Reproducibility of Dynamic MRI Pelvic Measures: A Multi-Site Study

Original Research

Advances in Knowledge:

1. The research utility of MRI soft tissue and pelvimetry parameters may be limited by the high measurement variability among readers at different institutions despite standardized training.
2. Interobserver agreement for bony parameters is better than for continuous soft tissue or categorical parameters.

Implications for Patient Care

1. High variability in pelvic floor MRI measurements may limit generalizability of research results.

2. Pelvic MRI is unlikely to influence management of anal sphincter tears, as measured parameters differ only slightly between normal and abnormal patients.

ABSTRACT

PURPOSE: To assess the reproducibility of bony and soft-tissue pelvimetry measurements obtained from dynamic MRI studies in primiparous women across multiple centers.

MATERIALS AND METHODS:

All subjects were prospectively consented for participation in this IRB-approved, HIPAA-compliant study. At six clinical sites, standardized dynamic pelvic MR imaging (1.5 T multiplanar T2-weighted) was performed in three groups of primiparous women at 6-12 months postpartum: (1) vaginal delivery with anal sphincter tear; (2) vaginal delivery without anal sphincter tear; and (3) cesarean delivery without labor. After standardized central training, blinded readers at separate clinical sites and a blinded central reader measured 9 bony and 10 soft tissue pelvimetry parameters. Subsequently, three readers underwent additional standardized training, then re-read 20 MRI studies. Measurement variability was assessed by intraclass correlation for agreement between the clinical site and central readers.

RESULTS:

There was adequate agreement (intraclass correlation (ICC) range 0.71-0.93) for 8 of 19 MRI parameters on initial readings of 198 subjects. Remaining parameters had neutral or poor agreement (ICC range 0.13-0.66). Additional training reduced measurement variability: 12 of 19 parameters had adequate agreement (ICC range 0.70-0.92). Correlations were greater for bony measurements [ICC (5/9 & 8/9 variables $\geq$0.70, initial reads and re-reads, respectively] than for soft tissue measures (3/10 & 4/10, ICC$\geq$0.70).

CONCLUSION:

Despite standardized centralized training, there is high variability of pelvic MRI measurements among readers, particularly for soft tissue structures. Although slightly improved with additional training, measurement variability adversely affects utility of many MRI measurements for multi-center pelvic floor disorder research.

INTRODUCTION

Pelvic floor symptoms are common in women after childbirth (1). Objective assessment of anatomic changes and structural pathology (Figure 1) is an important adjunct in characterization of pelvic floor symptoms resulting from childbirth. Dynamic magnetic resonance imaging (MRI) has been evaluated in the US and the European Union to assess pelvic organ prolapse (2-5). Correlation with physical examination and cystocolpoproctography is variable (5-7); defecography also has high interobserver variability (8). Measurement reproducibility is important to assess, especially in the research setting.

Previously, a few small single-center series or retrospective studies have shown variable interobserver reliability in characterizing specific anatomic findings of anal sphincter and pelvic structures demonstrated by MRI (9-14). Intraobserver correlation of pelvic organ prolapse has been weak even in single site studies (12). Poor interobserver correlation of external anal sphincter atrophy on endoanal and phased array coil MRI has been recently reported (15). Better interobserver correlation has been shown in other single-center studies (10, 11), but no large multi-institutional trials have evaluated this question.

The Childbirth and Pelvic Symptoms Study (CAPS) evaluated fecal and urinary incontinence symptoms at 6 weeks and 6 months after delivery in 3 cohorts of primiparous patients: after vaginal delivery with a clinically recognized anal sphincter tear, after vaginal delivery without a clinically recognized anal sphincter tear (vaginal controls), and those who underwent cesarean delivery without labor (cesarean control) (1). The Childbirth and Pelvic Symptom Imaging Study (CAPSIS) (16) is a multi-

institutional study in which endoanal ultrasound and dynamic MR imaging was performed 6 months postpartum in a subset of patients.

