STATISTICAL ANALYSIS PLAN

October 9, 2018

A Multicenter, Randomized, Double-Blind, Sham-Controlled Clinical Investigation of the EndoStim® Lower Esophageal Sphincter (LES) Stimulation System for the Treatment of Gastroesophageal Reflux Disease (GERD)

LESS GERD – Randomized Clinical Trial (RCT)

PROTOCOL NUMBER CS-100

EndoStim, Inc.
1609 Shoal Creek Boulevard, Suite 303
Austin, TX 78701 USA

TABLE OF CONTENTS

1 INTRODUCTION

This statistical analysis plan (SAP) outlines the data and procedures used for assessing the efficacy and safety endpoints of Protocol CS-100: "A *Multicenter, Randomized, Double-Blind, Sham-Controlled Clinical Investigation of the EndoStim® Lower Esophageal Sphincter (LES) Stimulation System for the Treatment of Gastroesophageal Reflux Disease (GERD).*"

2 SCOPE

This version of the plan has been developed with respect to the electronic case report forms (eCRF) and protocol Revision H, dated September 8, 2017. Any further changes to the protocol or CRFs may necessitate updates to the SAP.

3 DEFINITIONS

Table 1: Abbreviations

BMI		Body Mass Index
CRF		Case Report Form
DS		Delayed Stimulation Group
FDA		US Food and Drug Administration
GERD		Gastroesophageal Reflux Disease
GERD-HRQL		Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire
H2		Histamine-2 Receptor Antagonists
IPG		Implantable Pulse Generator
IS		Immediate Stimulation Group
LES		Lower Esophageal Sphincter
LES-EEP		Lower Esophageal Sphincter End-Expiratory Pressure
LES-MP		Lower Esophageal Sphincter Mean Pressure
LPR		Laryngopharyngeal Reflux
MedDRA		Medical Dictionary for Regulatory Activities
MI		Multiple Imputation
PI		Post Implant
PPI		Proton Pump Inhibitors
PSQI		Pittsburgh Sleep Quality Index
RSI		Reflux Symptom Index
WPAI:SHP		Work Productivity and Activity Impairment Questionnaire: Specific Health Problem V2.0

4 STUDY DESIGN

The LES GERD study is a multi-center prospective, randomized, double-blind sham-controlled study that will include up to 200 implanted subjects in up to thirty (30) sites across the United States, and Europe.

All subjects undergo screening and baseline visits. Those who meet the entrance criteria undergo Lower Esophageal Sphincter (LES) Stimulation System implantation and randomization to an immediate stimulation (IS) or 6-month delayed stimulation (DS) group.

Randomized subjects complete a 6-month double-blind phase. At the 6-month visit, all subjects are unblinded. The delayed stimulation group subjects begin receiving electrical stimulation therapy, and all subjects are followed for an additional open-label treatment phase (immediate stimulation group subjects undergo electrical stimulation for an additional 6 months

and delayed stimulation group subjects undergo electrical stimulation for 12 months, resulting in at least a total of 12 months of electrical stimulation therapy in both groups) during the initial open label period. Subjects continue stimulation treatment and an extended open-label follow-up phase includes an 18-month post-stimulation phone interview followed by annual visits through 5 years post-stimulation.

Study subjects will be followed until approval is granted from the FDA or an official study closure is issued (when regulatory requirements have been satisfied per the protocol or a decision by EndoStim and/or regulatory authority).

4.1 Randomization

Within each investigational site, subjects will be block randomized in a 1:1 ratio. Block sizes will be randomly varied between 2 and 4 to ensure double blinding. Randomization assignments were generated prior to the start of the trial by an unblinded statistician and entered into the electronic database's randomization assignment module.

4.2 Blinding

At the time of randomization, only the unblinded study technician will request and receive the randomization assignment from the password protected system. Both subject and physician will remain blinded to this assignment throughout the first six months of follow up. Additionally, all subjects will undergo the same test protocols and device setting sessions during the blinded phase of the study.

At the beginning of the 1 month and 6 month visits, subjects will be asked to identify to which treatment group they believe they have been assigned. This information will be documented in the medical record and submitted to EndoStim via a CRF.

4.3 Study Objectives

During the double-blind phase, the objective will be to compare the immediate stimulation arm versus 6-month delayed stimulation arm for the treatment of gastroesophageal reflux disease (GERD). The open-label phase of the trial will begin at 6 months post implantation, where patients will be unblinded and those in the delayed stimulation arm will begin receiving stimulation treatments. The objective during the open-label portion of the study, assuming the pooling of subjects is acceptable, will be to compare outcomes between pre-implant metrics to those post 6-months active stimulation up to 5 years.

The study objectives during the blinded phase are:

- To estimate the safety of IS and DS patients.
- To compare the mean percentage of time pH is <4.0 between IS and DS.
- To compare heartburn symptoms by frequency during day and evening hours along with worst episode severity between IS and DS.
- To compare regurgitation symptoms by frequency during day and evening hours along with worst episode severity between IS and DS.
- To compare acid-suppression medication utilization between IS and DS.
- To compare quality of life scores between IS and DS.

The study objectives during the open-label phase are:

- To compare the mean percentage of time pH is < 4.0 after 12 months of stimulation and prior to implant in the pooled IS and DS group (all implanted patients).
- To compare the heartburn symptoms by frequency during day and night and worst episode severity after 12 months of stimulation and prior to stimulation in all implanted patients.

- To compare the regurgitation symptoms by frequency and worst episode severity after 12 months of stimulation and prior to stimulation in all implanted patients.
- To compare acid suppression medication utilization after 12 months of stimulation and prior to stimulation in all implanted patients.
- To compare the number of subjects achieving symptom success after 12 months of stimulation as defined by a minimum 50% improvement in composite GERD-HRQL score compared to prior to simulation ***on-PPI*** and H2 blockers (for patients who are not considered PPI intolerant).
- To compare quality of life scores after 12 months of stimulation and prior to stimulation in all implanted patients.

4.4 Study Success Criteria

The primary efficacy analysis will be conducted at least 6 months after accrual to the trial is complete in order to allow all implanted patients to be followed for the primary endpoint. At the primary analysis, the proportion of patients achieving pH success will be compared between randomized groups. If the one-sided p-value comparing immediate stimulation to delayed stimulation is less than 0.022 in both the ITT and PP dataset, this trial will be considered a success.

4.5 Bayesian Adaptive Model

While the primary analysis is a frequentist test, a Bayesian model of the primary endpoint was built for the purpose of simulating the Bayesian quantities that may be reported during the execution of the trial, as described in Appendix A. The trial will implant a minimum of 120 and a maximum of 200 patients.

4.5.1 Schedule of Interim Analysis

Interim analyses are scheduled at 120, 140, 160 and 180 implanted patients and simulations were conducted at each of these time points.

4.5.2 Decision Rules and Stopping Boundaries

At each simulated interim analysis, the predictive probability of trial success at the final analysis for the currently implanted patients was calculated. This calculation assumes that the primary analysis would take place after all currently implanted patients complete their follow-up for the primary endpoint.

If during the actual trial, at an interim analysis, the predictive probability of trial success with the currently implanted patients is at least 95%, this trial may stop accrual early for expected success. The final analysis would be conducted after all implanted patients complete the follow-up for the primary endpoint.

Additionally, at each interim analysis, we calculated the predictive probability that the current trial will be successful should it continue to the maximum of 200 implanted patients and all patients are followed for the primary endpoint.

If during the actual trial, at an interim analysis, the predictive probability of trial success at the maximum sample size is less than 1%, this trial will stop early for futility.

4.5.3 Primary Efficacy Response Rate Estimates

In preliminary non-randomized studies, 73% (19/26) subjects achieved the pH success criteria. We assume that we will see some attenuation of this rate due to blinding and plan for a treatment group success rate of between 60% and 70%. We conservatively assume that the control group will experience somewhere between 20% and 40% success rate with the

resulting difference between groups between 20% and 50%. These estimates were used to develop the scenarios in the adaptive design simulation model below.

4.5.4 Simulated Model Operating Characteristics

The trial was simulated under 18 different response rate scenarios for the IS and DS groups in order to characterize and evaluate the performance of the adaptive design. There are nine null scenarios where the response rate in the IS and DS groups were estimated to be the same and nine alternative scenarios where the IS group has a higher response rate than the DS group. All scenarios are based on 10,000 simulations.

Table 2: Response Rate Scenarios

Scenarios	DS Response Rate	IS Response Rate	Prob. of Success	Mean Trial Size	Prob. of Stopping Accrual Early for Expected Success	Prob. of Stopping Early for Futility	Prob. Of Success to Futility Flip Flop
Null 1-1	0.1	0.1	0.021	179.0	0.004	0.512	0.004
Null 2-2	0.2	0.2	0.023	177.1	0.005	0.519	0.011
Null 3-3	0.3	0.3	0.023	176.6	0.006	0.515	0.013
Null 4-4	0.4	0.4	0.023	176.5	0.005	0.515	0.014
Null 5-5	0.5	0.5	0.024	175.8	0.007	0.526	0.015
Null 6-6	0.6	0.6	0.024	176.1	0.006	0.518	0.016
Null 7-7	0.7	0.7	0.025	176.7	0.006	0.520	0.013
Null 8-8	0.8	0.8	0.022	176.9	0.005	0.522	0.011
Null 9-9	0.9	0.9	0.020	178.7	0.003	0.519	0.004
Alternate 2-6	0.2	0.6	0.999	142.8	0.965	0.001	<0.0001
Alternate 2-7	0.2	0.7	>0.999	132.4	0.997	<0.0001	<0.0001
Alternate 3-6	0.3	0.6	0.982	161.1	0.760	0.003	0.006
Alternate 3-7	0.3	0.7	>0.999	143.8	0.956	<0.0001	<0.0001
Alternate 4-6	0.4	0.6	0.779	180.4	0.361	0.030	0.022
Alternate 4-7	0.4	0.7	0.983	161.2	0.758	0.003	0.004
Alternate 5-6	0.5	0.6	0.262	187.0	0.076	0.178	0.032
Alternate 5-6.5	0.5	0.65	0.535	186.0	0.186	0.079	0.032
Alternate 5-7	0.5	0.7	0.796	180.2	0.369	0.028	0.023

There is a simulated overall one-sided Type I error of between 0.020 and 0.025 across the null scenarios. There is a probability <0.007 that a null trial achieves an early success and a probability >0.512 that a null trial is stopped early for futility. The mean trial size ranges between 175.8 and 179.0 patients in the null scenarios.

In the "Alternate 4-7" case, where the response rate in the control group is 0.40 and the response rate in the treatment group is 0.70, the adaptive design has a power of 0.983 to declare superiority of the treatment. For this scenario, the probability of stopping trial accrual early for expected success is 0.758, and the probability of stopping a trial early for futility is 0.003. The probability of stopping trial accrual early for expected success but then going on to not be successful is 0.004. This scenario has a mean sample size of 161.2 patients. Power decreases with smaller treatment effects, as we see in the "Alternate 5-7" scenario, where the response rate in the control group is 0.50 and the response rate in the treatment group is 0.70. The power of the "Alternate 5-7" scenario is 0.796, with a 0.369 probability of stopping trial accrual early for expected success and a 0.028 probability of stopping early for futility. The probability of stopping trial accrual early for expected success and not achieving success at the final analysis in this scenario is 0.023.

The results in the alternative models have a maximum simulated Type I error of 2.5% with a probability greater than 0.50 of stopping early for futility under the null. If IS demonstrates superiority versus DS, the adaptive design offers the opportunity to identify this benefit with a sample size of less than 200 patients. The full text of the adaptive design report is included in appendix A.

5 STATISTICAL ANALYSES

5.1 General Statistical Considerations

5.1.1 Software

Version 9.2 or higher of SAS® statistical software package or other validated statistical software will be used to provide all summaries, listings, graphs, and statistical analyses.

5.1.2 Descriptive Statistics

Continuous data will be summarized using descriptive statistics: n, mean, standard deviation or standard error, median, minimum and maximum. The decision to use either standard deviation or standard error will be based upon the objective of the presentation: standard deviation will be used when the interest is the natural variability of the data; standard error will be used when comparing two or more means. Continuous variables that are recorded using approximate values (e.g., < or >) will be replaced by the closest exact value for the calculation of summary statistics.

Categorical variables will be summarized using frequency counts and percentages. For categorical and ordinal variables, percentages will be calculated based on non-missing data. When count data are presented, the percentage for zero counts may be suppressed in order to draw attention to the non-zero counts.

For ordinal-scaled variables, a combination of the above may be employed as appropriate: frequency and percentage of observations within a category and means and standard deviations of the scores of the categories.

5.1.3 P-values

Unless otherwise specified, statistical analyses will be performed using a one-sided hypothesis test at the overall 2.5% level of significance. P-values will be rounded to three decimal places. If a p-value is less than 0.001 it will be reported as "< 0.001." If a p-value is greater than 0.999 it will be reported as "> 0.999."

As we will be performing multiple hypothesis tests for the secondary endpoints, we will need to correct for multiple comparisons. We note that the simple Bonferroni correction (which would adjust all p-values by multiplying by a factor of 14 for the number of secondary endpoints) would be inappropriately punitive as the endpoints are likely to be strongly correlated. Instead, we propose to adjust the p-values by the Bootstrap method of Westfall, which controls the familywise error rate without assuming independence between the tests, as implemented in SAS Proc Multtest (as implemented in SAS Proc Multtest, option BOOTSTRAP). This will result in a level of adjustment that is appropriate for the correlation of the endpoints but will strictly preserve the family-wise type I error rate at the one-sided 0.025 level.

5.1.4 Missing Data

We anticipate no missing data for pre-implant pH in included subjects and no more than 5% to 10% missing data for month 6 pH. In the event of any missing values, we will use linear regression-based multiple imputation of the month 6 value based on pre-implant and month 3 visits (where available) for the primary analysis. The imputation will be implemented using Proc MI in SAS.

