STATISTICAL ANALYSIS PLAN

Treatment for Mixed Urinary Incontinence: Mid-urethral Sling vs. Botox A (MUSA)

Short title: Sling vs Botox for Mixed Incontinence

SAP VERSION: Version 1.0

SAP DATE: February 8, 2024

PREVIOUS VERSIONS: N/A

SPONSOR: NICHD

PREPARED BY: RTI International.
3040 Cornwallis Rd
Research Triangle Park, NC 27709-2104

AUTHOR (S): Lindsey Barden
Evan Rhodes
Ben Carper
Sonia Thomas

Contents

1 BACKGROUND AND PROTOCOL HISTORY 6
2 PURPOSE OF THE ANALYSES 6
3 STUDY AIMS AND OUTCOMES 6
3.1 STUDY AIMS 6
3.1.1 Primary Aims 7
3.1.2 Exploratory Aims 7
3.2 OUTCOMES 7
3.2.1 Primary Outcomes 7
3.2.2 Secondary Outcomes 7
3.2.3 Exploratory Outcomes 8
3.2.4 Safety Outcomes 8
4 STUDY METHODS 8
4.1 OVERALL STUDY DESIGN AND PLAN 8
4.2 STUDY POPULATION 9
4.3 STUDY ARM ASSIGNMENT AND RANDOMIZATION 10
4.4 CRITERIA FOR BOTOX RE-INJECTIONS 10
4.5 MASKING AND DATA LOCK 11
4.5.1 General Masking Procedures 11
4.5.2 Database Lock 11
4.6 STUDY FLOW CHART OF ASSESSMENTS AND EVALUATIONS 12
5 ANALYSIS POPULATIONS 14
5.1 MODIFIED INTENTION-TO-TREAT (MITT) POPULATION 14
5.2 PER-PROTOCOL (PP) POPULATION AND ANALYSIS 14
5.3 SAFETY (SAF) POPULATION 14
6 SAMPLE SIZE DETERMINATION 14
7 STATISTICAL AND ANALYTICAL ISSUES 15
7.1 GENERAL RULES 15
7.2 ADJUSTMENTS FOR COVARIATES 15
7.3 HANDLING OF DROPOUTS AND MISSING DATA 16
7.4 HANDLING OF CROSS-OVERS AND OFF-STUDY TREATMENT 16
7.5 INTERIM ANALYSES AND DATA MONITORING 16
7.6 MASKED DATA REVIEW 17
7.7 MULTICENTER STUDIES 17
7.8 MULTIPLE COMPARISONS AND MULTIPLICITY 17
7.9 EXAMINATION OF SUBGROUPS 17

7.10 ASSESSMENT WINDOWS 17
8 STUDY SUBJECT CHARACTERIZATION 17
8.1 SUBJECT DISPOSITION 17
8.2 PROTOCOL DEVIATIONS 17
8.3 DEMOGRAPHIC AND BASELINE CHARACTERISTICS 18
9 EFFICACY ANALYSES 18
9.1 OVERVIEW OF EFFICACY ANALYSES METHODS 18
9.2 VARIABLE DEFINITIONS 19
9.3 ANALYSIS METHODS FOR THE PRIMARY OUTCOME 48
9.4 ANALYSIS METHODS FOR SECONDARY AND EXPLORATORY OUTCOMES 49
9.5 OTHER ANALYSIS METHODS 49
9.5.1 Predictors of Treatment Response 49
9.5.2 Cost-effectiveness analysis 50
9.5.3 UDI MID 51
10 SAFETY ANALYSES 52
10.1 OVERVIEW OF SAFETY ANALYSIS METHODS 52
10.2 ADVERSE EVENTS/COMPLICATIONS/ADVERSE SYMPTOMS 52
10.3 DEATHS AND SERIOUS ADVERSE EVENTS 54
11 ANALYSIS OF OTHER OUTCOMES 54
12 REPORTING CONVENTIONS 54
13 CHANGES TO THE ANALYSIS PLANNED IN THE PROTOCOL 55
14 REFERENCES 55
15 LIST OF POTENTIAL DISPLAYS 55

LIST OF ABBREVIATIONS

BD	Bladder diary
BPTx	Behavioral/pelvic floor therapy
CISC	Clean intermittent self-catheterization
DCC	Data Coordinating Center
DO	Detrusor overactivity
DSMB	Data and Safety Monitoring Board
ESTEEM	Effects of Surgical Treatment Enhanced with Exercise for Mixed Urinary Incontinence trial
EQ-5D	European Quality of Life-5 Dimensions
IE	Incontinence episode
IIQ	Incontinence Impact Questionnaire
ITT	Intention-to-treat
MID	Minimum important difference
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
PFDN	Pelvic Floor Disorders Network
PGI-I	Patient Global Impression- Improvement
PGI-S	Patient Global Impression-Severity
PGSC	Patient Global Symptom Control
PISQ-IR	Pelvic Organ Prolapse/Urinary Incontinence Sexual Questionnaire
POPQ	Pelvic Organ Prolapse Quantification system
PRO	Patient reported outcomes
PVR	Postvoid residual
QoL	Quality of life
RCT	Randomized controlled trial
ROSETTA	Refractory Overactive Bladder: Sacral Neuromodulation v. Botulinum Toxin Assessment trial
SAE	Serious adverse event
SD	Standard deviation
SF-36	Short Form 36
SF-6D	Short Form 6D
SUI	Stress urinary incontinence

UDE	Urodynamic evaluation
UDI	Urogenital Distress Inventory
UI	Urinary incontinence
UIE	Urinary incontinence episode
UUI	Urgency urinary incontinence

1 BACKGROUND AND PROTOCOL HISTORY

Mixed urinary incontinence (MUI), defined as the presence of both stress urinary incontinence and urgency urinary incontinence, is a challenging condition for which clinicians frequently use multiple sequential treatments that have undergone limited head to head comparison in rigorous clinical trials.

Mid-**u**rethral **S**ling vs. Botox **A** (MUSA) is a randomized 2-arm clinical trial for women who have at least moderate bother from both stress and urgency incontinence and who have failed one or more conservative treatments. The primary objective is to estimate the effect of 100 units of intradetrusor injections of Botulinum toxin A (Botox A ®) compared to mid-urethral sling surgery for the treatment of MUI. Participants in the Botox A arm may receive one additional injection of 100 units Botox A between 3 months and 6 months after the initial injection if they have persistent urgency incontinence symptoms and meet the safety criteria. MUSA will randomize 146 women to treatment with Botox or Sling in a 1:1 ratio.

MUSA was originally designed as a 6-month trial and began enrolling participants in June 2020. Protocol amendment 4.0 approved by the DSMB in February 2021 added an observational 6-month follow-up period, with phone assessments at 9 and 12 months. At the time of transitioning to protocol version 4.0, 2 participants had already completed the study at 6 months and did not give consent to participate in the follow-up period.

At the completion of this study, we will better understand whether a surgical treatment that focuses on the urgency component (Botox A) is superior to a surgical treatment that focuses on the stress component (mid-urethral sling). The trial will provide clinically useful information for two treatments that are widely used to treat MUI but for which evidence-based data are not available.

2 PURPOSE OF THE ANALYSES

This statistical analysis plan (SAP) contains detailed information about statistical analyses to be performed to address the primary aim and selected exploratory aims of MUSA, including both the 6-month treatment period and the observational period to 12 months. 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.

Since the 6-month observational period was added to the protocol as an addendum, all design and analysis components for the observational period are described in the protocol addendum. In this SAP, the follow-up observational period design, aims, outcomes and analyses are fully integrated with the treatment period within each section.

Detailed methods for exploratory aim 2 (predictors of outcome) and exploratory aim 5 (cost effectiveness) will be described in separate analysis plans, as they will be covered in separate secondary manuscripts.

3 STUDY AIMS AND OUTCOMES

3.1 Study Aims

The purpose of MUSA is to compare treatment with either intradetrusor injections of Botulinum toxin A (Botox A ®) or mid-urethral sling for women with MUI and characterize patient characteristics associated with treatment response.

3.1.1 Primary Aims

The primary aim of this study is to compare the effectiveness of intradetrusor injection of 100 units of Botox A to mid-urethral sling for change in MUI symptoms 6 months following treatment.

3.1.2 Exploratory Aims

Exploratory aims of this study (consolidated from the main protocol and the 12-month addendum) are the following:

1. **12-month aim:** To compare the effectiveness of intradetrusor injection of 100 units of Botox A to mid-urethral sling for change in MUI symptoms 12 months following treatment
2. **Secondary urinary outcomes:** To compare treatment with Botox A to treatment with mid-urethral sling for improving the number of urinary incontinence episodes on bladder diary 6 and 12 months post-treatment.
3. **Predictors of poor treatment response:** To develop models to identify baseline predictors of change of MUI, OAB, and SUI outcomes measured using the UDI, between baseline and 6 and 12 months post-treatment.
4. **Quality of life and global impression:** To compare quality of life outcomes and Patient Global Impression-Improvement (PGI-I), Patient Global Impression-Severity (PGI-S) between groups randomized to Botox A versus mid-urethral sling 6 and 12 months post-treatment.
5. **Safety and additional treatments:** To describe rates of reoperation (sling revision) after mid-urethral sling and intermittent catheterization due to voiding dysfunction/partial urinary retention after Botox A detrusor injection, to compare the proportion of women in each group with UTI and recurrent UTI, rates of other serious and non-serious adverse events, and to compare the proportion of women in each group initiating additional (off protocol) treatment other than Botox A and mid-urethral sling for SUI and/or OAB.
6. **Cost-effectiveness analysis:** To determine the cost effectiveness of Botox A injection versus mid-urethral sling for the treatment of MUI symptoms on an intent-to-treat basis 6 and 12 months post-treatment.
7. **UDI MID:** To explore MIDs for UDI total score and stress and irritative subscores for this MUI population.

