	0.2816	0.2291	0.1372	0.1378	524
			0.0072	0.0114	0.0065	0.0053		
0.650	0.080	0.9672	0.2808	0.3377	0.2040	0.0987	0.0630	481
			0.0049	0.0054	0.0025	0.0030		
0.700	0.080	0.9772	0.4681	0.3212	0.1273	0.0457	0.0259	433
			0.0028	0.0037	0.0026	0.0027		
0.625	0.070	0.9356	0.2266	0.3015	0.2049	0.1221	0.1201	509
			0.0061	0.0100	0.0045	0.0042		
0.650	0.070	0.9809	0.3327	0.3460	0.1850	0.0825	0.0471	466
			0.0017	0.0025	0.0015	0.0010		
0.700	0.070	0.9945	0.5586	0.3139	0.0917	0.0257	0.0074	411
			0.0008	0.0009	0.0008	0.0002		

In the columns labeled N=350, N=450, N=550, and N=650, the first entry for each scenarios is the estimated probability that the trial will stop enrolling at that size for promising trends; the lower entry for each scenario is the estimated probability that the trial will stop enrolling at that size for futility. All estimates are based on 10,000 simulated trials (per scenario).

Table 6 Operating Characteristics for Various Scenarios, Fast Enrollment Rate (P_T = 0.650, Q_C = 0.08)

Scenario			Probability of Stopping Enrollment at …					
P_T	Q_T	Power	N=350	N=450	N=550	N=650	N=750	SSavg
0.625	0.100	0.7887	0.0283	0.1415	0.2062	0.1993	0.3116	606
			0.0258	0.0379	0.0235	0.0259		
0.650	0.100	0.8197	0.0399	0.1823	0.2404	0.1973	0.2377	586
			0.0205	0.0338	0.0254	0.0227		
0.700	0.100	0.8366	0.0750	0.2678	0.2494	0.1480	0.1686	552
			0.0188	0.0299	0.0213	0.0212		
0.625	0.080	0.9183	0.0592	0.2131	0.2702	0.1983	0.2261	580
			0.0077	0.0134	0.0072	0.0048		
0.650	0.080	0.9651	0.0842	0.2944	0.2997	0.1784	0.1257	545
			0.0043	0.0066	0.0037	0.0030		
0.700	0.080	0.9769	0.1541	0.4089	0.2746	0.1045	0.0463	497
			0.0021	0.0051	0.0021	0.0023		
0.625	0.070	0.9299	0.0770	0.2435	0.2656	0.1844	0.2020	567
			0.0094	0.0089	0.0048	0.0044		
0.650	0.070	0.9790	0.1155	0.3134	0.2908	0.1627	0.1078	532
			0.0039	0.0035	0.0018	0.0006		
0.700	0.070	0.9953	0.2095	0.4486	0.2489	0.0706	0.0188	474
			0.0014	0.0009	0.0007	0.0006		

In the columns labeled N=350, N=450, N=550, and N=650, the first entry for each scenarios is the estimated probability that the trial will stop enrolling at that size for promising trends; the lower entry for each scenario is the estimated probability that the trial will stop enrolling at that size for futility. All estimates are based on 10,000 simulated trials (per scenario).

15.8.2 Type I Error Rates for Effectiveness (Non-inferiority)

Table *7* displays the type I error rates for the primary effectiveness non-inferiority test under every combination of enrollment rate (slow, expected, fast), control success rate P_C (0.60, 0.65, 0.70), and data generation hazard (Methods Eff_1, Eff_2, and Eff_3). Also displayed is the average achieved sample size. Each of these scenarios is for the case that $P_T = P_C - 0.15$ and assumes that the safety sample size assessments would always stop enrolling for promise and the final safety analysis would always end by demonstrating non-inferiority. Each of the 27 scenarios is the result of running 10,000 simulated clinical trials. Among the 27 estimated type I error rates, the mean value is 0.0495, and the max is 0.0534. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.05 $\pm 1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043$, with upper bound 0.0543.

Average sample sizes tend to increase with increasing enrollment rate but do not meaningfully change as the other parameters are varied.

Table 7 Type I Error Rates for Primary Effectiveness Non-inferiority Test as Data Generation Parameters Vary (Safety data always indicate sample size stop)

Data Generation		Type I Error Rates (as accrual varies)			Average Sample Size (as accrual varies)		
Method	P_C	Slow	Expected	Fast	Slow	Expected	Fast
Eff_1	0.600	0.0503	0.0477	0.0481	530	555	589
Eff_1	0.650	0.0472	0.0487	0.0497	530	551	590
Eff_1	0.700	0.0491	0.0482	0.0484	531	555	599
Eff_2	0.600	0.0529	0.0534	0.0479	537	552	593
Eff_2	0.650	0.0498	0.0502	0.0441	531	557	596
Eff_2	0.700	0.0478	0.0515	0.0518	534	559	602
Eff_3	0.600	0.0526	0.0484	0.0503	518	537	575
Eff_3	0.650	0.0515	0.0472	0.0514	520	538	579
Eff_3	0.700	0.0497	0.0494	0.0482	519	544	585

All estimates are based on 10,000 simulated trials (per scenario).

The results in **Table** *7* are computed under the assumption that the primary safety objective results are so strongly favorable that the safety portion of the sample size assessments would always indicate a stop for promising trend. Table 8 presents results when this assumption is altered so that the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, as described in Section 12, and primary safety data are generated under the expected safety assumptions that $Q_T = Q_C = 0.080$ and Method Saf_1. Conservatively, it is further assumed that the safety endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case.

Table 8 Type I Error Rates for Primary Effectiveness Non-inferiority Test as Data Generation Parameters Vary (Safety data generated using $Q_T = Q_C = 0.080$, Method Saf_1)

Data Generation		Type I Error Rates (as accrual varies)			Average Sample Size (as accrual varies)		
Method	P_C	Slow	Expected	Fast	Slow	Expected	Fast
Eff_1	0.600	0.0473	0.0501	0.0454	533	555	594
Eff_1	0.650	0.0476	0.0511	0.0485	532	555	593
Eff_1	0.700	0.0534	0.0452	0.0480	530	552	601
Eff_2	0.600	0.0524	0.0488	0.0473	533	559	596
Eff_2	0.650	0.0509	0.0474	0.0444	535	555	598
Eff_2	0.700	0.0501	0.0477	0.0474	536	559	601
Eff_3	0.600	0.0494	0.0476	0.0480	520	541	577
Eff_3	0.650	0.0508	0.0511	0.0469	517	538	581
Eff_3	0.700	0.0476	0.0522	0.0453	521	544	585

All estimates are based on 10,000 simulated trials (per scenario).

