STATISTICAL ANALYSIS PLAN

PFDN Protocol Number: 25P01

Effects of Surgical Treatment Enhanced with Exercise for Mixed Urinary Incontinence (ESTEEM)

Short title: Combined treatment for mixed incontinence

SAP VERSION:	Version 2.2
SAP DATE:	September 12, 2018
PREVIOUS VERSIONS:	Version 1.0; August 7, 2017 Version 2.0; May 15, 2018 Version 2.1; July 13, 2018
SPONSOR:	NICHD
PREPARED BY:	RTI International. 3040 Cornwallis Rd Research Triangle Park, NC 27709-2104
AUTHOR (S):	Benjamin Carper Marie Gantz

Contents

SUMMARY OF CHANGES MADE IN SAP VERSION 2.0

1. Clarified that the intention-to-treat analysis population would include only eligible participants (Section 5.1).
2. Added detailed definition of the per protocol population (Section 5.2).
3. Clarified definition of retreatment (Attachment 16.1).
4. Added definition of postoperative complications based on the study manual of operations (Section 10.2).
5. Added sensitivity analysis requested by the DSMB (Section 13).

SUMMARY OF CHANGES MADE IN SAP VERSION 2.1

1. Added revised scoring instructions to the OAB-SAT-q (Section 9.5.1)
2. Added additional analysis endpoints for dryness based on diary responses, UDI-LF responses, and the combination of the two (Section 9.2).

SUMMARY OF CHANGES MADE IN SAP VERSION 2.2

1. Added description of minimally import difference (MID) calculations and analysis (Section 9.5.6)
2. Added change in methods to analyze improvement of voiding frequency (Section 13)
3. Added change in methods to analyze time to failure outcome (Section 13)

LIST OF ABBREVIATIONS

ABC	Anticholinergic versus Botox Comparison trial
ATLAS	Ambulatory Treatments for Leakage Associated with Stress Incontinence trial
BBUSQ	Birmingham Bowel Urinary Symptom Questionnaire
BD	Bladder diary
BE- DRI	Behavior Enhances Drug Reduction of Incontinence trial
BPTx	Behavioral/pelvic floor therapy
CDF	Cumulative distribution function
CST	Cough stress test
DCC	Data Coordinating Center
DO	Detrusor overactivity
DSMB	Data and Safety Monitoring Board
EQ-5D	European Quality of Life-5 Dimensions
HRQOL	Health related quality of life
IE	Incontinence episode
ICI	International Consultation on Incontinence
ICS	International Continence Society
IIQ	Incontinence Impact Questionnaire
IRB	Institutional Review Board
ITT	Intention-to-treat
IUGA	International Urogynecological Association
MESA	Medical, Epidemiologic, and Social Aspects of Aging
MID	Minimum important difference
MIMOSA	Mixed Incontinence: Medical or Surgical Approach trial
MSM	Medical Safety Monitor
MUI	Mixed urinary incontinence
MUS	Mid-urethral sling
OAB	Overactive bladder
OAB-q	Overactive Bladder Questionnaire
OAB-q-SS	Overactive Bladder Questionnaire-Symptom subscale
OAB-SAT-q	Overactive Bladder Questionnaire-Satisfaction with Treatment Questionnaire
OPTIMAL	Operations and Pelvic Muscle Training in the Management of Apical Support Loss trial
PFD	Pelvic floor disorder
PFDI	Pelvic Floor Disorder Inventory
PFDN	Pelvic Floor Disorders Network
PFMT	Pelvic floor muscle training
PGI-I	Patient Global Impression- Improvement
PGI-S	Patient Global Impression-Severity

PISQ	Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire
POPQ	Pelvic Organ Prolapse Quantification system
PVR	Postvoid residual
QoL	Quality of life
QUID	Questionnaire for Urinary Incontinence Diagnosis
RCT	Randomized controlled trial
RUBI	Refractory idiopathic urge incontinence and botulinum A injection trial
SAE	Serious adverse event
SD	Standard deviation
SISTEr	Stress Incontinence Surgical Treatment Efficacy Trial
SUI	Stress urinary incontinence
TOMUS	Trial of Mid-Urethral Slings
TOT	Transobturator tape sling
TVT	Tension-free vaginal tape sling
TVT-O	Tension-free vaginal tape obturator
UDE	Urodynamic evaluation
UDI	Urogenital Distress Inventory
UI	Urinary incontinence
UIE	Urinary incontinence episode
UITN	Urinary Incontinence Treatment Network
UUI	Urge urinary incontinence
ValUE	Value of Urodynamic Evaluation trial
VPFMC	Voluntary pelvic floor muscle contraction
3IQ	3-Incontinence Questions Assessment Tool

1 BACKGROUND AND PROTOCOL HISTORY

Mixed urinary incontinence (MUI), defined as both stress urinary incontinence (SUI) and urge urinary incontinence (UUI), is a challenging condition and there are limited trials evaluating interventions that can optimize treatment outcomes. The overarching goal of this randomized trial is to estimate the effect of combined midurethral sling (MUS) and peri-operative behavioral/pelvic floor therapy (BPTx) compared to MUS alone on successful treatment of MUI symptoms in 472 women. Secondary objectives include estimating the effect of combined treatment compared to MUS on improving OAB and SUI outcomes separately, need for additional treatment, time to failure and identifying predictors of poor outcomes in this MUI population.

2 PURPOSE OF THE ANALYSES

This statistical analysis plan (SAP) contains detailed information about statistical analyses to be performed to address the primary and secondary aims of ESTEEM. All analyses that will be included in the primary manuscript are described. Additional exploratory analyses may be performed to support further manuscript development. These analyses will not require an update to the SAP.

3 STUDY AIMS AND OUTCOMES

3.1 Study Aims

Mixed urinary incontinence (MUI), defined as both stress urinary incontinence (SUI) and urge urinary incontinence (UUI), is a challenging condition and there are limited trials evaluating interventions that can optimize treatment outcomes. The overarching goal of this randomized trial is to estimate the effect of combined midurethral sling (MUS) and peri-operative behavioral/pelvic floor therapy (BPTx) compared to MUS alone on successful treatment of MUI symptoms in 472 women. Secondary objectives include estimating the effect of combined treatment compared to MUS on improving OAB and SUI outcomes separately, need for additional treatment, time to failure and identifying predictors of poor outcomes in this MUI population.

3.1.1 Primary Aims

The primary aim of this study is to assess whether combined MUS and peri-operative BPTx is superior to MUS alone for improving MUI symptoms in women electing surgical treatment.

3.1.2 Secondary Aims

Secondary aims of this study include the following:

1. Assess whether combined MUS+BPTx is superior to MUS alone for improving change in OAB symptoms at 1 year in women electing surgical treatment.
2. Assess whether combined MUS+BPTx is superior to MUS alone for improving change in SUI symptoms at 1 year in women electing surgical treatment for MUI.

3.1.3 Exploratory Aims

Exploratory aims of this study include the following:

1. Assess whether combined MUS+BPTx is superior to MUS alone for improving the number of urgency and urge incontinence episodes on bladder diary at 1 year following MUS surgery.
2. Compare time-to-failure between MUS+BPTx versus MUS alone, where failure is defined as initiation of any additional treatment for lower urinary tract symptoms (SUI, UUI/OAB, or voiding dysfunction).
3. Develop models to identify predictors of change of MUI, OAB, and SUI outcomes measured using the UDI between baseline and 1 year post-treatment.
4. Compare quality of life outcomes, Patient Global Impression-Improvement (PGI-I), and Patient Global Impression-Severity (PGI-S) between groups
5. Describe rates of reoperation (sling revision) for worsening OAB symptoms after MUS and to compare the proportion of women in each group initiating additional treatment for SUI and/or OAB, and the types of additional treatment (BPTx, medications, other)
6. Determine MIDs and clinically meaningful definitions of MUI that predict clinical outcomes using cut-offs and combinations of standardized measures
7. Compare pelvic floor muscle strength changes between women randomized to combined MUS+BPTx versus MUS alone and to estimate associations between pelvic floor muscle strength improvement and UI symptoms. We will also explore predictors of unsuccessful pelvic floor muscle strengthening.
8. Determine the cost effectiveness of combined midurethral sling (MUS) and peri-operative behavioral/pelvic floor therapy (BPTx) compared to MUS alone for the treatment of MUI symptoms

3.2 Outcomes

3.2.1 Primary Outcomes

The primary outcome for this study is the mean change from baseline in UDI-total score at 1 year postoperative. The MID for the UDI-total score published by Dyer et al is estimated to be 35 points.[50]

3.2.2 Secondary Outcomes

The secondary outcomes for this study are the mean change from baseline in the UDI-irritative and -stress subscales. The MID for the UDI-irritative score published by Dyer et al is estimated to be 15 points[50] and the MID for the UDI-stress score published by Barber et al is estimated to be 8 points[51]

3.2.3 Exploratory Outcomes

Exploratory outcomes for this study include time-to-failure (defined as initiation of any additional treatment for lower urinary tract symptoms such as SUI, UUI/OAB, or voiding dysfunction), change from baseline quality of life outcomes at 1 year postoperative, global impression scales (PGI-I, PGI-S), change from baseline number of urgency and urge incontinence episodes at 1 year following MUS surgery (as determined by a bladder diary), reoperation rates for worsening OAB symptoms after MUS and additional treatments for SUI and/or OAB (both rates and types), and change from baseline measures of pelvic floor muscle strength at 1 yea post-operative (Peritron Perineometer readings).