The purpose of the current study is to assess reproducibility of both pelvic soft tissue and bony pelvimetry measurements across several centers.

METHODS

Study Design

Institutional review board approval was obtained at the six participating clinical sites and Data Coordinating Center (DCC). 256 primiparous subjects from the CAPS study were approached to participate in the CAPSIS study. A separate informed consent was obtained for the imaging portion of this HIPAA-compliant study.

Women in the anal sphincter tear group had a clinically diagnosed sphincter tear repaired at delivery (V-tear). One control group included women who delivered vaginally without a clinically recognized anal sphincter tear (VC). A second control group consisted of women who delivered by cesarean before onset of labor (CC). We restricted analysis to MRI studies interpreted by both the site and central reader.

MRI Initial Training and Data Acquisition

The participating radiologist from each of the 6 clinical sites attended a one-day training session with the expert consulting radiologist at the central site prior to study initiation. Training consisted of description of the desired measurements and review of measurement technique, including relevant images to be utilized, bony and soft tissue

landmarks, and use of measurement tools. In addition, participants viewed the acquisition of a pelvic MRI exam using the standardized protocol on a volunteer subject.

Study MRI exams were begun after test data sets submitted by each site were reviewed and approved for quality and protocol adherence by the expert central reader. Data collection required approximately 12 months. The "reference standard" was the expert central reader.

MRI Technique

After voiding, approximately 60 ml of inert ultrasound gel was placed into the rectum with the patient in the lateral decubitus position. The patient's position was changed to supine, and a pelvic phased array coil was placed around the lower pelvis. MR imaging was performed using a standardized protocol on 1.5T scanners. Because this study was carried out at 6 clinical sites, a variety of MRI equipment was used. Briefly, the protocol consisted of localizer images, sagittal ultra-fast T2-weighted images (TR 4400, TE 90, FOV 300, ST 10 mm, 128 x 256 matrix, NEX 1) at rest and at strain, transverse and coronal T2-weighted images at rest (TR 5000, TE 132, FOV 200, ST 3mm, 270 x 256 matrix, NEX 2), and oblique T2-weghted images parallel to the sacrum (TR 4400, TE 90, FOV 250, ST 5mm, 128 x 256 matrix, NEX 1). No intravenous contrast agents were used. Total imaging time was approximately 25 minutes.

Standardized measurements were made with electronic calipers on a workstation and recorded on standardized forms. The MRI examination was stripped of all measurements and protected health information, recorded on a compact disc with its appropriate research number as a sole identifier and sent to a central site for a second

interpretation by the expert reviewer. All completed data forms were sent to a central statistical center for analysis

MRI Retraining and Remeasurement of Subset Data

To assess whether additional measurement standardization training would improve interobserver variability, a subset of three radiologists were asked to reinterpret a mixed subgroup of MRI data sets from the initial study that had high interobserver variability. Approximately 18 months after the initial training session, each reproducibility reader underwent an additional 6 hours of interpretive training under the direction of the expert central reader, in conjunction with the project statistician. Specific pelvic MRI measurements were reviewed and practiced until satisfactory interobserver agreement had been achieved. The three radiologist "reproducibility readers" remained blinded to clinical cohort information and initial protocol outcomes.

Following completion of training, 20 selected MRI studies in compact disk format, devoid of personal health information and using new study identifiers, were sent for independent re-interpretation by the three readers at different geographic sites. The subset readings occurred over a 3-month period. Measurements were made using software embedded within each disk allowing digital caliper and angle measurements. Standardized data collection forms were again completed and sent to the central statistical center for analysis in a manner similar to the initial readings.

MRI Interpretation Parameters

On sagittal images, the pubococcygeal line was utilized to demarcate the pelvic floor. Rest and maximal strain images in the mid-sagittal plane (Figure 2) were obtained to evaluate the descent of the bladder and anorectal junction, anterioposterior length of the hiatus and angle of the levator plate with the pelvic floor. The presence or absence of rectocele defined as anterior protrusion of the rectal wall > 2 cm (Figure 3), or enterocele, defined as small bowel extending between the rectum and vagina, was recorded. Bony measurements included sacral length and depth, and obstetric conjugate.