Each of the 27 estimated Type I error rates in Table 8 is the result of running 10,000 simulated clinical trials. Among the 27 estimated type I error rates, the mean value is 0.0486, and the max is 0.0534. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval $0.05 \pm 1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043$, with upper bound 0.0543.

Although Table *7* and Table 8 generate data with varying methods, the method applied for the Treatment Group is the same as that applied for the Control Group. The below Table 9 presents a study of effectiveness type I error rates when the Treatment and Control groups are generated with differing methods. Specifically, 9 null scenarios of 10,000 iterations each were simulated, systematically and separately varying data generation methods $\{Eff_1, Eff_2, Eff_3\}$ in the Treatment group and $\{Eff_1, Eff_2, Eff_3\}$ in the Control group, under the null assumption of 50% success rate in the Treatment group and 65% success rate in the Control group. Safety data in these scenarios were generated using method Saf_1 for both treatment groups, with event rates of 8% in each group. As in Table 8, the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, but it is conservatively assumed that the safety endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case. Enrollment followed the Expected rate, as described in Section 15.7.

Table 9 Type I Error Rates for Effectiveness Non-Inferiority as data generation methods vary (when $P_T = 0.65 – 0.15$, $P_C = 0.65$)

Data Generation Method	*Control Group*		
Treatment Group	***Eff_1***	***Eff_2***	***Eff_3***
Eff1	0.0511	0.0466	0.0497
Eff2	0.0509	0.0474	0.0507
Eff3	0.0476	0.0476	0.0511

All estimates are based on 10,000 simulated trials (per scenario).

In Table 9, the maximum value is 0.0511. In Table 9 (as well as Table *7* and Table 8), no estimated type I error rates exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.05 ± 1.96×√(0.05×0.95/10000) = 0.05 ± 0.0043, with upper bound 0.0543. Therefore, the type I error rate of the effectiveness test is considered controlled at the level 0.05.

15.8.3 Type I Error Rates for Safety (Non-inferiority)

Table 10 displays the type I error rates for the primary safety non-inferiority test under every combination of enrollment rate (slow, expected, fast), control CSE rate Q_C (0.07, 0.08, 0.10), and data generation hazard (Methods *Saf_1* and *Saf_2*). Also displayed is the average achieved sample size. Each of these scenarios is for the case that $Q_T = Q_C + 0.08$ and assumes that the effectiveness sample size assessments would always stop enrolling for promise and the final effectiveness analysis would always end by demonstrating non-inferiority. Each of the 18 scenarios is the result of running 10,000 simulated clinical trials. Among the 18 estimated type I error rates, the mean value is 0.0453, and the max is 0.0540. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.05 ± 1.96×√(0.05×0.95/10,000) = 0.05 ± 0.0043, with upper bound 0.0543.

Table 10 Type I Error Rates for Primary Safety Non-inferiority Test as Data Generation Parameters Vary (Effectiveness data always indicate sample size stop)

Data Generation		Type I Error Rates (as accrual varies)			Average Sample Size (as accrual varies)		
Method	**Q_C**	**Slow**	**Expected**	**Fast**	**Slow**	**Expected**	**Fast**
Saf_1	0.100	0.0539	0.0491	0.0429	447	455	467
Saf_1	0.080	0.0507	0.0528	0.0442	449	457	465
Saf_1	0.060	0.0540	0.0515	0.0486	448	456	469
Saf_2	0.100	0.0407	0.0412	0.0378	468	475	498
Saf_2	0.080	0.0433	0.0414	0.0427	463	476	498
Saf_2	0.060	0.0438	0.0385	0.0379	465	475	495

All estimates are based on 10,000 simulated trials (per scenario).

The results in ***Table 10*** are computed under the assumption that the primary effectiveness objective results are so strongly favorable that the effectiveness portion of the sample size assessments would always indicate a stop for promising trend. **Table 11** presents results when this assumption is altered so that the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, as described in Section 12, and primary effectiveness data are generated under the expected effectiveness assumptions that $P_T = P_C = 0.650$ and Method *Eff_1*. Conservatively, it is further assumed that the effectiveness endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case.

Table 11 Type I Error Rates for Primary Safety Non-inferiority Test as Data Generation Parameters Vary (Effectiveness data generated using $P_T = P_C = 0.65$, Method *Eff*$_1$)

Data Generation		Type I Error Rates (as accrual varies)			Average Sample Size (as accrual varies)		
Method	Q_C	Slow	Expected	Fast	Slow	Expected	Fast
Saf$_1$	0.100	0.0500	0.0479	0.0405	451	460	470
Saf$_1$	0.080	0.0473	0.0449	0.0454	453	458	470
Saf$_1$	0.060	0.0484	0.0446	0.0418	451	456	470
Saf$_2$	0.100	0.0406	0.0405	0.0364	465	477	498
Saf$_2$	0.080	0.0401	0.0376	0.0337	451	460	470
Saf$_2$	0.060	0.0413	0.0333	0.0341	453	458	470

All estimates are based on 10,000 simulated trials (per scenario).

Each of the 18 estimated Type I error rates in **Table 11** is the result of running 10,000 simulated clinical trials. Among the 18 estimated type I error rates, the mean value is 0.0416, and the max is 0.0500. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.05 ± 1.96×√(0.05×0.95/10000) = 0.05 ± 0.0043, with upper bound 0.0543.

Although **Table 10** and **Table 11** generate data with varying methods, the method applied for the Treatment Group is the same as that applied for the Control Group. The below **Table 12** presents a study of safety type I error rates when the Treatment and Control groups are generated with different methods. Specifically, 4 null scenarios of 10,000 iterations each were conducted, systematically and separately varying data generations methods {*Saf*$_1$, *Saf*$_2$} in the Treatment group and {*Saf*$_1$, *Saf*$_2$} in the Control group, under the assumption of 16% success rate in the Treatment group and 8% success rate in the Control group. Effectiveness data in these scenarios were generated using method *Eff*$_1$ for both treatment groups, with the expected success rate of 65% in each group.

As in Table 8, the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, but it is conservatively assumed that the effectiveness endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case. Enrollment followed the Expected rate, as described in Section 15.7.

Table 12 Type I Error Rates for Safety Non-Inferiority as data generation methods vary, (when $Q_T = 0.08+0.08$, $Q_C = 0.08$)

Data Generation Method	*Control Group*	
Treatment Group	***Saf1***	***Saf2***
Saf1	0.0449	0.0463
Saf2	0.0380	0.0376

All estimates are based on 10,000 simulated trials (per scenario).

