PROCEPT BioRobotics

Official Title: WATER Study
Statistical Analysis Plan

NCT Number: NCT02505919

Document Date: 23-Sep-2016

PROCEPT BioRobotics Corporation · 900 Island Drive, Suite 210 · Redwood City, California 94065
office: +1.650.232.7200 · Fax: +1. 650.232.5782 · info@PROCEPT-BioRobotics.com

WATER Study
Statistical Analysis Plan

Reference: TP0038G, WATER Study Clinical Investigational Protocol
Date: September 23, 2016

Table of Contents

1 Background

WATER is a prospective double-blinded randomized controlled trial of aquablation using AQUABEAM system vs. standard transurethral resection of the prostate (TURP). See protocol TP0038 for details regarding eligibility, treatments and assessments. This document describes the

study's statistical analysis plan (SAP). In this document we use the abbreviations "A" and "T" to represent aquablation and TURP.

2 Electronic Data Capture

PROCEPT is using "iMedNet" (MedNet Solutions, Minnetonka, MN 55305) as the electronic data capture provider. iMedNet is a fully functional, 21 CFR 11-compliant web-based system for managing case report forms and downloading study-related data.

3 Data Analysis Software

Statistical analysis will be done using R,[*] an open source data analysis package. Analysis of the interim plan's performance characteristics, as well as the interim calculation itself, will be performed by Berry Consultants.[†] Selected output from statistical analyses will be shared with the study's DMC, which is being managed by Boston Biomedical Associates.[‡]

4 Analysis Cohorts

The following analytic cohorts are defined.

4.1 Modified Intent-to-Treat (mITT) Cohort

The mITT population includes all randomized subjects in whom the assigned device (resection loop for T or handpiece for A) is inserted into the penile urethra. A patient found at the time of the procedure to have a condition that results in study exclusion does not contribute to the mITT cohort. The mITT population is the primary analysis population for both the primary safety and effectiveness endpoints.

4.2 Per-protocol (PP) Cohort

The per-protocol (PP) population is all mITT subjects who:

1. meet critical study eligibility criteria;
2. have no significant protocol deviations that could affect the validity of data; and
3. have evaluable assessment for the endpoint of interest.

A significant protocol deviation means non-adherence on the part of the subject or investigator to clinically significant protocol-specific inclusion/exclusion criteria, primary objective variable criteria or critical study requirement that could affect the scientific validity of the observed data point or lead to bias. Missing data are not imputed for analyses with the PP cohort.

4.3 Safety Cohort

The safety analysis population includes all randomized subjects in whom the assigned BPH surgery is initiated. This cohort is used for most safety analyses.

5 Study Endpoints

5.1 Primary Safety Endpoint

The study's primary safety endpoint is the proportion of subjects with adverse events rated as probably or definitely related to the study procedure classified as Clavien-Dindo Grade 2 or higher or any Grade 1 event resulting in persistent disability (e.g. ejaculatory disorder or erectile dysfunction) evidenced through 3 months post treatment (see protocol for details). Note that the Clavien-Dindo classification scheme is for grading postoperative complications not events that

[*] See https://cran.r-project.org/
[†] http://www.berryconsultants.com/
[‡] See http://boston-biomedical.com/

reflect lack of effective treatment. The endpoint is adjudicated by the study's clinical events committee (CEC).

The primary safety endpoint is considered successful if the proportion of A subjects with the endpoint is non-inferior to the proportion of T subjects with a non-inferiority margin of 10%. Hypotheses are:

$$H_{0,S,N}: S=S_A-S_T \geq 10\%$$

$$H_{A,S,N}: S=S_A-S_T < 10\%$$

That is, if the difference in safety proportion (S) between A and T is statistically <10%, non-inferiority will be concluded. If non-inferiority is concluded, an additional test will be performed for statistical superiority:

$$H_{0,S,S}: S=S_A-S_T \geq 0\%$$

$$H_{A,S,S}: S=S_A-S_T < 0\%$$

That is, if the difference in safety proportion (S) between A and T is statistically <0%, superiority will be concluded. The last subscript in each of the 4 listed hypotheses is either N for non-inferiority and S for superiority. Therefore, the four safety hypotheses are:

- Null for non-inferiority
- Alternative for non-inferiority
- Null for superiority, and
- Alternative for superiority

As detailed below, an interim predictive probability of safety non-inferiority will be calculated and the study's enrollment may be enhanced to increase the chance of demonstrating superiority under certain conditions (described below). A 2-sided 95% confidence interval for the difference in proportions will be used to test for non-inferiority. Non-inferiority will be declared if the entire 2-sided 95% CI is less than 10%. A gate-keeping strategy is used such that if the non-inferiority test is positive, a superiority test will be performed. A 2-sided (1-α)% confidence interval will be used to test for superiority. Superiority will be declared if the entire 2-sided 95% CI is less than 0%. The sample size is fixed for the non-inferiority trial, but the sample size may increase (described below). For this reason, a multiplicity adjustment is required for the superiority analysis. The `testBinomial` function from the `gsDesign` package will be used to compare proportions and calculate non-inferiority and (if relevant) superiority p-values. The `ciBinomial` from `gsDesign` will be used to calculate 95% two-sided confidence limits.

FDA has suggested that the primary safety endpoint should include events deemed "possibly" related. PROCEPT believes that including such events could lead to bias, since investigators may be more likely to report an event as "possibly" related to A since A is a less familiar procedure. Independent adjudication by a CEC would not necessarily remove this reporting bias. Events that are only "possibly" related are of interest, but not as it regards an unbiased safety estimate. Based on FDA's feedback, an additional analysis will be performed including events meeting the primary safety endpoint that are "possibly" related.