3.2 Outcomes

3.2.1 Primary Outcomes

The primary outcome for this study is the change from baseline in UDI-total score at 6 months post-treatment.

3.2.2 Secondary Outcomes

The three secondary outcomes for this study are:

- the change from baseline in the UDI-stress and UDI-irritative subscales at 6 months post-treatment and
- change from baseline in UDI-total score at 3 months post-treatment.

3.2.3 Exploratory Outcomes

Exploratory efficacy outcomes for this study include:

- Change from baseline in UDI-total score, UDI-stress and UDI-irritative subscales at other collected timepoints (3, 9, 12 months)
- Change from baseline in number of urinary incontinence episodes on bladder diary at 3-, 6-, and 12-months post-treatment (not measured at 9 months),
- Change from baseline in overactive bladder symptoms and satisfaction scores and subscales (OAB-q, OAB-SAT-q) at 3-, 6-, 9- and 12-months post-treatment,
- Change from baseline in quality of life outcomes (IIQ, PISQ-IR) and global impression scales (PGI-I, PGI-S, PGSC) at 3, 6, 9, and 12 months post-treatment,
- Additional treatment for SUI and/or OAB (both rates and types) within 6 months of treatment and during the 6-month follow-up,
- a Cost-effectiveness analysis to compare the direct & indirect costs of urinary incontinence treatment by Botox A and mid-urethral sling. Effectiveness outcomes are: SF-6D, and EQ-5D

3.2.4 Safety Outcomes

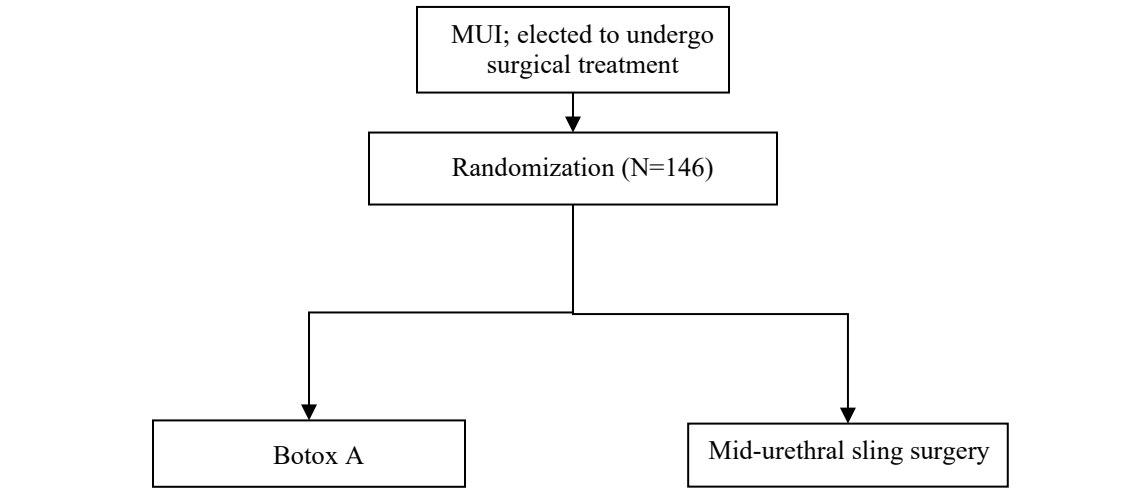
- Rates of reoperation or sling revision due to worsening OAB after MUS
- Rates of urinary retention / intermittent catheterization after Botox injection,
- Proportion of women in each group with UTI and recurrent UTI,
- Rates of other serious and non-serious adverse events,

Safety outcomes will be assessed in a descriptive manner at each DSMB meeting without formal statistical tests. 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 trial of women with MUI who have elected to undergo surgical treatment for MUI. Participants will be randomized to Botox A versus mid-urethral sling. At 6 months, the effect of treatment with Botox A or mid-urethral sling will be evaluated within a classic RCT model. Participants are followed for an additional 6-month observational period. The analysis will determine the effect of treatment on the primary outcome, change in Urogenital Distress Inventory (UDI) score at 6 months. A study schematic is shown below:



4.2 Study Population

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

Inclusion Criteria:

1. Reporting at least "moderate bother" from UUI item on UDI
    a. "Do you experience urine leakage associated with a feeling of urgency?"
2. Reporting at least "moderate bother" from SUI item on UDI
    a. "Do you experience urine leakage related to physical activity, coughing, or sneezing"
3. Diagnosis of SUI defined by a positive cough stress test (CST) or UDE within the past 18 months
    If participant does not demonstrate SUI during CMG they must demonstrate SUI through a cough stress test or other comparable valsalva maneuver to be eligible.
4. Presence of UUI on bladder diary with $\geq$ 4 Urgency IE/3-day diary
5. Urinary symptoms $\geq$3 months
6. Persistent symptoms despite at least one or more conservative treatments (e.g. supervised behavioral therapy, physical therapy) as determined adequate by the physician.
7. Inadequate response to oral overactive bladder medications (including anti-cholinergic and/or beta-mimetic medication) unless patient is
    a. intolerant of oral overactive bladder medications, or
    b. oral overactive bladder medications are contraindicated as determined by the treating provider
8. Urodynamics within past 18 months prior to enrollment or done after enrollment, prior to randomization.
9. Demonstrates understanding (or have caregiver demonstrate understanding) to perform clean intermittent self-catheterization.
    Provider and patient review CISC process and patient (and/or caregiver) demonstrates understanding to the satisfaction of the provider

Exclusion Criteria:

1. Anterior or apical compartment prolapse at or beyond the hymen ($\geq$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 other repair unrelated to anterior or apical compartment 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 6 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 urinary incontinence
    Women with known Burch or Marshall-Marchetti-Krantz (MMK) are excluded.
11. Spinal cord injury or advanced/severe neurologic conditions including Multiple Sclerosis, Parkinsons, Myasthenia Gravis, Charcot-Marie-Tooth
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 6 month follow-up
16. Diagnosis of and/or history of bladder pain or chronic pelvic pain
17. Women who had intravesical Botox injection within the past 12 months
18. Women who have undergone anterior or apical pelvic organ prolapse repair within the past 6 months

4.3 Study Arm Assignment and Randomization

The participant will be randomized to one of the two treatment arms (Botox A or MUS) using a web-based data management system. 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 age group (≥ 65, < 65). Once patients are enrolled and randomized, the intervention (either Botox A or mid-urethral sling) should be scheduled within 8 weeks from randomization. If treatment is not completed within 91 days of the baseline UDI, the baseline assessments are to be repeated.

4.4 Criteria for Botox re-injections

Participants in the Botox arm with persistent bothersome UUI at 3 months after their initial 100 unit Botox A injection may receive one additional injection of 100 units Botox A between 3 months and 6 months after initial injection. Eligibility for repeat Botox A injection are:

1. Persistent bothersome UUI as determined by reporting at least "moderate bother" from UUI item on UDI: "Do you experience urine leakage associated with a feeling of urgency, that is a strong sensation of needing to go to the bathroom?"
2. Continued UI bother based on the Patient Global Symptom Control (PGSC): "My current treatment is giving me adequate control of my urinary leakage" score ≤ 3
3. Participant desires additional treatment with Botox A
4. No UTI as determined by the health care provider on the day of the PGSC evaluation
5. PVR < 200 ml, per Botox A guidelines for UUI
6. No medical contraindication for the procedure as determined by the physician
7. Continuing to meet study inclusion/exclusion criteria (except inclusion items 2, 3, 4 7 and 8 and exclusion item 7).

4.5 Masking and Data Lock

4.5.1 General Masking Procedures

It is not feasible to mask the patients or surgeons to the intervention due to the nature of the interventions. While a placebo for Botox A is possible through saline injection during mid-urethral sling surgery, sham sling surgery would require suprapubic and vaginal incisions that would expose patients to the risks of sedation/anesthesia for the Botox A group.

Outcome assessors will be masked to treatment assignment. All post-randomization outcome measures will be assessed by masked outcome assessors. All patient-reported outcomes (PROs) will be administered prior to other clinical assessments or procedures.

Study Individual	**Masking**
Study participant	No
Study surgeon	No
Outcome assessors	Yes

The senior statistician who reviewed and approved the SAP (SMT) also remained fully masked until database lock, and all study PIs were masked to cumulative results.

4.5.2 Database Lock

Database lock and unmasking will occur when data collection has been completed and all data queries are resolved, and prior to the final data analysis. Prior to database lock a masked data review with the study PI (HH) will evaluate key data including study populations, protocol violations and completeness of key efficacy and safety data.

4.6 Study Flow Chart of Assessments and Evaluations

MUSA Study Visit	Screen/ Consent	Baseline/ Eligibility	Repeat Baseline*	Randomization	Operation / Procedure	2 wk	3 mo	Botox-Reinj (Botox only)	Botox Reinj 2wk FU	6 mo	9 mo	12 mo	Close-out	Add'nl /As nec.
MUSA Data Collection Forms	Durat. 1-4 w	Duration 1-4 w	Prior Tx >91 d after UDI	w/in 8 wks pre Tx	Time 0	14 ± 7 d	91 ± 30 d		14 ± 7 d	183 ± 30 d	274 ± 30 d	365 ± 30 d		
Consent	X													
§ UDI	X		X				X			X	X	X		
Bladder Diary		X	X				X			X		X		
Demographics		X												
Medical History / Medications		X												
Eligibility		X												
† Urodynamics		X												
† POPQ		X												
† PVR – Clinic Visit		X				X	X	X	X	X				
† Urine Dip		X	X		X	X	X	X	X	X				
Randomization				X										
† Pregnancy Test					X			X						
† Surgeon report/ Complications					X			X						
Medical – follow up						X	X		X	X	X	X		

§ OAB-q		X	X				X			X	X	X		
§ IIQ		X	X				X			X	X	X		
§ PISQ		X	X				X			X	X	X		
§ EQ-5D		X	X			X	X		X	X	X	X		
§ SF-36/ SF-6D		X	X			X	X		X	X	X	X		
§ PGI-S		X	X				X			X	X	X		
§ Lost Productivity / Costs						X	X		X	X	X	X		
§ OAB-SAT-q							X			X	X	X		
§ PGI-I							X			X	X	X		
§ PGSC							X			X	X	X		
Final Status													X	
Adverse Events														X
Serious AE														X
Protocol Viol. /Deviation														X
CISC/PVR – Patient Reported														X
Additional Therapies														X
Additional Office Visits														X

*Repeat baseline if not randomized and treatment not initiated within 90 days of the date UDI completed as part of the initial baseline visit.