In **Table 12**, the maximum value is 0.0463. In **Table 12** (as well as **Table 10** and **Table 11**), no estimated type I error rates exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.05 ± 1.96×√(0.05×0.95/10000) = 0.05 ± 0.0043, with upper bound 0.0543. Therefore, the type I error rate of the safety test is considered controlled at the level 0.05.

15.8.4 Type I Error Rates for Effectiveness (Superiority)

As described in Sections 6 and 9.2, superiority for the primary effectiveness endpoint will be tested as a secondary objective. The methods are identical to those for the primary effectiveness non-inferiority test (Sections 9.1 and 13.1) except that (1) there is no non-inferiority margin, and (2) the one-sided type I error rate is set at the level 0.025. Superiority is considered established if the posterior probability $\Pr(P_T > P_C \mid \text{data}) > \Psi_{\text{eff,sup}} = 0.977$ at the final analysis.

Table 13 displays the estimated Type I error rates under every combination of enrollment rate (slow, expected, fast), control success rate P_C (0.60, 0.65, 0.70), and data generation hazard (Methods Eff_1, Eff_2, and Eff_3). Each of these scenarios is for the case that $P_T = P_C$ and assumes that the safety sample size assessments would always stop enrolling for promise and the final safety analysis would always end by demonstrating non-inferiority. Furthermore, enrollment stopping is based on the predictive probabilities of demonstrating non-inferiority for this effectiveness endpoint (as specified in Section 12), without regard for any predictive probabilities of demonstrating superiority. Nevertheless, superiority is tested at the final analysis (12 months after enrollment stops).

Each of the 27 scenarios is the result of running 10000 simulated clinical trials. Among the 27 estimated type I error rates, the mean value is 0.0242, and the max is 0.0277. No values exceed the upper bound of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.025 ± 1.96×√(0.025×0.0975/10000) = 0.025 ± 0.0031 = (0.0219, 0.0281).

Table 13 Type I Error Rates for Primary Effectiveness Superiority Test as Data Generation Parameters Vary (Safety data always indicate sample size stop)

Data Generation		Type I Error Rates (as accrual varies)		
Method	**$P_C = P_T$**	**Slow**	**Expected**	**Fast**
Eff_1	0.600	0.0226	0.0227	0.0277
Eff_1	0.650	0.0221	0.0255	0.0248
Eff_1	0.700	0.0232	0.0247	0.0258
Eff_2	0.600	0.0238	0.0238	0.0245
Eff_2	0.650	0.0234	0.0274	0.0250
Eff_2	0.700	0.0224	0.0238	0.0243
Eff_3	0.600	0.0259	0.0245	0.0229
Eff_3	0.650	0.0239	0.0239	0.0245
Eff_3	0.700	0.0222	0.0223	0.0258

All estimates are based on 10,000 simulated trials (per scenario).

The results in **Table 13** are computed under the assumption that the primary safety objective results are so strongly favorable that the safety portion of the sample size assessments would always indicate a stop for promising trend, and the primary safety hypothesis test would always pass its non-inferiority criterion. **Table 14** presents results when this assumption is altered so that the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, as described in Section 12, and primary safety data are generated under the expected safety assumptions that $Q_T = Q_C = 0.080$ and Method Saf_1. Conservatively, it is further assumed that the safety endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case.

Table 14 Type I Error Rates for Primary Effectiveness Superiority Test as Data Generation Parameters Vary (Safety data generated using $Q_T = Q_C = 0.080$, Method Saf_1)

Data Generation		Type I Error Rates (as accrual varies)		
Method	**$P_C = P_T$**	**Slow**	**Expected**	**Fast**
Eff_1	0.600	0.0215	0.0230	0.0265
Eff_1	0.650	0.0221	0.0254	0.0238
Eff_1	0.700	0.0219	0.0248	0.0244
Eff_2	0.600	0.0245	0.0243	0.0227
Eff_2	0.650	0.0211	0.0224	0.0232
Eff_2	0.700	0.0240	0.0237	0.0263
Eff_3	0.600	0.0239	0.0238	0.0228
Eff_3	0.650	0.0225	0.0268	0.0253
Eff_3	0.700	0.0232	0.0234	0.0235

All estimates are based on 10,000 simulated trials (per scenario).

Each of the 27 estimated Type I error rates in **Table 14** is the result of running 10,000 simulated clinical trials. Among the 27 estimated type I error rates, the mean value is 0.0237, and the max is 0.0268. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.025 ± 1.96×√(0.025×0.975/10000) = 0.025 ± 0.0031, with upper bound 0.0281.

Although **Table 13** and **Table 14** generate data with varying methods, the method applied for the Treatment Group is the same as that applied for the Control Group. The below **Table 15** presents a study of type I error rates for the secondary superiority objective when the Treatment and Control groups are generated with different methods. Specifically, 9 null scenarios of 10,000 iterations each were conducted, systematically and separately varying data generations methods {Eff_1, Eff_2, Eff_3} in the Treatment group and {Eff_1, Eff_2, Eff_3} in the Control group, under the superiority null assumption of 65% success rate in each treatment group. Safety data in these scenarios were generated using method Saf_1 for both treatment groups, with expected rates of 8% in each group. As in **Table 14**, the sample size is jointly determined by trends in both the primary effectiveness and safety endpoints, but it is conservatively assumed that the safety objective would always pass its non-inferiority condition once enrollment had stopped. Enrollment followed the Expected rate, as described in Section 15.7.

Table 15 Type I Error Rates for Effectiveness Superiority (when $P_T = P_C = 0.65$), as data generation methods vary

Data Generation Method	*Control Group*		
Treatment Group	***Eff_1***	***Eff_2***	***Eff_3***
Eff1	0.0254	0.0226	0.0278
Eff2	0.0248	0.0224	0.0225
Eff3	0.0232	0.0238	0.0268

All estimates are based on 10,000 simulated trials (per scenario).

In **Table 15**, the maximum value is 0.0278. In **Table 15** (as well as **Table 13** and **Table 14**), no estimated type I error rates exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.025 ± 1.96×√(0.025×0.975/10000) = 0.025 ± 0.0031, with upper bound 0.0281. Therefore, the type I error rate of the secondary effectiveness superiority test is considered controlled at the (one-sided) level 0.025.