5.2 Primary Effectiveness Endpoint

The primary effectiveness endpoint is the IPSS change score from baseline to 6 months. IPSS is an accepted and commonly used measure for symptom severity in BPH.[1] IPSS varies from 0 (no symptoms) to 35 (maximal symptoms).

The difference in mean change scores will be calculated as $D=D_A-D_T$, where D_x signifies mean improvement from baseline, expressed as a positive number. D is the difference improvement; thus if the improvement is 14 points in A and 15 points in T, the difference D = -1 point. The primary hypothesis is that of non-inferiority of 6-month change scores:

$$H_{0,E,N}: D=D_A-D_T \leq NIM$$

$$H_{A,E,N}: D=D_A-D_T > NIM$$

For these hypotheses, NIM is expressed as a negative number (e.g., -5 points). If non-inferiority is concluded, an additional test will be performed for statistical superiority:

$$H_{0,E,S}: D=D_A-D_T \leq 0$$

$$H_{A,E,S}: D=D_A-D_T > 0$$

The same notation is used for the last subscript: N = non-inferiority and S = superiority. Therefore, the 4 effectiveness hypotheses are:

- Null for non-inferiority
- Alternative for non-inferiority
- Null for superiority
- Alternative for superiority

A 2-sided 95% confidence interval for the difference in means will be used to test for non-inferiority. Non-inferiority will be declared if the entire 2-sided 95% CI is greater than the NIM (where NIM is expressed as a negative number). A gate-keeping strategy is used such that if the non-inferiority test is positive, a superiority test will be performed. A 2-sided $(1-\alpha)$% confidence interval will be used to test for superiority. Superiority will be declared if the entire 2-sided 95% CI is greater than 0%. The sample size is fixed for the non-inferiority trial, but the sample size may increase (described below). For this reason, a multiplicity adjustment is required for the superiority analysis. Critical values will be chosen to ensure Type I error control.

The study will be deemed a non-inferiority success if both the safety and effectiveness primary endpoint null hypotheses are rejected. Additional analyses for the primary effectiveness endpoint will be performed using general linear models that incorporate potential predictors of IPSS change scores (e.g., baseline IPSS) as covariates. Subgroup analysis is described below.

5.3 Statistical Adjustment for Superiority Testing

As described below, an interim probability calculation will be performed and the study's enrollment may be enhanced to increase the likelihood of demonstrating superiority for either the safety or effectiveness primary endpoints (or both). If enrollment is enhanced, the nominal p-values for the safety and/or effectiveness superiority hypotheses will be adjusted to preserve overall Type 1 error rate. A description of the adjustment will be provided in a separate document. Adjustment is not required for safety and effectiveness non-inferiority testing since no planned trial adaptation is anticipated for the non-inferiority hypotheses.

5.4 Study-Specific NIM

WATER will use a study-specific non-inferiority margin based on baseline IPSS score, as detailed below.

5.4.1 Baseline Scores in TURP Trials

Baseline scores are important because of their relationship to change score that represent "slight improvement" (see discussion below). Marszalek[2] reported a meta-analysis of TURP trials (see Figure 1 and Table 1). The unweighted average mean baseline IPSS score in these TURP trials was 21, consistent with the American Urology Association's 2011 guidelines document, which notes baseline IPSS scores in TURP trials "between 20 and 24."[3]

5.4.2 Change Scores in TURP Trials

In Marszalek's group of TURP trials, the unweighted average IPSS change score at 12 months was 16 points. Because TURP has a large effect on IPSS scores, a "ceiling" effect is seen (wherein score changes reach their maximum). This results in a statistical relationship between baseline score and score change (Figure 4).

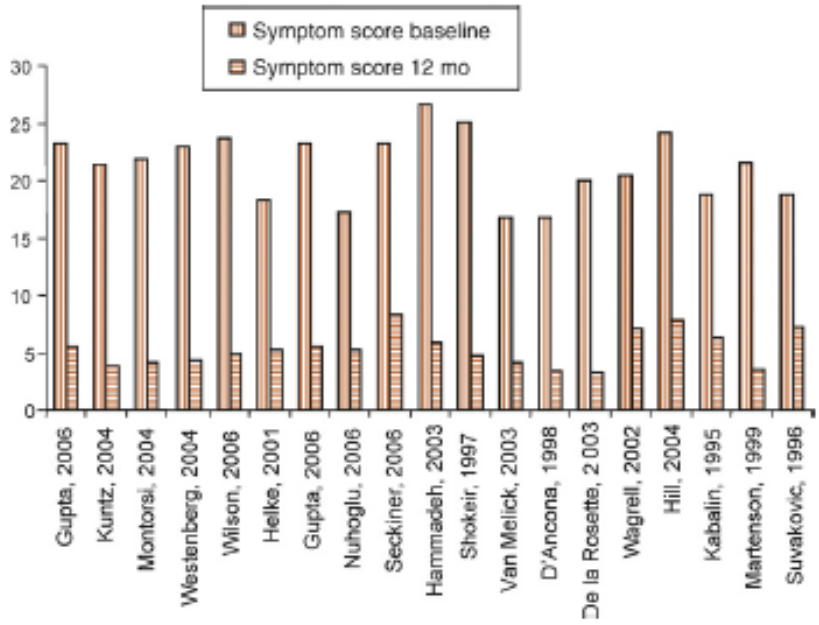

Figure 1. TURP trials reported in Marszalek.[2]

Table 1. Trials included in Marszalek et al.[2]