We anticipate no missing data for baseline secondary endpoints (with the exception of the GERD HRQL question 9 discussed in section 5.6) and no more than 5% to 10% missing data during follow up for subjects who remain in the study. In the event of any missing values, we will use linear regression-based multiple imputation of missing secondary outcomes value based on baseline demographics (site, sex and age), procedure details (crural status) and total percentage of time pH <4 as a measure of baseline severity for the secondary endpoint analyses. While it's not anticipated that there will be missing baseline demographic and procedure values, if such missingness does occur baseline values will be imputed prior to the outcomes. The imputations will be implemented using Proc MI in SAS.

We will also conduct tipping point analyses to examine the extent to which imputing systematically different values for missing delayed and immediate stimulation subjects would affect the trial results.

Data related to safety endpoints will not be imputed.

5.1.5 Partial Dates

In the case of partial dates, the dates of the event will be imputed. Imputation of partial dates is subject to the condition that the imputed date occurs on or after the procedure date and on or before the subject last contact date. In the case of adverse events with partial start and stop dates, the imputed dates are subject to the additional condition that the start date must occur on or before the stop date.

Table 3: Date Imputations

	Valid Portion	Missing Portion	Imputed Value for Missing Portion[1]
Start Date	Month, Year	Day	Set day to 15th day of the known month and year
	Year	Day, Month	Set date to June 30th of the known year
	None	Day, Month, Year	Date of implant procedure
Stop Date[2]	Month, Year	Day	Set day to 15th day of the known month and year
	Year	Day, Month	Set date to June 30th of the known year
	None	Day, Month, Year	None

[1]Imputed date must occur on or after the implant procedure date. For adverse events the start date must occur on or before the stop date.

[2]Date of death will be imputed per the imputation rules for a start date.

5.1.6 Visit Windows and Visit Definitions

For the purposes of monitoring protocol deviations, a visit will be considered in-window if it occurs within the intervals detailed below as specified in the protocol, and out of window otherwise. Endpoints outside of the window will be set to missing for the per-protocol dataset.

An annual visit is recommended for routine follow up. Such annual visits scheduled within 90 days (before or after) the implant anniversary will be used for annual outcome analysis. Other visits, either initiated by the patient or by the clinic will be recorded and included in the overall safety and efficacy analysis ("unscheduled visits"). In case more than one visit was done during the annual visit window then the later of those visits will be used for the annual data analysis.

Table 4: Visit Windows

Study Visit Name		Window	Visit Target[6]
Screening[1,2]		Any CRF entered in Screening visit	Any CRF entered in Screening visit
Baseline[3]		Any CRF entered in Baseline visit	Any CRF entered in Baseline visit
GERD-HRQL for inclusion		At least 7 days after resuming max PPI therapy	**Not used in final analysis*
Implant Procedure		≤6 Months After Baseline Visit	Implant Date
Week 2 / Randomization		7-14 Days After Implant Procedure	14 Days
Month 1		16-44 Days	30 Days (± 14 Days)
Month 2 Phone Interview [1,2]		46-74 Days	60 Days (± 14 Days)
3 Month		76-104 Days	90 Days (± 14 Days)
4 Month Phone Interview[1,2]		106-134 Days	120 Days (± 14 Days)
6 Month[4]		159-201 Days	180 Days (± 21 Days)
Open label (OL) portion of the study begins			
7 Month **(DS ONLY)**	*(1 Month OL)*	196-224 Days	210 Days (± 14 Days)
8 Month Phone Interview[1,2] **(DS ONLY)**	*(2 Month OL)*	226-254 Days	240 Days (± 14 Days)
9 Month[1,2,5] **(IS ONLY)**		240-300 Days	270 Days (± 30 Days)
9 Month[1,2] **(DS ONLY)**	*(3 Month OL)*	256-284 Days	270 Days (± 14 Days)
10 Month Phone Interview[1,2] **(DS ONLY)**	*(4 Month OL)*	286-314 Days	300 Days (± 14 Days)
12 Month[5] **(IS ONLY)**		330-390 Days	360 Days (± 30 Days)
12 Month[4] **(DS ONLY)**	*(6 Month OL)*	339-381 Days	360 Days (± 21 Days)
15 Month[1,2,5] **(DS ONLY)**	*(9 Month OL)*	420-480 Days	450 Days (± 30 Days)
18 Month[5] **(DS ONLY)**	*(12 Month OL)*	510-570 Days	540 Days (± 30 Days)
18 Month Phone Interview[5] (**IS ONLY**)		510-570 Days	540 Days (± 30 Days)
24 Month Phone Interview[5] (**DS ONLY**)	*(18 Month OL)*	690-750 Days	720 Days (± 30 Days)
24 Month[1,5] **(IS ONLY)**		690-750 Days	720 Days (± 30 Days)
30 Month[1,5] **(DS ONLY)**	*(24 Month OL)*	870-930 Days	900 Days (± 30 Days)
36 Month[1,5] **(IS ONLY)**		1050-1110 Days	1080 Days (± 30 Days)
42 Month[1,5] **(DS ONLY)**	*(36 Month OL)*	1230-1290 Days	1260 Days (± 30 Days)
48 Month[1,5] **(IS ONLY)**		1410-1470 Days	1440 Days (± 30 Days)
54 Month[1,5] **(DS ONLY)**	*(48 Month OL)*	1590-1650 Days	1620 Days (± 30 Days)
60 Month[1,5] **(IS ONLY)**		1770-1830 Days	1800 Days (± 30 Days)
66 Month[1,5] **(DS ONLY)**	*(60 Month OL)*	1950-2010 Days	1980 Days (± 30 Days)

[1]Subjects are instructed to start Daily Subject Diary ≥10 Days before Off-PPI and H2 Blockers Period.

[2]Subjects are instructed to stop PPI and H2 Blockers 10-14 Days Before next visit
[3]Pre-Implant is defined as the last measurement for the outcome of interest obtained before the activation of the device (except in the case when the intent is to compare to pre-implant "on-PPI" data, in which case measurements from the Screening visit would be utilized).
[4]Visit Window is ±21 days, rather than the initial ±14 days.
[5]Visit Window is ±30 days, rather than the initial ±14 days.
[6]Visit Windows are calculated as time from the Implant Date, unless otherwise noted.

5.1.7 Duration Variables

There are four possible dates of interest for this study; 1) date prior to implant where the majority of baseline pre-implant measurements were taken, 2) date of implant, 3) date of randomization, 4) date stimulation begins (on week 2 post implant visit/randomization for immediate stimulation and 6 months for delayed stimulation).

Due to the operational procedures, the time between baseline measurements and device implant will likely vary. This variable will be considered a demographic variable using the following formula.

Time Until Surgery = (Date of Implant – Date of Baseline Visit)

For tracking adverse events related to the surgery, time must be measured with respect to the implant. Duration of time from implant will be calculated using this formula.

AE Day = (Date of Event – Date of Device Implant)

For analyzing adverse events related to stimulation, time must be measured with respect to stimulation initiation.

AE Stimulation Day = (Date of Event – Date of Stimulation Initiation)

For all subjects, study day 0 is the day of randomization, which occurs approximately two weeks after the implantation of the LES. Study day is calculated relative to day 0 and will appear in the listings where applicable. However, since "baseline" measurements were not captured on this date this variable cannot be used to determine number of days between baseline and post measurements. Study day will be calculated as:

Study Day = (Date of Event – Date of Randomization)

5.2 Analysis Populations

All trial endpoints will be analyzed using the Intention-to-Treat (ITT) population, with the ITT analysis *a priori* designated as the primary analysis, however support from the agreement of the per-protocol is required for study success. If the safety set differs from the ITT set, sensitivity analyses may be conducted on the primary safety and efficacy endpoints using the safety set.

5.2.1 Intention-to-Treat Analysis Set

The Intention-to-Treat (ITT) set will be comprised of all subjects who successfully complete the preliminary qualification procedures and are subsequently randomized to receive immediate LES stimulation treatment (IS) or 6 month delayed LES stimulation (DS) regardless of stimulation duration and specifications.

5.2.2 Per-Protocol Set

The Per-Protocol (PP) population for the blinded portion of the study will consist of ITT subjects who had no major protocol deviations.

The Per-Protocol (PP) population for the open label portion of the study will consist of ITT subjects who were eligible to be included in the study at their pre-stimulation visit (IS: pre-implant, DS: 6 months) and had no major protocol deviations.

Prior to interim analyses, major protocol deviations will be identified by EndoStim and the DMC. Major deviations will include inclusion/exclusion violations that impact the ability to assess the safety and/or efficacy of the study device in the intended population, or procedural violations that change the intended treatment. Exclusions due to major protocol deviations will be defined prior to evaluation of outcomes and reasons for exclusion will be provided.

5.2.3 Safety Analysis Set

In the unlikely case where an implant procedure was attempted and not completed or where a subject undergoes the surgical procedure but they do not participate in the randomization visit (2 weeks post-procedure), said subject will not be included in the ITT set since they were not randomized to a treatment group, but included in the safety set since the subject underwent an attempt at implant. The safety set will be comprised of those subjects whom underwent an attempt at implanting a study device and will include subjects as they are treated in the case where stimulation duration and specifications differs from their randomization assignment.

5.3 Poolability of Data

All subjects will be unblinded to their treatment assignment at month 6. At this time the DS group will begin stimulation and be followed throughout the duration of the study. For outcomes related to a duration of stimulation IS and DS groups maybe be pooled if they are determined to be homogeneous. Further details regarding methods of pooling are in their relevant sections below.

5.4 Primary Safety Endpoint

The primary safety endpoint will estimate the rate of occurrence of device- and/ or procedure-related serious adverse events at the 12-month follow-up visit. To clarify, "procedure-related" refers to a surgical procedure to implant, revise or explant the study device.

Data provided by the site will include an event severity, relatedness, timing and MedDRA category. Safety data will then be adjudicated by the Clinical Events Committee. The outcome of these adjudications will be used for the primary safety endpoint. If at the time of interim analysis or DMC reporting, events have been reported by site but not yet adjudicated then the data reported by the site will be utilized.

5.4.1 Primary Analysis of the Primary Safety Endpoint

If data are allowed to be pooled after the month 12 visit, (12 months of stimulation for IS and 6 months of stimulation for DS), the percentage of patients with an AE will be calculated. If the chi-squared test does not allow pooling, then percentage of subjects with AEs will be stratified by randomization group and presented separately for each group for each MedDRA System Organ Class category.

In addition to the rate of occurrence, an exact 95% two-sided confidence interval will be computed.

5.4.2 Missing Data Analysis for Primary Safety Endpoint

If there is a difference between the ITT and the safety dataset, the safety dataset will be used for the primary safety analysis. Missing data related to the safety endpoint will not be imputed.

5.4.3 Poolability of Primary Safety Endpoint

A test for homogeneity of IS and DS patients' rate of overall serious adverse events at 6 months post implant and at 12 months post implant will be conducted. We will use a two sample chi-square tests to compare the group rates; poolability will be accepted if both the tests have p-values > 0.10.

To test for site/geographic poolability of the data, we will examine forest plots of the site-specific and geography-specific estimated adverse events. Formal tests for poolability will be conducted by including an interaction of (i) site and treatment arm and (ii) geography by treatment arm in our model. Sites with fewer than 5 subjects will be grouped together within country for analysis purposes.

5.5 Primary Efficacy Endpoint

The primary efficacy endpoint is the percentage (%) of subjects achieving pH success defined as normalization (pH < 4 for no more than 5.3% of monitoring time) or $\geq$ 50% improvement in their distal esophageal acid exposure at six months compared to their pre-implant distal esophageal pH both performed after at least 5 days ***off proton pump inhibitors*** (PPIs) and at least 2 days ***off histamine 2 receptor antagonists*** (H2 blockers) typically collected during the baseline assessment.

Minimum 18 hours of esophageal pH recording will be considered adequate for all pH analyses. Data collected beyond 48 hours of esophageal pH recording will not be evaluated or included in any analysis.

5.5.1 Hypothesis for Primary Efficacy Endpoint

The working hypothesis is that the immediate stimulation group will have a response rate superior to the delayed stimulation group.

$$logit(\pi_i) = \beta_0 + \beta_1 Active + \beta_2 Crural$$

H_0: $\beta_1 \leq 0$

H_1: $\beta_1 > 0$

5.5.2 Interim Analysis of the Primary Efficacy Endpoint

At each interim analysis (as described in section 4.5), we calculate the predictive probability of trial success at the final analysis for the currently implanted patients. This calculation assumes that the primary analysis would take place after all currently implanted patients complete their follow-up for the primary endpoint.

If, at an interim analysis, the predictive probability of trial success (as described in section 4.4) with the currently implanted patients is at least 95%, this trial may stop accrual early for expected success. The final analysis would be conducted after all implanted patients complete the follow-up for the primary endpoint.

5.5.3 Informational Analysis of the Primary Efficacy Endpoint

If the trial does not stop early by the interim analysis with 180 patients implanted, it will accrue to the maximum sample size of 200 implanted patients. Once the maximum number of patients has been implanted, an informational look at the data will be performed. The predictive probability of trial success at the final analysis when all 200 implanted patients have completed follow-up will be calculated.

This predictive probability will be communicated to the Data Monitoring Committee (DMC). The DMC will inform the trial sponsor whether or not the predictive probability of trial success is at least 80%. The actual predictive probability value will not be communicated to the sponsor. This informational look will have no impact on the stopping or continuation of this trial, instead it will aid sponsor planning for other projects in the clinical program.

5.5.4 Futility Analysis of the Primary Efficacy Endpoint

At each interim analysis, the predictive probability that the current trial will be successful given it is continued to the maximum of 200 implanted patients and all patients are followed for the primary endpoint.

If, at an interim analysis, the predictive probability of trial success at the maximum sample size is less than 1%, this trial will stop early for futility.

5.5.5 Final Analysis of the Primary Efficacy Endpoint

The final analysis for the primary efficacy endpoint will compare the response rates between immediate and delayed stimulation using a logistic regression. The primary predictor of interest will be a dummy variable for active vs. delayed stimulation. The model will also control for crural approximation as this is potentially an independent predictor of stimulation success.

Superiority for the final primary efficacy endpoint will be satisfied if the one-sided p-value for the primary predictor is less than 0.022.

We will conduct additional analyses adjusting for other important baseline characteristics, including those that may be significantly different at baseline despite randomization, age, gender, and hernia status.