16 Planned Sensitivity Analyses

16.1 Missing Data and Planned Sensitivity Analyses (Primary Objectives)

16.1.1 Sensitivity to Missing Data

Every effort will be undertaken to minimize missing data. However, some missingness is inevitable, and the simulations were conducted with the expectation that there may be up to 5% missing data for the primary safety analysis and 7.5% missing data for the primary effectiveness analysis. The reasons for missing data will be described in detail and evaluated for assessment of possible bias. The distribution of prognostic factors between subjects with data and those without data will be examined to evaluate any potential sources of bias.

By design, the analysis of the primary objectives will predict 12-month outcomes for any subjects without measured 12-month outcomes, based on available follow-up. At the final analysis, several additional analyses are planned that will explore the sensitivity of the main results to these predictions. Specifically:

- One analysis will exclude subjects known to be lost to follow-up without an event (or with final effectiveness outcome permanently missing). This is a completers-only analysis.
- A tipping point analysis will be conducted. This analysis will systematically consider each subject who is lost to follow-up (or has final effectiveness outcome permanently missing) as an Event or a Non-Event at 12 months, in all combinations. Posterior probabilities of non-inferiority (or superiority, as appropriate) for each combination will be presented.

16.1.2 Sensitivity to Imputation Model

In the principal effectiveness analysis, subjects without complete data will have their final outcomes imputed via the posterior predictive piecewise exponential model described in Section 13. Sensitivity to the piecewise exponential model (with 5 pieces) will be studied by instead implementing a piecewise exponential model with 6 pieces, partitioning the time axis beyond 90 days into 5 pieces with approximately equal numbers of events. Because the number and timing of events could differ between treatment arms, the partitions may be different in the two arms. As with the principal analysis, each interval of the partition will have an associated λ parameter that will be assigned a Gamma(0.5, 0.001) prior, and subjects with incomplete data will have their final outcomes imputed according to this model.

16.2 Sensitivity to COVID-19

The COVID-19 pandemic has the potential to affect both the primary effectiveness and primary safety endpoints. Therefore, several sensitivity analyses are planned for handling data from subjects with documented COVID-19. At a minimum, these will be conducted for the analyses of the primary endpoints.

1. Subjects with Severe COVID-19 (defined clinically at Protocol Section 6.3.16) will be censored as of the Onset Date of their COVID-19 infection.

2. Subjects with Severe COVID-19 will be censored as of the Onset Date of their COVID-19 infection, with the exception that primary safety events that are adjudicated to be

related to device or procedure will still be included in the analysis, regardless of the date of onset of COVID-19 infection.

3. A tipping point analysis will be conducted. This analysis will systematically consider each subject who has a primary endpoint event subsequent to the Onset Date of documented severe COVID-19 as an Event or a Non-Event at 12 months, in all combinations, while all other subjects will be analyzed according to the model of Section 13. Posterior probabilities of non-inferiority (or superiority, as appropriate) for each combination will be presented.

16.3 Other Sensitivity Analyses

- A sensitivity analysis of the Primary Effectiveness Endpoint will also be conducted, in which subjects requiring ablation for an accessory pathway, AVNRT, treatment-emergent left AFL, or incessant AT are counted as treatment failures.
- As indicated in Section 9, the primary analysis for effectiveness will be conducted in the mITT population, but supportive analyses will also be conducted in the PP population. Similarly, as indicated in Section 10, the primary analysis for safety will be conducted in the Safety Population, supportive analyses will also be conducted in the PP population.
- For the primary effectiveness endpoint, Kaplan-Meier estimates of event rates will also be presented.

17 Heterogeneity/Poolability

Incidence of the primary safety and primary effectiveness endpoints will be evaluated for poolability across key subgroups. In particular, the primary endpoints will be examined for differences between sexes and sites. A specific categorization of sites is those that use RF as the control therapy and those that use cryoablation as the control therapy. Poolability of treatment effect over this factor (site uses RF, site uses cryo) will also be studied.

Poolability tests will be conducted with frequentist methods. For each factor of interest (sex, site), a logistic regression will be fit, with predictors Treatment Group, the factor, and their interaction. If the interaction p-value is significant at the level 0.15, further analyses will be conducted. At a minimum, any baseline variables found to differ between treatment groups and any baseline variables found to significantly predict a primary event (Treatment Response, or occurrence of CSE) in univariate analyses will be candidates for inclusion in logistic regression adjustment to see whether the apparent heterogeneity is explained by these factors. If so, analyses that adjust for these factors will be presented. If not, further analyses will be conducted, including presenting results separately by sex / site.

Sites with fewer than 8 subjects will be considered small sites. Poolability analyses will be summarized descriptively for all sites, but inferential tests of poolability across sites will exclude small sites. If there are many small sites, or many sites with 0 events, inferential tests of poolability across sites may not be conducted.

18 Other Planned Analyses

18.1 Assessment of Subject Blinding

Subjects will complete a survey regarding their opinion as to which study group they were assigned to at the time of discharge and at the 12-month assessment. Self-assessment of treatment status will be formally assessed according to the procedures defined by Bang et al (2004)[7] and James et al (1996)[8].

18.2 Roll-in Subjects

Roll-in Subjects will be summarized with descriptive statistics and will be analyzed separately from any populations composed of randomized subjects.

18.3 Adverse Events

Adverse Events will be tabulated by treatment group, by time, by seriousness, and by relatedness to device and/or procedure. Relatedness to COVID-19 will also be reported.

18.4 Procedural Assessments

Procedural Assessments are identified in Protocol Section 7.7. These will be summarized with descriptive statistics and may be summarized with either Bayesian or frequentist inferential methods. Any p-values or posterior probabilities will be nominal and not adjusted for multiplicity.

18.5 Quality of Life Assessments

Quality of Life Assessments are identified in Protocol Section 7.8. These will be summarized with descriptive statistics and may be summarized with either Bayesian or frequentist inferential methods. Any p-values or posterior probabilities will be nominal and not adjusted for multiplicity.

19 Analysis Methods for Secondary and Additional Endpoints

This section describes methods intended to be used for the Additional Endpoints (Sections 9.3 and 10.3) or other two-sample analyses, such as baseline demographic comparisons. These are Bayesian analogs to common frequentist statistical procedures (*t*-test, Wilcoxon rank sum test, chi-squared or Fisher's exact test, and an extension to the latter that includes multiple imputation similar to that employed in the analyses of the co-primary objectives).