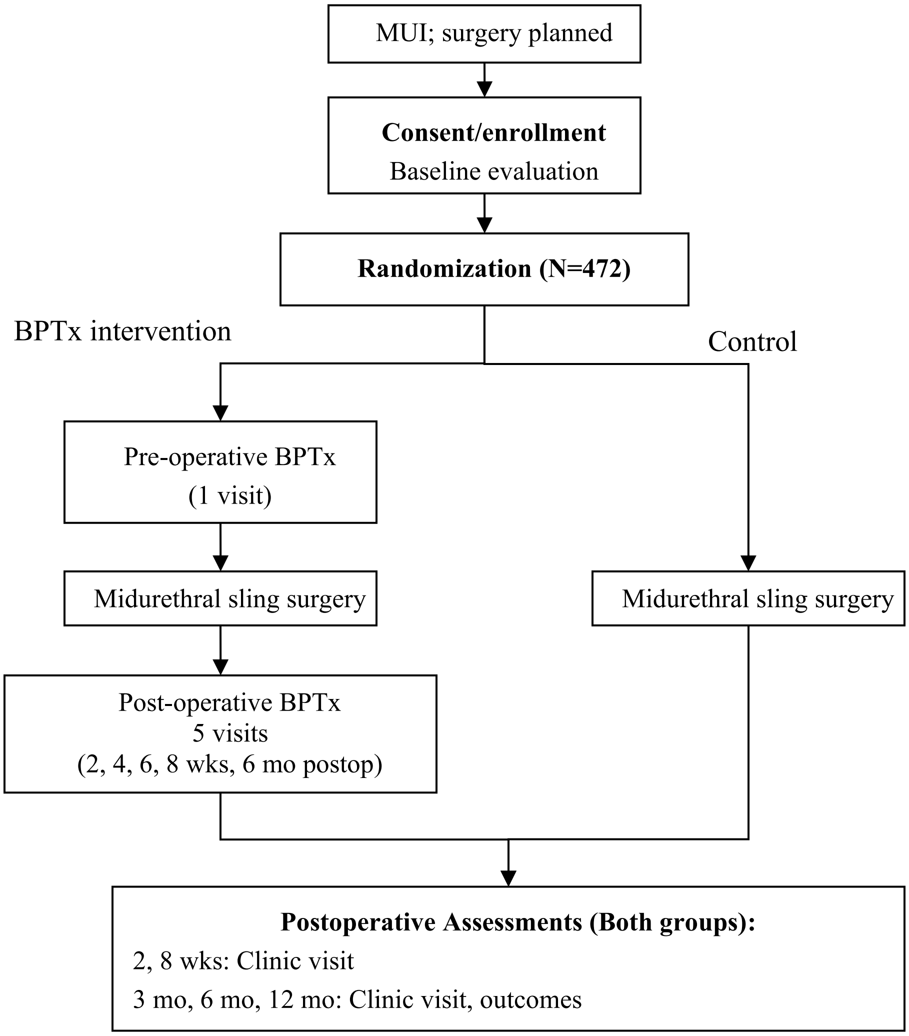
3.2.4 Safety Outcomes

Safety outcomes will be assessed in a descriptive manner at each DSMB meeting without formal statistical tests. This will include the need for sling revision due to worsening OAB symptoms. There is no established stopping rule to guide what sling revision rate is "appropriate" for worsening OAB symptoms in this population.

4 STUDY METHODS

4.1 Overall Study Design and Plan

The study is a multi-center, randomized, surgical trial of women with MUI who have elected to undergo surgical treatment for SUI. Participants will be randomized to a peri-operative BPTx program + MUS surgery versus MUS surgery alone. The purpose is to compare combined MUS+BPTx versus MUS alone (control) on improving MUI symptoms at 1 year. A study schematic is shown below:



4.2 Study Population

4.2.1 Subject Characteristics

Inclusion and exclusion criteria for the ESTEEM trial are as follows:

Inclusion Criteria:

1. Presence of both SUI and UUI on bladder diary; and > 2 IEs/3 days
   a. 1 Stress IE/3 days
   b. 1 Urge IE/3 days
2. Reporting at least "moderate bother" from UUI item on UDI: "Do you usually experience urine leakage associated with a feeling of urgency, that is a strong sensation of needing to go to the bathroom?"
3. Reporting at least "moderate bother" from SUI item on UDI: "Do you usually experience urine leakage related to coughing, sneezing, or laughing"
4. Diagnosis of SUI defined by a positive cough stress test (CST) or UDE within the past 18 months
5. Desires surgical treatment for SUI symptoms
6. Urinary symptoms >3 months
7. Subjects understand that BPTx is a treatment option for MUI outside of ESTEEM study protocol
8. Urodynamics within past 18 months

Exclusion Criteria:

1. Anterior or apical compartment prolapse at or beyond the hymen (>0 on POPQ), regardless if patient is symptomatic
   a. Women with anterior or apical prolapse above the hymen (<0) who do not report vaginal bulge symptoms will be eligible
2. Planned concomitant surgery for anterior vaginal wall or apical prolapse > 0
   a. Women undergoing only rectocele repair or another repair unrelated to anterior or apical compartment (i.e.: anal sphincter repair) are eligible
3. Women undergoing hysterectomy for any indication will be excluded
4. Active pelvic organ malignancy
5. Age <21 years
6. Pregnant or plans for future pregnancy in next 12 months, or within 12 months post-partum
7. Post-void residual >150 cc on 2 occasions within the past 6 months, or current catheter use
8. Participation in other trial that may influence results of this study
9. Unevaluated hematuria
10. Prior sling, synthetic mesh for prolapse, implanted nerve stimulator for incontinence
11. Spinal cord injury or advanced/severe neurologic conditions including Multiple Sclerosis and Parkinson's
12. Women on overactive bladder medication/therapy will be eligible after 3-week wash-out period
13. Non-ambulatory
14. History of serious adverse reaction to synthetic mesh
15. Not able to complete study assessments per clinician judgment, or not available for 12-month follow-up

16. Women who only report "other IE" on bladder diary, and do not report at minimum 1 stress and 1 urge IE/3 days
17. Diagnosis of and/or history of bladder pain or chronic pelvic pain
18. Women who had intravesical Botox injection within the past 12 months
19. Women who have undergone anterior or apical pelvic organ prolapse repair within the past 6 months

4.3 Study Arm Assignment and Randomization

Once patients are enrolled, surgery should be scheduled within 3 months from enrollment, and randomization should occur within 4-6 weeks prior to the booked surgical date. The participant will be randomized to one of the two treatment arms (MUS or MUS+BPTx) using a web-based randomization system. The system will supply the site coordinator with a randomization code.

Randomization (1:1 to the two treatment arms) will be performed using permuted blocks, with a block size that is known only to the DCC and will be stratified by site and severity of baseline UUI (with a cut-off of ≥4 urge incontinence episodes on 3-day diary).

4.4 Masking and Data Lock

4.4.1 General Masking Procedures

Study surgeons and outcome assessors will be masked to treatment assignment. All outcome measures will be collected by masked outcome assessors. Study coordinators / clinical staff performing objective measurement of PFM strength will be masked (Aim 7). All patient-reported outcomes (PROs) will be administered prior to other clinical assessments or procedures.

It is not feasible to mask the patients or interventionists to the BPTx intervention due to the nature of the treatment being studied.

Study individual	Masking
Study participant	No
Interventionist	No
Outcome assessors (includes clinical staff performing PFM measurement)	Yes
Study surgeon	Yes

Efforts will be made by unmasked research assistant/staff members to remind the patient that the surgeon is masked to her treatment assignment. If she desires additional treatment, it is likely the surgeon would offer BPTx as additional treatment and she will be reminded that she can decline additional BPTx without revealing to her surgeon that she received the BPTx intervention. Such methods have been effective for past PFDN trials (e.g. OPTIMAL trial[42]).

4.5 Database Lock

Database lock will occur when data collection has been completed prior to the final data analysis. Until it is agreed upon by the Steering Committee and the DCC that unmasking is not a risk to the study, only the DCC statistician(s) and data manager(s) working directly with the data will be unmasked to the treatment assignments of individual study participants. Thus, data collection of longer-term outcomes will not be compromised by unmasking.

4.6 Study Flow Chart of Assessments and Evaluations

	Baseline	**Random-ization visit (T0 -5 wks)**	**2 wk preop (range 1-4 wks preop)**	**Surg MUS (T0)**	**Call (3-5d post-op)**	**2 wk post-Clinic**	**4 & 6 wks post**	**8 wk post-**	**3 mo post-Clinic and QoL**	**6 mo post-Clinic and QoL**	**12 mo post-Clinic and QoL**
Window of time		± 7d	7-28d preop			± 7d	± 7d for each visit	± 7d	± 30d	± 30d	± 60d
Estimated duration of clinic and/or BPTx visit for each group		Both: 1.5-2hr	Control: N/A Interv: 1.5hr		Contr: N/A Interv: 15 min	Control: 1.5hr Interv: 2.5hr	Control: N/A Interv: 1hr	Control: 1hr Interv: 2hr	Both: 1.5hr	Control: 1.5hr Interv: 2.5 hr	Both: 1.5-2hr
All subjects											
Consent	X										
Coordinator visit	X	X				X		X	X	X	X
Masked clinical staff visit (for PFM measures)		X				X		X			X
Hx/PE (update)						X		X	X	X	X
Medication audit	X					X		X	X	X	X
UDE	X										
UDI (inclusion and primary outcome)	X								X	X	X
Other PRO questionnaires		X							X	X	X
Voiding diary	X*					X*	X	X*		X*	X*
PFM measures		X				X		X			X
Additional treatment**						X	X (both groups by phone)	X	X	X	X
Adverse events				X		X	X (both groups by phone)	X	X	X	X
Voiding function (PVR)	X					X					
Subjects randomized to intervention only											
BPTx visit			X			X	X	X		X	

BPTx self-efficacy questionnaire		X								X	X
BPTx Adherence / Barrier questionnaire						X	X	X		X	X

5 ANALYSIS POPULATIONS

5.1 Intention-to-Treat (ITT) Population

The primary analysis population and the population for all secondary analyses will be the intention to treat population, which includes all randomized and eligible subjects. Subjects that are determined to be ineligible after randomization will be excluded from the analysis. For the analysis, all subjects will be assigned to the arm to which they were randomized irrespective of treatment received. The analysis will include randomized and eligible patients who provided outcome data at any follow up time point from 3 to 12 months.