5.5.6 Missing Data Analysis of the Primary Efficacy Endpoint

A minimum 18 hours of esophageal pH recording after a minimum of 5 days off-PPI and a minimum of 2 days off of H2 will be considered an adequate pH study. We anticipate no missing data for baseline pH in included subjects and no more than 5% to 10% missing data for 6-month pH.

In the event of any missing values, we will use linear regression-based multiple imputation of the 6-month value based on baseline and 3-months (where available) for the primary analysis. The imputation will be implemented using multiple imputation in Proc MI.

Because we acknowledge the use of rescue medications in deviation with the protocol may have an impact on measured pH results in the ITT population. We will set the pH values of subjects with such deviations to missing and also impute these values using Proc MI.

5.5.7 Sensitivity Analysis of the Primary Efficacy Endpoint

In addition to the primary analysis that uses multiple imputation (MI) methods to account for missing data (described in section 6), a tipping point analysis will be employed to assess sensitivity to missing data imputation. This tipping point analysis will examine the extent to which imputing systematically different values for missing delayed and active stimulation subjects would affect the trial results. This analysis will be conducted on the ITT set.

5.6 Secondary Efficacy Endpoint (Patient Focused)

Number of subjects achieving GERD symptom success defined as an improvement in their composite GERD-HRQL score of 50% or more after 12-months of electrical stimulation therapy (12-month visit for immediate stimulation group and 18-month visit for the delayed stimulation group) compared to their baseline ***off-PPI*** and H2 blocker GERD-HRQL score.

Previous versions of this protocol did not include question #9 regarding bloating. Multiple imputation will be implemented for any missing values of this question, in order to calculate a composite GERD-HRQL score.

5.7 Blinded Endpoints

Table 5: Schedule for Collecting Blinded Endpoints

Visit Interval	Screen S	**Base** **B**	Impl Proc IMP	PI Week2 & Rand 0 days	1 M 30 days	2 M 60 days	3 M 90 days	4 M 120 days	**6 M** 180 days
Bravo Esophageal pH Monitoring		**X**					X		**X**
Subject Diary Collection		**X**					X		**X**
Quality of Life (SF-12)	X	**X**							**X**
Endoscopy		**X**[1]							**X**
Pittsburgh Sleep Quality Index (PSQI)	X	**X**							**X**
GERD-HRQL	X[2, 3]	**X**[3]		X	X	X	X	X	**X**

[1] Assess availability and acceptability of prior data collection within the last 6 months. If prior data collection is not used, one test required prior to implant

[2] This on-PPI GERD-HRQL assessment will not be used to assess inclusion criteria. This will be the ***on-PPI*** score that is used for pre-implant comparisons for relevant endpoint evaluations.

[3] Q9 was missing from the version utilized at the beginning of the study, once detected it was resolved. Any missing data around question #9 will be imputed using multiple imputation.

5.7.1 Secondary Blinded Endpoints

5.7.1.1 Hypotheses of the Secondary Blinded Endpoints

The working hypothesis for positive (high values are preferable) secondary scales is that the mean values in the immediate stimulation group (IS) will be greater than the mean values of the delayed stimulation group (DS). One sided p-values will be calculated and tested.

$$H_0: \mu_{IS} \leq \mu_{DS}$$

$$H_1: \mu_{IS} > \mu_{DS}$$

The working hypothesis for negative (lower values are preferable) secondary scales is that the mean values in the immediate stimulation group (IS) will be less than the mean values of the delayed stimulation group (DS).

$$H_0: \mu_{IS} \geq \mu_{DS}$$

$$H_1: \mu_{IS} < \mu_{DS}$$

5.7.1.2 Analyses of the Secondary Blinded Endpoints

For secondary blinded endpoints, we will compare the immediate stimulation arm to the delayed stimulation arm of either the mean change values (for continuous measures) from pre-stimulation to post using two-sample t-tests or the proportion of successful subjects (for binary measures) using Fisher's exact test unless otherwise defined.

5.7.1.3 Power Considerations for Secondary Blinded Endpoints

In order to support claims, secondary endpoints should be adequately powered. For the blinded-phase secondary endpoints, we consider the power for the case where the blinded-phase portion of the trial is stopped after 120 total subjects, at the first interim analysis. Power would of course increase if the study accrual is higher. Since many of the secondary endpoints are strongly related to pH, a large effect size for the primary endpoint likely corresponds to large effect sizes for the secondary endpoints. Therefore, the power for the secondary endpoints is expected to be adequate even if a large effect size for the primary endpoint results in stopping trial accrual at one of the smaller total sample sizes (e.g. 120 or 140).

For a total of 120 subjects divided evenly into delayed stimulation and immediate stimulation groups, we will be approximately 80% powered using the overly conservative one-sided level of 0.0018 to detect an effect size between the two groups equal to 70% of the standard deviation for continuous endpoints. Similarly, for binary endpoints, we will be approximately 80% powered to detect a difference in success rate of approximately 25 to 35 percentage points (e.g. Delayed Stimulation 68% vs. Immediate Stimulation 93%; Delayed Stimulation 33%, Immediate Stimulation Group 66%). If the trial does not stop accrual in the first two accrual phases and carries on to implant 160 subjects between the groups we would have 80% power with the 0.0018 alpha to detect an effect size between the groups equal to 63% of the standard deviation for continuous endpoints and 80% powered to detect a difference in success rate of approximately 22 to 29 percentage points for binary endpoints.

We note that many of the continuous endpoints improved from Baseline to Follow-up by 1 or more standard deviations. For example, mean percentage of time distal esophageal pH is <4.0 improved from 10.8% (5.1%) to 5.3% (4.9%) with the difference better than 1 SD. HRQL improved from 10.8 (7.3) to 3.1 (3.9), with the difference equal to 1 SD; heartburn symptom score from 1.9 (0.5) to 0.4 (0.6), better than 2 SD; similar magnitude in improvements seen for nocturnal heartburn and regurgitation and for levels of GERD medication use. This suggests that we would be adequately powered to see a difference between groups even if the placebo or crural repair alone effect on the change from baseline, which is expected to be equal in both groups, were substantial. For the binary endpoints of GERD improvement/control and nocturnal GERD improvement/control, we would be 80% powered if the true rates are 70% immediate stimulation (as in our preliminary data) vs. 33% delayed stimulation and 55% powered for 70% vs. 40%.

The power calculations above use a conservative worst-case alpha value determined by a Bonferroni adjustment for multiplicity. The actual power is expected to be higher since the alpha that will be used to assess p-values for secondary endpoints will be determined by the Westfall bootstrap method described in Section 5.1.3 that accounts for correlation among the secondary endpoints.

5.7.1.4 Measurements for Secondary Blinded Endpoints

5.7.1.4.1 Bravo Esophageal pH Monitoring - Time of pH<4.0

A fraction of time where patients' distal esophageal pH<4 (using Bravo Esophageal pH Monitoring performed after at least 5 days off-PPI and at least 2 days off H2 blockers) will be reported at each time point. Change of percentage from baseline to 6 months visit will be calculated for each patient, averaged within group and compared between group.

5.7.1.4.2 Acid Suppression Medication Use

Changes in the number of days where acid-suppression medication (minimum of 10 days to utilize diary) was reported between baseline and 6 months visit will be calculated within patient and summarized as group means and standard deviations and compared between groups.

5.7.1.4.3 Heartburn Symptoms

Heartburn symptoms severity will be assessed as measured by the Subject Diary. The daily highest level of severity will be averaged over the reported number of days (Max:14 days, Min: 10 days). The severity scale is constructed with no symptoms equal to 0, mild symptoms equal to 1, moderate equal to 2, and severe equal to 3.

Mean severity symptom changes and standard deviations will be calculated and compared between IS and DS.

In addition, frequency of days/nights with heartburn will be calculated as the percent of the reported time with any heartburn symptoms. Changes from baseline will be calculated within person and the mean compared between groups.

5.7.1.4.4 Regurgitation Symptoms

Regurgitation symptoms severity will be assessed as measured by the Subject Diary. The highest level of severity will be averaged over the number of reported time periods. (Max: 14 days, Min: 10 days) The severity scale is constructed with no symptoms equal to 0, mild symptoms equal to 1, moderate equal to 2, and severe equal to 3.

Mean symptom severity and standard deviations will be calculated and the mean compared between IS and DS.

In addition, frequency of days/nights with regurgitation will be calculated as the percent of reported time with any regurgitation symptoms. Changes from baseline will be calculated within person and the mean compared between groups.

5.7.1.4.5 Short Form Health Survey (SF-12)

Quality of life will be assessed by the SF-12. Higher composite scores are favorable.

Within person changes from pre-implant to 6 months visit will be calculated, summarized as means and standard deviations and compared between groups.

5.7.2 Other Endpoints of Interest during Blinded Phase

5.7.2.1 Bravo Esophageal pH Monitoring - Time of pH<4.0

Time, from a minimum of 18 hours of data collection, where the subject's distal supine esophageal pH <4.0 will be reported as a percentage of hours collected. Mean percentage change will be calculated within group and compared between groups.

A binary indicator will be calculated if subjects achieve supine pH success, defined as normalization (pH<4 for no more than 1.4% of supine period at 6 months visit) or ≥50% improvement in their supine distal esophageal acid exposure

compared to their baseline off-PPI and H2 blocker supine distal esophageal pH. A percentage of subjects achieving this definition of pH success will be created within each group and compared between them.

5.7.2.2 Bravo Esophageal pH Monitoring - Reflux Events

The total number of reflux events reported from the Bravo Esophageal pH monitoring report will be calculated. Changes from baseline to 6 months visit will be summarized and the mean compared between groups.

5.7.2.3 Bravo Esophageal pH Monitoring - DeMeester Score

The DeMeester Score is a composite score calculated by the Bravo Esophageal pH test results. The mean change from baseline to 6 months visit will be compared between groups.

5.7.2.4 Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire (*off-PPI* GERD-HRQL) score

Composite scores for GERD-HRQL will be calculated from data entered into CRFs. Pre-implant HRQL scores will utilize ***off-PPI*** scores captured during the baseline visit. Lower composite scores are favorable. Mean differences and standard deviations will be compared between groups.

Individual scores related to heartburn (question #1), difficulty swallowing (question # 7), bloating (question #9) and patient satisfaction (question #11) will be calculated as changes from baseline and compared between groups.

Previous versions of this protocol did not include question (question #9) regarding bloating. Multiple imputation will be implemented for any pre-implant missing values used in individual and composite scoring.

5.7.2.5 Pittsburgh Sleep Quality Index (PSQI)

The PSQI is a composite score calculated from data entered into CRFs measuring patients' self-reported quality of sleep. Mean changes and standard deviations from pre-implant to 6 months visit the mean will be compared between groups.

5.7.2.6 Mucosal Injury

The extent of mucosal injury from reflux will be evaluated during endoscopy and reported utilizing the LA classification. Binary indicators will be created for LA Classification improving, worsening, and holding from pre-implant to 6 months visit. Percentages will be compared between groups using chi squared test.

Note: Inclusion criteria required baseline LA classification to be less than or equal to Grade B. Subjects diagnosed via endoscopy with Grade C or D esophagitis at their month 6 endoscopy will be treated with the same pre-implant dose and brand of PPI that had previously been proven to heal the esophagitis in that particular patient. Patients will undergo a repeat endoscopy at their month 12 visit and those with persistent Grade C or D esophagitis will be exited from the study.

Additionally, for those patients in the delayed stimulation group who have a Grade C or D at 6 months when they begin stimulation will be excluded from the PP analysis for the open label portion of the study since they wouldn't have been eligible for the study (to receive stimulation) had that been their test result at baseline. These patients will remain in the ITT set until they are either exited at 12 months as described above or their LA classification reduces to Grade B.

5.8 Open Label Endpoints

Table 6: Schedule for Collecting Open Label Endpoints

Delayed Stimulation	Pre-Implant (Baseline or Screening)	7M	8M	9M	10M	12M	15M	18M
		210 days	240 days	270 days	300 days	360 days	450 days	720 days
Immediate Stimulation	Pre-Implant (Baseline or Screening)	1M	2M	3M	4M	6M	9M	12M
		30 days	60 days	90 days	120 days	180 days	270 days	540 days
Adverse Event Assessment	X	X	X	X	X	X	X	X
Bravo Esophageal pH Monitoring	X			X		X	X[3]	X
Subject Diary Collection	X			X		X	X	X
Endoscopy	X[1]					X[4]		X[4]
Pittsburgh Sleep Quality Index (PSQI)	X					X		X
GERD-HRQL Composite	X[2]	X	X	X	X	X	X	X
GERD-HRQL Q1	X[2]	X	X	X	X	X	X	X
GERD-HRQL Q7	X[2]	X	X	X	X	X	X	X
GERD-HRQL Q9	X[2]	X	X	X	X	X	X	X
GERD-HRQL Q11	X[2]	X	X	X	X	X	X	X
Esophageal Manometry (HRM) LES-MP & LES-EEP (Rev H)	X[1]					X		
Physical Exam	X	X		X		X	X	X
Reflux Symptom Index (RSI)	X	X	X	X	X	X	X	X
SF-12	X					X		X
Structured GI	X					X		X
WPAI:SHP	X					X		X
Healthcare Resource Utilization Questionnaire (Rev H)	X[2]					X		X

[1]Assess availability and acceptability of prior data collection within the last 6 months (See Section 6.1 of Rev H of the protocol). If prior data collection is not used, one test required prior to implant.

[2]All Pre-implant measurements are taken at the baseline visit no longer than 6 months prior to implant ***EXCEPT*** for the ***on-PPI*** GERD-HRQL and healthcare utilization questionnaire. If BMI is collected more than once pre-implant, the most recent measurement will be taken.

[3] pH monitoring is optional per investigator discretion based on subject symptoms. An endoscopy may be used to aid in placement of the Bravo capsule.

[4] A diagnostic endoscopy is required at 12 months only if esophagitis was present at the 6-month visit. An endoscopy may be used to aid in placement of the Bravo capsule.