† Performed in clinic only. MUS procedure urine dip collected per usual clinical practice.

§ Patient Reported Outcomes

Note; extended follow-up begins after the 6m visit.

5 ANALYSIS POPULATIONS

5.1 Modified Intention-to-Treat (mITT) Population

It is possible that some women in both groups may cancel their surgical procedure due to personal or other reasons. It is also possible that a treated participant might discontinue from the study after treatment while providing no efficacy data to be included in the analysis. The mITT population is defined as all randomized subjects, regardless of their final study status, that received any study treatment, and have some post-baseline efficacy data.

Unless otherwise specified, the mITT population will be used to analyze all efficacy outcomes. Subjects will be summarized according to the arm to which they were randomized, irrespective of the received treatment.

5.2 Per-Protocol (PP) Population and Analysis

The per-protocol population will be used in supportive sensitivity analyses of the primary and 3 secondary outcomes.

The Per-Protocol population is defined as a subset of the mITT population who received the intervention to which they were randomized, and had no major protocol violations, defined as any major inclusion or exclusion violations.

Receipt of any additional treatment for urinary incontinence before the 6-month visit assessments, including cross-over to the non-randomized study treatment or any other non-study UI treatment (other than a second Botox injection allowed by the protocol for the Botox arm) is also a major protocol violation. Thus, the per-protocol analysis will exclude (set value to missing) any data points at 3 or 6 months that are collected after a cross-over treatment or additional therapy. All data prior to the violation will be included in the analysis.

Note that if a second Botox injection is performed in the Botox arm before 6 months without meeting the protocol-specified criteria for a second injection, data after this second injection will not be excluded from the per-protocol analysis specifically due to this protocol deviation. In addition, since cross-over and additional UI treatments after the 6 month assessment were allowed by protocol, no values will be excluded from the 9 and 12 month data points for the per-protocol analysis.

5.3 Safety (SAF) Population

The safety population will comprise all subjects who were randomized and treated, grouped by the treatment to which they actually received (for any situations in which a participant was not initially treated with their randomized treatment assignment). If all patients received their randomized treatment and have some post-baseline efficacy data, then the safety population would be the same as the mITT.

6 SAMPLE SIZE DETERMINATION

MUSA will randomize 146 women to treatment with Botox or Sling in a 1:1 ratio. The sample size was derived to provide sufficient power for the primary Aim (comparison of Botox versus sling after 6 months) assessed for the primary outcome (change in UDI total score).

For the primary outcome (UDI total score change from baseline at 6 months), alpha was set at 0.05 and power was set at 90%.

Sample size estimates were based on an evaluation of minimum important differences (MIDs) and corresponding standard deviation (SD) for the UDI total score from the ESTEEM study of MUI, with support from other published populations of women with urge and stress incontinence; however, the populations on which those MIDs were based might differ from the target population of mixed incontinence for MUSA. Thus, our goal was to power the study to detect a statistically significant difference between groups in change from baseline in UDI total score at 6 months for the UDI MID determined from ESTEEM as that will be a clinically important difference in our population. To use the most accurate standard deviation applicable to MUSA, we calculated the SD based on a mixed effects repeated measures model with adjustment for baseline UDI score and clinical site rather than the SD of the observed data in order to account for improvements in variance estimation from the covariate adjustment planned for MUSA. To most closely match MUSA, this analysis model included all 3-month and 6-month ESTEEM data points after any retreatments and excludes patients with <4 Urge IEs/3 days. The MID for ESTEEM UDI-Total is 26, and the SD=46.5 from the analysis model, resulting in an effect size = 0.56.

Total number of randomized patients required to meet planned power estimates were adjusted for the assumed patient discontinuation rate at 6 months. Based on average estimates from ROSETTA and ESTEEM, we plan for 5% discontinuation rate by 6 months.

The evaluable sample size required at 6 months for 90% power for the primary total UDI outcome is 69 per arm. Because data from ROSETTA and ESTEEM suggests that follow-up at 6 months should be about 95%, a total sample size of 73 randomized per treatment arm is selected (69/0.95 =73), for a total sample size of 146 randomized participants.

With the sample size of 146 subjects and the planned standard deviation of 46.5, our study will have 82% power to detect a statistically significant difference in UDI-total score at $p<0.05$ if the true difference is as small as 23 points (0.50 effect size) between groups at 6 months. Thus, the planned sample size may allow for analyses to assess whether the true MID in this population is smaller than 26.

Assuming an additional 5% discontinuation rate between 6 and 12 month, the study will have 88% power for the 12 month UDI total score outcome based on the MID of 26, and the SD of 46.5.

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, and where appropriate, minimum, and maximum values. Non-normal continuous data will be summarized by only presenting median and interquartile range (25^{th} and 7^{th} percentile).

7.2 Adjustments for Covariates

Indicator variables for the study stratification factors of site and age group (≥ 65, < 65) will be included as covariates in most efficacy analyses performed for this study (details in section 9). The baseline value of the outcome variable will also be a covariate in models of change from baseline.

Additionally, demographic and baseline characteristics will be compared between study arms using a t-test 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 these analyses suggest that a substantial difference exists among arms for an important characteristic, supportive exploratory analyses of the primary efficacy outcome adjusting for the characteristic may be explored.

7.3 Handling of Dropouts and Missing Data

The repeated measures analysis for the primary outcome for the MITT population (see **Section 9.3**) assumes missing data due to a missed visit or early study discontinuation is missing at random. The impact of missing data on the primary efficacy MITT analysis results will be explored through multiple imputation sensitivity analyses using control-based imputation if more than 10% of participants in the mITT population are missing their 6-month UDI score and a treatment group difference is identified, in order to assess the robustness of the primary results to missing data. (See **Section 9.3**). Sensitivity analyses are not planned for secondary or other outcomes.

7.4 Handling of Cross-over and off-study treatment

It is a protocol deviation for participants to receive any other UI treatments during the first 6 months after treatment. After 6 months, participants may cross-over to receive their non-randomized treatment and may also receive additional off-study UI treatments.

For all primary mITT analyses, data for participants who cross-over and receive the other study treatment or receive an off-study treatment during the first 6 months will be included in all analyses in an intent-to-treat fashion.

The mITT analysis is conservative as it may minimize differences between groups or bias results to make a treatment appear more efficacious due to inclusion of data after a participant has received the other study treatment or additional treatments. Therefore, a supportive per-protocol analysis of the primary and 3 secondary outcomes will be based on the per-protocol population with data excluded (value set to missing) for any data points at 3 or 6 months that are collected after a cross-over treatment or additional therapy. Since crossing over and additional treatments were allowed after 6 months, data points after 6 months will not be excluded. See **Section 9.3**.

The per-protocol analysis is expected to remove bias in outcome caused by the off-protocol additional therapies, in order to provide a more accurate estimate of the true difference in UDI 6 months after the randomized treatment with no further interventions.

The repeated measures analysis assumes that the values set to missing after cross-over treatment are missing at random. However, since additional treatment was sought, it's possible that the outcome scores with no extra treatment received, had they been available, would reflect poorer outcome scores than other recorded data, and thus not be missing at random. Therefore, a multiple imputation tipping point sensitivity analysis of the per-protocol analysis will be conducted. Tipping point analysis will consist of a set of multiple imputation analyses in which the imputed UDI change from baseline scores are increased (made worse) by successively fixed amounts, in order to identify at what amount would have a change in the interpretation of the per-protocol analysis. See **Section 9.3**

7.5 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 symptom.

No formal interim analyses of efficacy outcomes will be performed. At each meeting, the DSMB will be presented with information about enrollment, participant adverse events, 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. Interim efficacy data can be reviewed by the DSMB upon their request.

7.6 Masked Data Review

A masked data review of the primary and secondary outcomes for this study will be performed by the protocol team. This review will occur prior to data lock analyses. This will include evaluation of missing data, descriptive statistics (e.g. means, standard deviations, percentiles for continuous variables and counts and percentages of categorical variables) of key efficacy and safety outcomes, protocol violations and model predictor variables. In the masked review, any summaries will be aggregated over both treatment groups.

7.7 Multicenter Studies

For this multicenter study, randomization of study participants was stratified within center and by age group (≥ 65, < 65). Consequently, for all model-based primary and secondary analyses, center and age group will be included as fixed effects in the models. There were 8 enrolling sites, each with at least 10 treated participants, so sites will not be pooled.

7.8 Multiple Comparisons and Multiplicity

The primary hypothesis will be tested at a nominal two-sided type I error of 0.05. The three secondary outcomes will also be tested at a nominal two-sided type 1 error of 0.05, with no adjustment for multiple comparisons, and thus p-values are descriptive. All p-values for any baseline and demographic characteristics, exploratory outcomes, and safety parameters will also be for descriptive purposes only.

7.9 Examination of Subgroups

We will evaluate the primary outcome for one pre-planned subgroup of interest, which is the type of mixed urinary incontinence based on the baseline UDI: stress-predominant, urgency-predominant, or balanced.