It is possible that some women in both groups may cancel their surgical MUS procedure due to personal or other reasons. It is also possible that women randomized to BPTx may cancel their surgical procedure if they receive preoperative BPTx treatment and experience improvement. These women will still be included from an ITT perspective.

5.2 Per-Protocol (PP) Population

Per-protocol analyses (in which participants are analyzed according to the treatment actually received and their degree of compliance with those treatments) will be considered exploratory secondary analyses. The per-protocol population will be defined in the following way:

1. Participants randomized to the MUS only treatment arm are considered adherent to the protocol if they undergo MUS surgery.
2. Participants randomized to the MUS plus behavioral therapy treatment arm will be considered adherent to the protocol if all the following are true:
    a. The participant undergoes MUS surgery.
    b. The participant attends at least 4 out of 5 of the behavioral therapy sessions between 2-weeks and 6-months post-surgery
    c. The participant reports performing exercises at least 4-5 times/week for weeks 2 and 6 or at least 2-3 times/week for week 8 at ≥ 75% of attended visits between 2 and 8 weeks post-surgery, inclusive
    d. The participant reports performing at least 30 contractions at ≥ 75% of attended visits between 2 and 8 weeks post-surgery, inclusive
    e. The participant reports no problems or, if participant reports problems with leakage, they also report using the "Knack" at ≥ 75% percent of attended visits between 2 weeks and 6 months post-surgery, inclusive
    f. The participant reports no problems or, if participant reports problems with urgency, they also report using the urgency suppression strategy at ≥ 75% percent of attended visits between 2 weeks and 6 months post-surgery, inclusive
    g. The participant reports no problems or, if participant reports problems with urgency or frequency, they also report using any other bladder training strategies at ≥ 75% percent of attended visits between 2 weeks and 6 months post-surgery, inclusive

5.3 Safety (SAF) Population

The safety population will comprise all subjects who were randomized, grouped by the treatment to which they were randomized.

6 SAMPLE SIZE DETERMINATION

This study is designed to compare the relative effectiveness of MUS and MUS combined with BPTx. Power and sample size calculations were generated to determine the sample size needed to test for treatment difference favoring MUS combined with BPTx across the study arms for a variety of assumptions about effective size and follow-up time. Sample size estimates are based on simulations using analysis methods accounting for both the rate of additional treatment in the two groups as well as UDI total score or subscore values over the 12 month follow up. We assumed that 30% of women in the MUS only group and 20% of women in the MUS+BPTx group would request additional treatment. In TOMUS, 10-12% of women who had baseline MUI had persistent UUI postoperatively based on MESA responses and/or initiation of anticholinergic treatment.[24] In Barber's TVT vs TOT equivalence trial, 70% reported baseline MUI and postoperatively, 30% of all women reported bothersome UUI with 16% of subjects on anticholinergic treatment postoperatively.[31] In Abdel-Fattah's transobturator MUS trial, 25% reported worsening OAB and almost all of these women were on anticholinergic treatment postoperatively.[34] In Palva's TVT vs TVT-O trial, 174 women reported preoperative UUI and of these, 7 women (4%) had tried anticholinergics postoperatively after 3 years. Therefore, based on existing MUS trials, the rate of additional treatment for OAB ranges from 4-25%, supporting our conservative assumption that 30% of women will request additional treatment in the MUS only group.

6.1 Primary Outcome

The MID for the UDI-total score published by Dyer et al. is estimated to be 35 points.[50] Assuming a two-sided alpha of .05, SD of 50.4, and true difference in mean change from baseline in UDI-total scores at 1 year between treatment groups of 35, 75 women per group would provide 90% power to detect a statistically significant difference between groups.

6.2 Secondary Outcomes

For the UDI-irritative subscale, the published MID estimate is 15 points.[50] Assuming a two-sided alpha of 0.05, SD of 25.6, and true difference in mean change from baseline in UDI-irritative scores at 1 year between treatment groups of 15, 92 women per group would provide 90% power. For the UDI-stress subscale, the published MID is 8 points.[51] Assuming a two-sided alpha of 0.05, SD of 21.5, and true difference in mean change from baseline in UDI-stress scores at 1 year between treatment groups of 8, 200 women per group would provide 90% power to detect a statistically significant difference between groups.

6.3 Final Sample Size

Using 200 per group as our base estimate and adjusting for 15% dropout post-operatively results in a total sample size of 472 randomized to treatment. Additionally, this sample size will provide approximately 90% power to detect a difference as small as 19 between treatment groups for the UDI-total score, and a difference as small as 16.5 points with 80% power.

7 STATISTICAL AND ANALYTICAL ISSUES

7.1 General Rules

All statistical computations will be performed and data summaries will be created using SAS 9.4 or higher. If additional statistical packages are required, these will be discussed in the study report. For summaries of study data, categorical measures will be summarized in tables listing the frequency and the percentage of subjects in each study arm; continuous data will be summarized by presenting mean, standard deviation, median, minimum, and maximum; and ordinal data will be summarized by only presenting median, minimum, and maximum.

7.2 Adjustments for Covariates

Indicator variables for the study stratification factors of site and severity of baseline UUI (with a cut-off of ≥4 urge incontinence episodes on 3-day diary) will be included as covariates in most efficacy analyses performed for this study (details in section 9). Additionally, demographic and baseline characteristics for subjects and clinicians will be compared between study arms using analysis of covariance techniques for continuous measures, Mantel-Haenszel mean score test using standardized midrank scores for ordinal measures, and Cochran Mantel-Haenszel chi-square tests for general association for categorical measures. If sample sizes allow, these analyses will control for the study stratification factors. If these analyses suggest that substantial differences exist among arms, the use as covariates of these parameters on which the arms differ will be explored in secondary exploratory analyses of the efficacy data.

7.3 Handling of Dropouts and Missing Data

Missing data mechanisms will be explored, and sensitivity analyses will be conducted on primary outcomes to assess the robustness of the described analyses. Methods employed for sensitivity analyses may include multiple imputation or inverse probability weighting methodology. Imputation is not planned for secondary analyses. For further details on sensitivity analyses involving missing data see Attachment 16.2.

7.4 Interim Analyses and Data Monitoring

Safety outcomes will be assessed at each DSMB meeting. This will include the need for sling revision due to worsening OAB symptoms. Rates of sling revision and other safety outcomes will be compared between treatment groups using Fisher's exact tests and provided to the DSMB. There is no established guidance regarding what sling revision rate is "appropriate" for worsening OAB symptoms in this population: this is one of the exploratory aims of this study.

Since we expect to enroll ESTEEM within 2 years, and since the primary outcome is attained at 12 months following surgery, we propose that no interim analyses of outcomes will be performed. Thus, reports to the DSMB will not include outcome data until primary outcomes have been attained for all participants. At each meeting, the DSMB will be presented with information about enrollment and outcome data attainment (for example, the percent of expected clinic visits that have been completed) to allow them to determine that the study is making reasonable progress.

7.5 Masked Data Review

A masked data review of the primary outcome and secondary outcomes for this study will be performed by the protocol team. This review will occur prior to data lock and completion of the 1-year analyses. This will include a presentation of descriptive statistics (e.g. means, standard deviations, percentiles for continuous variables and counts and percentages of categorical variables) of the selected outcomes and model predictor variables. In the masked review, all data will be aggregated over both treatment groups.

7.6 Multicenter Studies

For this multicenter study, randomization of study participants was stratified within center and by severity of baseline UUI (with a cut-off of ≥4 urge incontinence episodes on 3-day diary). Consequently, for all model-based primary and secondary analyses, center and UUI severity group will be included as fixed effects in the models.

7.7 Multiple Comparisons and Multiplicity

The primary hypothesis and two secondary hypotheses will be tested at a nominal two-sided type I error of 0.05. All p-values for any baseline and demographic characteristics, secondary outcomes, and safety parameters will be for descriptive purposes only.

7.8 Examination of Subgroups

No subgroup analyses are planned.

7.9 Assessment Windows

Baseline assessments were to be completed no longer than 3 months prior to surgery with assessments repeated if participant surgery is delayed for over 3 months. All other visits were completed at 3-month intervals with a ± 1 month window around the visit. Coordinators attempted to complete all follow-up visits, even if they couldn't be completed within window. For the primary analysis, decisions about how to treat out-of-window visits will be made prior to unmasking data. For secondary analyses, all available data will be used.