5.8.1 Secondary Open Label Endpoints

5.8.1.1.1 Hypotheses for Secondary Open Label Efficacy Endpoints

The working hypothesis for secondary open label endpoints that are of positive scales, is that the average difference between the 12-month post stimulation and pre-implant values will be equal to or greater than 60% improvement unless otherwise noted.

$$H_0: \mu_{12} < .6$$

$$H_1: \mu_{12} \geq .6$$

Where μ_{12} equals the difference between the 12-month post stimulation and pre-implant. One sided p-values will be calculated and tested and adjusted by the Multtest procedure.

The working hypothesis for secondary open label endpoints that are of negative scales, is that the average difference between the 12-month post stimulation and pre-implant value will be less than or equal to 60% reduction.

$$H_0: \mu_{12} > -.6$$

$$H_1: \mu_{12} \leq -.6$$

5.8.1.1.2 Analyses for Secondary Open Label Efficacy Endpoints

For secondary open label endpoints assuming that these data are judged to be poolable between the two arms (see section 5.8.1.1.4 below), we will compare either the mean change values (for continuous measures) from pre-stimulation to 12-months post stimulation (the 12-month time point for the immediate stimulation arm and the 18-month time point for the delayed stimulation arm, after 12 months of stimulation and 6 month of delayed stimulation) using one-sample t-tests or the proportion of successful subjects (for binary measures) using one sample test of proportion, unless otherwise defined.

Additionally, unadjusted two-sided 95% CI will be calculated solely for descriptive summary purposes.

5.8.1.1.3 Power Considerations for Secondary Open Label Efficacy Endpoints

At trial design it was conservatively assumed that each endpoint is to be tested at a one-sided alpha of 0.025/14 = 0.0018. For a particular secondary endpoint with true success rate of 76%, we would have 80% power to reject the null hypothesis that the success rate was 60% or lower. This suggests the study would be strongly powered for 9 of the 10 open-label secondary endpoints and better than 50% powered for the remaining endpoints. For the open-label endpoints, we will implement the Westfall bootstrap which will account for correlation between the endpoints while maintaining the familywise error rate of 0.025.

5.8.1.1.4 Poolability for Secondary Open Label Efficacy Endpoints

For all endpoints, we will examine:

(i) for the immediate stimulation (IS) subjects, the change from pre-implant (T1) to the 12-month visit (T2).

(ii) for the delayed stimulation (DS) subjects, the change from pre-implant (T1) to the 18-month visit (T2), and

As the DS subjects are unblinded for full duration of their 12-month treatment and the IS subjects are unblinded for half (6 month) duration of their 12-month treatment, there may be a potential concern as to the poolability of the 2 groups for describing 12-month therapy effect, especially on outcomes that may be susceptible to the sham effect. To address any such concerns, we propose the following steps:

(a) A test for homogeneity of IS and DS patients' outcomes at T1 and T2 (baseline and post 12 month of active stimulation). We will use a two-sample t-test to compare the group means; poolability will be accepted if both the two-sided t-tests have p-values ≥ 0.10.

(b) if poolability accepted in (a), then pool subjects to compute estimate and one-sided p-value for that specific outcome after 12 months on stimulation. P-values will be adjusted using the Westfall bootstrap method for multiple testing. Should the adjusted p-value be <.025 then the endpoint will be met.

(c) if poolability not accepted in (a), then compute estimates and one-sided p-value for that specific outcome after 12 months on stimulation separately in each group. Each p-value will be included in the Westfall bootstrap method and adjusted for multiple testing. Should the p-value be <0.025 then the endpoint will be met.

5.8.1.2 Measurements for Secondary Open Label Endpoints

The open label portion of the study includes all the outcomes described below. Outcomes will utilize the same scoring manuals and algorithms as in the blinded portion of the study; however, statistics in the open label analysis will be evaluating average within person pre-implant to 12 months post stimulation changes. LES-MP and LES-EEP collected through esophageal manometry are the only two exceptions since they will be measured from pre-implant to 6 months post stimulation.

5.8.1.3 Bravo Esophageal pH Monitoring - Time of pH<4.0

Time, from a minimum of 18 hours of data collection, where the subject's distal supine esophageal pH <4.0 will be reported as a percentage of hours collected. Mean percentage change will be compared to the performance goal of 60% improvement in the group.

5.8.1.4 Heartburn Symptoms

Heartburn symptoms severity will be assessed as measured by the Subject Diary. The severity of the worst episode over the last 24 hours will be averaged over the reported number of days (Max:14 days, Min: 10 days). The severity scale is constructed with no symptoms equal to 0, mild symptoms equal to 1, moderate equal to 2, and severe equal to 3. Changes from baseline will be calculated.

A binary indicator will be created to determine if a patient achieves a 50% or more reduction in the mean worst heartburn symptom severity from baseline to 12 months post stimulation. This proportion will then be compared to see if it is greater than or equal to a 60% improvement in the group.

A binary indicator will be created to determine if a patient achieves a 50% or more reduction in frequency of nocturnal symptoms of heartburn. This proportion will then be compared to see if it is greater than or equal to a 60% improvement in the group.

5.8.1.5 Regurgitation Symptoms

Regurgitation symptoms severity will be assessed as measured by the Subject Diary. The severity of the worst episode over the last 24 hours will be averaged over the number of reported time periods. (Max: 14 days, Min: 10 days). The severity scale is constructed with no symptoms equal to 0, mild symptoms equal to 1, moderate equal to 2, and severe equal to 3. Changes from baseline will be calculated.

A binary indicator will be created to determine if a patient achieves a 50% or more reduction in the mean worst regurgitation symptom severity from baseline to 12 months post stimulation. This proportion will then be compared to see if it is greater than or equal to a 60% improvement in the group.

A binary indicator will be created to determine if a patient achieves a 50% or more reduction in frequency of nocturnal symptoms of regurgitation. This proportion will then be compared to see if it is greater than or equal to a 60% improvement in the group.

5.8.1.6 Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire (*on-PPI* GERD-HRQL) score

Composite scores for GERD-HRQL will be calculated from data entered into CRFs. Pre-implant HRQL scores will utilize on PPI scores captured during the screening visit. Lower composite scores are favourable. Mean differences and standard deviations will be compared to the performance goal of 60% improvement in the group.

5.8.1.7 Acid Suppression Medication

Binary indicators will be calculated for patients whose daily diary reports 50% or more of diary days (minimum of 10 days to utilize diary) without regular use of acid-suppression medication. Percentage of patients meeting criteria will be calculated and compared to the performance goal of 60% improvement in the group.

Additionally, a binary indicator will be calculated for patients reporting 100% diary days (minimum of 10 days to utilize diary) without regular use of acid- suppression medication. Percentage of patients meeting criteria will be compared to the performance goal of 60% improvement in the group.

5.8.1.8 Short Form Health Survey (SF-12)

Quality of life will be assessed by the SF-12. Higher composite scores are favourable.

Within person changes from pre-implant to 12 months post stimulation will be summarized as within person percent improvement from pre-implant to 12 months post stimulation. Percentage improvement in the group will be compared to the performance goal of 60%.

5.8.2 Other Endpoints of Interest during Open Label Phase

5.8.2.1 Other Safety Endpoints

The event rate at each time point will be determined by the adverse event start date of the first event meeting the criteria for each device or procedure related complication. Rates will be reported as proportions and stratified by MedDRA category. Device and/or procedure related complications will be classified following final CEC adjudication and reported separately by MedDRA System Organ class.

Frequencies and percentages will be calculated and summarized.

5.8.2.2 Bravo Esophageal pH Monitoring - Time of pH<4.0 - Overall

A binary indicator will be calculated if subjects achieving pH success defined as normalization (pH < 4 for no more than 5.3% of monitoring time) or >50% improvement in their distal esophageal acid exposure compared to their baseline ***off-PPI and H2 blocker*** distal esophageal pH. A percentage of subjects achieving this definition of pH success will be created from baseline to post stimulation and compared to the performance goal of 60% improvement.

5.8.2.3 Bravo Esophageal pH Monitoring - Time of pH<4.0 - Supine

A binary indicator will be calculated if subjects achieve supine pH success, defined as normalization (pH<4 for no more than 1.4% of supine period at post) or ≥50% improvement in their supine distal esophageal acid exposure compared to their baseline ***off-PPI and H2 blocker*** supine distal esophageal pH. A percentage of subjects achieving this definition of pH success will be created from baseline to post stimulation and compared to the performance goal of 60% improvement.

5.8.2.4 Bravo Esophageal pH Monitoring - Reflux Events

The total number of reflux events reported from the Bravo Esophageal pH monitoring report will be calculated. Percent change from baseline to 12 months post stimulation will be compared within group.

5.8.2.5 Bravo Esophageal pH Monitoring - DeMeester Score

The DeMeester Score is a composite score calculated by the Bravo Esophageal pH test results. The mean change from baseline to 12 months will be compared to the performance goal of 60% improvement.

5.8.2.6 Pittsburgh Sleep Quality Index (PSQI)

The PSQI is a composite score calculated from questions recorded in the CRF measuring patients self-reported quality of sleep. Number of subjects reporting 50% or more improvement in sleep related quality of life will be calculated and compared to the performance goal of 60% improvement.

Mean change and standard deviations will be summarized for descriptive purposes only.

5.8.2.7 Mucosal Injury

The extent of mucosal injury from reflux will be evaluated during endoscopy and reported utilizing the LA classification. Indicators will be created for LA Classification holding, improving, and worsening from pre-implant to 12 months post stimulation. The percentage of patients improving will be compared to the performance goal of 60%.

Note: Inclusion criteria required baseline LA classification to be less than or equal to Grade B. Subjects diagnosed via endoscopy with Grade C or D esophagitis at their month 6 endoscopy will be treated with the same pre-implant dose and brand of PPI that had previously been proven to heal the esophagitis in that particular patient. Patients will undergo a repeat endoscopy at their month 12 visit and those with persistent Grade C or D esophagitis will be exited from the study.

Additionally, for those patients in the delayed stimulation group who have a Grade C or D at 6 months when they begin stimulation will be excluded from the PP analysis for the open label portion of the study since they wouldn't have been eligible for the study (to receive stimulation) had that been their test result at baseline. These patients will remain in the ITT set until they are either exited at 12 months as described above or their LA classification reduces to Grade B.

5.8.2.8 Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire (*off-PPI* GERD-HRQL) score

Composite scores for GERD-HRQL will be calculated from data entered into CRFs. Pre-implant HRQL scores will utilize ***off-PPI*** scores captured during the baseline visit. Lower composite scores are favorable. Mean differences and standard deviations will be compared to the performance goal of 60% improvement.

Individual change scores related to heartburn (question #1), difficulty swallowing (question # 7), bloating (question #9) and patient satisfaction (question #11) will be compared within groups to the hypothesized 60% improvement.

Previous versions of this protocol did not include question (question #9) regarding bloating, multiple imputation will be implemented for any pre-implant missing values used in individual and composite scoring.

5.8.2.9 Heartburn Symptoms

Heartburn symptoms severity will be assessed as measured by the Subject Diary (Max:14 days, Min: 10 days). The severity scale is constructed with no symptoms equal to 0, mild symptoms equal to 1, moderate equal to 2, and severe equal to 3.

The percent change in the number of days/nights with severe symptoms reported will be calculated and compared to the performance goal of 60%.

Changes from baseline in the frequency of days and nights with reported heartburn episodes will be compared to the performance goal of 60%.

5.8.2.10 Regurgitation Symptoms

Regurgitation symptoms severity will be assessed as measured by the Subject Diary (Max: 14 days, Min: 10 days). The severity scale is constructed with no symptoms equal to 0, mild symptoms equal to 1, moderate equal to 2, and severe equal to 3.

The percent change in the number of days/nights with severe symptoms reported will be calculated and compared to the performance goal of 60%.

Changes from baseline in the frequency of days and nights with reported heartburn episodes will be compared to the performance goal of 60%.

5.8.2.11 Structured GI Questionnaire

GERD symptoms will be assessed by the Structured GI questionnaire, a self-report questionnaire evaluating the duration of 13 symptoms, including: heartburn, chest pain, regurgitation, lung problems, swallowing, overall pain, nausea/vomiting, bloating, gas and bowel movements. There is no algorithm for a composite score, therefore individual binary assessments for symptoms not including heartburn and regurgitation will be assessed and compared. Changes from baseline will be evaluated within person.

5.8.2.12 Work Productivity and Activity Impairment Questionnaire: Specific Health Problem (WPAI:SHP)

How patients' GERD symptoms affect their work and activity levels are calculated from a health problem specific version of the WPAI. Mean changes from pre-implant to 12-months post stimulation will be calculated and evaluated in comparison to the 60% performance goal.

Scales including missing data have no validated imputation procedures and will not be imputed.

5.8.2.13 Lower Esophageal Sphincter Mean Pressure (LES-MP)

Lower esophageal dysfunction will be measured by calculating the mean difference in lower esophageal pressure from pre-implant to 6-months post stimulation using high resolution manometry (HRM) (Rev H). A binary indicator will be created if a patient is within the category defined as "normal" (defined as ≥8mmHg), resulting in the proportion of patients at both pre and post stimulation which are falling within a normal range.

Pre-implant levels of LES-MP can be collected from evaluable (criteria listed in section 6.1 of Rev H of the protocol) HRM results captured 6-months prior to baseline evaluations. If no other results are present, then an HRM test is required prior to implant.

Change in proportion of patients that move from low into normal range will be reported and compared to 60% performance goals, where increased proportions are favorable. Mean differences and standard deviations in pressure will be reported for descriptive purposes.

5.8.2.14 Lower Esophageal Sphincter End Expiratory Pressure (LES-EEP)

Lower esophageal dysfunction will be measured by calculating the change in expiratory pressure from pre-implant to 6-months post stimulation using high resolution manometry (HRM). (Rev F) A binary indicator will be created if a patient is within the "normal" range (defined as ≥8mmHg). The proportion of patients in the normal range will be summarized at both pre and post stimulation.

Pre-implant levels of LES-EEP can be collected from evaluable (criteria listed in section 6.1 of Rev H of the protocol) HRM results captured 6-months prior to baseline evaluations. If no other results are present, then an HRM test is required prior to implant.

Change in proportion of patients that move from low into normal range will be reported and compared to 60% performance goals, where increased proportions are favorable. Mean differences and standard deviations in pressure will be reported for descriptive purposes.