7.10 Assessment Windows

Baseline assessments were to be completed no longer than 3 months prior to intervention with assessments repeated if participant treatment is delayed for over 3 months. All other visits were completed at 2 weeks with a ± 1 week window around the visit and 3-month intervals (3, 6, 9, 12 months) 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 all analyses, all available data will be used regardless of if the assessment was completed within window or not.

8 STUDY SUBJECT CHARACTERIZATION

8.1 Subject Disposition

Disposition of study participants will be described using a standard Consort diagram, including the number of subjects screened, consented, randomized, treated, receiving protocol specified botox at 3-6 months, and completing or discontinuing from the treatment period prior to 6 months and from the 6-month follow-up period. Reasons for study withdrawal will be presented.

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. For DSMB reports, protocol deviations will be listed by site with information such as type of deviation, time of occurrence, and reason, and incidence rate of protocol deviations will be summarized overall and for each

protocol deviation category. Incidence rate of protocol deviations will be calculated as: number of deviations divided by the number of subjects at the site.

Key protocol deviations are violations of inclusion/exclusion criteria, receipt of any off-protocol treatments for UI prior to the 6 month assessment, including switching arms, re-treatment of botox before 3 months, and any non-study UI treatment.

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, education level (classified as binary variable as having some education greater than high school), smoking status (current yes or no), menopausal status, and estrogen use. Also included is time from baseline to treatment and percent of participants with > 91 days to treatment.

Baseline levels of all OAB and QOL measures will be presented on efficacy tables.

9 EFFICACY ANALYSES

9.1 Overview of Efficacy Analyses Methods

All efficacy analyses will be performed on the mITT population unless otherwise specified. All primary, secondary and exploratory efficacy variables were collected longitudinally across this study. Consequently, all efficacy analyses will be conducted using appropriate models for these correlated data collected across time. Specifically, continuous outcome variables will be analyzed using linear mixed models and binary outcome variables will be analyzed using generalized linear mixed models. Models will include terms for treatment and for the stratification variables included in the study design (site and age group (≥ 65, < 65)). Additional details are provided in the specific sections below (See **section 9.3**).

9.2 Variable Definitions

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

Note that for any patient who had a repeat baseline, their repeat measures will be used in the definition of baseline (their updated baseline scores were entered into the repeat baseline CRF by the sites).

Variable	Type	Definition
Primary Outcomes		
type of urinary incontinence	**Categorical** stress-predominant urgency-predominant balanced	The definition of the type of urinary incontinence will be based on 2 questions of the Urogenital Distress Inventory, 1. "Do you experience urine leakage associated with a feeling of urgency?" 2. "Do you experience urine leakage related to physical activity, coughing, or sneezing?" Based on our inclusion criteria, patients must report at least "moderate" bother on both questions. If the subject reports greater bother on the stress question, she will be classified as stress-predominant MUI. If the subject reports greater bother on the urgency question, she will be classified as urgency-predominant MUI. If the subject reports equal bother on both questions, she will be classified as "balanced" MUI.
Change from baseline in UDI Total Score at 6 months (and at 3, 9, and 12 months) 6 months is primary,	Continuous	The Urogenital Distress Inventory-Total Score will be computed at baseline and 6 months (and 3, 9, and 12 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 the score will be calculated as the sum of the non-missing scales. The outcome will then be computed as the difference in Total score at 6 months (and 3, 9, and 12 months) and the Total score at baseline. If data for an assessment time point are missing, the outcome variable will be coded as missing.

Variable	Type	Definition
Secondary Outcomes		
Change from baseline in UDI Obstructive Score at 6 months (and at 3, 9 and 12 months)	Continuous	The Urogenital Distress Inventory-Obstructive Score will be computed as baseline and 6 months (and 3, 9, and 12 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 be computed as the difference in score at 6 months (and 3, 9, and 12 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 Stress Score at 6 months (and at 3, 9 and 12 months)	Continuous	The Urogenital Distress Inventory-Stress Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 6 months (and at 3, 9 and 12 months)	Continuous	The Urogenital Distress Inventory-Irritative Score will be computed at baseline and 6 months (and 3, 9, and 12 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. The outcome will then be computed as the difference in score at 6 months (and 3, 9, and 12 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
Exploratory Outcomes		
Normalization of Voiding Frequency at 6 (and at 3 and 12 months)	Dichotomous (Yes/No)	The total number of daytime and nighttime voids will be taken from the bladder diary and summed and then divided by the number of diary days (<=3) to get a voiding frequency. The indicator for normalization at 6 months (and 3 and 12 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 of 50% or more in Voiding Frequency at 6 months (and at 3 and 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 6 months (and 3 and 12 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.
Worsening of Voiding Frequency at 6 months (and at 3 and 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 an increase of any amount between baseline and the time point, the worsening indicator at 6 months (and 3 and 12 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.

Variable	Type	Definition
		The variable is summarized as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.
Change from baseline in Urge Incontinence Episodes at 6 months (and at 3 and 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 6 months (and 3 and 12 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 6 months (and at 3 and 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 6 months (and 3 and 12 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 Mixed or Don't Know Incontinence Episodes at 6 months (and at 3 and 12 months)	Continuous	The daily frequency of mixed urge and stress incontinence episodes will be taken from the bladder diary. The outcome will then be computed as the difference in daily frequency of mixed incontinence (noted on diary as "both stress/urge") or "don't know" episodes at 6 months (and 3 and 12 months) and the daily frequency of mixed 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 6 months (and at 3 and 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 6 months (and 3 and 12 months) and the daily frequency of total incontinence episodes 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 Nocturnal voids at 6 months (and at 3 and 12 months)	Continuous	The daily frequency of total 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 6 months (and 3 and 12 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 Diurnal voids at 6 months (and at 3 and 12 months)	Continuous	The daily frequency of total 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 6 months (and 3 and 12 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 OAB-q Symptom Severity Score at 3, 6, 9, 12 Months	Continuous	The OAB-q Symptom Severity Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 Months	Continuous	The OAB-q HRQL Coping Score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 9, 11, 16, 21, 22, 26, 32, and 33 (1= None of the time, 2=A little bit of the time, 3=Some of the time, 4=A good bit of the time, 5=Most of the time, 6=All of the time), 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 6 months

Variable	Type	Definition
		(and 3, 9, and 12 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 Concern Score at 3, 6, 9, 12 Months	Continuous	The OAB-q HRQL Concern Score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 12, 13, 14, 19, 23, 25, and 29 (1= None of the time, 2=A little bit of the time, 3=Some of the time, 4=A good bit of the time, 5=Most of the time, 6=All of the time), 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 Months	Continuous	The OAB-q HRQL Sleep Score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 10, 15, 17, 24, and 30 (1= None of the time, 2=A little bit of the time, 3=Some of the time, 4=A good bit of the time, 5=Most of the time, 6=All of the time), 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 Months	Continuous	The OAB-q HRQL Social Score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 18, 20, 27, 28, and 31 (1= None of the time, 2=A little bit of the time, 3=Some of the time, 4=A good bit of the time, 5=Most of the time, 6=All of

Variable	Type	Definition
		the time), 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 6 months (and 3, 9, and 12 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 Total Score at 3, 6, 9, 12 Months		The OAB-q HRQL Total Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 months) and the score at baseline. If any of the listed 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 3, 6, 9, 12 Months	Continuous	NOTE: The scoring manual for the OAB-SAT-q states that higher scores should correspond to better outcomes and 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 adjusted 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. The OAB-SAT-q Satisfaction Score will be computed at 6 months (and 3, 9, and 12 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 the lowest possible score (3) from the raw score and divide by the range of possible scores (18-3=15), then multiply by 100. If less than half of the

Variable	Type	Definition
		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 3, 6, 9, 12 Months	Continuous	NOTE: The scoring manual for the OAB-SAT-q states that higher scores should correspond to better outcomes and 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 adjusted 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. The OAB-SAT-q Side Effects Score will be computed at 6 months (and 3, 9, and 12 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, …, 5=Not at all), then subtract the lowest possible score (3) from the raw score 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.
OAB-SAQ-q Endorsement Score at 3, 6, 9, 12 Months	Continuous	NOTE: The scoring manual for the OAB-SAT-q states that higher scores should correspond to better outcomes and 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 adjusted the scoring algorithm instead to be: 100*(raw score – lowest possible

Variable	Type	Definition
		score)/score range. This calculation results in higher scores corresponding to better outcomes and an overall scale of 0-100. The OAB-SAT-q Endorsement Score will be computed at 6 months (and 3, 9, and 12 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 the lowest possible score (3) 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 3, 6, 9, 12 Months	Continuous	NOTE: The scoring manual for the OAB-SAT-q states that higher scores should correspond to better outcomes and 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 adjusted 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. The OAB-SAT-q Convenience Score will be computed at 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to use the response to question 4 (1=Extremely Inconvenient, …, 6=Extremely Convenient), then subtract the lowest possible score (1) and divide by the range of possible scores (6-1=5), 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.