8 STUDY SUBJECT CHARACTERIZATION

8.1 Subject Disposition

Participant eligibility status will be summarized and listed by study arm and overall disposition of study participants will be described using a standard cohort diagram. The number of subjects randomized; completing or discontinuing from study therapy; completing each follow-up visit will be summarized by study arm. Reasons for study treatment discontinuation and study withdrawal will be listed.

8.2 Protocol Deviations

Protocol deviations are identified via automated checks of the clinical database and reported by site study coordinators in the study data management system. Protocol deviations will be listed by site with information such as type of deviation, time of occurrence, and reason. Incidence rate of protocol deviations will also be summarized overall and for each protocol deviation category by site. Incidence rate of protocol deviations will be calculated as: number of deviations divided by the number of subject months at the site

8.3 Demographic and Baseline Characteristics

Demographic and baseline clinical characteristics for the study participants will be summarized by study arm using the general analysis rules describe above. Variables of interest include: age (years), parity, gravidity, race and ethnicity, BMI, marital status, education level (classified as binary variable as having some college or greater or no college education), health insurance status (private only, Medicare/Medicaid only, combination of both), smoking status (never, previous, current), menopausal status, prior prolapse surgery, estrogen use, prior pelvic floor therapy or treatments, prior use of OAB medications, and baseline levels of all QOL measures.

9 EFFICACY ANALYSES

9.1 Overview of Efficacy Analyses Methods

- All efficacy analyses will be performed on the ITT population.
- All efficacy variables will be summarized by treatment group at baseline and at the 3, 6, and 12-month time points. N, mean, standard deviation, minimum, and maximum will summarize continuous efficacy variables, whereas number and percent will summarize categorical efficacy variables.
- Unless otherwise noted, all analyses of dichotomous outcomes, measured at respective endpoints, will be performed using a generalized linear mixed model. Models will be adjusted for stratification by clinical site and UUI severity group. If there are not enough patients in every clinical site to include the variable in the models as a fixed effect,

clinical site will be included as random intercepts to account for correlation between outcomes of patients treated by the same clinical site. Consistent with the description of the primary analysis in the protocol, analyses of primary and secondary outcomes will also include independent variables for time and request for additional treatment. If differences between treatment groups are to be assessed at multiple time points (for example, 6 and 12 months), then longitudinal modeling will be used and the interaction between time and treatment groups will be included. Unless otherwise noted, all analyses of continuous efficacy variables (e.g., QOL scales) will be performed using general linear mixed models. Variables with distributions substantially different from normal will be transformed prior to analysis. Models will be adjusted for clinical site and UUI severity group. If there are not enough patients per clinical site to include the variables in the models as fixed effects, clinical site will be included as a random intercept to account for correlation between outcomes of patients treated by the same clinical site. If differences between treatment groups are to be assessed at multiple time points (for example, 6 and 12 months), then longitudinal modeling will be used and the interaction between time and treatment groups will be included. Under the assumption that any missing outcome data will be missing at random (thus, missing UDI scores at 12 months may be related to both 3 and 6-month outcomes and covariates), this model will produce more accurate estimates in the presence of missing data than one that models only outcomes at 12 months. For the primary outcome and for secondary outcomes, models will include 2- and 3-way interactions between treatment assignments, time, and request for additional treatment.

9.2 Efficacy Variables

Primary and secondary efficacy variables as well as exploratory and safety outcomes are described in the table below.

Variable	**Type**	**Definition**
Primary Outcomes		
Change from baseline in UDI Total Score at 12 months	Continuous	The Urogenital Distress Inventory-Total Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, the instructions are to sum the Stress, Irritative, and Obstructive subscales of the UDI (see below). If any subscale scores are missing, then no Total score will be calculated. The outcome will then be computed as the difference in Total score at 12 months (and 3 and 6 months) and the Total score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Secondary Outcomes		
Change from baseline in UDI Stress Score at 12 months	Continuous	The Urogenital Distress Inventory-Stress Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, the instructions are to average responses across questions D and F and their respective sub-questions (0=No or Yes and Not at all; 1=Yes and Slightly; 2=Yes and Moderately; 3=Yes and Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in UDI Irritative Score at 12 months	Continuous	The Urogenital Distress Inventory-Irritative Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, the instructions are to average responses across questions A, B, C, G, H, and I and their respective sub-questions (0=No or Yes and Not at all; 1=Yes and Slightly; 2=Yes and Moderately; 3=Yes and Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses.

Variable	**Type**	**Definition**
		The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in UDI Obstructive Score at 12 months	Continuous	The Urogenital Distress Inventory-Obstructive Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, the instructions are to average responses across questions E, J, K, L, M, N, O, P, Q, R, and S and their respective sub-questions (0=No or Yes and Not at all; 1=Yes and Slightly; 2=Yes and Moderately; 3=Yes and Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Exploratory Outcomes		
Change from baseline in Urge Incontinence Episodes at 12 months	Continuous	The daily frequency of urge incontinence episodes will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of urge incontinence episodes at 12 months (and 6 months) and the daily frequency of urge incontinence episodes at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in Stress Incontinence Episodes at 12 months	Continuous	The daily frequency of stress incontinence episodes will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of stress incontinence episodes at 12 months (and 6 months) and the daily frequency of stress incontinence episodes at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in Total Incontinence Episodes at 12 months	Continuous	The daily frequency of total incontinence episodes will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of total incontinence episodes at 12 months (and 6 months) and the daily frequency of total

Variable	**Type**	**Definition**
		incontinence episodes at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in Urgency Episodes at 12 months	Continuous	The daily frequency of urgency episodes will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of urgency episodes at 12 months (and 6 months) and the daily frequency of urgency episodes at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in Nocturia Episodes at 12 months	Continuous	The daily frequency of nighttime voids will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of nighttime voids at 12 months (and 6 months) and the daily frequency of nighttime voids at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in Daytime voids at 12 months	Continuous	The daily frequency of daytime voids will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of daytime voids at 12 months (and 6 months) and the daily frequency of daytime voids at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in total voids at 12 months	Continuous	The daily frequency of total voids will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of total voids at 12 months (and 6 months) and the daily frequency of total voids at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in total voids without incontinence at 12 months	Continuous	The number of voids without incontinence will be taken from the bladder diary. The outcome will then be computed as the difference in number of voids without incontinence at 12 months (and 6 months) and the number of voids without incontinence at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.

Variable	**Type**	**Definition**
Change from baseline in Number of Pads per Day at 12 months	Continuous	The daily average number of pads will be taken from the bladder diary. The outcome will then be computed as the difference in daily average number of pads at 12 months (and 6 months) and the daily average number of pads at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from baseline in Number of Wet Pads per Day at 12 months	Continuous	The daily average number of wet pads will be taken from the bladder diary. The outcome will then be computed as the difference in daily average number of wet pads at 12 months (and 6 months) and the daily average number of wet pads at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Normalization of Voiding Frequency at 12 months	Dichotomous (Yes/No)	The total number of daytime and nighttime voids will be taken from the bladder diary and summed and divided by the number of diary days (<=3) to get a voiding frequency. The indicator for normalization at 12 months (and 6 months) will be defined for each subject as follows: If > 8 voids/24 hours at baseline: If > 8 voids/24 hours at time x: No If ≤ 8 voids/24 hours at time x: Yes If ≤ 8 voids/24 hours at baseline: Missing If data for the assessment time point are missing, the normalization indicator will be coded as missing. The variable is summarized as the percentage of participants with indicator equal to “Yes” among all non-missing indicators.
Improvement in Voiding Frequency at 12 months	Dichotomous (Yes/No)	The total number of daytime and nighttime voids will be taken from the bladder diary and summed and divided by the number of diary days (<=3) to get a voiding frequency. If there has been at least a 50% reduction in the voiding frequency between baseline and the time point, the improvement indicator at 12 months (and 6 months) will be set to “Yes”, otherwise it will be set to “No”. If data for the assessment time point are missing, the improvement indicator will be coded as missing. The variable is summarized as the percentage of participants with indicator equal to “Yes” among all non-missing indicators.

Variable	**Type**	**Definition**
High Voiding Frequency at 12 months	Dichotomous (Yes/No)	The total number of daytime and nighttime voids will be taken from the bladder diary and summed and divided by the number of diary days (<=3) to get a voiding frequency. The indicator for high frequency at 12 months (and 6 months) will be defined for each subject as follows: If > 8 voids/24 hours at time x: Yes If ≤ 8 voids/24 hours at time x: No If data for the assessment time point are missing, the high frequency indicator will be coded as missing. The variable is summarized as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.
Dryness at 12 Months (Diary)	Dichotomous (Yes/No)	The total number of incontinence episodes will be taken from the bladder diary and summed and divided by the number of diary days (<=3) to get an average daily number of episodes. The indicator for dryness at 12 months will be defined for each subject as follows: If 0 episodes/24 hours at time x: Yes If > 0 episodes/24 hours at time x: No If data for the assessment time point are missing, the dryness indicator will be coded as missing. The variable is summarized as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.
Dryness at 12 Months (UDI-LF)	Dichotomous (Yes/No)	The response to questions C (Do you experience urine leakage related to the feeling of urgency?) and D (Do you experience urine leakage related to physical activity, coughing, or sneezing?) will be taken from the UDI-LF. The indicator for dryness at 12 months will be defined for each subject as follows: If question C = No and question D = No at time x: Yes If question C = Yes or question D = Yes at time x: No If data for the assessment time point are missing, the dryness indicator will be coded as missing. The variable is summarized as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.