Author	Baseline IPSS	IPSS at 12 months	Improvement
Gupta	23.5	6	17.5
Kuntz	21	2.4	18.6
Montorsi	22	3.8	18.2
Westernberg	23	4	19
Wilson	24	5	19
Helke	18	5.5	12.5
Gupta	23.5	6	17.5
Nuhoglue	17.5	5	12.5
Seckiner	24	8	16
Hemmadeh	27	6	21
Shokeir	25.2	5	20.2
Van Melick	17	3.5	13.5
D'Ancona	17	3	14

De la Rosette	20	3	17
Wagrell	20.5	7.5	13
Hill	24.5	8	16.5
Kabalin	18	6.5	11.5
Martenson	22	3.5	18.5
Suvakovic	18.5	7.5	11
Non-weighted average	**21.4**	**5.2**	**16.2**

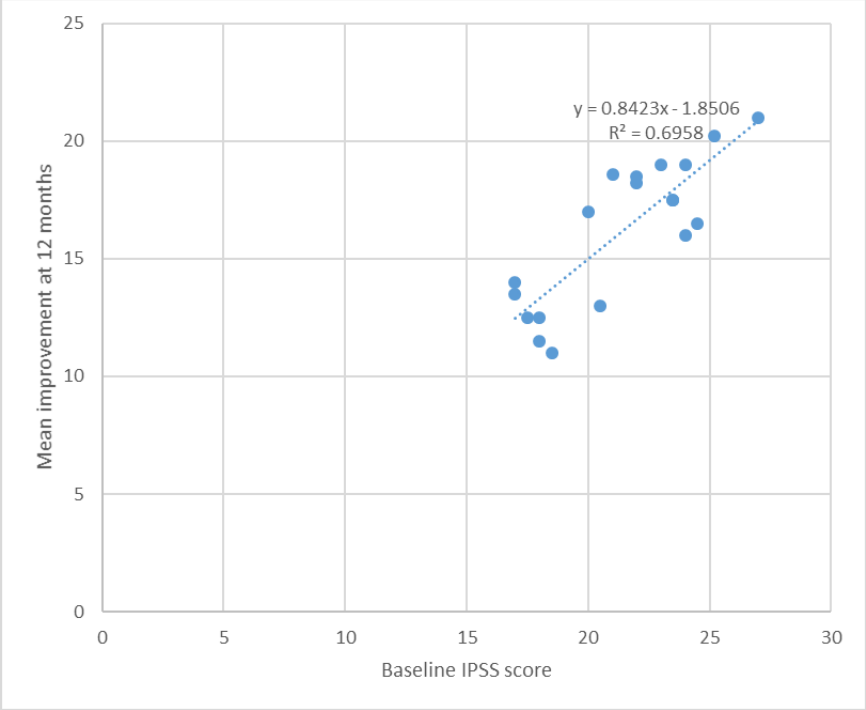

Figure 2. Relationship between baseline IPSS score and change score in TURP studies summarized in AUA 2003 document.

5.4.3 Effect Size with No Treatment

FDA's draft guidance document on non-inferiority margins (NIMs) in equivalence trials* notes that NIMs should take into account not only the comparator treatment but also the treatment effect size, M1, i.e., the difference between treatment and no treatment. Non-inferiority trials are not supposed to use a NIM value larger than M1, the treatment effect for the comparator device.

For TURP, values for M1 are not available: there are no published placebo-controlled (or sham-controlled) trials of TURP. While in theory, the IPSS change score with no treatment in patients with moderate-to-severe BPH seeking surgical treatment is zero, several factors could contribute to non-zero changes, including regression to the mean, test-to-test variability, placebo effect, or spontaneous remission. Proxies for this "surgical placebo" change score are available. In placebo-controlled trials of alfusozin, a commonly used alpha blocker, baseline scores were in the range of 17-19 and available change scores with placebo were 1.6 points, 4.2 points, 4.7 points and 4.9 points (see Figure 3). Meta-analysis of sham-controlled trials of endoscopic treatments for BPH (e.g., transurethral microwave thermotherapy)[4] has shown a mean IPSS change score of about 6 points. Because subjects in sham trials are unaware of treatment, their expectations for improvement may be substantially higher than what would be observed with no treatment in

* See http://www.fda.gov/downloads/Drugs/.../Guidances/UCM202140.pdf

standard clinical practice.* We conclude that the IPSS change score in men with BPH seeking surgical treatment is likely to fall between 0 and 6 points. Therefore, the difference between the TURP change score and the no-treatment change score, i.e., M1 in FDA's guidance document, is 10-16 points.

Note that the IPSS improvement resulting from BPH medication use is not relevant to WATER since WATER eligibility criteria require either failure of medical treatment or refusal of medical treatment.

Table 3.1b. Efficacy and effectiveness outcomes in alfuzosin randomized, controlled trials