5.8.2.15 Reflux Symptom Index (RSI)

Laryngopharyngeal reflux (LPR) symptoms will be measured by the validated RSI scale from pre-implant to 12-months post stimulation. The RSI is a 9 item self-report scale, each question ranging from 0 to 5, with scores ranging (0,45), lower scores are favourable. An RSI greater than 10 recommends further evaluation for GERD, while 13 indicates LPR. Percent changes from baseline will be evaluated within person and averaged over the group, then compared to the 60% performance goal. Mean changes and standard deviations will be calculated for descriptive purposes only.

5.8.2.16 Body Mass Index

Pre-implant measurements were part of eligibility criteria since patients with BMI greater than 35 were excluded at either screening or baseline. Pre-implant measurements will be determined to be prior and closest to the implant. Post stimulation BMI will be calculated from height and weight measurements obtained during office visits. Changes from baseline will be summarized as individual percentage changes from baseline and averaged over groups.

5.8.2.17 Healthcare Resource Utilization Questionnaire

Healthcare utilization outside of study indicating GERD symptoms and healthcare costs are captured in an 11 item self-report questionnaire. Questions related to the number of Emergency Department Visits (for any non-GERD reason) and Inpatient Hospitalizations (for any non-GERD reason) will be analyzed and compared to pre-implant responses captured at patients' screening visits compared to responses at 12 months post stimulation. Mean and standard deviations will be calculated but no hypothesis testing will be performed.

5.9 Extended Open Label Follow Up

Subjects continue on stimulation treatment and an extended open-label follow-up phase includes an 18-months post-stimulation phone interview followed by annual visits through 5 years post stimulation.

Table 7: Schedule for Collecting Extended Open Label Endpoints

Delayed Stimulation	Base <= 6 months prior to Implant	24 M 720 days	30 M 900 days	42 M 1275 days	54 M 1640 days	66 M 2005 days
Immediate Stimulation	**Base <= 6 months prior to Implant**	**18 M 540 days**	**24 M 720 days**	**36 M 1095 days**	**48 M 1460 days**	**60 M 1825 days**
GERD-HRQL	X[1]	X	X	X	X	X
Subject Diary Collection	X		X	X	X	X
Pittsburgh Sleep Quality Index (PSQI)	X		X	X	X	X
Structured GI Symptoms Questionnaire	X		X	X	X	X
WPAI:SHP	X		X	X	X	X
SF-12	X		X	X	X	X
Reflux Symptom Index (RSI)	X	X	X	X	X	X
Physical Exam	X[2]		X	X	X	X
Healthcare Resource Utilization Questionnaire	X[1]		X	X	X	X

[1] All Pre-Implant measurements are taken at a baseline visit no longer than 6 month prior to implant ***EXCEPT*** for the on-PPI GERD-HRQL and healthcare resource utilization questionnaire. See protocol for further details.

[2] One physical exam is required prior to determining eligibility. May be done at Screening or Baseline visit.

5.9.1 Hypothesis of Extended Open Label Follow Up Efficacy Endpoints

The working hypothesis for extended open label endpoints that are of positive scales, is that the average changes from pre-implant at 18-month (where applicable), 24-month, 36-month, 48-month, 60-month (respectively) will be equal to or greater than 60% increase.

$$H_0: \mu_a < .6$$

$$H_1: \mu_a \geq .6$$

Where μ_a equals the mean value at the follow up time points. One sided p-values will be constructed and tested.

The working hypothesis for extended open label endpoints that are of negative scales, is that the average difference between pre-implant values and 18-month (where applicable), 24-month, 36-month, 48-month, 60-month (respectively) will be less than or equal to 60% decrease.

$$H_0: \mu_a > .6$$

$$H_1: \mu_a \leq .6$$

Where μ_a equals the mean value at the follow up time points. One sided p-values will be constructed and tested.

5.9.2 Estimation of Extended Open Label Follow Up Safety Endpoint

The event rate at each time point will be determined by the adverse event start date of the first event meeting the criteria for each device or procedure related complication. Rates will be reported as proportions and stratified by MedDRA System Organ class reported and verified by the CEC. If groups are determined to not be poolable then events will also be stratified by arm.

5.9.3 Analysis of Extended Open Label Follow Up Efficacy Endpoints

For extended open label endpoints assuming that these data are judged to be poolable between the two arms (see section 5.8.1.1.4), we will compare either the mean change values (for continuous measures) from pre-stimulation to the extended time point using one-sample t-tests or the proportion of successful subjects (for binary measures) using one sample test of proportion, unless otherwise defined. Extended time points occur at 18, 24, 36, 48 and 60 months post stimulation. One sided p-values generated from these tests will be adjusted for multiple comparisons using the Westfall bootstrap method to account for correlation between endpoints and maintain the familywise error rate of 0.025.

Additionally, unadjusted two-sided 95% CI will be calculated solely for descriptive summary purposes.

5.9.3.1 Extended Open Label Outcomes

The outcomes collected in the extended open label portion of the study are listed in table 6 (above), but also they are described in greater detail in section 5.7 and will be calculated in the same ways.

5.10 Subgroup Analyses

An important subgroup analysis will be to compare treatment effects in subjects who underwent crural approximation versus those who did not. The estimates of treatment differences from these stratified analyses will be of clinical interest. Formal tests for subgroup heterogeneity will be conducted by including an interaction of the treatment arm indicator and the subgroup factor in models for the primary and secondary outcomes.

In addition, we will compare several pre-specified subgroups of clinical interest to examine the heterogeneity of treatment effect including age (<65 vs >65), gender (male vs. female), race /ethnicity (White vs. non-White), BMI (<25, 25-30, >30), hiatal hernia status (hernia, no hernia). Formal tests for subgroup heterogeneity will be conducted by including an interaction of the treatment arm indicator and the subgroup factor in a logistic regression model with pH success as the outcome.

5.10.1 Screening Failures

Those patients who fail screening will be tabulated within the text of the Clinical Study Report (CSR) with their reason for screening failure. This ensures that the protocol did not systematically and inadvertently bias the resultant patient population sets.

5.10.2 Subject Disposition

The number of subjects that screened, were implanted, and complete or discontinue from the study will be presented.

5.10.3 Changes in Planned Analyses

Deviations or changes from this SAP deemed necessary due to violation of critical underlying statistical assumptions, data characteristics, or missing data will be clearly described with justification and rationale.

There will be no imputation to the safety endpoint and the safety dataset will be utilized if there is a difference between the ITT set and the safety dataset. The secondary endpoints related to adverse events were also separated from the efficacy endpoints and determined to be secondary safety endpoints since the hypotheses do not apply and no formal tests were going to be done on these endpoints.

In order to examine the robustness of the secondary efficacy endpoints, missing outcomes will be imputed and tipping points analyses will be conducted.

Additional details were added regarding secondary blinded endpoints described in the protocol Rev F, where no specifications were indicated if the outcomes were related to post stimulation or mean changes. It was determined that all metrics would be regarding changes from baseline and graphical depictions could be created for pre and post measurements for publication at a later date.

Rev F of the protocol, on which this SAP was initially created, had an interim analysis beginning at 110 patients. After updating the interim analysis plan and procedures, the first look was scheduled to wait until 120 and sample size justification was rerun at 120 for verification; however, given that the initial protocol supported 80 percent power at 110, it seems reasonable to assume increased power at 120. In order to incorporate the interim analysis updates, the protocol was updated to Rev H at the conclusion of Version 1 of the SAP (November 10, 2017). (Rev G was created to update a site IRB requirement and had no impact on the analysis plan.)

There was an instance in the protocol where clarity around the one sided or two-sided nature of the secondary outcomes was contradicting other language utilized in the document. After much consideration and the attempt to explore superiority in the primary efficacy, it was determined that one sided tests would also be appropriate in the secondary outcomes. We did however allow two-sided tests in relationship to tests for pooling groups to explore heterogeneity in the open label portion of the study.

Given the one-sided nature of the secondary outcomes the alpha was declared to be .025, rather than .05. The protocol was approved at .05, however for the sake of publication and improved rigor it was reduced during the writing of the SAP and a priori to any interim analysis involving outcomes.

We acknowledge that the use of rescue medications without an appropriate wash-out period has the potential to impact the measured pH values and thereby hamper the interpretation of efficacy within the ITT population. In our initial protocol we collected a value intended to show how many days subjects had abstained from taking PPI/H2 medications. However, based on the structure of the database these data were collected in only one field despite PPI and H2 having different wash out periods. As such these data will now be extracted from subject reported diaries which record each type of medication taken. We believe the diaries to be more reliable than the conmeds data as they are assessed on a daily basis and are collected on all subjects regardless of PPI intolerance. Therefore, in the primary analysis subjects in the ITT population who have taken a rescue medication in deviation with the protocol for pH testing will have their pH values set to missing and imputed using multiple imputation. We will also perform a pre-specified sensitivity analysis in the ITT population (described in further detail below) to assess the number of days from the pH test in which the rescue medication was used and what impact these medications could have on pH results.

Prior to device programming, there is no reason to suspect that the reason for continuing to take rescue medications could be related to treatment efficacy. One might expect that the deviations prior to the date of device programming will be just as likely to occur in the control and the active arms. If this is indeed the case, the impact on the primary outcome should be negligible, especially given the low proportion of the overall cohort non-compliance. In order to assess the impact of

potentially biased baseline measurements, a sensitivity analysis will be performed that excludes patients that were not off PPIs for at least 5 days and off H2 blockers for at least 2 days.

We recognize that it is quite possible that deviations occurring after assignment to treatment group by device programming may be different between the study arms, since a patient's use of rescue medications may be related to a need to manage symptoms. In that case, use of rescue medications is potentially informative of treatment efficacy. Using pH measurements taken after recent use of rescue medications could bias the estimate of efficacy. A descriptive analysis will be performed to compare the proportion of patients in each arm that used a PPI within 5 days or an H2 blocker within 2 days of the 6 month measurement.

In addition, in order to understand the potential bias due to rescue medication use, we propose an additional sensitivity analysis in the ITT population that imputes an endpoint failure for patients using rescue medications within a specified number of days of the 6-month pH measurement. The sensitivity analysis will be repeated for the range of days that a patient needs to refrain from PPI/H2 use in order for their observed measurement to be used in analysis rather than an imputed failure. At the high end of the spectrum, a patient would have an imputed failure if the patient used a PPI within 5 days or of the pH measurement. At the low end of the spectrum, a patient would have an imputed failure if the patient used a PPI or H2 within 1 day of the pH measurement. The range of days required to refrain from PPI/H2 is intended to reflect a range in the degree to which rescue medication use is likely due to symptoms. For example, refraining from PPI use for 4 days does not meet the protocol washout period of 5 days for the primary endpoint, but still suggests that symptoms were sufficiently well-managed that the patient for the patient to tolerate being off PPIs and providing a substantial wash-out. In contrast, imputing failure for patients that took a PPI/H2 within 1 day of the 6-month measurement will exclude patients with almost no washout period and those that could not tolerate refraining from PPI/H2 use even for one day.

The protocol was written with adjustments for multiple testing declared as the "Boostrap" method of Proc Multtest. While this method does control for correlation among the datasets (by resampling from the actual data) its implementation has been modified under the current protocol for the following reasons. There is no guidance from SAS or in the literature about how to use the multiple imputation procedure and the bootstrap method together. The use of multiple imputation creates multiple datasets which are then analyzed together to output a single p-value along with descriptions of within and between imputation variation. This is contrary to the raw data input required with the bootstrap method. Based on this issues that were unknown in the original protocol and the initial submission of this SAP, the decision was made to take the mean of all imputed values in the dataset to be reported as raw data into the bootstrap method. Sensitivity analysis will be performed to determine the influence of the missing data.

Additionally, it was thought that the multtest procedure would return the adjusted alpha values used the in the p-value adjustments, however this is not possible therefor the previous version of this SAP determine that adjusted confidence intervals would be created for the secondary endpoints, instead unadjusted 95% CI will be displayed for descriptive purposes only and both raw p-values and adjusted p-values will be displayed.

5.10.4 Eligibility Criteria

Inclusion and exclusion criteria will be summarized by subject for the ITT set. A summary of any pre-approved inclusion and exclusion criteria waivers will be presented.

5.10.5 Demographic and Pre-Implant Characteristics

Subject demographic and pre-implant characteristics will be summarized overall and by treatment arm. Descriptive summaries will include the number of subjects, mean, standard deviation, median, minimum and maximum for continuous

parameters, frequencies and percentages for categorical parameters. Wilcoxon rank sum tests will be used to compare continuous characteristics between arms and chi-squared tests to compare categorical characteristics between arms.

5.10.6 Maintenance of Blinding

Subjects will be asked at 1 month and at 6 months to guess their treatment allocation. The percentage of subjects that guess correctly will be tabulated overall and by arm. A comparison between arms will be made using a chi-squared test.

6 MULTIPLE IMPUTATION ANALYSIS

Primary and secondary analyses of outcomes, with the exclusion of the safety endpoints, will be performed using multiple imputation (MI), whereby each missing datum is replaced by multiple values in multiple datasets unless otherwise noted. The datasets are conventionally analyzed and the multiple results are combined to yield statistically valid inferences with estimated uncertainty.

A single imputation model will be used for the two components of the primary endpoint; achieving pH success defined as normalization (pH < 4 for no more than 5.3% of monitoring time) or $\geq$ 50% improvement in their distal esophageal acid exposure. In addition, the imputation model will be run separately for each analysis population as subjects that are not part of the analysis population may have a different predictive relationship with the outcome.

For each of the primary outcome imputation models, covariates will include but may not be limited to:

- Baseline variables: site number, sex, age (years), fraction of time pH <4.0% total (% time spent in reflux) at baseline, fraction time pH<4.0% -supine at baseline, number of reflux events (total) at baseline, number of reflux events greater than five minutes at baseline, duration of longest reflux event baseline (HH:MM), DeMeester score at baseline.
- Crural status during implant, randomization assignment
- Follow-up variables (Month 3): fraction of time pH <4.0% total Month 3, fraction time pH<4.0% -supine at Month 3, number of reflux events (total) at Month 3, number of reflux events greater than five minutes at Month 3, duration of longest reflux event Month 3 (HH:MM), DeMeester score Month 3, endpoint success at Month 3 criteria 1, endpoint success at Month 3.