On transverse images, width of the levator hiatus, presence or absence of muscle tears, width and signal intensity of the puborectalis muscle, and vaginal shape were reported (Figure 4). Bony measurements included intertuberous distance, interspinous distance and angle of the pelvic arch.

In the coronal plane, the shape of the iliococcygeus muscle was evaluated for loss of the normal superior bowing (Figure 5). The transverse diameter of the pelvic inlet was measured at the level of the femoral head fovea. On oblique coronal images obtained in the plane of the sacrum, the maximum transverse inlet diameter was measured again.

Thirty individual pelvic MRI measurements were made by two readers (one site reader and one from the QA reader) on each subject during the initial trial. Of the 30 measurement parameters, 22 were continuous variables and 8 were categorical. Because of inconsistencies in the definitions used by the readers, three continuous variables, distance from bladder neck to PCL with straining, angle of levator plate with PCL at rest and with straining, were omitted from this analysis. As the QA reader could not read the two measurements involving signal intensities they were also omitted from the analysis.

Difference between measurements at rest and strain were calculated for both H line and M line, creating two new variables. Therefore the analyses on initial readings include 19 continuous variables (9 bony and 10 soft tissue dimensions) and 8 categorical variables

For the subset of 20 re-read MRI's, 25 continuous and 6 categorical MRI measurements were collected by three readers. Anal sphincter integrity analysis was not re-evaluated because of poor results in the initial trial. Levator symphysis gap was defined differently during the second round of training prohibiting comparison between the two sets of readings. Six new continuous variables were evaluated, including right and left minimal gap distances if a levator symphysis gap was present, urethral angle at rest and with straining, and vagina length at rest and with straining. We compared the 19 common continuous variables between the initial readings and the rereads.

Statistical Analysis

Intraclass correlation (ICC) (17) was calculated for each parameter. The ICC can be conceptualized as the ratio of variance between images to total variance. This ratio is high when the values from readings of each image clusters in a narrow range compared to the range over which all the images are measured. A high ICC value indicates good reliability.

"SD ratio", the ratio of the standard deviation (SD) computed between the readings of the same image (within image SD) to the SD computed from all the data (which is similar to the SD between images), was also calculated. A small SD ratio indicates a good reliability. A large SD ratio indicates that variability in measurements

between readers is similar to the variability between images. The SD ratio and ICC are related: ICC is approximately 1 – ratio^2.

Since measurements were repeated on the same image, a high correlation between readers was expected. Therefore, an ICC threshold of 0.85 was considered reliable and a lower limit of 0.7 was considered acceptable.

For the 8 categorical variables in the initial trial, four are dichotomous. The others - vaginal shape, iliococcygeus contour, and two analyses of anal sphincter tears – were dichotomized for statistical analysis. Most responses regarding vaginal shape were "normal H" or "butterfly", and most iliococcygeus muscles categorized as "superiorly bowed" (Table 4a). These responses were counted as "Yes" responses, minority responses counted as "No". For the two anal sphincter tear questions, "cannot visualize" was also treated as a "No" response.

Kappa statistic was calculated (18) for all 8 dichotomous variables (Table 4b). The 6 categorical variables in the re-reads were analyzed similarly (Table 5b). Since there were more than two readers, generalized kappa (19) was calculated for rereads.

RESULTS

Demographic characteristics

Table I summarizes the demographic characteristics of the subjects. Most participants were Caucasian (133/196, 67.9%) and young (15-43 years). The three cohorts included 93 V-tear women, 79 VC women, and 26 women in the CC group.