Variable	Type	Definition
OAB-SAT-q Preference Indicator at 3, 6, 9, 12 Months	Continuous	The OAB-SAT-q Preference Score will be computed at 6 months (and 3, 9, and 12 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 or the response is "Never been treated before for overactive bladder", then the indicator is set to missing. 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.
Change from Baseline IIQ Physical Activity Score at 3, 6, 9, 12 Months	Continuous	The Incontinence Impact Questionnaire-Physical Activity Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 Months	Continuous	The Incontinence Impact Questionnaire-Travel Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 Social Relationships Score at 3, 6, 9, 12 Months	Continuous	The Incontinence Impact Questionnaire-Social Relationships Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 Months	Continuous	The Incontinence Impact Questionnaire-Emotional Health Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 Total Score at 3, 6, 9, 12 Months	Continuous	The Incontinence Impact Questionnaire-Total Score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 PISQ-IR not sexually active - partner related subscale score at 3, 6, 9, 12 months	Continuous	The PISQ-IR not sexually active – partner related subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q2a, Q2b using reverse scores (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

Variable	Type	Definition
		be computed as the difference in score at 6 months (and 3, 9, and 12 months) and the score at baseline. If data for a time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants not sexually active.
Change from Baseline PISQ-IR not sexually active – condition specific subscale score at 3, 6, 9, 12 months	Continuous	The PISQ-IR not sexually active – condition specific subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q2c, Q2d, Q2e using reverse scores (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 6 months (and 3, 9, and 12 months) and the score at baseline. If data for a time point are missing, the outcome variable will be coded as missing. Scores should only be calculated for participants not sexually active.
Change from Baseline PISQ-IR not sexually active – global quality subscale score at 3, 6, 9, 12 months	Continuous	The PISQ-IR not sexually active – global quality subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q4a, Q4b, Q5a, and Q6 using reverse scores for only 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 6 months (and 3, 9, and 12 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 at 3, 6, 9, 12 months	Continuous	The PISQ-IR not sexually active – condition subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q3, Q5b, Q5c using reverse scores for Q5b and Q5c (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

Variable	Type	Definition
		items answered. The outcome will then be computed as the difference in score at 6 months (and 3, 9, and 12 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 sexually active – arousal, orgasm subscale score change from baseline at 3, 6, 9, 12 months	Continuous	The PISQ-IR sexually active – arousal, orgasm subscale score will be computed at baseline and 6 months (and 3, 9, and 12 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). 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 months	Continuous	The PISQ-IR sexually active – condition specific subscale score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 months	Continuous	The PISQ-IR sexually active – partner related subscale score will be computed at baseline and 6 months (and 3, 9, and 12 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;

Variable	Type	Definition
		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 6 months (and 3, 9, and 12 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 and have a sexual partner.
Change from Baseline PISQ-IR sexually active – desire subscale score at 3, 6, 9, 12 months	Continuous	The PISQ-IR sexually active – desire subscale score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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 3, 6, 9, 12 months	Continuous	The PISQ-IR sexually active – condition impact subscale score will be computed at baseline and 6 months (and 3, 9, and 12 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 6 months (and 3, 9, and 12 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.

Variable	Type	Definition
Change from Baseline PISQ-IR sexually active – global quality rating subscale score at 3, 6, 9 12 months	Continuous	The PISQ-IR sexually active – global quality rating subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using standard scoring algorithms. Specifically, instructions are to sum the scores for questions Q19a-Q19c, Q20a using reverse scores for Q19a-Q19c (Q19a: 1=satisfied, …, 5=dissatisfied; Q19b: 1=adequate, …, 5=inadequate; Q19c: 1=confident, …, 5=not confident; Q20a: 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 6 months (and 3, 9, and 12 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 – average score at 3, 6, 9 12 months	Continuous	The PISQ-IR sexually active – average score will be computed at baseline and 6 months (and 3, 9, and 12 months) using standard scoring algorithms. For participants that have a sexual partner (Q12 = 1), instructions are to sum the scores for questions Q7, Q8a-8c, Q9, Q10, Q11, Q13, Q14a-b, Q15, Q16, Q17, Q18, Q19a-c, Q20a-d using reverse scores for Q8b-c, Q9, Q11, Q14a-b, Q16, Q17, Q18, Q19a-c. If there are more than 10 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. Scores should only be calculated for participants that are sexually active. For participants that do not have a sexual partner (Q12 = 2), instructions are to sum the scores for questions Q7, Q8a-8c, Q9, Q10, Q11, Q15, Q16, Q17, Q18, Q19a-c, Q20a-d using reverse scores for Q8b-c, Q9, Q11, Q16, Q17, Q18, Q19a-c. If there are more than 9 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. Scores should only be calculated for participants that are sexually active.
Change from Baseline EQ-5D Total Score at 3, 6, 9,12 Months	Continuous	The EQ-5D Total Score will be computed at baseline and 6 months (and 3, 9, and 12 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

Variable	Type	Definition
		difference in Total score at 6 months (and 3, 9, and 12 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 3, 6, 9, 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 6 months (and 3, 9, and 12 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 SF-36 Physical Component Summary at 3, 6, 9, 12 Months	Continuous	The SF-36 Physical Component Summary score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to standardize the 8 SF-36 subscales scores using a set of means and standard deviations from Ware et al. (https://www.researchgate.net/publication/292390260_SF-36_Physical_and_Mental_Health_Summary_Scales_a_User%27s_Manual), then multiply each subscale z-score by its respective physical/mental factor score coefficient, and sum the result from each subscale for the aggregate summary score. Taft et al. have conveniently created a table of reference values for scoring the SF-36 Physical Component Summary (https://www.researchgate.net/publication/11595008_Do_SF-36_Summary_Component_Scores_Accurately_Summarize_Subscale_Scores): **Table 1.** US subscale means, standard deviations and factor score coefficients for PCS and MCS Subscale / Mean / STD / PCS coefficients / MCS coefficients PF 84.52404 22.89490 0.42402 −0.22999 RP 81.19907 33.79729 0.35119 −0.12329 BP 75.49196 23.55879 0.31754 −0.09731 GH 72.21316 20.16964 0.24954 −0.01571 VT 61.05453 20.86942 0.02877 0.23534 SF 83.59753 22.37642 −0.00753 0.26876 RE 81.29467 33.02717 −0.19206 0.43407 MH 74.84212 18.01189 −0.22069 0.48581 Reproduced from Ware et al. [2].

<table>
<tr><th>Variable</th><th>Type</th><th>Definition</th></tr>
<tr><td>Change from Baseline SF-36 Mental Component Summary at 3, 6, 9, 12 Months</td><td>Continuous</td><td>The SF-36 Mental Component Summary score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to standardize the 8 SF-36 subscales scores using a set of means and standard deviations from Ware et al. (https://www.researchgate.net/publication/292390260_SF-36_Physical_and_Mental_Health_Summary_Scales_a_User%27s_Manual), then multiply each subscale z-score by its respective physical/mental factor score coefficient, and sum the result from each subscale for the aggregate summary score.

Taft et al. have conveniently created a table of reference values for scoring the SF-36 Physical Component Summary (https://www.researchgate.net/publication/11595008_Do_SF-36_Summary_Component_Scores_Accurately_Summarize_Subscale_Scores):
Table 1. US subscale means, standard deviations and factor score coefficients for PCS and MCS
<table><tr><th>Subscale</th><th>Mean</th><th>STD</th><th>PCS coefficients</th><th>MCS coefficients</th></tr><tr><td>PF</td><td>84.52404</td><td>22.89490</td><td>0.42402</td><td>−0.22999</td></tr><tr><td>RP</td><td>81.19907</td><td>33.79729</td><td>0.35119</td><td>−0.12329</td></tr><tr><td>BP</td><td>75.49196</td><td>23.55879</td><td>0.31754</td><td>−0.09731</td></tr><tr><td>GH</td><td>72.21316</td><td>20.16964</td><td>0.24954</td><td>−0.01571</td></tr><tr><td>VT</td><td>61.05453</td><td>20.86942</td><td>0.02877</td><td>0.23534</td></tr><tr><td>SF</td><td>83.59753</td><td>22.37642</td><td>−0.00753</td><td>0.26876</td></tr><tr><td>RE</td><td>81.29467</td><td>33.02717</td><td>−0.19206</td><td>0.43407</td></tr><tr><td>MH</td><td>74.84212</td><td>18.01189</td><td>−0.22069</td><td>0.48581</td></tr></table>Reproduced from Ware et al. [2].</td></tr>
<tr><td>Change from Baseline SF-36 Physical Functioning subscale score at 3, 6, 9, 12 Months</td><td>Continuous</td><td>The SF-36 Physical Functioning subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 3 through 12 (1 = Yes, limited a lot, …, 3 = No, not limited at all), then subtract the minimum possible score (10) and divide by the range of possible score (30–10=20), 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 6 months</td></tr>
</table>

Variable	Type	Definition
		(and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 Role-Physical subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 Role-Physical subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 13 through 16 (1 = Yes, 2 = No), then subtract the minimum possible score (4) and divide by the range of possible score (8–4=4), 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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 Bodily Pain subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 Bodily Pain subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 21 and 22 with specific recoded values (Q21: 1 = None, …, 6 = Very severe; Q22: 1 = Not at all, …, 5 = Extremely), then subtract the minimum possible score (2) and divide by the range of possible score (12-2=10), 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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 General Health Perception subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 General Health Perception subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 1 and 33 through 36 with specific recode values for Q1 and using reverse scores for Q34 and Q36 (Q1: 1 = Excellent, …, 5 = Poor; Q33 – Q36: 1 = Definitely True, …, 5 = Definitely False), then subtract the minimum possible score (5) and divide by the range of possible score (25-5=20), then multiply by 100. If less than half of the responses are missing, then the score will be

Variable	Type	Definition
		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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 Vitality subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 Vitality subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 23, 27, 29, and 31 using reverse scores for Q23 and Q27 (1 = All of the time, …, 6 = None of the time), then subtract the minimum possible score (4) and divide by the range of possible score (24-4=20), 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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 Social Functioning subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 Social Functioning subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 20 and 32 using reverse scores for Q20 (Q20: 1 = Not at all, …, 5 = Extremely; Q32: 1 = All of the time, …, 5 = None of the time), then subtract the minimum possible score (2) and divide by the range of possible score (10-2=8), 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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 Role-Emotional subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 Role-Emotional subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 17 through 19 (1 = Yes, 2 = No), then subtract the

Variable	Type	Definition
		minimum possible score (3) and divide by the range of possible score (6-3=3), 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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-36 Mental Health subscale score at 3, 6, 9, 12 Months	Continuous	The SF-36 Mental Health or Emotional Wellbeing subscale score will be computed at baseline and 6 months (and 3, 9, and 12 months) using the standard scoring algorithm. Specifically, instructions are to sum responses across questions 24 through 26, 28, and 30 using reverse scores for Q26 and Q30 (1 = All of the time, …, 6 = None of the time), then subtract the minimum possible score (5) 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 6 months (and 3, 9, and 12 months) and the score at baseline. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Change from Baseline SF-6D at 3, 6, 9, 12 Months	Continuous	The SF-6D will be computed at baseline and 6 months (and 3, 9, and 12 months). The SF-6D is a classification for describing health derived from a selection of SF-36 items, composed of six multi-level dimensions. The instrument is managed by QualityMetric (https://www.qualitymetric.com/) where the licensed algorithm for the MUSA study was obtained. The algorithm comes with a set of preference weights obtained from a sample of the general population using the recognized valuation technique of standard gamble. The algorithm will generate 6 dimension scores (physical functioning, role limitations, social functioning, bodily pain, mental health, & vitality) and a utility value ranging from 0 – 1 (0=”Dead”, 1=”Full Health”).