Variable	**Type**	**Definition**
Dryness at 12 Months (UDI-LF and Diary)	Dichotomous (Yes/No)	The responses to Dryness at 12 Months based on the UDI-LF and the diary will be combined. The indicator for dryness at 12 months will be defined for each subject as follows: If Dry (Diary) = Yes and Dry (UDI-LF) = Yes at time x: Yes If Dry (Diary) = No or Dry (UDI-LF) = No at time x: No If data for the assessment time point are missing, the dryness indicator will be coded as missing. The variable is summarized as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.
Change from Baseline OAB-q Symptom Severity Score at 12 Months	Continuous	The OAB-q Symptom Severity Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 1 through 8 (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=A great deal, 6=A very great deal), then subtract the minimum possible score (8) and divide by the range of possible score (48-8=40), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline OAB-q HRQL Coping Score at 12 Months	Continuous	The OAB-q HRQL Coping Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 9, 11, 16, 21, 22, 26, 32, and 33 (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=A great deal, 6=A very great deal), then subtract from the maximum possible score (48) and divide by the range of possible score (48-8=40), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.

Variable	**Type**	**Definition**
Change from Baseline OAB-q HRQL Concern Score at 12 Months	Continuous	The OAB-q HRQL Concern Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 12, 13, 14, 19, 23, 25, and 29 (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=A great deal, 6=A very great deal), then subtract from the maximum possible score (35) and divide by the range of possible score (42-7=35), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline OAB-q HRQL Sleep Score at 12 Months	Continuous	The OAB-q HRQL Sleep Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 10, 15, 17, 24, and 30 (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=A great deal, 6=A very great deal), then subtract from the maximum possible score (30) and divide by the range of possible score (30-5=25), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline OAB-q HRQL Social Score at 12 Months	Continuous	The OAB-q HRQL Social Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 18, 20, 27, 28, and 31 (1= Not at all, 2=A little bit, 3=Somewhat, 4=Quite a bit, 5=A great deal, 6=A very great deal), then subtract from the maximum possible score (30) and divide by the range of possible score (30-5=25), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.

Variable	**Type**	**Definition**
Change from Baseline OAB-q HRQL Total Score at 12 Months	Continuous	The OAB-q HRQL Total Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum the OAB-q HRQL Coping, Concern, Sleep, and Social prior to subtracting from the respective maximum possible scores, then subtract from the maximum possible sum score (150) and divide by the range of possible scores ((18+42+30+10=150)-(8+7+5+5=25)=125), then multiply by 100. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If any of the subscales are missing, the Total score will also be set to missing. If data for an assessment time point are missing, the outcome variable will be coded as missing.
OAB-SAT-q Satisfaction Score at 12 Months	Continuous	The OAB-SAT-q Satisfaction Score will be computed at 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 1, 2, and 3 (1=Extremely Dissatisfied, …, 6=Extremely Satisfied), then subtract from the maximum possible sum score (18) and divide by the range of possible scores (18-3=15), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. If data for an assessment time point are missing, the outcome variable will be coded as missing.
OAB-SAT-q Side Effects Score at 12 Months	Continuous	The OAB-SAT-q Side Effects Score will be computed at 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, if the response to question 5 is Never then the score is 100, regardless of the responses to questions 6 and 7. If the response to question 5 is not Never, then sum the responses across questions 5 (reverse coded: 1=All of the time, …, 5=A little of the time), 6 (1=Extremely bothersome, …, 5=Not at all bothersome), and 7 (1=A great deal, …, 7=Not at all), then subtract from the maximum possible sum score (15) and divide by the range of possible scores (15-3=12), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. If data for an assessment time point are missing, the outcome variable will be coded as missing.

Variable	Type	Definition
OAB-SAT-q Endorsement Score at 12 Months	Continuous	The OAB-SAT-q Endorsement Score will be computed at 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 9 (1=Definitely would not use the same treatment again, …, 4=Definitely would use the same treatment again), 10 (1=Definitely would not recommend, …, 4=Definitely would recommend), and 11 (1=Extremely Dissatisfied, …, 6=Extremely Satisfied), then subtract from the maximum possible sum score (14) and divide by the range of possible scores (14-3=11), then multiply by 100. If less than half of the responses are missing, then the score will be calculated by substituting the missing responses with the average of the non-missing responses, otherwise no score is calculated. If data for an assessment time point are missing, the outcome variable will be coded as missing.
OAB-SAT-q Convenience Score at 12 Months	Continuous	The OAB-SAT-q Convenience Score will be computed at 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to use the response to question 4 (1=Extremely Inconvenient, …, 6=Extremely Convenient), then subtract from the maximum possible sum score (6) and divide by the range of possible scores (6-1=1), then multiply by 100. If the response to question 4 is missing then no score is calculated. If data for an assessment time point are missing, the outcome variable will be coded as missing.
OAB-SAT-q Preference Indicator at 12 Months	Dichotomous	The OAB-SAT-q Preference Score will be computed at 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, the indicator is set to 1 if the response to question 8 is either “Slight preference for the treatment I am receiving now” or “Definitely prefer the treatment I am receiving now”, otherwise it is set to 0. If the response to question 8 is missing then indicator is calculated. If data for an assessment time point are missing, the outcome variable will be coded as missing. The score is calculated as the percentage of participants with Indicator=1 among all non-missing Indicators.

Variable	**Type**	**Definition**
Retreatment at 12 months	Categorical and Dichotomous	The types of retreatment for SUI or UUI will be identified and categorized from the Additional Therapy and Concomitant Medications forms. Further details can be found in Attachment 16.1. An indicator of retreatment will be calculated for each subject where subjects with retreatment identified above will be assigned "Yes", otherwise subjects are assigned "No". The rate of retreatment is calculated as the percentage of participants with indicator="Yes" among all non-missing indicators.
Time-to-Failure	Semi-continuous number of days/years until failure	Failure = The first time at which a subject seeks any additional treatment for SUI or UUI/OAB symptoms as determined by the Additional Therapy and Concomitant Medications forms. Further details can be found in Attachment 16.1. Success = Any subject for whom no additional treatment for SUI or UUI/OAB symptoms is identified, censored at 12 months if the subject completed the entire follow-up period. Subjects lost to follow-up will be censored at the time of their last visit (scheduled or unscheduled).
Sling Revision for Worsened OAB Symptoms at 12 months	Dichotomous	An indicator of sling revision will be calculated for each subject where those who have a sling revision for worsening OAB listed on the Additional Therapy form will be assigned a "Yes", otherwise subjects are assigned "No". The rate of sling revision is calculated as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.
Change from Baseline IIQ Physical Activity Score at 12 Months	Continuous	The Incontinence Impact Questionnaire-Physical Activity Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to average responses across questions A, B, C, D, E, and U (0=Not at all, 1=Slightly, 2=Moderately, 3=Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses. The outcome will then be computed as the difference in score at 12 months

Variable	Type	Definition
		(and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline IIQ Travel Score at 12 Months	Continuous	The Incontinence Impact Questionnaire-Travel Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to average responses across questions F, G, H, I, J, and M (0=Not at all, 1=Slightly, 2=Moderately, 3=Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline IIQ Social Relationships Score at 12 Months	Continuous	The Incontinence Impact Questionnaire-Social Relationships Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to average responses across questions K, L, N, O, P, Q, R, S, W, and X (0=Not at all, 1=Slightly, 2=Moderately, 3=Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline IIQ Emotional Health Score at 12 Months	Continuous	The Incontinence Impact Questionnaire-Emotional Health Score will be computed at baseline, 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to average responses across questions T, V, Y, Z, AA, BB, CC, and DD, (0=Not at all, 1=Slightly, 2=Moderately, 3=Greatly), and then multiply by 100/3. If any responses are missing, then the score will be calculated as the average of the non-missing responses. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.

Variable	Type	Definition
Change from Baseline IIQ Total Score at 12 Months	Continuous	The Incontinence Impact Questionnaire-Total Score will be computed at baseline and 12 months (and 3 and 6 months) using the standard scoring algorithm. Specifically, instructions are to sum the Physical Activity, Travel, Social Relationships, and Emotional Health subscale scores of the IIQ. If any subscale scores are missing, then no Total score will be calculated. The outcome will then be computed as the difference in Total score at 12 months (and 3 and 6 months) and the Total score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline EQ-5D Total Score at 12 Months	Continuous	The EQ-5D Total Score will be computed at baseline and 12 months (and 3 and 6 months) using the algorithm obtained from https://archive.ahrq.gov/professionals/clinicians-providers/resources/rice/EQ5Dscore.html. The outcome will then be computed as the difference in Total score at 12 months (and 3 and 6 months) and the Total score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline EQ-5D Visual Analogue Scale (VAS) at 12 Months	Continuous	The EQ-5D VAS is the response to a single question from the EQ-5D. The outcome will then be computed as the difference in EQ-5D VAS at 12 months (and 3 and 6 months) and the EQ-5D VAS at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline PISQ-IR not sexually active – partner related subscale score at 12 months	Continuous	The PISQ-IR not sexually active – partner related subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q2a, Q2b (1=strongly agree, …, 4=strongly disagree). If there is more than 1 missing response then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are not sexually active.