Author, Year Study duration	Intervention (no. of patients assessed)	Baseline [mean (SD)]	Endpoint [mean (SD)]	Within group difference (within group P-value)	Between group difference (P-value)
Total International Prostate Symptom Score (I-PSS)					
Alfuzosin compared with placebo					
McNeill SA, 2005 6m	Alfuzosin 10mg QD (82)	NR	8.75(NR)	NR	Vs placebo: NR (P=0.012)
	Placebo QD (83)	NR	11.45(NR)	NR	NR
Roehrborn CG, 2001 3m	Alfuzosin 10mg QD (170)	18.2(6.3)	NR	-3.6 (NR)	Vs placebo: -2.0 (P=0.001)
	Alfuzosin 15mg QD (165)	17.7(5.7)	NR	-3.4 (NR)	Vs placebo: -1.8 (P=0.004)
	Placebo QD (167)	18.2(6.4)	NR	-1.6 (NR)	NR
Roehrborn CG, 2003 3m	Alfuzosin 10mg QD (473)	18.7(4.6)	12.7(6.1)	-6.0 (NR)	Vs placebo: -1.8 (P<0.001)
	Placebo QD (482)	18.8(4.4)	14.6(6.8)	-4.2 (NR)	NR
Roehrborn CG, 2006 2y	Alfuzosin 10mg QD (749)	19.2(4.7)	NR	-5.9 (NR)	Vs placebo: -1.2 (P=0.0017)
	Placebo QD (757)	19.2(4.7)	NR	-4.7 (NR)	NR
Van Kerrebroeck P, 2000 3m	Alfuzosin 10mg QD (137)	17.3(3.5)	10.4(4.7)	-6.9 (NR)	Vs placebo: -2.0 (P=0.002)
	Alfuzosin 2.5mg TID (147)	16.8(3.7)	10.5(6.1)	-6.4 (NR)	Vs placebo: -1.5 (P=0.02)
	Placebo QD (152)	17.7(4.1)	12.8(6.7)	-4.9 (NR)	NR
Alfuzosin compared with doxazosin					
De Reijke TM, 2004 14w	Alfuzosin 2.5mg BID/TID (87)	18.0(4.8)	NR	-7.5 (P<0.00)	Vs doxazosin: 1.7 (P<0.05)
	Doxazosin 1-8mg/day (93)	19.1(5.2)	NR	-9.2 (P<0.001)	NR

Figure 3. Alfuzosin trials summarized in AUA 2010 guideline.

5.4.4 Barry's Estimate of Slight Change

NIMs for BPH trials have historically relied on 1995 data from Barry et al,[5] which describes men with mild-to-moderate BPH participating in a clinical trial of BPH medications. The specific purpose of Barry's report was to note the correlation between IPSS change score and a global measure of improvement on a 5-point scale: marked, moderate, slight or no improvement, or worse. For each category of perceived improvement, the mean change score associated with that degree of global improvement was linearly related to baseline IPSS score (Figure 4). For the purposes of

* A patient in a sham trial may think that he has a 50% (if 1:1 randomization) chance of having received the active treatment (vs. sham). This may result in change scores that are biased upwards compared to a patient who knows he is receiving non-active treatment only.

a non-inferiority clinical trial, the mean IPSS change representing "slight improvement" is most relevant and can serve as an appropriate NIM.

The mean IPSS improvement representing "slight change" was approximately 1 point for subjects with a score of 10 and 8 points for a baseline score of 30. Barry's regression line describing this relationship is:

$$\text{Change score representing slight change} = -0.3636 * IPSS_{base} + 2.9091$$

where $IPSS_{base}$ is the baseline IPSS score. (Roehrborn et al reported a very similar relationship in another BPH trial.[6]) Barry's finding makes clinical sense: a 1-point improvement in a patient with a baseline score of 30 would not be perceptible, where as one-point improvement in a patient with a baseline score of 10, while small, is probably detectable.

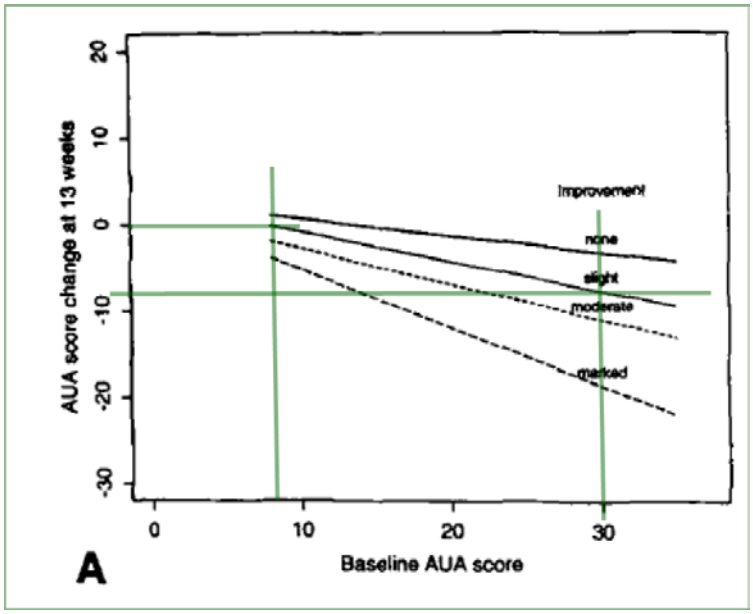

Figure 4. Relationship between baseline IPSS and change score constituting various degrees of global improvement. Drawn on the figure are horizontal and vertical lines used to calculate the formula corresponding to the line for "slight improvement": score change = $-0.3636 * IPSS_{base} + 2.9091$

Barry et al highlighted that "the range of patient baseline scores must be considered when attaching a point value to a slight or moderate improvement rating." Moreover, investigators "must describe patient baseline scores, which clearly influence how they perceive the subsequent score changes." The strong and clear implication of these statements and Figure 4 is that interpretation of IPSS change scores needs to take into account baseline severity, as reflected in baseline IPSS scores. Moreover, because of the marked differences in target patient populations between Barry and WATER (Table 2), using Barry's "slight improvement" mean IPSS of 3 points for WATER is scientifically invalid.