A single imputation model will be used for each of the secondary endpoints, regardless of whether the endpoint is a composite or subscale of each questionnaire.

For each of the secondary outcome imputation models, covariates will include but may not be limited to:

- Baseline variables: site number, sex, age (years), total percentage of time pH <4% as a measure of baseline disease severity
- Crural status during implant, randomization assignment

Missing data will be imputed using the PROC MI procedure in SAS. PROC MI imputes missing covariates in the order variables are listed on the VAR statement. The variables will be listed on the VAR statement in the order listed above.

Continuous missing covariates will be imputed using the FCS REG statement, while categorical covariates will be imputed with the FCS DISCRIM statement. For each endpoint success criterion, the FCS LOGISTIC method will be used to impute

the missing binary success criterion. The imputation will be executed in one steps in SAS. PROC MI will impute the covariates first, followed by the success criterion. The final endpoint will be calculated by determining if an imputed subject reaches success on either of the criterion.

Explorations to omit predictors may be conducted if the primary safety or efficacy multiple imputation models will not converge. In practice, 3-5 imputations have been shown to produce valid inference for most purposes, but since the only cost of more imputations is in computing time, 10 imputed datasets will be created for each of the models. The imputed datasets will then be combined for inference using standard methods such as those available in SAS PROC MIANALYZE or other valid statistical software.

7 REFERENCES

1. Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. J. Wiley & Sons, New York. (2002) Second edition.
2. Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons. (2004) Classic edition.
3. Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

8 APPENDIX A: BAYESIAN ADAPTIVE DESIGN

Bayesian Adaptive Design for the Sham-Controlled Study of the
EndoStim Lower Esophageal Sphincter (LES) Stimulation System
for the Treatment of Gastroesophageal Reflux Disease (GERD)

Submitted to EndoStim

By Berry Consultants

November 8, 2017

1.0 Introduction

This document describes the Bayesian adaptive trial design for a sham-controlled study of EndoStim's Lower Esophageal Sphincter (LES) Stimulation System for the treatment of gastroesophageal reflux disease (GERD). Features of the trial design, including the statistical models, timing of the interim analyses, decision rules, and the trial's simulated operating characteristics are detailed.

This is a two-armed, randomized, Bayesian adaptive trial designed to evaluate the safety and efficacy of the EndoStim LES Stimulation System. Patients will be randomized 1:1 to active treatment or sham control. This trial will implant a minimum of 120 and a maximum of 200 patients. Sample size selection interim analyses are planned to determine the trial's actual sample size between this minimum and maximum.

The primary efficacy endpoint is pH success ("response") at 6 months. pH success is defined as the normalization of distal esophageal pH (distal esophageal pH < 4 for no more than 5.3% of monitoring time) or ≥50% improvement in distal esophageal pH at 6 months as compared to baseline. The primary hypothesis of interest is to show the superiority of active LES treatment to the sham control.

2.0 Primary Analysis

The primary efficacy analysis will be conducted at least 6 months after accrual to the trial is complete in order to allow all implanted patients to be followed for the primary endpoint. At the primary analysis, the proportion of patients achieving a response will be compared between randomized groups. If the one-sided p-value comparing active treatment to sham control is less than 0.022, this trial will be considered a success. A significance threshold of 0.022 is selected to control for overall Type I error. Specifically, the null and alternative hypotheses are:

$$H_0: p_T = p_C$$
$$H_1: p_T > p_C$$

where p_T is the response rate for the active treatment arm and p_C is the response rate for the sham control arm.

3.0 Statistical Modeling

3.1 Primary Endpoint

While the primary analysis is a frequentist test, we build a Bayesian model of the primary endpoint for the purpose of determining the Bayesian quantities that may be reported during the execution of the trial such as the predictive probabilities of trial success described below.

Each trial arm is modeled independently, but identically. We assume that response (Y) for patient *i* on arm *d* has a Bernoulli distribution where p_d is the response rate for arm *d:*

$$Y_{id} \sim Bernoulli(p_d).$$

We model the log-odds of response

$$\theta_d = \log\left(\frac{p_d}{1-p_d}\right),$$

We assume non-informative prior distributions for θ_d. We use the same prior distribution for both arms:

$$\theta_d \sim N(0,2^2).$$

3.2 Predictive Probabilities

3.2.1 Predictive Probability of Trial Success at the Current Sample Size

We determine the predictive probability of trial success by simulating data from the predictive distribution of the data and observing the proportion of simulated data sets that would produce a statistically significant result. This approach accounts for the inherent variability in future data as well as uncertainty about the model parameters.

The treatment arms are modeled independently, so the predictive distribution of data for each treatment arm can be simulated separately. Let n_d be the total number of patients enrolled on treatment arm *d*. Among these patients, let x_d be the number of patients on treatment arm *d* with a known outcome who have achieved success and n_d^* be the number of patients on treatment arm *d* with an unknown outcome. Among the n_d^* enrolled patients with an unknown outcome, x_d^* of those patients will ultimately achieve success.

The posterior distribution of θ_d given the known outcomes is simulated via Markov chain Monte Carlo (MCMC). At each step in the MCMC, a value of θ_d is drawn. That value is used to simulate the unknown outcomes for the n_d^* patients. Conditional on the drawn value of θ_d, x_d^* is a binomial random variable:

$$x_d^* | \theta_d \sim Binom(n_d^*, \text{logit}(\theta_d)),$$

which is simulated for each step in the MCMC.

Each step in the MCMC thus generates simulated outcomes for the currently enrolled patients: ($x_c + x_c^*$, $x_T + x_T^*$), where C represents the control (sham treatment) and T represents the experimental treatment. A z-test for a difference in proportions then determines whether this set of data would result in statistical significance (i.e. trial success or trial failure). The predictive probability of trial success is then calculated as the proportion of MCMC steps that resulted in simulated outcomes resulting in trial success. This value is denoted P_n.

3.2.2 Predictive Probability of Success at the Maximum Sample Size

A similar calculation to that above is performed. However, the predictive distribution is now updated to include both patients who are enrolled and patients who are not yet enrolled. Patients yet to be enrolled are allocated to control and treatment in a 1:1 ratio, giving each treatment arm m_d additional subjects whose outcomes are unknown, effectively adjusting n_d^* above to $n_d^* + m_d$. The predictive probability of success at the maximum sample size is denoted as P_{max}.

4 Study Design

4.1 Timing of Interim Analyses

Interims will occur based on the number of patients who have been implanted with the study device (not the number of patients with complete data). The schedule for the 4 interims is specified in Table 1. If accrual to the trial is not stopped at any of these interim analyses, accrual to the trial will continue to the maximum sample size of 200 patients.

Table 1: Interim Analysis Schedule

Interim	Patients Implanted
1	120
2	140
3	160
4	180

A trial flag analysis will occur if the trial enrolls to the maximum sample size of 200 patients.

4.2 Allocation

The trial will implant a maximum of 200 patients that will be randomized to the treatment arms in a fixed ratio of 1:1.

4.3 Predictive Probability of Trial Success at the Current Sample Size and Early Stopping for Expected Success

At each interim analysis, we calculate the predictive probability of trial success at the final analysis for the currently implanted patients. This calculation assumes that the primary analysis would take place after all currently implanted patients complete their follow-up for the primary endpoint.

If, at an interim analysis, the predictive probability of trial success with the currently implanted patients is at least 95%, this trial may stop accrual early for expected success. The final analysis would be conducted after all implanted patients complete the follow-up for the primary endpoint.

4.4 Predictive Probability of Trial Success at the Maximum Sample Size and Early Stopping for Futility

At each interim analysis, we calculate the predictive probability that the current trial will be successful should it continue to the maximum of 200 implanted patients and all patients are followed for the primary endpoint.

If, at an interim analysis, the predictive probability of trial success at the maximum sample size is less than 1%, this trial will stop early for futility.

4.5 Flag at Maximum Sample Size

If the trial enrolls to the maximum sample size of 200 patients, we calculate the predictive probability that the current trial will be successful when all 200 patients are followed for the primary endpoint.

If this predictive probability of trial success at 200 patients enrolled is greater than 80%, the trial will be flagged as a predicted success. Otherwise it will be flagged as a predicted failure. The data monitoring committee will communicate the results of this flag at 200 patients enrolled to the trial sponsor. The trial will continue and follow up all patients for the primary endpoint.

5 Example Trials

In this section, we show single simulations of the trial to illustrate the adaptive features.

5.1 Example Trial 1

Table 2: Example Trial 1

Interim Look	Pts Complete[1]	Pts Remaining	Arm	Observed Proportion (observed/ complete)	95% CI	Pr Superiority	P_n	P_{max}
120	52	68	Ctrl	0.38 (10/26)	0.21,0.57	*0.959*	*0.667*	*0.776*
			Trt	0.62 (16/26)	0.43,0.78			
140	82	58	Ctrl	0.37 (15/41)	0.24,0.52	*0.992*	*0.872*	*0.882*
			Trt	0.61 (25/41)	0.47,0.75			
160	108	52	Ctrl	0.37 (20/54)	0.24,0.51	*0.996*	*0.969*	*0.956*
			Trt	0.63 (34/54)	0.50,0.74			
Final Analysis	160	0	Ctrl	0.39 (31/80)	0.29,0.49	*0.9996*	***P-value = 0.001***	
			Trt	0.64 (51/80)	0.53,0.74			

[1] The number of patients complete that is presented here is based on a single simulation result. This example trial illustrates early stopping rules should such data be observed. The actual number of completed patients available at each interim will vary depending on recruitment.

In example trial 1 (**Table 2**), the first interim analysis occurs when 120 patients are implanted and 52 of those patients have complete 6-month data. The observed response rate is 0.38 for the control group and 0.62 for the treatment group. The predictive probability of success at the current sample size (P_n) is 0.667, which is less than the 95% required to stop accrual early for expected success. The predictive probability of success at the maximum sample size (P_{max}) is 0.776, which is greater than the 1% required to stop the trial early for futility. The trial continues.

At the second interim analysis with 140 patients implanted, results are similar. With now 82 patients with complete data, the predictive probability of success at the current sample size increases to 0.872. This is still less than the 95% required to stop accrual early for expected success.

At the third interim analysis, with 160 patients implanted, 108 patients have complete 6-month data. The observed response rate is 0.37 for the control and 0.63 for the treatment arm. The predictive probability of success at the current sample size (P_n) is 0.969, which is now greater than the 95% required to stop the trial early for expected success. The trial stops accrual for expected success and follows the implanted patients for their primary endpoint. The final analysis includes all 160 patients with complete data. At the final analysis, the one-sided p-value is 0.001 and the active treatment arm is considered superior to the sham control.

5.2 Example Trial 2

Table 3: Example Trial 2

Interim Look	Pts Complete	Pts Remaining	Arm	Observed Proportion (observed/complete)	95% CI	Pr Superiority	P_n	P_{max}
120	28	92	Ctrl	0.36 (5/14)	0.15, 0.64	*0.500*	*0.122*	*0.198*
			Trt	0.36 (5/14)	0.15, 0.64			
140	62	78	Ctrl	0.49 (15/31)	0.32, 0.66	*0.209*	*0.002*	*0.0096*
			Trt	0.39 (12/31)	0.24, 0.57			

In example trial 2 (**Table 3**), the first interim analysis occurs when 120 patients are implanted and 28 of those patients have complete 6-month data. The observed response rate is 0.36 for the control and for the treatment. The predictive probability of success at the current sample size (P_n) is 0.122 and the predictive probability of success at the maximum sample size (P_{max}) is 0.198. Neither early stopping criterion is satisfied and the trial continues. At the second interim analysis, with 140 patients implanted, 62 patients have complete 6-month data. The treatment is observed to have a lower response rate than the control group. The predictive probability of success at the maximum sample size (P_{max}) is 0.0096, which is less than the futility stopping boundary of 0.01. This trial stops early for futility.

6 Simulation Scenarios and Operating Characteristics

We simulated the trial exactly as described above under 18 different response rate scenarios for the control and treatment arms in order to characterize and evaluate the performance of the adaptive design. Simulations were completed using version 6.0.4 of FACTS (Berry Consultants Austin TX). FACTS is a validated commercially available software package. The specific simulated scenarios are detailed in section *6.1*. Simulations assume the patient accrual as described in the accrual profile in section *6.2*.

For each of these scenarios, we generate virtual patient data according to the assumed truths and execute the design as specified above. We repeat this process to create multiple "virtual trials" and we track the behavior of each trial. More specifically, the trial simulation proceeds as follows:

1. Define a rate of accrual of patients into the trial.

2. Define the response rates for each arm.

3. Simulate the arrival of the first 120 virtual patients into the trial according to the defined accrual rate.

4. Assign virtual patients to control or treatment according to the trial's 1:1 randomization ratio.

5. Simulate a final outcome for these virtual patients according to the defined response rates for their assigned treatment arm.

6. Based on the accrual times of the virtual patients into the trial, determine which patients have known final outcomes and which patients have an unknown final outcomes at the time the interim analysis is scheduled to occur. Conduct the interim analysis with the available data at the time of the planned interim analysis as specified.

7. Based on the interim analysis result, determine if the criteria for expected success or early futility is met.

8. If an early stopping boundary is met, record the trial's final result, including sample size at the interim and the reason for stopping. For trials that stop early for expected success, conduct the final analysis with complete information for the currently enrolled patients and record the final analysis result as well.

9. If an early stopping boundary is not met, simulate the enrollment of the next 20 patients in the trial and repeat steps 4 – 8 above. Continue this process until the trial meets an early stopping boundary or until the maximum sample size is attained.

The process described above creates a single simulated trial with a particular sample size and particular final result. We repeat this process 10,000 times per scenario, thus producing 10,000 simulated trials per scenario. We calculate operating characteristics for a particular scenario as the average final trial results across the corresponding 10,000 simulated trials.

6.1 Virtual Patient Response Profiles

We consider 18 response rate scenarios including 9 null scenarios where the control and treatment are the same and 9 alternative scenarios where the treatment has a higher response rate than the control. Virtual patient outcomes are simulated in terms of the probability of response on each arm. Each virtual patient in the trial receives an outcome that is simulated as an independent Bernoulli random variable with the probability of success corresponding to the patients treatment arm.