Initial MRI Reading

Table 2 summarizes the results for the initial MRI reading. Two of 19 continuous variables had good reliability, obstetric conjugate (ICC=0.93, SD ratio=0.26) and sacral length (ICC=0.86, SD ratio=0.37). Six had acceptable reliability, including interspinous distance (ICC=0.75, SD ratio=0.51), intertuberous diameter (ICC=0.73, SD ratio=0.52), distance from bladder neck to PCL at rest (ICC=0.71, SD ratio=0.54), H line (ICC=0.77, SD ratio=0.48), difference between M line at rest and strain (ICC=0.74), and anterior-posterior outlet (ICC=0.78, SD ratio=0.47), and 11 variables had poor reliability (range from 0.13 to 0.66). The M line at rest had ICC=0.13, indicating extremely poor reliability.

Interobserver variability for measurements based upon soft tissue landmarks was greater than for bony elements. Five of 9 bony pelvimetry measurements showed reliable or acceptable interobserver correlation based on the initial training. Only 3 of 10 soft tissue measurements are considered acceptable.

There was disagreement between paired readers for the 8 categorical variables, (Table 4b) particularly for the two sphincter tear measures, with poor kappa values of -0.023 and -0.019. The other kappa values vary from 0.12 to 0.54. The small number of enteroceles in this sample precluded adequate statistical evaluation of this parameter.

Repeat MRI Readings and Outcomes

Table 3 summarizes the results for the repeat MRI readings after retraining, and Figure 6 graphically illustrates comparative ICC's and SD ratios. Among the 11 variables with unacceptable reliability in the initial trial, retraining improved the reliability of 6

variables to the acceptable level (ICC> 0.7), including width of levator hiatus, angle of pubic arch, H line with straining, difference between H line at rest and with straining, depth of sacral hollow, transverse inlet, and transverse diameter. Two other variables were also improved, although their ICC still did not reach 0.7. The reliability of the remaining three variables did not improve.

Among the 6 variables with acceptable reliability in the initial trial, retraining improved the ICC value for intertuberous diameter from 0.73 to 0.91. Two other variables, interspinous distance and distance from bladder neck to PCL at rest, also improved slightly.

The ICC values of H line with straining and the difference between M line at rest and with straining decreased slightly. Measurements of anterior-posterior outlet became less reliable in rereads (ICC decreased from 0.78 to 0.62). Obstetric conjugate and sacral length, which had good reliability in the initial trial, both had acceptable reliability in the rereads (ICC=0.75 and 0.81, respectively).

Overall, 7 measurements improved by at least one category of ICC and 2 measurements deteriorated by one category. In the rereads, 12 out 19 measurements had adequate or good rating of correlation, compared to 8 in the initial reads.

Table 3 gives the overall mean and SD for the 60 reads of the 20 MRI's. The difference between readers is demonstrated by the difference between each reader and the mean of all three readers.

The categorical variable re-reads show disagreement (generalized kappa: -0.34 ~ 0.35) among the three readers (Table 5b). As with initial readings, categorical variables such as vaginal shape or presence of levator tear continued to show poor agreement.

DISCUSSION

Bony pelvimetry measurements were more consistent than soft tissue measures at initial and repeated readings. Additional training increased measurement consistency for soft tissue parameters more than bony measurements. However, some soft tissue variables such as resting sagittal measures of hiatus (M-line and H-line) and posterior levator plate angles showed less improvement than other variables, with poor correlation despite additional training. Poor delineation of soft tissue interfaces despite optimized pelvic phased array imaging technique likely contributes to the greater interobserver variability of soft tissue parameter measurements.

Continuous parameters with large values, such as bony pelvimetry measurements, showed the highest overall agreement as a group. Parameters with small values demonstrated high variability. The relative lack of improvement for pelvimetry measures after additional training is expected given that these measures already had high consistency and therefore less room for improvement. Bony parameters tended to have better defined margins and greater contrast with adjacent soft tissue structures, particularly fat, enhancing readers' ability to produce reliable measurements. Some variability for bony measurements likely resulted from limited contrast between cortical bone and contiguous hypointense structures, for example tendons, in areas such as the ischial tuberosities.