19.1 Continuous Data: Bayesian version of a t-test.

Assuming that the quantity of interest *Y* follows a $N(\mu,\sigma^2)$ distribution, and placing a uniform prior distribution on (μ, $\log(\sigma)$), the posterior distribution of μ has the form[6]

$$\left[\left. \frac{\mu - \bar{y}}{s / \sqrt{n}} \right| Y \right] \sim t_{n-1}$$

Employing this for both μ_t and μ_c, credible intervals for μ_t and μ_c can be individually calculated as the 2.5th and 97.5th percentiles from their respective (scaled t) posterior distributions. The posterior mean of the difference $\mu_t - \mu_c$ is simply the difference in means of the respective scaled-t posterior distributions, which is the difference in sample means. The distribution of the difference $\mu_t - \mu_c$ can be easily estimated by any of the following approaches:

Full Monte Carlo Sampling

Employing the above for both μ_t and μ_c, the distribution of $(\mu_t - \mu_c)$ can be estimated in Monte Carlo fashion by drawing a large number of observations (eg., 100,000) from a t_{n-1} distribution for $n = n_t$ and $n = n_c$, back-transforming these observations to the μ_t- and μ_c-scales, and subtracting the sample values. The posterior probability $P(\mu_t - \mu_c > \delta \mid \text{data})$ is estimated as the proportion of observed values of $(\mu_t - \mu_c)$ that exceed δ. A 95% equal-tail Bayesian credible interval (BCI) can be calculated from the 2.5th and 97.5th percentiles of the sampled distribution of $\mu_t - \mu_c$.

Monte Carlo Sampling and Integration: More precise estimates can be achieved by sampling from only one posterior distribution (say, for μ_c). The cumulative distribution function $F_{T\text{–}C}$ of the difference μ_t - μ_c can be written as

$$F_{T\text{–}C}(\delta) = P(\mu_t - \mu_c \le \delta) = P(\mu_t \le \mu_c + \delta) = \int P(\mu_t \le \mu_c + \delta \mid \mu_c)\times f(\mu_c)\, d\mu_c$$

$$= \int F_T(\mu_c + \delta \mid \mu_c)\, f(\mu_c)\, d\mu_c\,,$$

where $f(\mu_c)$ represents the density of μ_c and F_T represents the c.d.f. of μ_t.

By Monte Carlo Integration, the last integral above is approximated by $1/M \sum F_T(\mu_c+\delta \mid \mu_c)$ after first drawing and conditioning on a large number M (e.g., M=100,000) of observations from the scaled t posterior distribution for $f(\mu_c)$. The resulting quantity is a mixture of M (scaled-t distributed) cumulative distribution functions for μ_t (conditional on the sampled μ_c values). Posterior quantities such as $P(\mu_t - \mu_c < 0)$ can be easily calculated from $F_{T\text{–}C}(0)$. Quantiles of the distribution of μ_t - μ_c can be easily computed by setting the desired quantile equal to $F_{T\text{–}C}(\delta) = 1/M \sum F_T\,(\mu_c+\delta \mid \mu_c)$ and using a numeric root-finder to solve for δ. In this manner the posterior median and a 95% BCI (calculated from the 0.025 and 0.975 quantiles) are found.

Numerical Integration: Nearly exact estimates can be achieved as follows: The cumulative distribution function $F_{T\text{–}C}$ of the difference μ_t - μ_c can be written as

$$F_{T\text{–}C}(\delta) = P(\mu_t - \mu_c \le \delta) = P(\mu_t \le \mu_c + \delta) = \int P(\mu_t \le \mu_c + \delta \mid \mu_c)\times f(\mu_c)\, d\mu_c$$

$$= \int F_T(\mu_c + \delta \mid \mu_c)\, f(\mu_c)\, d\mu_c\,,$$

where $f(\mu_c)$ represents the density of μ_c and F_T represents the c.d.f. of μ_t.

For any given value of δ, $F_{T\text{–}C}(\delta)$ can be computed by one-dimensional numerical integration, and posterior quantities such as $P(\mu_t - \mu_c < 0)$ can be easily calculated as $F_{T\text{–}C}(0)$. Quantiles of the distribution of μ_t - μ_c can be easily computed by setting the desired quantile equal to $F_{T\text{–}C}(\delta)$ and using a numeric root-finder to solve for δ. In this manner the posterior median and a 95% BCI (calculated from the 0.025 and 0.975 quantiles) are found. This approach does not require any Monte Carlo sampling.

The above 3 methods do not require large sample sizes.

Normal Approximation

For large sample sizes, the *t*-distribution is well approximated by a Normal distribution, and the difference of two *t*-distributions is thus the difference of 2 Normal distributions, which is Normally distributed.

More specifically, approximating a t_{n-1} distribution by a Normal distribution with the same mean and variance,

$$\left[\frac{\mu - \bar{x}}{s/\sqrt{n}}\right] \sim \mathrm{t}_{n-1} \approx \mathrm{N}\left(0, \frac{n-1}{n-3}\right),$$

and thus $(\mu_T - \mu_C)$ is approximated by a Normal distribution with mean $(\bar{x}_\mathrm{T} - \bar{x}_\mathrm{C})$ and variance $\left(\frac{(n_T-1)s_T^2}{n_T(n_T-3)} + \frac{(n_C-1)s_C^2}{n_C(n_C-3)}\right)$. Posterior probabilities and quantiles for $(\mu_T - \mu_C)$ can then be computed from this Normal distribution.

Unless otherwise specified, all implementations of this Bayesian t-test will use the (exact) Numerical Integration approach, above. Specifically, the test of PV Dimensional Change: Aggregate PV Cross-Sectional Area (Section 10.2.1) will use the Numerical Integration approach.

19.2 Continuous Data (Distribution-free): Bayesian Version of a Wilcoxon Rank-Sum Test

For this test, data are ranked. Analogous to the Wilcoxon Rank-Sum test, observations in each group are replaced by their ranks in the combined order statistic. Then the two groups of rank-transformed data are compared via the Bayesian version of a t-test, above. It has been shown that conducting a two-sample t-test on rank-transformed data is approximately equivalent to the Wilcoxon Rank-Sum test[4,5].

Non-inferiority can be evaluated by first subtracting a margin (δ) from each observation in one treatment group prior to the rank transformation..