Table 4 contains the response rates for the 18 different simulated scenarios.

Table 4: Virtual Patient Response

Scenarios	Control Response Rate	Treatment Response Rate
Null 1-1	0.1	0.1
Null 2-2	0.2	0.2
Null 3-3	0.3	0.3
Null 4-4	0.4	0.4
Null 5-5	0.5	0.5
Null 6-6	0.6	0.6
Null 7-7	0.7	0.7
Null 8-8	0.8	0.8
Null 9-9	0.9	0.9
Alternate 2-6	0.2	0.6
Alternate 2-7	0.2	0.7
Alternate 3-6	0.3	0.6
Alternate 3-7	0.3	0.7
Alternate 4-6	0.4	0.6
Alternate 4-7	0.4	0.7
Alternate 5-6	0.5	0.6
Alternate 5-6.5	0.5	0.65
Alternate 5-7	0.5	0.7

6.2 Accrual

We simulate the random arrival of patients into the trial from a Poisson process with the mean (peak) monthly rate specified in **Table 5**. The arrival of patients into the trial is modeled by a Poisson process with specified piecewise constant rates specified by month. When transitions between recruitment rates occur, linear increases or decreases are assumed across the month, and the mean rate is used for the month.

Note that only the mean rate is shown in Table 5. Actual accrual into any single simulated trial is derived

from these mean accrual profiles but given variability in simulated arrival times, some simulated trials recruit more quickly than this and some more slowly.

Table 5: Patient Accrual

Month(s)	Peak Monthly Rate
1	4
2	4
3	4
4	4
5	4
6	8
7 to 12	16
13 +	10

6.3 Dropout

We assume no dropout in these simulations.

6.4 Operating Characteristics

We report the following operating characteristics in **Table 6**: 1) The overall probability of success (power/Type I error), 2) the mean trial size, 3) the total probability that the trial achieves an early success, 4) the total probability that the trial stops early for futility, and 5) the probability that the trial is stopped for expected success and yet does not achieve success at the final analysis with complete follow-up data ("Flip Flop"). All scenarios are based on 10,000 simulations. Specifically, we define

1. Prob. of Success: The proportion of simulated trials that meet statistical significance for the primary analysis, regardless of whether they stopped early for expected success or continued to the maximum sample size

2. Mean Trial Size: The mean sample size for all simulated trials.

3. Prob. of Stopping Early for Success: The proportion of simulated trials that stop enrollment early for predicted success, i.e. the proportion of trials with $P_n > 0.95$ at an interim analysis.

4. Prob. of Stopping Early for Futility: The proportion of simulated trials that stop early for futility, i.e. the proportion of trials with $P_{max} < 0.01$ at an interim analysis.

5. Prob. of Success to Futility Flip Flop: The proportion of simulated trials that stop enrollment early for predicted success ($P_n > 0.95$ at an interim analysis) but then fail to meet statistical significance ($p < 0.022$) at the primary analysis.

Table 6: Operating Characteristics

Scenario	Prob. of Success	Mean Trial Size	Prob. of Stopping Early for Success	Prob. of Stopping Early for Futility	Prob. Of Success to Futility Flip Flop
Null 1-1	0.021	179.0	0.004	0.512	0.004
Null 2-2	0.023	177.1	0.005	0.519	0.011
Null 3-3	0.023	176.6	0.006	0.515	0.013
Null 4-4	0.023	176.5	0.005	0.515	0.014
Null 5-5	0.024	175.8	0.007	0.526	0.015
Null 6-6	0.024	176.1	0.006	0.518	0.016
Null 7-7	0.025	176.7	0.006	0.520	0.013
Null 8-8	0.022	176.9	0.005	0.522	0.011
Null 9-9	0.020	178.7	0.003	0.519	0.004
Alternate 2-6	0.999	142.8	0.965	0.001	<0.0001
Alternate 2-7	>0.999	132.4	0.997	<0.0001	<0.0001
Alternate 3-6	0.982	161.1	0.760	0.003	0.006
Alternate 3-7	>0.999	143.8	0.956	<0.0001	<0.0001
Alternate 4-6	0.779	180.4	0.361	0.030	0.022
Alternate 4-7	0.983	161.2	0.758	0.003	0.004
Alternate 5-6	0.262	187.0	0.076	0.178	0.032
Alternate 5-6.5	0.535	186.0	0.186	0.079	0.032
Alternate 5-7	0.796	180.2	0.369	0.028	0.023

We see an overall one-sided Type I error of between 0.020 and 0.025 across the null scenarios. There is a probability <0.007 that a null trial achieves an early success and a probability >0.512 that a null trial is stopped early for futility. The mean trial size ranges between 175.8 and 179.0 patients in the null scenarios.

In the "Alternate 4-7" case, where the response rate in the control group is 0.40 and the response rate in the treatment group is 0.70, the adaptive design has a power of 0.983 to declare superiority of the treatment. For this scenario, the probability of stopping a trial early for expected success is 0.758, and the probability of stopping a trial early for futility is 0.003. The probability of stopping a trial early for

expected success but then going on to not be successful is 0.004. This scenario has a mean sample size of 161.2 patients. Power decreases with smaller treatment effects, as we see in the "Alternate 5-7" scenario, where the response rate in the control group is 0.50 and the response rate in the treatment group is 0.70. The power of the "Alternate 5-7" scenario is 0.796, with a 0.369 probability of stopping early for expected success and a 0.028 probability of stopping early for futility. The probability of stopping early for expected success and not achieving success at the final analysis in this scenario is 0.023.

7.0 Covariate Adjustment

The primary analysis as described in the trial protocol is a frequentist logistic regression that includes adjustment for a covariate of interest. This covariate will be considered as dichotomous (yes/no). For the purpose of trial simulation and demonstrating the operating characteristics of the trial design, we did not consider this covariate. However, during the conduct of the trial, the primary analysis will be covariate adjusted as described in the protocol. As such, the predictive probability calculations performed at each interim during the conduct of the trial will take the covariate into consideration and will predict the probability of trial success based on the covariate adjusted frequentist logistic regression final analysis.

Specifically the trial will be executed exactly as described. However, the predictive probability at the current sample size as described in Section 3.0 will model each treatment arm by covariate group independently. The predictive probability at the maximum sample size described in Section 3.0 will additionally assign future patients enrolled into the trial a covariate value. Covariate "yes" versus "no" will be simulated from a Beta Binomial distribution reflecting the interim uncertainty in the rate of Covariate "yes" as with as the sampling variability in the future covariate values.

8.0 Summary

This trial design will provide high quality evidence for the study of the superiority of the EndoStim LES Stimulation System to a control. The design has a maximum simulated Type I error of 2.5% with a probability greater than 0.50 of stopping early for futility under the null. If the EndoStim LES Stimulation System treatment demonstrates superiority versus a control, the design offers the opportunity to identify this benefit with a sample size of less than 200 patients.

STATISTICAL ANALYSIS PLAN ADDENDUM

Addendum to SAP version October 9, 2018

Modifications v1.1 on 28Dec2024

A Multicenter, Randomized, Double-Blind, Sham-Controlled Clinical Investigation of the EndoStim® Lower Esophageal Sphincter (LES) Stimulation System for the Treatment of Gastroesophageal Reflux Disease (GERD)

LESS GERD – Randomized Clinical Trial (RCT) PROTOCOL

NUMBER CS-100

EndoStim, Inc.
1609 Shoal Creek Boulevard, Suite 303
Austin, TX 78701 USA

SIGNATURES

_____	_____
VK Sharma, MD	Date
Chief Medical Officer	
_____	_____
Tara Maddala, PhD	Date
Independent Statistician	

1 INTRODUCTION

This statistical analysis plan (SAP) addendum adds to the approved SAP (October 9, 2018) for Protocol CS-100: "A *Multicenter, Randomized, Double-Blind, Sham-Controlled Clinical Investigation of the EndoStim® Lower Esophageal Sphincter (LES) Stimulation System for the Treatment of Gastroesophageal Reflux Disease (GERD)."* On 15 October 2019, CS-100 was terminated early due to a DMC decision that was based on an incorrect analysis of the primary endpoint. Specifically, the incorrect dataset was used and pooling by geography was not assessed. Hence, this document outlines the analysis plan using the correct dataset and analysis methods as detailed in the original SAP while accounting for consequences of erroneously stopping the trial early.

2 SCOPE

This SAP addendum should be interpreted and implemented within the context of the SAP dated October 9, 2018, and Protocol CS-100 dated May 30, 2018. Analyses described herein build on the original SAP and highlight any important differences between this addendum and the original SAP.

3 DEFINITIONS

Table 1: Abbreviations

BMI		Body Mass Index
CRF		Case Report Form
DS		Delayed Stimulation Group
FDA		US Food and Drug Administration
GERD		Gastroesophageal Reflux Disease
GERD-HRQL		Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire
H2		Histamine-2 Receptor Antagonists
IPG		Implantable Pulse Generator
IS		Immediate Stimulation Group
LES		Lower Esophageal Sphincter
LES-EEP		Lower Esophageal Sphincter End-Expiratory Pressure
LES-MP		Lower Esophageal Sphincter Mean Pressure
LPR		Laryngopharyngeal Reflux
MedDRA		Medical Dictionary for Regulatory Activities
MI		Multiple Imputation
PI		Post Implant
PPI		Proton Pump Inhibitors
PSQI		Pittsburgh Sleep Quality Index
RSI		Reflux Symptom Index
WPAI:SHP		Work Productivity and Activity Impairment Questionnaire: Specific Health Problem V2.0

4 STUDY DESIGN

With the exception of early termination of enrollment of the study, the study design was unchanged. Rather than randomize up to 200 planned subjects, the study randomized only 161 subjects (144 in the US and 17 in the EU).

4.1 Randomization

Randomization was unchanged from the original SAP, with the exception of early termination of enrollment of the study.

4.2 Blinding

No details of blinding of the original study were changed. The original SAP was produced and signed before unblinding. This document was produced after the termination of the study. Supplemental monitoring was conducted on the data not previously monitored (i.e., data collected since the time of the last monitoring visit that was performed prior to closure of the study/IDE). This mainly included adverse events and concomitant medication. Additionally, open queries were resolved, and source data verification (SDV) was performed on previously collected data from various visits, including screening/baseline, implant, and follow-up. While supplemental monitoring was performed after the study was no longer blinded and necessarily resulted in either new data entries or changes to previously entered data to ensure data completeness and accuracy, it was entirely done only by the independent monitors and was completed prior to the final data analysis, thereby mitigating the potential for bias.

4.3 Study Objectives

Analyses will be conducted in accordance with the study objectives outlined in the SAP.

The design of the study was such that an analysis of the 12-month endpoint can be analyzed using a delayed start design (Liu-Seifert, et al., 2015). A delayed start design was used in certain trials of disease-modifying agents to assess the effect of the agent in altering the natural course of a condition.

4.4 Study Success Criteria

The trial terminated enrollment early due to an erroneous DMC decision, so the primary analysis was not conducted on 200 patients, as planned. Additionally, there was a data entry error in pH measurements. Specifically, sites erroneously entered the worst of two days of esophageal acid exposure data rather than the average of two days for the primary outcome variable, which caused an incorrect analysis to be performed in place of the originally planned interim analysis. A subsequent re-evaluation by the DMC after the data entry error was discovered, suggested that the DMC would not have recommended early termination if the expected analysis had been performed. Furthermore, the interim analysis did not mimic the methodology of the final analysis, which required a poolability assessment by geographic region per FDA before assessing the primary endpoint.

Hence, the analyses will be performed on the corrected data (*i.e.,* the average time esophageal pH was below 4.0 over the two days recording). In addition, analysis of trial success had the trial continued to the end with 200 patients will be executed. This analysis will be for informational purposes only and should not be interpreted as a definitive statement on the efficacy of the EndoStim LES system.

4.5 Bayesian Adaptive Model

The Bayesian adaptive model as described in Appendix A of the SAP was run as planned except that the incorrect pH values were used and poolability by geographic region was not evaluated.

4.5.1 Schedule of Interim Analysis

Interim analyses were scheduled at 120, 140, 160 and 180 implanted patients and simulations were conducted at each of these time points. The interim analysis at 160 implanted subjects led to a decision to terminate enrollment of the study. Thus, the interim analysis at 180 implanted subjects was not conducted.

4.5.2 Decision Rules and Stopping Boundaries

At the 160-subject interim analysis, the futility boundary was crossed on an analysis that incorrectly used the worst day esophageal acid exposure measurement over two days rather than the average esophageal acid exposure measurement over two days. In addition, the analysis did not include a poolability assessment by geographic region which the primary outcome analyses required.

4.5.3 Primary Efficacy Response Rate Estimates

The assumptions and planning for the trial were described in the original SAP and are not altered in this addendum.

4.5.4 Simulated Model Operating Characteristics

Operating characteristics of the trial as planned are described in the original SAP and are not altered in this addendum.

5 STATISTICAL ANALYSES

5.1 General Statistical Considerations

General statistical considerations will not change from the original SAP, with the exception that updated software (*e.g.,* SAS™ Version 9.4) or comparable package will be considered for production of outputs and analyses.

5.1.1 Software

Version 9.4 or higher of SAS® statistical software package or other validated statistical software will be used to provide all summaries, listings, graphs, and statistical analyses.

5.1.2 Descriptive Statistics

Strategies for calculating and displaying descriptive statistics are not altered in this addendum.

5.1.3 P-values

The p-values of the primary endpoints will be calculated per the original SAP. Due to the early termination of the study, the study will be underpowered. Hence, multiple testing adjustment will not be performed for secondary efficacy endpoints.

The original SAP specified parametric methods (e.g., two-sample t-test and one-sample t-test) for all endpoints. However, acid exposure time (pH data) and manometry data are typically evaluated using non-parametric methods such as the Wilcoxon rank sum (Man-Whitney U) test or the Wilcoxon signed-rank test. Therefore, for these two endpoints, nonparametric tests will be employed.

5.1.4 Missing Data

Multiple imputation as specified in the original SAP will only be used in the analysis of the endpoint of subjects achieving pH success.