Variable	**Type**	**Definition**
Change from Baseline PISQ-IR not sexually active – condition specific subscale score at 12 months	Continuous	The PISQ-IR not sexually active – condition specific subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q2c, Q2d, Q2e (1=strongly agree, …, 4=strongly disagree). If there is more than 1 missing response then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are not sexually active.
Change from Baseline PISQ-IR not sexually active – global quality subscale score at 12 months	Continuous	The PISQ-IR not sexually active – global quality subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q4a, Q4b, Q5a, and Q6 using reverse scores for all but Q5a (Q4a and Q4b are Likert scales of 1 to 5; Q5a: 1=strongly agree, …, 4=strongly disagree; Q6: 1=not at all, …, 4=a lot). If there are more than 2 missing responses then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are not sexually active.
Change from Baseline PISQ-IR not sexually active – condition impact subscale score 12 months	Continuous	The PISQ-IR not sexually active – condition subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q3, Q5b, Q5c using reverse scores for Q3 (Q3: 1=not at all, …, 4=a lot; Q5b, Q5c: 1=strongly agree, …, 4=strongly disagree). If there is more than 1 missing response then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are not sexually active.

Variable	**Type**	**Definition**
Change from Baseline PISQ-IR sexually active – arousal, orgasm subscale score change from baseline at 12 months	Continuous	The PISQ-IR sexually active – arousal, orgasm subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q7, Q8a, Q10, and Q11 using reverse scores for Q11 (Q7, Q8a, Q11: 1=never, …, 5=[almost] always; Q10: 1=much less intense, …, 5=much more intense; check box response to Q1=1). If there are more than 2 missing responses then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are sexually active.
Change from Baseline PISQ-IR sexually active – condition specific subscale score at 12 months	Continuous	The PISQ-IR sexually active – condition specific subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q8b, Q8c, Q9 using reverse scores for all (Q8b, Q8c, Q9: 1=never, …, 5=[almost] always). If there is more than 1 missing response then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are sexually active.
Change from Baseline PISQ-IR sexually active – partner related subscale score at 12 months	Continuous	The PISQ-IR sexually active – partner related subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q13, Q14a, Q14b using reverse scores for Q14a and Q14b (Q13: 1=all of the time, …, 4=hardly ever/rarely; Q14a, Q14b: 1=very positive, …, 4=very negative). If there is more than 1 missing response then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be

Variable	Type	Definition
		coded as missing. Scores should only be calculated for participants that are sexually active and have a sexual partner.
Change from Baseline PISQ-IR sexually active – desire subscale score at 12 months	Continuous	The PISQ-IR sexually active – desire subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q15, Q16, Q17 using reverse scores for Q16 and Q17 (Q15: 1=never, …, 5=always; Q16: 1=daily, …,5=never; Q17: 1=very high, …, 5=very low or none at all). If there is more than 1 missing response then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are sexually active.
Change from Baseline PISQ-IR sexually active – condition impact subscale score at 12 months	Continuous	The PISQ-IR sexually active – condition impact subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q18, Q20b-d using reverse scores for Q18 (Q18: 1=not at all, …, 4=a lot; Q20b-d: 1=strongly agree, …, 4=strongly disagree). If there are more than 2 missing responses then a total score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are sexually active.
Change from Baseline PISQ-IR sexually active – global quality rating subscale score at 12 months	Continuous	The PISQ-IR sexually active – global quality rating subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q19a-Q19c, Q20a using reverse scores for Q19a-Q19c (Q19a-c: 1=satisfied, …, 5=dissatisfied; Q20a: 1=strongly agree, …, 4=strongly disagree). If there are more than 2 missing responses then a total

Variable	**Type**	**Definition**
		score is not calculated. To handle missing values, the final score is obtained by dividing the sum by the number of items answered. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants that are sexually active.
Change from Baseline Adaptation Index Hygiene subscale as 12 months	Continuous	The AI Hygiene subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to average the responses for questions Q1-Q4, Q9, and Q14 (0=never, 25=rarely, 50=sometimes, 75=often, 100=always). If there are at least 5 non-missing responses, then the score is the average of the non-missing responses, otherwise, the score is not calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline.
Change from Baseline Fecal Incontinence Adaptation Index Avoidance subscale as 12 months	Continuous	The AI Hygiene subscale score will be computed at baseline and 12 months (and 3 and 6 months) using standard scoring algorithms. Specifically, instructions are to average the responses for questions Q5-Q8, Q10-Q13, and Q15-17 (0=never, 25=rarely, 50=sometimes, 75=often, 100=always). If there are at least 9 non-missing responses, then the score is the average of the non-missing responses, otherwise, the score is not calculated. The outcome will then be computed as the difference in score at 12 months (and 3 and 6 months) and the score at baseline.
Patient Global Impression of Improvement at 12 months	Dichotomous	For defining improvement, Yes="much better" or "very much better" on the PGI-I scale. The outcome will be computed at 12 months (and 3 and 6 months)
Patient Global Impression of Severity at 12 months	Categorical	For defining severity, Normal/mild="Normal" or "Mild" on the PGI-S scale. The outcome will be computed at 12 months (baseline, 3 months, and 6 months)

Variable	**Type**	**Definition**
Change from Baseline Brink Score at 12 months	Continuous	The Brink Score will be calculated from the pelvic floor muscle assessment form. It is calculated by summing the Brink pressure, duration, and displacement metrics if the assessment could be performed (i.e. not unable to perform due to pain). The outcome will then be computed as the difference in Brink Score at 12 months (and 6 months) and the Brink Score at baseline. If data for the assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline Maximum Pelvic Floor Muscle Contraction Pressure (Maximum Amplitude) at 12 months	Continuous	The maximum amplitude from the Peritron device will be averaged across the valid squeeze measures (out of 3 possible). The outcome will then be computed as the difference in average maximum amplitude at 12 months (and 8 weeks) and the average maximum amplitude at baseline.

9.3 Primary Analysis Methods

The mean change from baseline in UDI scores will be compared between groups at 1 year. As explained previously, participants will be permitted to seek additional treatment for SUI and/or OAB after 3 months following MUS. Because such treatment is expected to impact the participant's UDI score at 1 year, we will use an analysis method that accounts for the impact of additional treatment. Specifically, a general linear mixed model will be constructed to model change from baseline in UDI scores using scores recorded at time points up to 1 year following MUS. For participants who request additional treatment, only UDI measurements up to the time of additional treatment will be included in the model, and measurements taken between additional treatment and 1 year will be considered missing for the purpose of the primary analysis. The model will include fixed effects for treatment group, time, request for additional treatment, and interactions between those variables. It will also be adjusted for the design effects of stratification by center and by baseline urge IE group. Thus, the models will allow for different trajectories of change for women who are or are not randomized to BPTx and for those who do or do not request additional treatment. A statistical test based on the model will be conducted to assess whether mean changes from baseline in UDI scores at 1 year are significantly different between the two treatment groups, accounting for the percent of women in each group who request additional treatment. Sensitivity analysis will be conducted to test the robustness of test results to model specifications.

We will report whether change in total UDI score between baseline and one year is significantly different in the two groups. If the difference is statistically significant, the potential clinical significance of the difference will be discussed. We recognize that our sample size would allow us to find a difference between groups that is statistically significant yet smaller than published MIDs for total UDI score for women with MUI. However, published MIDs were calculated based on populations that may be somewhat different from the one targeted for enrollment in ESTEEM, and a secondary aim of ESTEEM is to explore whether the true MID in this population differs from previously published values.

9.4 Secondary Analysis Methods

The mean change from baseline in UDI-irritative and UDI-stress scores at 1 year will be compared between groups using the same analysis methods described for the primary analysis. If the difference is statistically significant, the potential clinical significance of the difference will be discussed. Additional analyses will be conducted to determine whether the MIDs in this MUI population differ from previously published MIDs.

9.5 Other Analysis Methods

9.5.1 Other UUI/OAB Outcomes

Bladder diary

We will compare change in number of urge IEs and urgency-episodes and nocturia episodes between groups from baseline to 6 and 12 months. Of note, not all four symptoms of OAB (frequency, urgency, nocturia, and UUI) are required to be present at baseline for eligibility into this trial (only UUI required). Changes from baseline in bladder diary outcomes will be calculated and analyzed using the methods described for the analysis of the primary outcome.

For urinary frequency, women reporting on average >8 voids/24 hours at baseline will be considered symptomatic, and normalization of voiding frequency will be defined as < 8 voids/24 hours at 1 year. A 50% improvement will be defined as a reduction by half in the number of voids that patients had at baseline. The number of women who had normalization of voiding frequency and 50% improvement will be compared between groups separately and collectively. We will also assess the proportion of women who had worsening of urinary frequency (includes women who developed de novo frequency and those who worsened). These dichotomous outcomes will be analyzed using logistic regression, controlling for the design effects of stratification by center and by baseline urge IE group. To assess the impact of additional treatment prior to 1 year, sensitivity analyses will be conducted in which women who request additional treatment will be assigned the less-favorable outcome.

OAB-SAT-q and OAB-q

For these scales and associated subscales, differences from baseline will be calculated for the OAB-q, and methods described for analysis of the primary outcome will be used to test for differences between treatment groups at 12 months. For the OAB-SAT-q, differences in post-treatment scores will be compared between groups.

The scoring manual for the OAB-SAT-q states that higher scores should correspond to better outcomes and also details that final composite scores for each dimension should be obtained by subtracting raw scores from the highest possible score and dividing by range. However, based on the direction of score for individual responses, this derivation would result in lower scores corresponding with better outcomes. We have adjusting the scoring algorithm instead to be: 100*(raw score – lowest possible score)/score range. This calculation results in higher scores corresponding to better outcomes and an overall scale of 0-100.