19.3 Discrete (Dichotomous) Data: Bayesian Version of a Comparison of Proportions

Given proportions p_t and p_c for a dichotomous outcome, let each be assigned a non-informative Beta(0.5, 0.5) prior. Then the posterior distributions for p_t and p_c follow Beta($0.5+Y_t$,$0.5+N_t–Y_t$) and Beta($0.5+Y_c$,$0.5+N_c–Y_c$) distributions, respectively, where Y_t and Y_c represent the number of "successes" and N_t and N_c represent the number of "tries" (e.g., subjects) in their respective treatment groups. Credible intervals for p_t and p_c can be individually calculated as the 2.5th and 97.5th percentiles from their respective (beta) posterior distributions. The posterior mean of the difference $p_t - p_c$ is simply the difference in means of the respective beta posterior distributions. The distribution of the difference $p_t - p_c$ can be summarized by any of the following approaches:

Full Monte Carlo Sampling: Drawing a large number of observations (eg., 100,000) from each Beta posterior, and subtracting the sampled values. The posterior probability P($p_t - p_c < \delta$ | data) is estimated as the proportion of these observations where the difference is less than δ. A 95% equal-tail Bayesian credible interval (BCI) can be calculated from the 2.5th and 97.5th percentiles of the sampled distribution of $p_t - p_c$.

Monte Carlo Sampling and Integration: More precise estimates can be achieved by sampling from only one posterior distribution (say, for p_c). The cumulative distribution function $F_{T\text{–}C}$ of the difference $p_t - p_c$ can be written as

$F_{T\text{–}C}(\delta) = P(p_t - p_c < \delta) = P(p_t < p_c + \delta) = \int P(p_t < p_c + \delta \mid p_c) \times f(p_c)\, dp_c$

$= \int F_T(p_c + \delta \mid p_c)\, f(p_c)\, dp_c$,

where $f(p_c)$ represents the density of p_c and F_T represents the c.d.f. of p_t.

By Monte Carlo Integration, the last integral above is approximated by $1/M \sum F_T(p_c + \delta \mid p_c)$ after first drawing and conditioning on a large number M (e.g., M=100,000) of observations from the beta posterior distribution for $f(p_c)$. The resulting quantity is a mixture of M (beta-distributed) cumulative distribution functions for p_t (conditional on the sampled p_c values). Posterior quantities such as $P(p_t - p_c < 0)$ can be easily calculated from $F_{T\text{–}C}(0)$. Quantiles of the distribution of $p_t - p_c$ can be easily computed by setting the desired quantile equal to $F_{T\text{–}C}(\delta) = 1/M \sum F_T(p_c + \delta \mid p_c)$ and using a numeric root-finder to solve for δ. In this manner the posterior median and a 95% BCI (calculated from the 0.025 and 0.975 quantiles) are found.

Numerical Integration: Nearly exact estimates can be achieved as follows: The cumulative distribution function $F_{T\text{–}C}$ of the difference $p_t - p_c$ can be written as

$F_{T\text{–}C}(\delta) = P(p_t - p_c < \delta) = P(p_t < p_c + \delta) = \int P(p_t < p_c + \delta \mid p_c) \times f(p_c)\, dp_c$

$= \int F_T(p_c + \delta \mid p_c)\, f(p_c)\, dp_c$,

where $f(p_c)$ represents the density of p_c and F_T represents the c.d.f. of p_t.

For any given value of δ, $F_{T\text{–}C}(\delta)$ can be computed by one-dimensional numerical integration, and posterior quantities such as $P(p_t - p_c < 0)$ can be easily calculated as $F_{T\text{–}C}(0)$. Quantiles of the distribution of $p_t - p_c$ can be easily computed by setting the desired quantile equal to $F_{T\text{–}C}(\delta)$ and using a numeric root-finder to solve for δ. In this manner the posterior median and a 95% BCI (calculated from the 0.025 and 0.975 quantiles) are found.

This is the method that was used in the simulation when comparing completed datasets.

The above 3 methods do not require large sample sizes.

Normal Approximation

For large sample sizes, the beta distribution is well approximated by a Normal distribution with the same mean and variance, and the difference of two beta distributions is thus approximately the difference of 2 Normal distributions, which is Normally distributed.

19.4 Discrete (Dichotomous) Data: Bayesian Version of a Comparison of Proportions (with predictions)

The analysis method in Section 19.3 applies only to subjects with observed data at the time point of interest (e.g., complete data at 12-months). A modification is required to take into account subjects with partial follow-up. The modeling is analogous to the predictions/imputations used in the Primary Objective analyses (Section 13).

For effectiveness assessment, subjects without outcome data for the time point of interest T_0 (e.g., T_0 = 12 months) will have their outcome imputed on the basis of the same model described in Section 13.1, with the possible exception that the piecewise exponential model may use a different partition of the time axis. If there are < 5 events in a "piece" of the piecewise exponential model (using the partition described in Section 13), pieces may be combined. Other partitions (e.g., created so that there are approximately equal numbers of events per piece) may be explored as sensitivity analyses.

For safety assessment, subjects without outcome data for the time point of interest T_0 (e.g., T_0 = 12 months) will have their outcome imputed on the basis of the same model described in Section 13.2, with the possible exception that the piecewise exponential model will use a different partition of the time axis. The default safety analysis (Section 13.2) uses a piecewise exponential model using the time axis partition 0–7 days, 7–360 days. If there are > 10 events in the time window from 7 to 360 days, other partitions (e.g., created so that there are approximately equal numbers of events per piece, with a minimum of 5 events per piece) may be explored as sensitivity analyses.

By predicting binary outcomes at T_0 for each subject without an observed outcome, and combining with the data from subjects with observed outcomes at Day T_0, multiple completed datasets are obtained, and from each of these a posterior probability of non-inferiority (or superiority) is calculated. By averaging the posterior probabilities that result from each completed dataset, we are integrating over the posterior predictive distribution of imputed data. Statistical summaries (e.g., probability of non-inferiority/superiority, posterior quantiles) will be computed from this averaged distribution. This is a form of Bayesian multiple imputation.