Differences between measurements at rest and strain would be expected to show less variability, since the landmark of each static measurement should be consistent for the individual reader, including the M line measurements. This is supported by our data,

as the ICC for M line improved on assessment of the difference between rest and strain (Table 2b, final entry).

The literature regarding variability of MRI evaluation of pelvic organs is limited (9-13), underscoring the importance of assessment of measurement reliability within a diagnostic study. Even in single site studies, there may be unacceptable variability in MRI measurements. In a study of 10 volunteers, unacceptable variability was due mainly to high intraobserver variability. There was also high interobserver variability, and the authors recommended strategies to reduce sources of measurement error (12). In the current multi-institutional study, the process of training, initial interpretation, retraining, and rereads provided an excellent opportunity to evaluate the variability of pelvic MRI measurements among readers with specialized training from different institutions. Our data demonstrates that reproducibility of pelvic MRI measurement is improved by targeted training that includes clear agreement about measurement landmarks.

Different measurement software among sites, different MRI scanners, inconsistent choice of the same image for measurement among a series of slices, and variations in the understanding of image landmarks over time could each contribute to the variability. Comparison of overall variability of initial pelvic MRI measurements and repeated measurements suggests the existence of technical limitations that extend beyond training of image interpreters. Despite additional training of readers using techniques to improve interobserver reliability, there was still wide disparity in effectiveness of the additional training. Further analysis identified potential underlying reasons.

Reasonably anticipated greater interobserver consistency among binary variables compared with continuous variables was not demonstrated in our study. In general, the

categorical and binary variables showed poor correlation between readers. The inconsistency of these measurements despite additional training may be due to limitations of the technique rather than interpretive errors.

Variability by site was a significant problem. Two sites recruited predominantly African-American subjects while the others recruited mostly Caucasians, possibly accounting for some of this variability. Variations in study acquisition could also lead to differences in interpretations. An interesting finding of our analysis was persistent variability between readers on the rereads.

Limitations of our study include the lack of inclusion of all potential subjects in the imaging trial; thus, the full spectrum of primiparous women may not be represented. Another potential limitation is selection bias of the subset of MRI studies chosen for rereads. However, these studies were designated by the DCC to include a spectrum of normal versus abnormal subjects. Further, it is theoretically possible, although highly unlikely, that the rereaders remembered studies from the initial interpretation.

Finally, there was some inconsistency between readers in the definitions of measurement parameters despite additional training. Three parameters, distance from bladder neck to PCL with straining, angle of levator plate with PCL at rest and with straining, were excluded from statistical analysis mainly due to inconsistent use of positive and negative signs for the measurements thus skewing the means of the affected parameters.

In conclusion, our study demonstrated excessive variability of specific pelvic MRI measurements performed at separate institutions by different readers. These results have significant implications that may limit the utility of certain MRI measurements for the evaluation and treatment of pelvic floor dysfunction. The evolution of MRI techniques with improved distinction of landmarks and greater spatial and contrast resolution, particularly between contiguous soft tissue structures, will hopefully increase its utility in the future.

REFERENCES:

1. Borello-France D, Burgio KL, Richter HE, et al. Fecal and urinary incontinence in primiparous women. Obstet Gynecol 2006; 108(4):863-872.
2. Hodroff MA, Stolpen AH, Denson MA, Bolinger L, Kreder KJ. Dynamic magnetic resonance imaging of the female pelvis: the relationship with the Pelvic Organ Prolapse quantification staging system. J Urol 2002; 167(3):1353-1355.
3. Singh K, Jakab M, Reid WM, Berger LA, Hoyte L. Three-dimensional magnetic resonance imaging assessment of levator ani morphologic features in different grades of prolapse. Am J Obstet Gynecol 2003; 188(4):910-915.
4. Fletcher JG, Busse RF, Riederer SJ, et al. Magnetic resonance imaging of anatomic and dynamic defects of the pelvic floor in defecatory disorders. Am J Gastroenterol 2003; 98(2):399-411.
5. Healy JC, Halligan S, Reznek RH, Watson S, Phillips RK, Armstrong P. Patterns of prolapse in women with symptoms of pelvic floor weakness: assessment with MR imaging. Radiology 1997; 203(1):77-81.
6. Kelvin FM, Maglinte DD, Hale DS, Benson JT. Female pelvic organ prolapse: a comparison of triphasic dynamic MR imaging and triphasic fluoroscopic cystocolpoproctography. Am J Roentgenol 2000; 174(1):81-88.
7. Vanbeckevoort D, Van Hoe L, Oyen R, Ponette E, De Ridder D, Deprest J. Pelvic floor descent in females: comparative study of colpocystodefecography and dynamic fast MR imaging. J Magn Reson Imaging 1999; 9(3):373-377.

8. Dobben AC, Wiersma TG, Janssen LW, et al. Prospective assessment of interobserver agreement for defecography in fecal incontinence. Am J Roentgenol 2005; 185(5):1166-1172.

9. Terra MP, Beets-Tan RG, van Der Hulst VP, et al. Anal sphincter defects in patients with fecal incontinence: endoanal versus external phased-array MR imaging. Radiology 2005; 236(3):886-895.

10. Beets-Tan RG, Morren GL, Beets GL, et al. Measurement of anal sphincter muscles: endoanal US, endoanal MR imaging, or phased-array MR imaging? A study with healthy volunteers. Radiology 2001; 220(1):81-99.

11. Keller TM, Rake A, Michel SC, et al. Obstetric MR pelvimetry: reference values and evaluation of inter- and intraobserver error and intraindividual variability. Radiology 2003; 227(1):37-43.

12. Morren GL, Balasingam AG, Wells JE, Hunter AM, Coates RH, Perry RE. Triphasic MRI of pelvic organ descent: sources of measurement error. Eur J Radiol 2005; 54(2):276-283.

13. Hetzer FH, Andreisek G, Tsagari C, Sahrbacher U, Weishaupt D. MR defecography in patients with fecal incontinence: imaging findings and their effect on surgical management. Radiology 2006; 240(2):449-457.

14. Morgan DM, Umek W, Stein T, Hsu Y, Guire K, DeLancey JO. Interrater reliability of assessing levator ani muscle defects with magnetic resonance images. Int Urogynecol J Pelvic Floor Dysfunct 2007; 18(7):773-778.

15. Terra MP, Beets-Tan RG, van der Hulst VP, et al. MRI in evaluating atrophy of the external anal sphincter in patients with fecal incontinence. Am J Roentgenol 2006; 187(4):991-999.

16. Richter HE, Fielding JR, Bradley CS, et al. Endoanal ultrasound findings and fecal incontinence symptoms in women with and without recognized anal sphincter tears. Obstet Gynecol 2006; 108(6):1394-1401.

17. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979; 86:420-428.

18. Cohen J. A coefficient of agreement for nominal scales. In:Educational and Psychological Measurement, 1960; 37-46.

19. Fleiss JL. Statistical methods for rates and proportions, Second Edition. In: Wiley J, ed. New York, 1981; 38-46.

Table 1. Subject Demographics

		Mean (SD)	N (%)
Age (years)		26.6 (6.2)	
Cohort	V-tear	27.2 (6.4)	93/198 (47%)
	VC	25.8 (5.5)	79/198 (40%)
	CC	27.3 (7.4)	26/198 (13%)
Race	White		133/196 (68%)
	African American		53/196 (27%)
	Other		10/196 (5%)
Pre-pregnancy Weight (kg)		75.7 (19.5)	191
Pre-delivery Weight (kg)		93.6 (20.6)	193
BMI (pre-pregnancy) (kg/m^2)		27.8 (6.9)	189
BMI (pre-delivery) (kg/m^2)		34.5 (7.9)	189