Additionally, it was observed that the scoring of the OAB-SAT-q Endorsement subscale could have unintended behavior when dealing with missing data. The range of possible responses for component questions 9 and 10 (1 through 4) is different from the range of possible responses for component question 11 (1 through 6), so if any question is missing, the range of possible values for the imputed value of the missing response does not match the range of the original question. For example, if Q9=4, Q10=Missing, and Q11=6, then the imputed value for Q10 is 5, which is outside of the 1-4 range possible for Q10. This further causes problems when transforming the score as the transformed score can exceed 100. From the example provided, the transformed score is 100*(sum(4,5,6) - 3)/11 = 109.09.

We have adjusted the missing value handling instead to be: average the non-missing responses as proportions of their maximum possible values (divide each non-missing response by its maximum possible value) and then scaling that average against the maximum possible value of the missing response. For example, if Q9=4, Q10=Missing, and Q11=6, then the imputed value for question 10 is as follows [(4/4 + 6/6)/2] * 4 = 4 and the transformed Endorsement score is 100*(sum(4,4,6) – 3)/11 = 100, so the score does not exceed 100.

9.5.2 Time-to-Failure

Although our primary outcome is at 12 months, the team was interested in whether perioperative BPTx may be associated with a delayed time to failure compared to Control. In other words, is BPTx associated with a significant effect, but the effect is not sustained at the 12-month time point? For example, if BPTx could delay the need for anti-muscarinics for up to 9 months, this would be relevant information for counseling women and perhaps clinically recommending perioperative BPTx. As described previously, failure will be defined as initiation of any additional treatment for either SUI or UUI/OAB symptoms.

A class of survival model which can account for interval censoring (outcomes measured at pre-planned time points as opposed to continuously over time) will be used to determine if combined MUS+BPTx is associated with a decrease time to failure compared to MUS alone between 3-12 months. Depending on the distribution of the observed data, an accelerated failure time frailty model or a Bayesian survival model may be used. The model will be adjusted for the design effects of stratification by center and by baseline urge IE group.

9.5.3 Predictors of Treatment Success and Failure

Regression models will be created to identify predictors of change from baseline to 1 year for UDI total score and stress and irritative subscale scores. Participants who request additional treatment prior to 1 year will not be included in the predictive models. Potential predictors will include age, diary parameters such as number of UUI episodes/3 days, functional bladder capacity, bother severity at baseline. The relationship between potential predictors and outcomes will be explored in models that include one predictor plus stratification factors (center and baseline urge IE group). Predictive models will be constructed using backward selection of predictors. The impact of collinearity between predictors will be assessed and the final model modified as necessary.

9.5.4 Quality of Life/Global Impression

For these scales and associated subscales, differences from baseline will be calculated and methods described for analysis of the primary outcome will be used to test for differences between treatment groups from baseline and 6 and 12 months.

9.5.5 Safety and Initiation of Additional Treatment

We will describe rates of sling revision due to worsening OAB symptoms and rates of additional treatment.

9.5.6 Determine MIDs and Clinically-Meaningful MUI Definitions

We will explore potential MIDs for UDI total score and stress and irritative subscores for this MUI population. MIDs will be calculated using anchor- and distribution-based approaches. Potential anchors include global impression of change, incontinence episodes from the bladder diary, and request for additional treatment. The change from baseline UDI total score and stress and irritative subscores for each ESTEEM participant will be compared to the MIDs estimated from the ESTEEM data, and the percentage of participants who meet or exceed the MID will be compared between treatment groups..

We will attempt to create threshold definitions, based on baseline measures of the UDI, IIQ, OAB-q, UDE, and baseline bladder diary parameters in isolation and in combination, that are predictive of clinical success at 1 year. Definitions of success will be based on a change from baseline in total UDI score, UDI-irritative score or UDI-stress score at least as large as the MID for this MUI population.

9.5.7 Compare Pelvic Floor Muscle Strength

As mentioned above, all women will undergo PFM strength measurements using the Peritron device by masked coordinators at baseline, postoperative at 2 weeks, 8 weeks (end of intervention), and 12 months (primary endpoint). The difference in the maximum pelvic floor muscle contraction pressure (maximum amplitude) will be compared between the BPTx and the control groups. A table of comparative studies using the Peritron device to measure PFM strength changes with PFM therapy is provided in Table 12 of the protocol.

Based on the existing comparative studies using the Peritron, continent women have a maximum amplitude PFM contraction between 36-45 cm H2O. Incontinent women have significantly lower maximum contractions, ranging from 15.5 to 26.5 cm H2O, with most studies showing a maximum contraction of 25 cm H2O. In these studies, incontinent women can improve their maximum contraction pressure up to 34-41 cm H2O with PFM training, which is comparable to continent women. In addition, these studies report women experience significant improvement in UI symptoms, although there is limited information on the direct specific relationship between PFM strength changes and UI symptom changes.

Assuming that women in ESTEEM will have a mean baseline PFM maximum contraction amplitude of 25 cm H2O, and that women randomized to control will not demonstrate significant improvement postoperatively (no change from mean maximum amplitude of 25 cm H2O (SD 13), and that women randomized to BPTx will demonstrate improvement to 35 (SD 13) to 40 (SD 16) cm H2O at 6-12 months, the power to detect a difference between the groups with the current ESTEEM sample size of 400 women would be greater than 0.99. Also, the difference from 25 (SD 13) that we could detect with 80% power is 3.66 cm H2O between groups and with 90% power we could detect a difference as small as 4.23 cm H2O.

For analyses, we will compare the mean change from baseline in PFM maximum contraction strength between the BPTx and control groups at 8 weeks and at 12 months. General linear mixed modeling will be used, controlling for stratification factors and time (8 weeks and 12 months). We will test whether there is significant interaction between treatment group and time. Because additional treatment is not expected to impact this outcome, it will be ignored for the purpose of this analysis. We will estimate the correlation between PFM strength and UI symptoms at baseline and at 12 months. Using regression models, we will also explore potential predictors of unsuccessful pelvic floor muscle strengthening and urge suppression and their effects on urinary outcomes. We will assess the effect of self-efficacy, adherence, and barriers to performing pelvic muscle contractions and behavioral therapy.

10 SAFETY ANALYSES

10.1 Overview of Safety Analysis Methods

All safety analyses will be performed using all participants who were randomized. Descriptive p-values comparing the study arms will be provided on most safety table summaries and will be obtained using chi-square tests or Fisher’s exact tests (in the case of small numbers of events) for binary outcomes. If the number of events allow, a 2-sided Cochran Mantel-Haenszel test controlling for strata defined by site and baseline urge IE group will be used to obtain the p-values.

10.2 Adverse Events

Per the protocol, participants were asked to report any adverse events from initiation of treatment through 12 months follow-up. All adverse events were collected on an adverse event log and coded using MedDRA (V17.0).

Postoperative complications noted as adverse symptoms and events in the ESTEEM manual of procedures (MOP) (see table below) will be proactively reviewed and reported. ESTEEM complications are divided into "adverse symptoms" and "adverse events". To help standardized the data collection process for complications, adverse symptoms and events will be captured as follows:

Adverse symptoms will be captured using either an active approach using validated questionnaires or an open-ended approach.

a. Active capture for adverse symptoms: Specific patient-reported adverse symptoms will be identified by responses on validated questionnaires (denoted by * in table below). These 3 adverse symptoms include a patient report of new or worsening: 1) pelvic/vaginal pain, 2) sensation of incomplete bladder emptying (based on UDI responses) and 3) dyspareunia (based on PISQ-IR response). These are captured at 3, 6, and 12 months. To aid with retrospective collection of these AEs, the DCC will provide each site with a report listing longitudinal responses to the applicable UDI and PISQ-IR questions for participants who reported one of these symptoms at 3-12 months. The clinical site will report these adverse symptoms on the adverse events CRF per the usual process described in the MOP.

   The clinical site will be responsible for reviewing the UDI and PISQ-IR responses for these 3 items and identifying any new or worsening symptoms. Once a new or worsening symptom is identified, the clinical site will track these adverse symptoms per the usual process in the MOP. Sites should review the UDI and PISQ-IR CRFs at the time the participant completes them so that any additional information needed to complete the adverse event CRF can also be collected at that time.

b. Open-ended capture for adverse symptoms: 2 adverse symptoms will be captured using the open-ended approach because they are not captured in any of the validated questionnaires administered in ESTEEM. These include: 1) New/worsening partner dyspareunia and 2) New/worsening constipation. Using the open-ended approach, symptoms are captured if reported by the participant.

Adverse events will be captured using the standard capture for AEs as described in the MOP and manually reviewed to determine which events meet the criteria for post-operative complication listed in the table below. This is an open-ended capture approach using chart review, physician report, or patient report corroborated by medical documentation.