Table 2. Comparison of Barry and WATER

	Barry	**WATER**
IPSS entry criteria	≥8	≥12
What patients are seeking	Medications for BPH	Surgery for BPH*
Baseline IPSS score, mean	Approximately 16	Estimated 22**

* Note that WATER excludes patients failing medical therapy or refusing medical therapy
** Baseline IPSS scores in PROCEPT's Phase 2 studies were >20

It is easily demonstrated that a cohort's mean IPSS change score perceived as "slight change" is mathematically directly related to the population's mean baseline IPSS score. Using Barry's regression line, we can calculate the expected mean score change representing slight change for a hypothetical group of men with BPH as:

$$(1)\ \text{NIM} = \text{Mean Change for Slight Improvement} = \frac{\sum_{i=1}^{N}(mBase_i+b)}{N}$$

i.e., the sum of each man's threshold change score calculated from Barry's regression line (m = -.3636, b = 2.9091) and the man's baseline IPSS ($Base_i$) divided by the sample size. Expanding (1):

$$(2)\ \text{NIM} = \frac{\sum_{i=1}^{N}(mBase_i+b)}{N} = \frac{\sum_{i=1}^{N} mBase_i}{N} + \frac{\sum_{i=1}^{N} b}{N} = \frac{m\sum_{i=1}^{N} Base_i}{N} + b = m\overline{Base} + b$$

Equation (2) directly demonstrates that the mean IPSS change score representing "slight improvement" is linearly related to the mean baseline score ($\overline{Base}$). The median baseline IPSS in Barry's cohort was 16, which, based on (2), corresponds to a change score representing slight improvement of approximately 3, i.e., the value reported by Barry (and applicable, we would argue, to his trial only). In TURP trials, with mean baseline scores of 20-24 (see Figure 1), the calculated change score for slight improvement would be 4.4 points for a 20-point baseline mean and 5.8 points for a 24-point baseline mean.

We will therefore calculate a study-specific NIM for WATER using (2) and the overall (both groups) baseline mean IPSS score. With an expected baseline mean IPSS score of about 23 (the mean baseline values in PROCEPT's phase 2 studies), the study-specific NIM would be approximately 5 points. This value represents the best literature-based estimate of the change score representing slight improvement in men with BPH seeking surgical treatment.

Note that FDA's guidance document on NIMs suggests choosing a NIM value (M2) that is substantially less than that of M1 such that the new treatment preserves most of the effect size compared to the standard treatment. Clearly, an M2 of approximately 5 points preserves most of TURP's effect size (10-16 points, see Section 5.4.3).

Based on FDA's feedback of August 26, 2016, an additional calculation will be made using a fixed NIM value of 4.7 points.

5.5 Sample Size Calculation

Sample size was reported in the original study protocol. Version G proposed an NIM of 4.7 points (based on an estimated baseline score distribution). With a sample size of 177 randomized subjects (118 A and 59 T), an effect size of -1.5 points, a standard deviation of 6 and a NIM of 4.7 points, the study has adequate power for the primary effectiveness endpoint. Power for the safety endpoint was also high assuming endpoint rates of 65% in T and 40% in A, including a 12% loss to follow-up rate.

5.6 Bayesian Predictive Probability Calculation

5.6.1 Overview

The study will include an interim predictive probability calculation. The calculation will take place after the 177th randomized patient is enrolled and treated. At this time point, it is estimated that 60 subjects will have 6-month data and 132 will have 3-month data, and 45 will have less than 3-

month data. The goal of the calculation is to determine the predictive probability of achieving non-inferiority and/or superiority for the primary safety and effectiveness endpoints at final analysis given all available trial data at the interim time point. The analysis will incorporate both observed data and predicted data.

If the predictive probability of non-inferiority for both the safety and effectiveness endpoints is ≥0.99, a report summarizing primary and secondary endpoints will be submitted to FDA for consideration for device clearance. Primary endpoint claims from this submission would be based on non-inferiority only. PROCEPT will provide an update to FDA when all final data are available.

If the predictive probability of non-inferiority trial success is <0.99, the trial will continue to the pre-planned sample size (177 randomized) and a single calculation (described in sections 5.1 and 5.2) performed when all study data are available.

Additionally, the predictive probability of superiority for the primary safety and effectiveness endpoints will be calculated. See the attached report from Berry Consultants for further details. In certain situations trial sample size may be increased to 300 randomized subjects with the goal of showing statistical superiority. This additional enrollment is for showing superiority only and will not affect FDA's consideration of the device for market approval based on non-inferiority safety and effectiveness claims. It is expected that much of this additional enrollment, follow-up and analysis would take place in the post-market setting (i.e., after device clearance). Superiority claims will be based on a single final analysis when all data from all enrolled and treated subjects are available.Interim Predictive Calculation

At the interim time point, we will calculate the Bayesian predictive probability of rejecting the safety and effectiveness non-inferiority hypotheses ($H_{0,S,N}$ and $H_{0,E,N}$) for 177 subjects so as to declare non-inferiority for the safety and effectiveness endpoints. We will use a similar approach to calculate the predictive probability of rejecting the safety and effectiveness superiority hypotheses ($H_{0,S,S}$ and $H_{0,E,S}$) for 177 as well as for total sample size of 300 subjects. Note that although interim calculations take a Bayesian approach, final calculations will be frequentist.

Predictive Probability of Safety

To calculate the predictive probability of non-inferiority or superiority, we use non-informative Jeffreys' priors for the primary safety endpoints for the A and T groups:

$$S_A, S_T \sim \text{Beta}(½, ½)$$

Where Beta is the beta distribution. The posterior distributions are therefore:

$$S_A \sim \text{Beta}(1/2 + X_A, ½+N_A-X_A) \text{ and } S_T \sim \text{Beta}(1/2 + X_T, ½+N_T-X_T)$$

where X_A and X_T represent the number of safety events in the A and T groups, respectively, and N_A and N_T represent the number of A and T subjects, respectively with assessed 3-month safety outcomes.