Variable	Type	Definition
Patient Global Impression of Improvement at 3, 6, 9, 12 months	Dichotomous 1 = Improved 0 = No Improvement /Worsened	The Patient Global Impression of Improvement score will be computed at 6 months (and 3, 9, and 12 months). Improvement will be defined as a response of "Much better" or "Very much better", any other response will be indicated as no improvement or worsened. If data at an assessment time point are missing, the outcome variable will be coded as missing.
Patient Global Impression of Severity at 3, 6, 9, 12 months	Dichotomous 1 = Normal/Mild 0 = Moderate/Severe	The Patient Global Impression of Severity score will be computed at baseline and 6 months (and 3, 9, and 12 months). For defining binary severity of urinary tract condition, a scale response of "Normal" or "Mild" will be set to "Normal/Mild". Scale responses of "Moderate" or "Severe" will be set to "Moderate/Severe". If data at an assessment time point are missing, the outcome variable will be coded as missing.
Patient Global Symptom Control at 3, 6, 9, 12 months	Dichotomous 1=adequate leakage control 0 = inadequate leakage control	The Patient Global Symptom Control score will be computed at 6 months (and 3, 9, and 12 months). The statement "My current treatment is giving me adequate control of my urine leakage" is scored on a scale from 1 (strongly disagree) to 5 (strongly agree). If the response is scored 1, 2, or 3, coded as "inadequate leakage control". If the response is scored 4 or 5, coded as "adequate leakage control". A score ≤ 3 is required to meet the criteria for additional Botox injection.
Had protocol-specified additional Botox between 3-6 months	Dichotomous 1=yes 0=no	= yes for women who received Botox as the main study treatment and also received the additional 2nd allowed Botox treatment prior to 6 months, = no otherwise For participants who received MUS as their main study treatment, this indicator will be coded to missing.
Additional treatment up to 6 months, and	Dichotomous 1=yes 0=no	The types of additional treatment for SUI and/or OAB will be identified and categorized from the Additional Therapy form.

Variable	Type	Definition
Additional treatment between 6 and 12 months		An indicator of additional treatment will be calculated for each subject where subjects with additional treatments identified below will be assigned "Yes", otherwise subjects are assigned "No". The rate of receiving additional treatment is calculated as the percentage of participants with indicator ="Yes" among all non-missing indicators. This includes anyone who had a crossover treatment (MUS to Botox, or Botox to MUS), in addition to any other type of UI/OAB treatment or medication. If a participant is missing the visit required to derive additional treatment timing (6 month or 12 month), the visit window close date (6 or 12 months from treatment + 30 days) was used in place of the missing visit date. Indicators will be created for overall additional treatment and the following individual categories: **Up to 6 months (up to the day before the 6 months visit):** MUS for Botulinum toxin A arm Botulinum toxin A for MUS arm OAB/UUI medication Behavioral Therapy Physical Therapy / Biofeedback PTNS Continence pessary / POISE Impressa Sacral neuromodulation Bulking injections **6-12 months (starting the day of the 6 months visit)** Mid-urethral Sling placement

Variable	Type	Definition
		Botulinum toxin A injection OAB/UUI medication Behavioral Therapy Physical Therapy / Biofeedback PTNS Continence pessary / POISE Impressa Sacral neuromodulation Bulking injections
Safety outcomes		Adverse events are reported for a participant from the date they are randomized to the date that they exit the study. There were 9 adverse events that were collected after randomization but occurred before treatment. The majority of these were UTIs (n=8), with one "Fall" event that is important documentation leading into a subsequent serious adverse event.
Sling Revision or removal at any time	Dichotomous 1=yes 0=no	An indicator of sling revision/removal will be calculated for each subject where those who have a sling revision/removal listed on the Additional Therapy form will be assigned a "Yes", otherwise subjects are assigned "No". At final database lock reviews, one "sling removal" was identified from the Additional Therapy form. Adverse events were further reviewed for implemented actions of "explant" or "surgical intervention" but yielded no further cases identified. An adverse event form was requested to be entered for the single reported removal. The rate of sling revision is calculated as the percentage of participants with indicator equal to "Yes" among all non-missing indicators.

Variable	Type	Definition
Sling mesh exposure	Dichotomous 1=yes 0=no	An indicator of sling mesh exposure will be calculated for each subject (yes/no) using the MEDRA coded preferred term of reported adverse events. Subjects will be coded to "Yes" if they have any adverse event with a MEDRA preferred term of "Medical device site erosion", or if the verbatim adverse event term includes text strings "Mesh", "Mesh exposure", or "MUS mesh exposure". Two cases were identified from adverse event forms at final database lock review.
Rate of Urinary Tract Infection (UTI) Events	Dichotomous 1=yes 0=no	The frequency of acute UTI events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported acute UTIs over their study duration, identified by MEDRA coding preferred term of "Urinary tract infection" and "recurrent" is NOT specified by verbatim adverse event term or comments. If no acute UTI events were reported for a participant, the outcome variable will be coded with a value of "No". Final database lock review took efforts to query sites for participants with questionable urine test results and get PI confirmation that all necessary events had been reported.
Rate of recurrent Urinary Tract Infection (UTI) Events	Dichotomous 1=yes 0=no	The frequency of recurrent UTI events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported recurrent UTIs over their study duration, identified by MEDRA coding preferred term of "Urinary tract infection" and "recurrent" is specified by verbatim adverse event term or comments. If no recurrent UTI events were reported for a participant, the outcome variable will be coded with a value of "No". Final database lock review took efforts to query sites for participants with questionable urine test results and get PI confirmation that all necessary events had been reported.
Rate of Urinary Retention Events	Dichotomous 1=yes 0=no	The frequency of urinary retention events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported urinary retention over their study duration,

Variable	Type	Definition
		identified by MEDRA coding preferred term of "Urinary retention". If no urinary retention events were reported for a participant, the outcome variable will be coded with a value of "No". Final database lock review took efforts to query sites for participants with reported self-catheterization logs or questionable post void residual results (>200mL, aside from visits immediately post-surgery) and get PI confirmation that all necessary events had been reported.
Rate of AE of "residual Urine Volume"	Dichotomous 1=yes 0=no	The frequency of residual urine volume events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported residual urine events over their study duration, identified by MEDRA coding preferred term of "Residual urine volume". If no residual urine volume events were reported for a participant, the outcome variable will be coded with a value of "No". Final database lock review took efforts to query existing adverse events with terms such as "incomplete bladder emptying", "sensation of incomplete bladder emptying" for any post void residual assessments completed around the time of the event. Details from query responses helped with MEDRA coding the relevant level of severity ("Urinary Retention" vs. "Residual urine volume") accurately.
Complications after study intervention And Complication after cross-over study intervention Bladder Injury Ureteral Injury	Dichotomous 1=yes 0=no	The frequency of complications after study intervention will be taken from Surgeon's Report forms submitted for a participant, for both the main study treatment and allowed Botox reinjection if applicable. The outcome variable for each specific complication will be coded with a value of "Yes" if ever reported over their study duration, otherwise it will be coded with a value of "No". If a participant experienced any complication after main study interventions, the composite indicator will be coded with a value of "Yes", otherwise coded to "No". The frequency of complications after cross-over study intervention will be taken from any Off Protocol Surgeon's Report forms submitted for a participant The outcome

Variable	Type	Definition
Urethral Injury Vaginotomy		variable for each specific complication will be coded with a value of "Yes" if ever reported over their study duration, otherwise it will be coded with a value of "No". If a participant experienced any complication after a crossover study intervention, the composite indicator will be coded with a value of "Yes", otherwise coded to "No". Note that all reported intervention complications are expected to have an accompanying adverse event form entered but may not be recorded as an adverse event to reduce database redundancy.
Any catheter placement or intermittent self-catheterization 2 or more weeks post-procedure	Dichotomous 1=yes 0=no	The frequency of catheter placement or intermittent self-catheterization will be identified from self-catheterization (CISC) logs submitted for a participant. Additionally, adverse events that occurred 2 or more weeks after procedure reported with a "Catheter placement" action or have any text detailing self-catheterization in the comments field will also identify these cases. The outcome variable will be coded with a value of "Yes" if the participant has any reported CISC log or adverse events detailing self-catheterization, otherwise it will be coded with a value of "No". Note that the Surgeon's Report form completed at time of study intervention contains questions regarding self-catheterization, but this outcome is specifically limited to timing greater than 2 weeks from study procedure. Final database lock review took efforts to query sites for participants with reported self-catheterization logs or questionable post void residual results (>200mL, aside from visits immediately post-surgery) and get PI confirmation of self-catheterization in relevant adverse event comments.
Any AE that is unexpected and Related to intervention	Dichotomous 1=yes 0=no	Participants will be coded to "Yes" if any adverse event was reported over their study duration that had the following fields indicated on the form: Expected = "No" Relationship = "Related" Otherwise, the participant will be coded to "No".