References

1. Fleiss, J. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research* 2: 121-145.
2. Fleiss, J., B. Levin, and M. Paik. (2003). *Statistical Methods for Rates and Proportions (Third Edition)*. John Wiley and Sons, Inc., New York.
3. Kuck, H., J. Brugada, A. Fürnkranz, A. Metzner, F. Ouyang, K. Julian Chun, A. Elvan, T. Arentz, K. Bestehorn, S. Pocock, J. Albenque, and C. Tondo for the FIRE AND ICE Investigators*. (2016). Cryoballoon or radiofrequency ablation for paroxysmal atrial fibrillation. *N Engl J Med*;374:2235-45. DOI: 10.1056/NEJMoa1602014.
4. Berry D & Lindgren B, Statistics: Theory and Methods, 2ed. Belmont, CA: Duxbury, 1996, p 505.
5. Conover WJ & Iman RL, "Rank transformations as a bridge between parametric and nonparameteric statistics," *American Statistician* 35 (1981), 124-129
6. Gelman A, Carlin JB, Stern HS, Rubin DB, *Bayesian Data Analysis*, 1st ed. London: Chapman & Hall, 1995, p. 69.
7. Bang, H., L. Ni, and C.E. Davis, Assessment of blinding in clinical trials. *Control Clin Trials*, 2004. 25(2): p. 143-56.
8. James, K.E., et al., An index for assessing blindness in a multi-centre clinical trial: disulfiram for alcohol cessation--a VA cooperative study. *Stat Med*, 1996. 15(13): p. 1421-34.
9. Ibrahim JG, Chen M-H, Sinha D. *Bayesian Survival Analysis*. New York, NY: Springer, 2001, p 48.
10. Borenstein M, Hedges LV, Higgins JP, Rothstein HR (2010): A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, **1**, 97—111.
11. PASS 13 Power Analysis and Sample Size Software (2014). NCSS, LLC. Kaysville, Utah, USA, ncss.com/software/pass.
12. Gelman A, Carlin JB, Stern HS, Rubin DB, *Bayesian Data Analysis*, 1st ed. London: Chapman & Hall, 1995, p. 69.
13. Albert J, *Bayesian Computation with R*, 1st ed. New York: Springer, 2007.
14. Reddy VY, Dukkipati SR, Neuzil P, Natale A Albenque J-P, Kautzner J, Shah D, Michaud G, Wharton M, Harari D, Mahapatra S, Lambert H, Mansour M (2015), Randomized, Controlled Trial of the Safety and Effectiveness of a Contact Force-Sensing Irrigated Catheter for Ablation of Paroxysmal Atrial Fibrillation. *Circulation* 2015;**132**:907-915.
15. Broglio KR, Connor JT, Berry SM., “Not too big, not too small: a goldilocks approach to sample size selection,” *J Biopharm Stat*. 2014;**24**(3):685-705.
16. Dukkipati SR, Cuoco F, Kutinsky I, et al. Pulmonary Vein Isolation Using the Visually Guided Laser Balloon: A Prospective, Multicenter, and Randomized Comparison to Standard Radiofrequency Ablation. *J Am Coll Cardiol* 2015;66:1350-60.
17. Packer DL, Kowal RC, Wheelan KR, et al. Cryoballoon ablation of pulmonary veins for paroxysmal atrial fibrillation: first results of the North American Arctic Front (STOP AF) pivotal trial. *J Am Coll Cardiol* 2013;61:1713-23.
18. Wilber DJ, Pappone C, Neuzil P, et al. Comparison of antiarrhythmic drug therapy and radiofrequency catheter ablation in patients with paroxysmal atrial fibrillation: a randomized controlled trial. *JAMA* 2010;303:333-40.

Changes are enclosed in quotes and/or shown as underline (added) or ~~strikethrough~~ (deleted) text.

Section	January 5, 2020 Version	October 1, 2021 Version	Rationale
Version throughout	Date January 5, 2020 Issued Revision B	Date October 1, 2021 Issued Revision C	Version control procedure
Version throughout	Header and date revisions	[Corrected and/or updated]	Version control procedure
Version throughout	Corrections	[Minor clarifications and typographical errors corrected, shown in redline]	Version control procedure
Section 9.3.9	“The endpoint is Treatment Success with PVI/CTI Only.”	"The endpoint is Treatment Success with PVI/CTI Only, excluding subjects receiving extra PVI/CTI ablation."	Changes are made to address FDA Study Design Considerations
Section 15.8.2	[None]	“The results in Table 7 are computed under the assumption that the primary safety objective results are so strongly favorable that the safety portion of the sample size assessments would always indicate a stop for promising trend. Table 8 presents results when this assumption is altered so that the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, as described in Section 12, and primary safety data are generated under the expected safety assumptions that $Q_T = Q_C = 0.080$ and Method Saf_1. Conservatively, it is further assumed that the safety endpoint would always pass its noninferiority criterion once follow-up is complete, even though this will not always be the case.” Added Table 8: Type I Error Rates for Primary Effectiveness Non-inferiority Test as Data Generation Parameters Vary (Safety data generated using $Q_T = Q_C = 0.080$, Method Saf_1) “Each of the 27 estimated Type I error rates in Table 8 is the result of running 10,000 simulated clinical trials. Among the 27 estimated type I error rates, the mean value is 0.0486, and the max is 0.0534. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval $0.05 \pm 1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043$, with upper bound 0.0543.	Changes are made to address FDA Study Design Considerations

		Although Table 7 and Table 8 generate data with varying methods, the method applied for the Treatment Group is the same as that applied for the Control Group. The below Table 9 presents a study of effectiveness type I error rates when the Treatment and Control groups are generated with differing methods. Specifically, 9 null scenarios of 10,000 iterations each were simulated, systematically and separately varying data generation methods {Eff_1, Eff_2, Eff_3} in the Treatment group and {Eff_1, Eff_2, Eff_3} in the Control group, under the null assumption of 50% success rate in the Treatment group and 65% success rate in the Control group. Safety data in these scenarios were generated using method Saf_1 for both treatment groups, with event rates of 8% in each group. As in Table 8, the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, but it is conservatively assumed that the safety endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case. Enrollment followed the Expected rate, as described in Section 15.7." Added Table 9: Type I Error Rates for Effectiveness Non-Inferiority as data generation methods vary (when $P_T = 0.65 – 0.15$, $P_C = 0.65$) "In Table 9, the maximum value is 0.0511. In Table 9 (as well as Table 7 and Table 8), no estimated type I error rates exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval 0.05 ± $1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043$, with upper bound 0.0543. Therefore, the type I error rate of the effectiveness test is considered controlled at the level 0.05."	
Section 15.8.3	[None]	"The results in Table 10 are computed under the assumption that the primary effectiveness objective results are so strongly favorable that the effectiveness portion of the sample size assessments would always indicate a stop for promising trend. Table 11 presents results when this assumption is altered so that the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, as described in Section 12, and primary effectiveness data are generated under the expected effectiveness assumptions that $P_T = P_C = 0.650$ and Method Eff1. Conservatively, it is further assumed that the effectiveness endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case." Added Table 11: Type I Error Rates for Primary Safety Non-inferiority Test as Data Generation Parameters Vary (Effectiveness data generated using $P_T = P_C = 0.65$, Method Eff_1)	Changes are made to address FDA Study Design Considerations