The number of future successes is therefore calculated from a beta-binomial distribution. For the observed number of successes (X_A and X_T) and every possible observed number of future successes (X_A' and X_T'), we calculate the probability of each outcome $Pr(X_A')\times Pr(X_T')$, then determine if the combination ($X_A + X_A'$)/ (N_A and N_A') vs. ($X_T + X_T'$)/ (N_T and N_T') would produce a statistically significant result for the non-inferiority and superiority tests. Predictive probability is the sum of probabilities that result in positive trials, as described in Saville et al.[7]

Predictive Probability of Effectiveness

At the time of the predictive probability calculation, some patients will have complete six-month outcome data (as well as 1- and 3-month data), some will have only 3-month data, some will have only 1-month data, and the most recently enrolled patients will not yet provide any efficacy data. Statistical models for predicting 1, 3 and 6-month IPSS scores, as well as a calculation of the predictive probability of effectiveness non-inferiority or superiority, are described in the attached report from Berry Consultants and not repeated here.

5.6.2 Augmenting Enrollment for Superiority

As noted above, predictive distributions will be calculated for 177 as well as for a total sample size of 300 subjects. For the superiority-focused calculations, there are 3 possible outcomes:

- Predicted probability of superiority for both outcomes with augmenting trial to 300 subjects size is low (<50%). In this case, the trial's sample size will NOT be increased. As noted above, a trial report will be sent to FDA, with an update when all subjects have 3-month safety and 6-month effectiveness data.

- Predicted probability of success at the interim sample size in both outcomes is high (>80%). As in the above scenario, the trial's sample size will NOT be increased and a trial report sent to FDA, along with the above-described update when 3- and 6-month safety and effectiveness data are available.

- Predicted probability of safety or effectiveness superiority after augmentation to 300 randomized subjects is between 50 and 80%. In this case, trial size will be increased to 300 subjects. If trial size is increased, it is expected that much of the follow-up phase and a single final analysis for these additional subjects will take place in the post-market setting when all data are available.

6 Secondary Endpoint Assessments

Six secondary endpoints, described below, are intended for marketing claims: 3 are based on 6-month results and 3 are based on perioperative results. Statistical testing for the 6-month secondary endpoints will only occur after data collection is complete.

Tests based on 6-month data:

1. Reoperation or re-intervention within 6 months. Reoperation means any surgical procedure on the lower urinary tract to treat problems potentially related to BPH. Re-intervention means any invasive procedure (e.g., cystoscopy) to evaluate problems potentially related to BPH. Re-intervention excludes TRUS and urodynamics, which are required study evaluations during follow-up. Re-intervention also excludes catheterization (which is not a surgical procedure) for acute urinary retention. The proportion of subjects with this endpoint will be compared using Fisher's exact test. Differences in proportions will be calculated with `testbinomial` from the gsDesign package. A 10% non-inferiority hypothesis (like the primary safety endpoint, of which this is a component) will be used and superiority will be explored after non-inferiority is shown.

2. The proportion of sexually active subjects reporting a worsening of sexual function through 6 months on either the IIEF or the MSHQ-EjD questionnaires. A subject is counted as having worse sexual function if IIEF-5 (part of IIEF-15) is decreased by at least 6 points[8] or MSHQ is decreased by at least 2 points.[9] The two-point threshold for MSHQ is close to

the value that distinguishes affected from unaffected men.* Both IIEF and MSHQ assume that a man is sexually active. If the subject reports that he is not sexually active for the time period assessed, these threshold decreases cannot be met by definition. Differences in proportions will be calculated as described above. Superiority will be tested.

3. The proportion of subjects with major adverse urologic events (MAUE). MAUE is defined as any subject with a serious device- or serious procedure-related AE judged to be urologic through 6 months. Whether an event meets this endpoint is judged using the study's SAE definition and relatedness, both of which are adjudicated by the CEC. Differences in proportions will be calculated as described above. A 10% non-inferiority hypothesis (like the primary safety endpoint) will be used.

Tests based on perioperative data:

4. Length of hospital stay (days) in the treatment groups. Proportional odds logistic regression will be used. A NIM of ½ a day will be used. Additional analyses will be done stratifying for geography (e.g., Europe vs. US vs. Australia/New Zealand), as it is known that LOS is generally shorter in the US.

5. Length of operative time (minutes) in the treatment groups, defined as time from pre-treatment visualization to insertion of indwelling catheter (IDC), will be compared using a t test. Simple superiority will be tested.

6. Length of resection time (minutes) in the treatment groups, defined as start of first pedal activation to end of last pedal use for either A or T, will be compared using a t test. Simple statistical superiority will be evaluated.

The 6 endpoints listed above will be tested in sequential fashion using the Holm step-down procedure for type-I error rate correction.[10] The testing will be done only if the primary safety and effectiveness endpoints meet the study's non-inferiority goals. Secondary endpoints listed above will be ordered sequentially from most significant to least significant. The endpoints will then be tested in order at adjusted levels of significance (i.e. E1: 0.05/k, E2: 0.05/k-1; E3: 0.05/k-2, etc., where k=number of specified endpoints). Once an endpoint fails, all subsequent secondary endpoints will not be statistically tested.

7 Additional Endpoint Assessments

The following additional endpoints, evaluated using the PP cohort, are pre-specified but are not intended to support product labeling. Comparisons of proportions described below will be done using methods similar to those described for the reoperation/reintervention secondary endpoint. Comparisons of continuous variables will use a t test unless otherwise described. Testing of continuous variables will examine superiority, and in some cases, non-inferiority. Superiority tests will be 2-sided (alpha=.05). Non-inferiority tests will be one-sided (alpha=.025).

1. Proportion of subjects with Clavien-Dindo classification of Grade 2 or higher or any Grade 1 with persistent disability (e.g. ejaculatory disorder or erectile dysfunction) at 30 days, 6 months, and 1 year. 6-month and 1-year calculations will be performed only when full datasets are available. A 10% NIM will be used.