Variable	Type	Definition
Any Hospital admission or return to OR related to intervention	Dichotomous 1=yes 0=no	The frequency of hospital readmissions / returns to the operating room related to study intervention will be identified by searching for applicable terms and comments submitted across all adverse events for a participant. If the participant has an event indicates hospital readmission or a return to the operating room and is reported as "Related", this indicator will be coded to "Yes", otherwise it will be coded to "No". At final database lock review, there were no adverse events reported as "related" that indicated a hospital readmission or return to the operating room.
Pelvic Pain	Dichotomous 1=yes 0=no	The frequency of pelvic pain events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported pelvic pain over their study duration, identified by MEDRA coding preferred term of "Pelvic pain". If no pelvic pain events were reported for a participant, the outcome variable will be coded with a value of "No".
Dyspareunia	Dichotomous 1=yes 0=no	The frequency of dyspareunia events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported dyspareunia over their study duration, identified by MEDRA coding preferred term of "Dyspareunia". If no dyspareunia events were reported for a participant, the outcome variable will be coded with a value of "No".
Levator Syndrome	Dichotomous 1=yes 0=no	The frequency of levator spasm events will be taken from all adverse event reporting forms submitted for a participant. The outcome variable will be coded with a value of "Yes" if the participant had any reported levator spasms over their study duration, identified by MEDRA coding preferred term of "Levator syndrome". If no levator spasm events were reported for a participant, the outcome variable will be coded with a value of "No".
Self-report New/worsening abdominal/genital pain (UDI item N)	Dichotomous 1=yes 0=no	Self-report of new or worsening abdominal/genital pain is captured in the UDI form item N at baseline and 6 months (and 3, 9, and 12 months). The outcome variable will be coded with a value of "Yes" if any post-baseline UDI assessment reported a worse bother than at baseline (ordered categories are "No", "Yes" with bother not specified / "Not at all", "Yes" and "Slightly", "Yes" and "Moderately, "Yes" and "Greatly"). For

Variable	Type	Definition
		participants that did not have any post-baseline visits, the outcome variable will be set to missing. UDI item N: Do you experience pain in the lower abdominal or genital area? ☐ Yes ☐ No (Skip to O.) If yes, how much does it bother you? ☐ Not at all ☐ Slightly ☐ Moderately ☐ Greatly
Self-report New/worsening dyspareunia (PISQ item C5)	Dichotomous 1=yes 0=no	Self-report of new or worsening dyspareunia is captured on the PISQ-IR form item C5 at baseline and 6 months (and 3, 9, and 12 months). The outcome variable will be coded with a value of "Yes" if any post-baseline PISQ-IR assessment reported a worse frequency of pain than at baseline (ordered categories are "Never", "Rarely", "Sometimes", "Usually", "Always"). For participants that did not have any post-baseline visits or did not indicate they were sexually active at baseline, the outcome variable will be set to missing. PISQ item C5: (completed only for sexually active) How often do you feel pain during sexual intercourse? (If you don't have intercourse check this box ☐ and skip to the next item.) 1 ☐ Never 2 ☐ Rarely 3 ☐ Sometimes 4 ☐ Usually 5 ☐ Always

Variable	Type	Definition
Self-report New/worsening difficulty emptying bladder (UDI item J)	Dichotomous 1=yes 0=no	Self-report of new or worsening difficulty emptying bladder is captured on the UDI form item J at baseline and 6 months (and 3, 9, and 12 months). The outcome variable will be coded with a value of "Yes" if any post-baseline UDI assessment reported a worse bother than at baseline (ordered categories are "No", "Yes" with bother not specified / "Not at all", "Yes" and "Slightly", "Yes" and "Moderately, "Yes" and "Greatly"). For participants that did not have any post-baseline visits, the outcome variable will be set to missing. UDI item J: Do you experience difficulty emptying your bladder? ☐ Yes ☐ No (Skip to K.) If yes, how much does it bother you? ☐ Not at all ☐ Slightly ☐ Moderately ☐ Greatly
Other AEs or symptoms of interest	Dichotomous 1=yes 0=no	We will also look to see if there are any additional AEs to report on the AE table, covering the following: These will be identified via evaluation of AE preferred terms, verbatim descriptions/comments and AE relationship at least possibly related. - Other MUS surgical wound healing problems >6 weeks - New/worsening vaginal infection - Other infection possibly related to intervention - New/worsening partner dyspareunia At final database lock review, 2 additional event terms of interest were identified: "pyelonephritis", and "Gonococcal pelvic inflammatory disease". Appropriate indicator variables will be created for each additional term of interest, coded to "Yes" if the participant has any reported events coded to the respective MEDRA preferred term. The outcome variable will be set to "No", otherwise.

9.3 Analysis Methods for the Primary Outcome

The mean change from baseline in UDI total score will be compared between groups at 6 months at the $p<0.05$ statistical significance level using a mixed effects analysis of covariance model for repeated measures (MMRM) with adjustment for baseline UDI score, and randomization stratification factors site and age group. The model will include fixed effect categorical factors for treatment group, visit (3 6 months, 9, 12), and treatment × visit interaction with an unstructured covariance pattern for the within-subject repeated measures separately within the two treatment groups. Estimates, p-values and 95% confidence intervals will be presented for treatment group comparisons at 6 moths (primary) and 3 months (secondary), and 9 and 12 months (other efficacy). (SM Davis, 2014)

Example code from SAS to produce this analysis is as follows:

```
proc mixed covtest;
    class subject site trtgp VISIT;
    model DELTA_VAR = BL_VAR site trtgp VISIT trtgp*VISIT / ddfm=kenwardroger;
    repeated VISIT / subject=subject type=UN group=trtgrp R rcorr;
    title "Mixed model for change from baseline";
    lsmeans trtgp*VISIT / cl pdiff=all;
run;
```

We will report whether change in total UDI score from baseline to 6 months is statistically significantly different between groups. If the difference is statistically significant, the potential clinical significance of the difference will be discussed. We recognize that our sample size may allow us to find a difference between groups that is statistically significant yet smaller than published MID 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 MUSA, and an exploratory aim of MUSA is to explore whether the MID in this population differs from previously published values (see section 9.5.3).

Sensitivity analysis: This model assumes missing data due to a missed visit or early study discontinuation is missing at random (MAR). If more than 10% of participants are missing their 6-month score and a statistical difference is seen between groups at 5 months, then a sensitivity analysis in order to assess the robustness of the primary results to missing data will be performed using multiple imputation under a not -missing-at-random (NMAR) as follows: missing data from discontinuations in the arm found to be superior will be imputed based on data from the less superior arm ("control-based imputation"). (Ratitich et al 2013, 2,14). This sensitivity analysis essentially takes the conservative assumption that missing data after discontinuation in the superior arm have outcome values as if they were in the non-superior arm.

A supportive per-protocol analysis of the primary outcome will be based on the per-protocol population with data excluded (set value to missing) for any data points at 3 or 6 months that are collected after a cross-over treatment or additional therapy. Since crossing over and additional treatments were allowed after 6 months, data points after 6 months are not planned to be excluded in any sensitivity analyses.

Sensitivity of the per-protocol analysis: the data set to missing due to off-protocol treatments is assumed by the model to be missing at random. However, it is possible that if these measures were assessed without additional therapy, they might have reflected poor outcomes. To assess the impact of the results from the missing data, we will conduct a tipping point multiple imputation

analysis, in which the missing values are increased (worsened) by a fixed value, in successive intervals across successive analyses until a point at which statistical interpretation of the p-value at 6 months between treatment groups is changed. (Ratitich et al 2013, 2,14).

9.4 Analysis Methods for secondary and exploratory outcomes

There are three secondary outcomes for the primary aim. The mean change from baseline in UDI-stress and UDI-irritative subscores will be compared between groups at 6 months, and the UDI total will be compared at 3 months (see above). The statistical analysis of the three secondary outcomes will be identical to the primary analysis described above for the UDI total score.

In addition, the per-protocol analysis will be repeated for the secondary outcomes.

For all continuous exploratory measures, groups will be compared using the same model as the primary analysis described above for the UDI total score.

For any dichotomous outcomes assessed across time, groups will be compared for the odds of the event, and an odds ratio will be estimated from a repeated measures logistic generalized linear regression model with adjustment for baseline value, site and baseline age category, and with treatment, visit and treatment by visit interaction with an unstructured covariance pattern separately within each treatment group for the within-subject repeated measures.

For PGI-I and PGSC, the baseline PGI-S score will be the baseline covariate.

Example code from SAS to produce this analysis is as follows:

```
Proc glimmix;
   class subjectID trtgrp VISIT site;
   model dep_var = bbaseline_var trtgrp site visit trgrp*visit /
         solution link=logit dist=binary oddsratio DDFM=KR;
   random time / type=UN Group=trtgrp subject=subjectID RSIDE;
   lsmeans trtgrp*time / ILINK CL; /*estimates event rates*/
   lsmeans trtgrp*time / oddsratio diff CL; /*estimates ORs*/
   estimate 'trt a vs trt b at time 2' trtgrp 1 -1 trtgrp*time 0 1 0 0 0  -10 0  / EXP CL;
Run;
```

Groups will be compared for receipt of additional treatment with chi-squared tests or Fisher's exact test for small event rates.