Postoperative complications	
Adverse symptom	**Definition**

New/worsening abdominal/genital pain at or beyond 3 month visit	UDI item N captured at 3, 6, 12 months and per patient report, chart review, physician report
New/worsening dyspareunia at or beyond 3 month visit	PISQ item C5 captured at 3, 6, 12 months and per patient report, chart review, physician report
New/worsening report of difficulty emptying bladder at or beyond 3 months	UDI item J captured at 3, 6, 12 months and per patient report, chart review, physician report
New/worsening partner dyspareunia at or beyond 3 month visit	Per patient report, chart review, physician report
New/worsening urge incontinence at or beyond 3 month visit	UDI item C captured at 3, 6, 12 months and per patient report, chart review, physician report
Adverse Event	
Postop need for catheter and/or ISC at or beyond 2 weeks	Captured by chart review, physician report, exam, or patient report corroborated by medical documentation
Vaginal mesh exposure	Captured by chart review, physician report, exam, or patient report corroborated by medical documentation
Vaginal mesh erosion into organ	Captured by chart review, physician report, exam or patient report corroborated by medical documentation
Other wound healing problems >6 weeks	Captured by chart review, physician report, exam or patient report corroborated by medical documentation
New/worsening vaginal infection	Captured by chart review, physician report, exam, or patient report corroborated by medical documentation
Urinary tract infection beyond 2 weeks	UTI based on clinical judgment or confirmation of culture proven, also includes empiric antibiotic treatment for symptoms thought to be due to UTI, or patient report corroborated by medical documentation. If a culture is sent that turns out to be negative, this would not be classified as a UTI. Captured by chart review, physician report, or patient report corroborated by medical documentation
Other infection possibly related to intervention (sling and/or BPTx)	Infection diagnosed using clinical or radiologic indicators – not including vaginal infection, UTI Captured by chart review, physician report, or patient report corroborated by medical documentation
Non-study health care provider visit due to complication related to intervention (sling and/or BPTx)	Captured by chart review, physician report, or patient report corroborated by medical documentation

Emergency room visit due to complication related to intervention (sling and/or BPTx)	Captured by chart review, physician report, or patient report corroborated by medical documentation
Hospital readmission related to intervention (sling and/or BPTx)	Captured by chart review, physician report, or patient report corroborated by medical documentation
Reoperation related to surgery (any return to OR for issue related to intervention (e.g. recurrent SUI, UUI/OAB)	Captured by chart review, physician report, or patient report corroborated by medical documentation
An adverse event that is “unexpected” and “possibly, probably or definitely related”	Captured by chart review, physician report, or patient report corroborated by medical documentation

AEs will be listed and summarized by system organ class and preferred event term. Summaries will be of the number of events and number of individuals experiencing events by treatment group and will be created for all AEs, and AEs by severity. Any events starting outside of the reportable timeframe will be included in separate listings and will be excluded from summary tables. If a complete onset date is unknown and it cannot be confirmed that the event occurred during this time period, then the event will be considered a treatment-emergent AE.

Adverse events will be individually reviewed to determine if they meet the conditions specified for each post-operative complication. The review will be conducted via keyword searches through reported event terms and comment for specific events (e.g. vaginal mesh exposure, ER visits, etc.) and by preferred term and system/organ class for more general complications (e.g. new/worsening vaginal infection, other infections, etc.). Worsening symptoms will be assessed by identifying worsening responses (compare to baseline) to the specific patient-questionnaire items identified at post-baseline collection times. Post-operative complications will be summarized by complication type. Summaries will be the number of individuals experiencing complications by treatment group.

10.3 Deaths and Serious Adverse Events

A serious adverse event (SAE) is any event that is life threatening, results in death, causes or prolongs hospitalization, leads to a disability or birth defect, or requires an intervention to prevent a disability. SAEs will be listed and SAEs and treatment-related SAEs will be summarized in the manner mentioned in Section 10.2 if there are enough events to summarize. Deaths will be listed.

11 ANALYSIS OF OTHER OUTCOMES

No analyses of outcomes other than efficacy and safety outcomes are planned.

12 REPORTING CONVENTIONS

Unless required otherwise by a journal, the following rules are standard:

- Moment statistics including mean and standard deviation will be reported at 1 more significant digit than the precision of the data.
- Order statistics including median, min and max will be reported to the same level of precision as the original observations. If any values are calculated out to have more significant digits, then the value should be rounded so that it is the same level of precision as the original data.

- Following SAS rules, the median will be reported as the average of the two middle numbers if the dataset contains even numbers.
- Test statistics including t and z test statistics will be reported to two decimal places.
- P-value will be reported to 3 decimal places if > 0.001. If it is less than 0.001 then report '<0.001'. Report p-values as 0.05 rather than .05.
- No preliminary rounding should be performed, rounding should only occur after analysis. To round, consider digit to right of last significant digit: if < 5 round down, if >=5 round up.

13 CHANGES TO THE ANALYSIS PLANNED IN THE PROTOCOL

At the request of the DSMB, a sensitivity analysis will be conducted to assess the impact of including vs. excluding data from a clinical site at which potential data quality concerns were raised during conduct of the trial.

At the request of the writing team, additional endpoints describing the dryness status of participants based on the UDI and bladder diary responses will be added. For further details on how these diary indicators will be calculated, see Section 9.2.

Due to the relatively small number of participants who met failure (additional treatment) criteria, Kaplan Meier methods were used instead of the more complex survival modeling described in the protocol to compare MUS+BPTx vs. MUS alone.

Due to the relatively small number of women who met the improvement in voiding frequency definition (at least a 50% reduction in voiding frequency), the difference in improvement in voiding frequency between treatment groups could not be analyzed as described in Section 9.4 nor could the analysis that examined women who met both normalization and improvement in voiding frequency be performed.

14 REFERENCES

To be completed

15 LIST OF POTENTIAL DISPLAYS

Data displays may be added, deleted, rearranged or the structure may be modified after finalization of the SAP. Such changes require no amendment to the SAP as long as the change does not contradict the text of the SAP.

Tables
Participant Eligibility
Participant Disposition
Protocol Deviations
Demographic and Baseline Characteristics
Primary Efficacy Model Results
Secondary Efficacy Outcome Measures

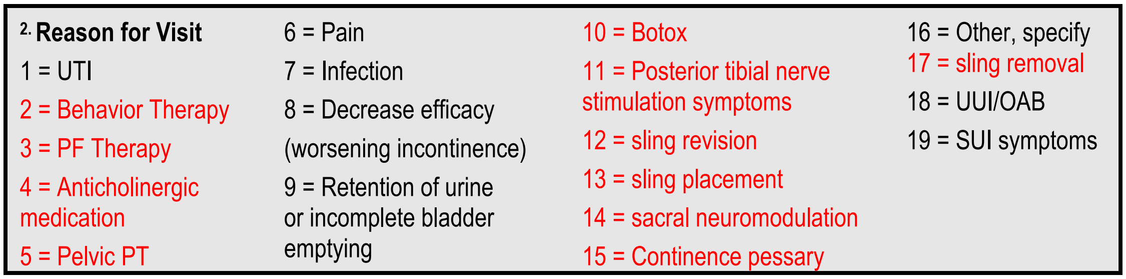
Safety Summary Adverse Events
Figures
Consort diagram of participant disposition Initiation of Efficacy Kaplan Meier Curves (per outcome) Duration of Treatment Success Kaplan Meier Curves
Data Listings
Subject Eligibility Subject Disposition Protocol Deviations Adverse Events Serious Adverse Events

16 ATTACHMENTS

16.1 Defining Treatment Failures and Retreatment Rates

Treatment failure is defined as initiation of any additional treatment for either SUI or UUI/OAB symptoms. These will be determined through an analysis of the Additional Therapy Form (CRF #31) and Concomitant Medications Form (CRF #28).

a. Additional Therapy Form – Any subject with a reported reason for visit of 2, 3, 4, 5, 10, 11, 12, 13, 14, 15, or 17 will be considered to have treatment failure at the time of the reported treatment. Similarly, each subject will be considered to have had retreatment.



2. **Reason for Visit**
1 = UTI
2 = Behavior Therapy
3 = PF Therapy
4 = Anticholinergic medication
5 = Pelvic PT
6 = Pain
7 = Infection
8 = Decrease efficacy (worsening incontinence)
9 = Retention of urine or incomplete bladder emptying
10 = Botox
11 = Posterior tibial nerve stimulation symptoms
12 = sling revision
13 = sling placement
14 = sacral neuromodulation
15 = Continence pessary
16 = Other, specify
17 = sling removal
18 = UUI/OAB
19 = SUI symptoms

b. Concomitant Medications – Any subject with medications or therapies that are known treatments of OAB, UUI, or SUI (Detrol, Ditropan, Vesicare, Hyoscyamine, Oxybutynin) or for whom UUI/OAB, SUI Therapy, or Voiding Dysfunction indicators are marked will be considered to have treatment failure at the start date of the medication if it is after randomization. Similarly, each subject will be considered to have had retreatment.

Medication Name	Start Date *(DD/MMM/YYYY)*	Stop Date *(DD/MMM/YYYY)*	PRN	UUI/OAB Therapy	SUI Therapy	Voiding dysfunction	Continuing at End of Study
	— —/— — — —/— — — — —	— —/— — — —/— — — — —	□	□	□	□	□

In general, any adverse events that could be considered treatment failures will be requested to be removed and re-entered on the additional therapies form if applicable.

16.2 Missing Data and Multiple Imputation

Sensitivity analysis to see the impact of missing data on the primary outcome (UDI Total score) and the dryness and MID indicators will be performed using multiple imputation. Missing data at 3, 6, and 12 months will be imputed using non-missing data from prior time points (including baseline) to inform the imputed values as well as age, treatment, race, ethnicity, BMI, smoking status, estrogen use (oral/patch), site, incontinence episode group (from randomization), vaginal delivery count, and private insurance indicator. Missingness will first be assured to monotone missing (i.e. ensure not missing prior time points when later time points are present) 25 time using MCMC methods, then all other missing data will be imputed for each of the 25 monotone-missing datasets using regression and logistic imputation methods. Each imputed dataset will be fit using the models described in Section 9.3. Differences between groups for each imputed dataset will then be combined to estimate average effects and standard errors and compared to the model without imputation. All imputation will be performed using SAS v9.4.