* According to personal communication Ray Rosen, creator of MSHQ, no published MCID for MSHQ exists.

2. Evaluation of the proportion of subjects with dysuria through the day 30 visit. Dysuria is defined as burning sensation while urinating of at least "about ½ the time" on the dysuria questionnaire. An additional comparison using proportion odds logistic regression will be performed. A 15% NIM will be used.

3. Duration of bladder catheterization (i.e., intraoperative catheter placement to removal prior to discharge) after the assigned study procedure. Duration will be compared with a t test. An NIM of ½ a day will used.

4. Change in hemoglobin (gm/dl) (i.e., laboratory test) at discharge from baseline.

5. Reoperation or Re-intervention within 30 days, 12 months, 24 months and 36 months. For the 30-day time point, proportions will be compared with a 10% NIM. For longer time points, due to the expectation of study withdrawal, Kaplan-Meier survival analysis will be performed and a log rank test calculated.

6. Changes in the proportion of subjects using medications for BPH symptoms at month 6. Two analyses will be done. 1) an analysis of the proportion of subjects who increased doses of BPH medications or started a new BPH medication, and 2) the proportion who were able to decrease or stop doses of BPH medications. Relevant medications include alpha blockers, prostaglandin inhibitors, and phosphodiesterase inhibitors. A 10% NIM will be used.

7. IPSS change score during follow-up (1 week, 1, 3, 6, 12, 24 and 36 months). Repeated measures analysis of variance (RMANOVA) will be used, with additional models that include baseline IPSS as a covariate. This analysis accounts for all available measurements and correlation of measurements within subjects. The mean differences across groups will be determined and the confidence limits from the regression model compared with the study-specific NIM. A treatment × time interaction will be explored as well. The purpose of this analysis is to evaluate whether the IPSS change score across all time points is non-inferior in A compared to T. The same NIM will be used as described for the primary effectiveness endpoint. If non-inferiority is shown, superiority will be tested. All available time points will be used up to month 36.

8. IPSS-QoL change scores (1 week, 1, 3, 6, 12, 24 and 36 months). A similar method as that with IPSS change score will be used and a 1-point NIM will be assumed.

9. Qmax change scores at 1, 3, 6, 12, 24 and 36 months. RMANOVA will be used.

10. PVR change scores at 1, 3, 6, 12, 24 and 36 months. RMANOVA will be used.

11. Proportion of subjects with device- or procedure-related adverse event. Proportions compared as described for primary safety endpoint.

12. Pelvic pain intensity level, population mean and change score mean, measured on a 0-10 numeric rating scale. Population score distributions will be compared with a Wilcoxon signed rank test. Change scores will be compared with a Wilcoxon test or t test. A non-inferiority margin of 2 points will be used.

13. EQ-5D time trade-off index (TTO) change scores during follow-up (7 days, and 1, 3, 6, 12, 24 and 36 months). TTO will be calculated using study data combined with available country-specific norms. RMANOVA will be used.

14. ISI at baseline, at post-op, 7 days, 1, 3, and 6 months. RMANOVA will be used.

15. IIEF-15 at baseline, and 7 days, 1, 3, 6, 12, 24 and 36 months. RMANOVA will be used.

16. MSHQ-EjD at baseline, 7 days, 1, 3, 6, 12, 24 and 36 months. RMANOVA will be used.

17. WPAI:US at baseline, 7 days, 1, 3, 6, 12, 24 and 36 months. Specifically, among those subjects who are currently working, the number of hours missed due to urinary symptoms will be compared. Given a large expected number of zeroes, a non-distributional method such as Wilcoxon's test will be used.

18. Pdet@Qmax at baseline and 6 months. This continuous outcome will be evaluated with a t test.

19. Change in categorization of subjects from obstructed to unobstructed or unobstructed to obstructed at month 6 in those subjects in whom urodynamics was performed. (Urodynamics is an optional test.) The determination of "obstructed" and "unobstructed" is based on urodynamics values and Chapple.[11] Specifically, AG (Abrams-Griffith) number is calculated as AG=Pdet@Qmax – 2*Qmax. AG is interpreted as:

    If AG≥40: obstructed
    If AG <20: not obstructed
    If 20≤AG < 40: calculate slope = (Pdet@Qmax-Pvoid0)/Qmax. If slope >2, obstructed, else unobstructed.

    Changes from baseline to follow-up within a treatment group will be calculated with a McNemar's test. Conditional odds ratio will be calculated according to Suzuki.[12]

20. Use of cautery immediately post Aquablation. The proportion in which cautery was used and the distribution of cautery time will be calculated. Cautery use will be assessed as minutes from cautery start to stop and total amount of "on time".

21. Proportion of subjects in whom re-catheterization was needed between discharge after the index procedure and month 3. Re-catheterization is defined as the need to place a urinary catheter in the bladder for symptoms related to BPH. This excludes re-catheterization for study purposes or for purposes unrelated to LUTS.

22. Amount of irrigation fluid used intraoperatively (liters). A t test or Wilcoxon's test will be used to evaluate mean differences.

23. Proportion of subjects in whom postoperative bladder irrigation was started.

24. Prostate size reduction from baseline to 3 months as measured TRUS. A t test will be used for the comparison.

25. Relationship between prostate size reduction and change in various measurements (IPSS, etc.).

26. Relationship between prostate size and procedure or resection times. It is expected that procedure times will be related to prostate size in T but not A. Linear models will be run for each treatment group. In addition, a linear model will be run for all subjects that includes an interaction term.