		“Each of the 18 estimated Type I error rates in Table 11 is the result of running 10,000 simulated clinical trials. Among the 18 estimated type I error rates, the mean value is 0.0416, and the max is 0.0500. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval $0.05 \pm 1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043$, with upper bound 0.0543. Although Table 10 and Table 11 generate data with varying methods, the method applied for the Treatment Group is the same as that applied for the Control Group. The below Table 12 presents a study of safety type I error rates when the Treatment and Control groups are generated with different methods. Specifically, 4 null scenarios of 10,000 iterations each were conducted, systematically and separately varying data generations methods $\{Saf_1, Saf_2\}$ in the Treatment group and $\{Saf_1, Saf_2\}$ in the Control group, under the assumption of 16% success rate in the Treatment group and 8% success rate in the Control group. Effectiveness data in these scenarios were generated using method Eff1 for both treatment groups, with the expected success rate of 65% in each group. As in Table 8, the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, but it is conservatively assumed that the effectiveness endpoint would always pass its noninferiority criterion once follow-up is complete, even though this will not always be the case. Enrollment followed the Expected rate, as described in Section 15.7.” Added Table 12: Type I Error Rates for Safety Non-Inferiority as data generation methods vary, (when $Q_T = 0.08+0.08$, $Q_C = 0.08$) “In Table 12, the maximum value is 0.0463. In Table 12 (as well as Table 10 and Table 11), no estimated type I error rates exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval $0.05 \pm 1.96\times\sqrt{(0.05\times0.95/10000)} = 0.05 \pm 0.0043$, with upper bound 0.0543. Therefore, the type I error rate of the safety test is considered controlled at the level 0.05.”	
Section 15.8.4	[None]	“The results in Table 13 are computed under the assumption that the primary safety objective results are so strongly favorable that the safety portion of the sample size assessments would always indicate a stop for promising trend, and the primary safety hypothesis test would always pass its non-inferiority criterion. Table 14 presents results when this assumption is altered so that the sample size is jointly determined by trends in both the primary safety and effectiveness endpoints, as described in Section 12, and primary safety data are generated under the expected safety assumptions that $Q_T = Q_C = 0.080$ and Method Saf_1. Conservatively, it is further assumed that the safety endpoint would always pass its non-inferiority criterion once follow-up is complete, even though this will not always be the case.”	Changes are made to address FDA Study Design Considerations

		Added Table 14: Type I Error Rates for Primary Effectiveness Superiority Test as Data Generation Parameters Vary (Safety data generated using $Q_T = Q_C = 0.080$, Method Saf$_1$) "Each of the 27 estimated Type I error rates in Table 14 is the result of running 10,000 simulated clinical trials. Among the 27 estimated type I error rates, the mean value is 0.0237, and the max is 0.0268. No values exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval $0.025 \pm 1.96\times\sqrt{(0.025\times0.975/10000)} = 0.025 \pm 0.0031$, with upper bound 0.0281. Although Table 13 and Table 14 generate data with varying methods, the method applied for the Treatment Group is the same as that applied for the Control Group. The below Table 15 presents a study of type I error rates for the secondary superiority objective when the Treatment and Control groups are generated with different methods. Specifically, 9 null scenarios of 10,000 iterations each were conducted, systematically and separately varying data generations methods {Eff$_1$, Eff$_2$, Eff$_3$} in the Treatment group and {Eff$_1$, Eff$_2$, Eff$_3$} in the Control group, under the superiority null assumption of 65% success rate in each treatment group. Safety data in these scenarios were generated using method Saf1 for both treatment groups, with expected rates of 8% in each group. As in Table 14, the sample size is jointly determined by trends in both the primary effectiveness and safety endpoints, but it is conservatively assumed that the safety objective would always pass its non-inferiority condition once enrollment had stopped. Enrollment followed the Expected rate, as described in Section 15.7." Added Table 15: Type I Error Rates for Effectiveness Superiority (when $P_T = P_C = 0.65$), as data generation methods vary "In Table 15, the maximum value is 0.0278. In Table 15 (as well as Table 13 and Table 14), no estimated type I error rates exceed the range of what is expected in a simulation of 10,000 iterations, calculated as a 95% probability interval $0.025 \pm 1.96\times\sqrt{(0.025\times0.975/10000)} = 0.025 \pm 0.0031$, with upper bound 0.0281. Therefore, the type I error rate of the secondary effectiveness superiority test is considered controlled at the (one-sided) level 0.025."	
Section 16.3	• As indicated in Section 9, the primary analysis for effectiveness will be conducted in the mITT	• A sensitivity analysis of the Primary Effectiveness Endpoint will also be conducted, in which subjects requiring ablation for an accessory pathway, AVNRT, treatment-emergent left AFL, or incessant AT are counted as treatment failures.	Changes are made to address FDA Study Design Considerations

	population, but supportive analyses will also be conducted in the PP population. Similarly, as indicated in Section 10, the primary analysis for safety will be conducted in the Safety Population, supportive analyses will also be conducted in the PP population. • For the primary effectiveness endpoint, Kaplan-Meier estimates of event rates will also be presented.	• As indicated in Section 9, the primary analysis for effectiveness will be conducted in the mITT population, but supportive analyses will also be conducted in the PP population. Similarly, as indicated in Section 10, the primary analysis for safety will be conducted in the Safety Population, supportive analyses will also be conducted in the PP population. • For the primary effectiveness endpoint, Kaplan-Meier estimates of event rates will also be presented.	
Section 19.1	**Normal Approximation** For large sample sizes, the t-distribution is well approximated by a Normal distribution, and the difference of two t-distributions is thus the difference of 2 Normal distributions, which is Normally distributed.	**Normal Approximation** For large sample sizes, the t-distribution is well approximated by a Normal distribution, and the difference of two t-distributions is thus the difference of 2 Normal distributions, which is Normally distributed. More specifically, approximating a t_{n-1} distribution by a Normal distribution with the same mean and variance, $\left[\frac{\mu - \bar{x}}{s/\sqrt{n}}\right] \sim t_{n-1} \approx N\left(0, \frac{n-1}{n-3}\right)$ and thus $(\mu_T - \mu_C)$ is approximated by a Normal distribution with mean $(\bar{x}_T - \bar{x}_C)$ and variance $\left(\frac{(n_T-1)s_T^2}{n_T(n_T-3)} + \frac{(n_C-1)s_C^2}{n_C(n_C-3)}\right)$.Posterior probabilities and quantiles for $(\mu_T - \mu_C)$ can then be computed from this Normal distribution. Unless otherwise specified, all implementations of this Bayesian t-test will use the (exact) Numerical Integration approach, above. Specifically, the test of PV Dimensional Change: Aggregate PV Cross-Sectional Area (Section 10.2.1) will use the Numerical Integration approach.	Changes are made to address FDA Study Design Considerations