5.1.5 Partial Dates

Handling of partial dates will be performed as specified in the original SAP.

5.1.6 Visit Windows and Visit Definitions

Visit windows and definitions are as specified in the original SAP.

5.1.7 Duration Variables

Definitions of duration variables are as specified in the original SAP.

5.2 Analysis Populations

Analysis populations are as specified in the original SAP.

5.2.1 Intention-to-Treat Analysis Set

The definition of the ITT analysis population is as specified in the original SAP.

5.2.2 Per-Protocol Set

The definition of the per-protocol analysis population is as specified in the original SAP.

5.2.3 Safety Analysis Set

The safety analysis set is as specified in the original SAP.

5.3 Poolability of Data

Poolability types are as specified in the original SAP. Poolability by geography (US and OUS) and poolability of IS and DS groups for safety analyses and outcomes related to duration of stimulation.

The test of poolability for geography will be performed on the 6-month primary efficacy outcome data. If data are not poolable, efficacy and safety endpoints will be presented by geographical region (US and outside US [OUS]). If poolable, blinded and open label endpoint data will be combined by geographical region.

The test of poolability for duration of stimulation (IS/DS) will be performed on the 12-months visit and the 12-months on stimulation safety data (rate of occurrence device or procedural related serious adverse events). If data are not poolable, safety endpoints will be presented by duration of stimulation (IS/DS). If poolable, safety data will be combined.

The test of poolability for duration of stimulation (IS/DS) will be performed on the 12-months on stimulation for the secondary efficacy outcome (proportion of patients with patients achieving symptom success as measured by GERD-HRQL). If data are not poolable, secondary efficacy endpoints will be presented by duration of stimulation (IS/DS). If poolable, open label endpoint data will be combined.

5.4 Primary Safety Endpoint

The primary safety endpoint is as specified in the original SAP.

5.4.1 Primary Analysis of the Primary Safety Endpoint

The analysis of the primary safety endpoint will be performed as specified in the original SAP.

5.4.2 Missing Data Analysis for Primary Safety Endpoint

Missing data related to the safety endpoint will not be imputed as specified in the original SAP.

5.4.3 Poolability of Primary Safety Endpoint

See Section 5.3.

5.5 Primary Efficacy Endpoint

The primary efficacy endpoint is as specified in the original SAP.

5.5.1 Hypothesis for Primary Efficacy Endpoint

The hypothesis for the primary efficacy endpoint is as specified in the original SAP.

5.5.2 Interim Analysis of the Primary Efficacy Endpoint

Interim analyses were conducted according to the SAP. However, the endpoint analyzed from erroneous data and poolability by geography was not assessed.

5.5.3 Informational Analysis of the Primary Efficacy Endpoint

This section is not applicable as the trial was erroneously halted early.

5.5.4 Futility Analysis of the Primary Efficacy Endpoint

Since the trial was erroneously halted early, no further interim assessments will apply.

5.5.5 Final Analysis of the Primary Efficacy Endpoint

The final analysis for the primary efficacy endpoint will be performed as specified in the original SAP.

5.5.6 Missing Data Analysis of the Primary Efficacy Endpoint

The missing data strategy (use of multiple imputation) is as specified in the original SAP.

5.5.7 Sensitivity Analysis of the Primary Efficacy Endpoint

The tipping point analysis is as specified in the original SAP.

5.6 Secondary Efficacy Endpoint (Patient Focused)

GERD related quality of life outcome is unchanged from the SAP. Multiple imputation will not be implemented for this endpoint.

5.7 Blinded Endpoints

Blinded endpoints are as specified in the original SAP.

5.7.1 Secondary Blinded Endpoints

5.7.1.1 Hypotheses of the Secondary Blinded Endpoints

The hypotheses of the secondary blinded endpoints are as specified in the original SAP.

5.7.1.2 Analyses of the Secondary Blinded Endpoints

Analysis techniques of the secondary blinded endpoints will be performed as specified in the original SAP. To clarify for the blinded analyses, we will compare the immediate stimulation (IS) arm to the delayed stimulation (DS) arm by evaluating median change values and percent change (for continuous measures) from pre-stimulation to post using the Wilcox rank-sum test (non-parametric test where appropriate) or the proportion of successful subjects (where appropriate) using Fisher's exact test. Where specified in the SAP, the median percent change from pre-implant to post-implant will be compared to a prespecified percent change. Also, where specified in the SAP, the proportion of patients meeting success criteria will be compared to the specific performance goal.

5.7.1.3 Power Considerations for Secondary Blinded Endpoints

Because the study stopped early, the power calculations in the SAP represent a point of reference only. Hence, multiple testing adjustment will not be performed.

5.7.1.4 Measurements for Secondary Blinded Endpoints

5.7.1.4.1 Bravo Esophageal pH Monitoring - Time of pH<4.0

The time of pH<4.0 endpoint remains is as specified in the original SAP.

5.7.1.4.2 Acid Suppression Medication Use

Analysis of medication use will be performed as specified in the original SAP.

5.7.1.4.3 Heartburn Symptoms

Analysis of heartburn symptoms will be performed as specified in the original SAP.

5.7.1.4.4 Regurgitation Symptoms

Analysis of regurgitation symptoms will be performed as specified in the original SAP.

5.7.1.4.5 Short Form Health Survey (SF-12)

Analysis of the quality-of-life SF-12 quality of life endpoints will be performed as specified in the original SAP.

5.7.2 Other Endpoints of Interest during Blinded Phase

5.7.2.1 Bravo Esophageal pH Monitoring - Time of pH<4.0

Time, from a minimum of 18 hours of data collection, where the subject's distal supine esophageal pH <4.0 is as specified in the original SAP.

5.7.2.2 Bravo Esophageal pH Monitoring - Reflux Events

The analysis of the total number of reflux events reported from the Bravo Esophageal pH monitoring report will be performed as specified in the original SAP.

5.7.2.3 Bravo Esophageal pH Monitoring - DeMeester Score

The analysis of the DeMeester Score calculated by the Bravo Esophageal pH test results will be performed as specified in the original SAP.

5.7.2.4 Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire (*off-PPI* GERD-HRQL) score

Computation of composite scores and analysis of the GERD-HRQL quality of life outcome will be performed as specified in the original SAP. Multiple imputation will not be used.

5.7.2.5 Pittsburgh Sleep Quality Index (PSQI)

The computation and analysis of the PSQI will be performed as specified in the original SAP.

5.7.2.6 Mucosal Injury

Analysis of the extent of mucosal injury from reflux will be performed as specified in the original SAP.

The strategy for selecting subjects for the PPS for the open label part (*i.e.,* excluding subjects with LA classification C or D at 6 months) is as specified in the original SAP.

5.8 Open Label Endpoints

Analysis of open-label endpoints will be performed as specified in the original SAP.

5.8.1 Secondary Open Label Endpoints

5.8.1.1.1 Hypotheses for Secondary Open Label Efficacy Endpoints

Hypotheses for secondary open label endpoints are as specified in the original SAP with the exception of the comparison to 60% for the expected improvement in any given endpoint at 12 months. The rationale is that intervention in this population of patients is not expected to result in 60%+ improvement across various symptom measurements and is only relevant to the proportion of patients achieving success (defined as 50% or more improvement) on GERD-HRQL Composite Score. The 60% criterion was erroneously specified in the original SAP for other endpoints. Therefore, the comparison to 60% will only be analyzed on the change from baseline to 12-months for GERD-HRQL.

5.8.1.1.2 Analyses for Secondary Open Label Efficacy Endpoints

Analyses for secondary open label endpoints are as specified in the original SAP.

5.8.1.1.3 Power Considerations for Secondary Open Label Efficacy Endpoints

Because the study stopped early, the power calculations in the SAP represent a point of reference only. Hence, multiple testing adjustment will not be performed.

5.8.1.1.4 Poolability for Secondary Open Label Efficacy Endpoints

Poolability, by geography, of open label efficacy endpoints is based on the poolability of the 6-month primary efficacy endpoint. To clarify, if the tests for poolability for IS and DS indicate that pooling criteria is met, then open label efficacy endpoints will be combined for presentation.

5.8.1.2 Measurements for Secondary Open Label Endpoints

The open label portion of the study includes the outcomes described below and are as specified in the original SAP.

5.8.1.3 Bravo Esophageal pH Monitoring - Time of pH<4.0

The time of pH<4.0 endpoint is as specified in the original SAP.

5.8.1.4 Heartburn Symptoms

Heartburn symptoms severity analysis will be performed as specified in the original SAP.

5.8.1.5 Regurgitation Symptoms

Regurgitation symptoms severity endpoint definitions and analyses will be performed as specified in the original SAP.

5.8.1.6 Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire (*on-PPI* GERD-HRQL) score

Composite scores for GERD-HRQL analysis will be performed as specified in the original SAP.

5.8.1.7 Acid Suppression Medication

Acid suppression medication analysis remains will be performed as specified in the original SAP.

5.8.1.8 Short Form Health Survey (SF-12)

Quality of life analysis, assessed by the SF-12 will be performed as specified in the original SAP.

5.8.2 Other Endpoints of Interest during Open Label Phase

5.8.2.1 Other Safety Endpoints

Frequencies and percentages will be calculated and summarized as specified in the original SAP.

5.8.2.2 Bravo Esophageal pH Monitoring - Time of pH<4.0 - Overall

Analysis of time of pH<4.0 overall will be performed as specified in the original SAP.

5.8.2.3 Bravo Esophageal pH Monitoring - Time of pH<4.0 - Supine

Analysis of time of pH<4.0 supine will be performed as specified in the original SAP.

5.8.2.4 Bravo Esophageal pH Monitoring - Reflux Events

The analysis of the total number of reflux events reported from the Bravo Esophageal pH monitoring report will be performed as specified in the original SAP.

5.8.2.5 Bravo Esophageal pH Monitoring - DeMeester Score

The analysis of the DeMeester Score calculated by the Bravo Esophageal pH test results will be performed as specified in the original SAP.

5.8.2.6 Pittsburgh Sleep Quality Index (PSQI)

The computation and analysis of the PSQI will be performed as specified in the original SAP.

5.8.2.7 Mucosal Injury

Analysis of the extent of mucosal injury from reflux will be performed as specified in the original SAP.

The strategy for selecting subjects for the PPS for the open label part (*i.e.,* excluding subjects with LA classification C or D at 6 months) is as specified in the original SAP.

5.8.2.8 Gastroesophageal Reflux Disease-Health Related Quality of Life Heartburn Questionnaire (*off-PPI* GERD-HRQL) score

Computation of composite scores and analysis of the GERD-HRQL quality of life outcome will be performed as specified in the original SAP. Multiple imputation will not be used.

5.8.2.9 Heartburn Symptoms

Analysis of heartburn symptoms will be performed as specified in the original SAP.

5.8.2.10 Regurgitation Symptoms

Regurgitation symptoms severity endpoint definitions and analyses will be performed as specified in the original SAP.

5.8.2.11 Structured GI Questionnaire

Analysis of GERD symptoms assessed by the Structured GI questionnaire will be performed as specified in the original SAP.

5.8.2.12 Work Productivity and Activity Impairment Questionnaire: Specific Health Problem (WPAI:SHP)

Analysis of these endpoints will be performed as specified in the original SAP.

5.8.2.13 Lower Esophageal Sphincter Mean Pressure (LES-MP)

Endpoints and analyses of LES-MP will be performed as specified in the original SAP.

5.8.2.14 Lower Esophageal Sphincter End Expiratory Pressure (LES-EEP)

Endpoints and analyses of LES-EEP will be performed as specified in the original SAP.

5.8.2.15 Reflux Symptom Index (RSI)

Analysis of laryngopharyngeal reflux (LPR) symptoms measured by the validated RSI scale will be performed as specified in the original SAP.

5.8.2.16 Body Mass Index

No further analysis of BMI as described in the original SAP was performed or deemed clinically necessary.

5.8.2.17 Healthcare Resource Utilization Questionnaire

Analysis of the healthcare resource utilization questionnaire will be performed as specified in the original SAP.

5.9 Extended Open Label Follow Up

Due to the early termination of the study, the data collected from the extended open label follow-up period of the study is scarce. Hence, analyses are limited to only listings for any data collected in the extended open label follow-up period.

5.9.1 Hypothesis of Extended Open Label Follow Up Efficacy Endpoints

No hypotheses for extended open label efficacy endpoints will be tested.

5.9.2 Estimation of Extended Open Label Follow Up Safety Endpoint

No estimation for extended open label safety endpoints will be performed.

5.9.3 Analysis of Extended Open Label Follow Up Efficacy Endpoints

Analysis of extended open label follow up efficacy endpoints will not be performed.

5.9.3.1 Extended Open Label Outcomes

The outcomes collected in the extended open label portion of the study will be listed only

5.10 Subgroup Analyses

Subgroup analyses will be performed as specified in the original SAP.

5.10.1 Screening Failures

Screening failures will be tabulated as specified in the original SAP.

5.10.2 Subject Disposition

Subject disposition definition is as specified in the original SAP.

5.10.3 Changes in Planned Analyses

This addendum serves to clarify the intent of the SAP given the erroneous early termination of the study resulting from an incorrectly entered/analyzed primary efficacy endpoint and omission of US and OUS poolability analysis. In addition, specification of multiple poolability assessments across outcomes pertaining to geography led to the decision to use only one poolability assessment with respect to geography, which is performed on the 6-month primary efficacy endpoint. This poolability assessment will serve as the basis for pooling or not pooling, by geography, for all other endpoints.

5.10.4 Eligibility Criteria

Inclusion and exclusion criteria summaries are as specified in the original SAP.

5.10.5 Demographic and Pre-Implant Characteristics

Subject demographic and pre-implant characteristics summaries are as specified in the original SAP.

5.10.6 Maintenance of Blinding

Analysis of subject guesses of treatment allocation will be performed as specified in the original SAP.

6 MULTIPLE IMPUTATION ANALYSIS

The (multiple) imputation strategy (where performed) is as specified in the original SAP.

7 REFERENCES

1. Little, R.J.A. and Rubin, D.B. (1987) Statistical Analysis with Missing Data. J. Wiley & Sons, New York. (2002) Second edition.
2. Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons. (2004) Classic edition.
3. Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall, London.