8 Subgroup Analyses

Subgroup analyses of primary safety and effectiveness endpoints, as well as selected secondary endpoints, will be performed for the following subgroups:

- baseline IPSS scores of <20 vs. ≥20
- baseline prostate size of <50g vs. ≥50g
- Age <65 vs. ≥65 years at baseline

9 Pooling

Data will be pooled across sites when performing statistical analysis. The justification for pooling includes the following:

- Study sites will be following the same Protocol
- Study sites use the same device system
- Study sites follow the same instructions for use document
- Study subjects are enrolled using identical criteria across sites
- Randomization within site is designed to ensure balance within sites, minimizing site-to-site variation of treatment effect
- Frequent contact with sites and monitoring of study data

Potential heterogeneity of results will be examined. For the primary effectiveness endpoint, analysis of variance will be used to determine heterogeneity of effect sizes across site. Heterogeneity will be assumed to be present if the site × treatment interaction p-value is <.05. Additional models may incorporate potential predictors of IPSS change scores, including baseline demographic factors (age, race), baseline IPSS scores, procedure-related variables (ablation time, resection time). If these variables do not adequately explain heterogeneity across sites, mixed models will be used that assume variation of change scores or treatment effect across sites.

For the primary safety endpoint, heterogeneity will be assessed using a Mantel-Haenzel odds ratio test or similar test. Heterogeneity will be assumed to be present if the site × treatment interaction p-value is <.05. If evidence of non-poolability is found, baseline and procedural variables found to be different between sites will serve as predictors in a logistic or linear regression that also includes as predictors the treatment assignment, site, and site-by-treatment interaction. If these variables do not adequately explain heterogeneity across sites, mixed models will be used that assume variation of event rates across sites. Mixed models treat some factors (e.g., site ID) as a random effect and are often used in this situation.

Variation may be different across procedures. That is, variation may be smaller in A vs. T, since A is automated. Variation in IPSS change scores and the primary safety endpoint across sites will therefore be examined within procedure.

10 Missing Data Imputation

Missing data will be minimized through careful study monitoring. To evaluate bias associated with missing data, characteristics of subjects with missing data will be compared, when relevant, to those of subjects with study data.

The impact of missing data on the final analysis of the primary safety endpoint analysis will be evaluated using the following models:

- Ignore missing data
- Assume missing values have met the primary safety endpoint
- Assume missing values have NOT met the primary safety endpoint
- Regression analysis. Logistic regression will be used to determine baseline predictors, if any, of the occurrence of the primary safety endpoint. If predictors are found and have biologic relevance, logistic models will be used to calculate the probability of each missing subject having had a primary safety endpoint. The sum of these probabilities across all missing subjects will be used. Covariates to be examined in logistic regression models include: age, site ID, prostate volume, assigned procedure, procedure length, preoperative sexual function, baseline IPSS score, baseline MSHQ score, baseline IIEF score.
- Finally, a tipping point analysis will be performed for the primary safety endpoint. This is the number of additional events occurring in the A group (or fewer events in the T group) that would cause the study to fail to conclude safety non-inferiority.

The impact of missing data on the final analysis of the primary effectiveness endpoint analysis will be evaluated using the following models:

- Predictive models, as described above for the interim analysis effectiveness predictive calculation
- Last observation carry forward
- Assume relatively low change score, i.e., 8 points
- Assume relatively high change score, i.e., 18 points
- Assume changes consistent with the group's median change score
- Assume zero change score (though this is highly unlikely given the long history of large mean change scores with TURP and similarly large change scores in Phase II data)

11 Citations

1. Barry, M. J. *et al.* The American Urological Association symptom index for benign prostatic hyperplasia. The Measurement Committee of the American Urological Association. *J. Urol.* **148,** 1549–1557; discussion 1564 (1992).
2. Marszalek, M., Ponholzer, A., Pusman, M., Berger, I. & Madersbacher, S. Transurethral Resection of the Prostate. *Eur. Urol. Suppl.* **8,** 504–512 (2009).
3. McVary, K. T. *et al.* Update on AUA guideline on the management of benign prostatic hyperplasia. *J. Urol.* **185,** 1793–1803 (2011).
4. Welliver, C., Kottwitz, M., Feustel, P. & McVary, K. Clinically and Statistically Significant Changes Seen in Sham Surgery Arms of Randomized, Controlled Benign Prostatic Hyperplasia Surgery Trials. *J. Urol.* **194,** 1682–1687 (2015).
5. Barry, M. J. *et al.* Benign prostatic hyperplasia specific health status measures in clinical research: how much change in the American Urological Association symptom index and the benign prostatic hyperplasia impact index is perceptible to patients? *J. Urol.* **154,** 1770–1774 (1995).
6. Roehrborn, C. G., Wilson, T. H. & Black, L. K. Quantifying the contribution of symptom improvement to satisfaction of men with moderate to severe benign prostatic hyperplasia: 4-year data from the CombAT trial. *J. Urol.* **187,** 1732–1738 (2012).

7. Saville, B. R., Connor, J. T., Ayers, G. D. & Alvarez, J. The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clin. Trials Lond. Engl.* **11,** 485–493 (2014).
8. Rosen, R. C., Allen, K. R., Ni, X. & Araujo, A. B. Minimal clinically important differences in the erectile function domain of the International Index of Erectile Function scale. *Eur. Urol.* **60,** 1010–1016 (2011).
9. Rosen, R. C. *et al.* Male Sexual Health Questionnaire (MSHQ): scale development and psychometric validation. *Urology* **64,** 777–782 (2004).
10. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **6,** 65–70 (1979).
11. Chapple, C. R. *Urodynamics Made Easy*. (Elsevier, 2009).
12. Suzuki, S. Conditional relative odds ratio and comparison of accuracy of diagnostic tests based on 2 x 2 tables. *J. Epidemiol. Jpn. Epidemiol. Assoc.* **16,** 145–153 (2006).