9.5 Other Analysis Methods

9.5.1 Predictors of Treatment Response

Predictors of response will be identified in a secondary manuscript, and methods for analysis will be documented in a separate SAP prior to beginning of the analysis. Both traditional prediction model as well as machine learning methods will be used. Methods specified in the protocol are as follows:

Regression models will be created to identify predictors of change from baseline to 6 months for UDI total score and stress and irritative subscale scores.

Potential predictors will include age, diary parameters such as number of UUI episodes/3 days, functional bladder capacity, stress and irritative bother severity at baseline, type of urinary incontinence (stress-predominant or urgency-predominant, balanced), and co-existing anterior vaginal wall prolapse.

The relationship between potential predictors and outcomes will be explored in models that include one predictor plus stratification factors (treatment group, site and age 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.

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 6 months. 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 (see section 9.5.3).

9.5.2 Cost-effectiveness analysis

The EQ-5D and the SF-6D will be used to calculate QALYs. Methods for cost-effectiveness analysis from the protocol are included here and will be expanded upon is a cost-effectiveness analysis SAP that will be developed prior to analysis of the cost effectiveness data, which will be reported in a secondary manuscript.

Differential mean costs and differential mean QALYs between the two treatment groups will be estimated using multiple regression analysis. Specifically, a generalized linear model with appropriate link function (e.g., log-link) and response probability distribution (e.g., gamma distribution) will be used to analyze costs due to the potential skewness and heteroscedasticity of medical expenditure data, while an ordinary least squares regression will be used for analyzing QALY data. The models will account for stratification factors of study site and age group, as well as other characteristics of the subjects that are found to differ significantly between the groups. When estimating QALYs, we will also adjust for subjects' baseline utility scores to account for potential imbalance in baseline utility between the two treatment groups.

We will calculate the incremental cost-effectiveness ratio (ICER), which is the differential mean costs divided by the differential mean QALYs between the two groups, to assess the additional costs associated with each additional QALY gained. Our base case analysis will be conducted based on subjects with complete data. Sensitivity analysis will be conducted to include subjects with incomplete data using the multiple imputation method. Non-parametric bootstrapping resampling technique will be used to derive the 95% confidence interval for the ICER. In addition, cost-effectiveness acceptability curve (CEAC) will be generated to illustrate the likelihood that one treatment is more cost-effective than the other with various ceiling cost-effectiveness ratios.

We plan to conduct supplemental analyses using alternative outcome measures, such as incremental cost per reduction in UIE/day.

The cost-effectiveness evaluations will be conducted as within-trial comparisons. A decision analytic model will also be developed from trial data to evaluate the trajectory of the cost-effectiveness ratio.

9.5.3 UDI MID

We will explore potential minimally important differences (MIDs), the smallest change that can be regarded as clinically meaningful, 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 the PGSC, patient global impression of improvement (PGI-I), and incontinence episodes from the bladder diary.

Anchor-based methods will examine the relationship between the UDI scores and independent external measure (or anchor) to elucidate the meaning of a particular change in the score. The anchor-based MID approach will evaluate the change from baseline at 6 months in the UDI score and 2 subscores corresponding to a self-reported small, but important, change on the PGSC, PGI-I and number of incontinence episodes from the diary.

- The PGSC "My current treatment is giving me adequate control of my urine leakage" is scored on a scale from 1 (strongly disagree) to 5 (strongly agree). Responses of 1, 2, or 3 are considered inadequate leakage control, responses of 4 or 5 are considered adequate leakage control. The MID will be defined as the difference between the mean change in UDI scores for patients whose PGSC rating at 6 months was "**4**" and the mean change in UDI scores for patients whose PGSC rating at 6 months was "**3**.
- The PGI-I "the one that best describes how your urinary tract condition is now, compared with how it was before treatment in this study" is scored on a scale from Very much better, Much better, A little better, No change, A little worse, Much worse, Very much worse. . The MID will be defined as the difference between the mean change in UDI scores for patients whose PGI-I rating at 6 months was "**a little better**" and the mean change in UDI scores for patients whose PGI-I rating at 6 months was "**no change**".
- Change in average incontinence episodes per day from baseline to 6 months will be calculated from the 3-day diary. "No change" in episodes will be defined for a person as a percent change in average episodes across the 3 days of **from 0 to 25%**. A "small improvement" in average episodes will range from **25% to 50%**. The MID will be defined as the difference between the mean change in UDI scores for patients with a small improvement and the mean change in UDI scores for patients with no change.

Distribution-based MID calculations will focus on variability, using the baseline standard deviation of the UDI scores. The distribution-based MID will be defined as 0.5 × baseline SD (i.e., medium effect size) and 0.2 × baseline SD (i.e., small effect size).

MID estimates of the anchor- and distribution-based approaches will be compared, and recommendations for the MID of the UDI and iterative and stress subscales will be made by consensus, consistent with the recommendations of Revicki et al. For reference, ESTEEM used the anchor-based method and the PGI-I (selected as the most clinically relevant anchor), and recommends a MID of -26 for the UDI total, -10.2 for the UDI-irritative subscale, and -5.4 for the UDI-stress subscale.(Sung et al)

The change from baseline UDI total score and stress and irritative subscores for each participant will be compared to the MIDs estimated from the MUSA data, and the percentage of participants who meet or exceed the MID will be compared between treatment groups.

10 SAFETY ANALYSES

10.1 Overview of Safety Analysis Methods

All safety analyses will be performed using the Safety population unless otherwise specified, consisting of all randomized and treated participants, summarized by the treatment actually received in the case of patients not receiving their randomized treatment. 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.

10.2 Adverse Events/Complications/Adverse Symptoms

MUSA complications are divided into "adverse symptoms" and "adverse events".

Participants were asked to report any adverse events or adverse symptoms from randomization through 12 months follow-up. All adverse events were collected on an adverse event log and coded using MedDRA (V24.0). In addition, a few adverse symptoms are captured as part of the PRO questionnaires.

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

a. Active capture for specific patient-reported adverse symptoms will be identified by responses on validated questionnaires These 3 adverse symptoms are 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, 9, and 12 months.

b. Open-ended capture for adverse symptoms will also be captured using the open-ended approach and recorded on the AE log. Symptoms of interest include: 1) New/worsening partner dyspareunia and 2) New/worsening constipation. Using the open-ended approach, symptoms are captured if reported by the participant and documented on the Adverse Event log.

Adverse events will be captured using the standard AE open-ended capture approach using chart review, physician report, or patient report corroborated by medical documentation and recorded on the Adverse Event log.

A table summarizing the Adverse Events and Adverse symptoms of interest is as follows (all variable definitions are included in Section 9.2):

Postoperative complications	
Adverse symptom	**Definition**
*New/worsening abdominal/genital pain at or beyond 3 month visit	UDI item N captured at 3, 6, 9, and 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, 9, and 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, 9, and 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

Adverse Event	
Postop need for catheter and/or ISC at or beyond 2 weeks (sling or Botox)	Captured by chart review, physician report, exam, or patient report corroborated by medical documentation
Vaginal mesh exposure (sling only)	Captured by chart review, physician report, exam, or patient report corroborated by medical documentation
Vaginal mesh erosion into organ (sling only)	Captured by chart review, physician report, exam or patient report corroborated by medical documentation
Other wound healing problems >6 weeks (sling only)	Captured by chart review, physician report, exam or patient report corroborated by medical documentation
New/worsening vaginal infection (sling or Botox)	Captured by chart review, physician report, exam, or patient report corroborated by medical documentation
Urinary tract infection beyond 2 weeks (sling or Botox)	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 or Botox)	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 or Botox)	Captured by chart review, physician report, or patient report corroborated by medical documentation
Emergency room visit due to complication related to intervention (sling or Botox)	Captured by chart review, physician report, or patient report corroborated by medical documentation
Hospital readmission related to intervention (sling or Botox)	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) (sling only)	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. 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.

Reported adverse events will be reviewed to identify terms meeting definitions for events or symptoms of interest The review will be conducted via keyword searches through reported event terms and comments 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 (compared 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. Overall number percent of patients with an SAE will be presented, and 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

The primary repeated measures mixed model analysis planned in the protocol specified a heterogeneous Toeplitz covariance pattern for the within-subject repeated measures. The SAP updates this covariance pattern to instead be unstructured separately within each treatment arm. The unstructured pattern within group is preferable for the MUSA design since the Botox arm can receive a second treatment within the first 6 months but the sling arm does not, and additional treatments are allowed in either arm after 6 months. It is therefore best to specify the most general correlation structure possible.

14 REFERENCES

Davis, S., Chapter 5. Mixed models for repeated measures with categorical time effects (MMRM). in Clinical trials with missing data : a guide for practitioners, M. O'Kelly and B. Ratitch, Editors. 2014, John Wiley & Sons Inc.: Chichester, West Sussex.

Ratitch, B., M. O'Kelly, and R. Tosiello, *Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models.* Pharm Stat, 2013. **12**(6): p. 337-47.

Ratitch, B., *Chapter 6: Multiple Imputation*, in *Clinical trials with missing data : a guide for practitioners*, M. O'Kelly and B. Ratitch, Editors. 2014, John Wiley & Sons Inc.: Chichester, West Sussex.

Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102–9.

Sung VW, et al. Effect of Behavioral and Pelvic Floor Muscle Therapy Combined with Surgery vs Surgery Alone on Incontinence Symptoms Among Women With Mixed Urinary Incontinence: The ESTEEM Randomized Clinical Trial. JAMA. 2019 Sep 17;322(11):1066-1076.

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
Demographic and Baseline Characteristics
Primary and secondary Efficacy Model Results
Exploratory Efficacy and Quality of Life Outcomes Outcome Measures
Additional Treatments received.
MID estimates for UDI
Adverse Events, Symptoms, and SAEs
Figures
Consort diagram of participant disposition
Primary and secondary measures over time by treatment arm