V-tear – Vaginal Delivery with Clinically Diagnosed Anal Sphincter Disruption
VC – Vaginal Control
CC – Cesarean Control

Table 2a. Difference between site readings and QA readings in the initial trial (bony tissues)

Variables (mm)	QA (N=198) Mean (SD)	Range of QA readings	Difference between QA and Site (N=198) Mean (SD)	Range of difference	ICC	SD ratio
B11. Interspinous distance	104.21 (8.11)	85, 125	3.67 (3.53)	1.01, 6.11	0.75	0.51
B12. Intertuberous diameter	121.79 (9.5)	99.6, 146	-1.69 (5.22)	-6.52, 5.04	0.73	0.52
B13. At same level as in B12, measure angle of pubic arch with symphosis as apex	83.53 (6.96)	62, 101	1.98 (4.97)	-5.05, 5.67	0.55	0.68
C7. Obstetric conjugate (measured from sacral promontory to superior aspect of symphysis	122.34 (10.63)	99, 150.5	-0.8 (2.67)	-2, 2.33	0.93	0.26
C8. Anterior-posterior outlet (measured from last vertical joint of coccyx to inferior aspect of symphysis)	111.59 (10.72)	86.6, 147	-1.66 (4.75)	-4.17, 3.94	0.78	0.47
C9. Sacral length (measured from the sacral promontory to the tip of the coccyx)	116.66 (14.4)	74, 166	1.15 (5.27)	-1.45, 2	0.86	0.37
C10. Depth of sacral hollow (measured perpendicular from the line in C9 to the anterior context of the deepest portion of the sacral curvature)	39.68 (7.5)	17, 60.3	-2.48 (4.17)	-7.14, 0.54	0.63	0.62
D2. Transverse inlet (measured at the level of the fovea of femoral heads from inner aspect of ischial cortex on right to left)	104.79 (7.46)	85, 128.9	7.33 (6.81)	0.35, 27.59	0.29	0.9
E1. Transverse diameter (measured as maximal transverse width of pelvis along plane of sacrum)	123.34 (8.2)	103.5, 146.5	5.16 (4.22)	2.76, 10.45	0.64	0.63

Table 2b. Difference between site readings and QA readings in the initial trial (soft tissue)

Variables (mm)	QA (N=198) Mean (SD)	Range of QA readings	*Difference between QA and Site (N=198) Mean (SD)	Range of difference	ICC	**SD ratio
B3. Width of levator hiatus at superior most aspect symphysis	38.22 (5.48)	27, 56	3.78 (4.49)	0.99, 7.75	0.47	0.76
B4. Width of right levator sling muscle	4.25 (1.62)	1, 9	0.5 (1.54)	0.2, 2.49	0.4	0.78
B5. Width of left levator sling muscle	5.02 (1.95)	2, 18	0.93 (1.73)	0.63, 3.3	0.42	0.77
C1. Distance from bladder neck to PCL -- Rest	21.76 (5.21)	0, 35	-0.81 (2.74)	-1.77, 0.25	0.71	0.54
C3. H Line (leviator hiatus AP) -- Rest	43.56 (7.66)	3, 69.1	5.37 (5.33)	4.22, 8.39	0.46	0.77
C3. H Line (leviator hiatus AP) -- Strain	52.64 (11.33)	28.1, 90.4	2.17 (5.78)	0.28, 6.62	0.77	0.48
Difference between H Line rest and strain	9.01 (9.39)	-11.9, 45.5	-3.17 (5.21)	-5.17, -1.77	0.66	0.59
C4. M Line (PCL to H Line) – Rest	12.71 (5.75)	2, 28	5.77 (7.71)	-4.95, 18.4	0.13	0.98
C4. M Line (PCL to H Line) -- Strain	24.58 (12.4)	3, 65	7.52 (7.24)	1.25, 16.86	0.61	0.65
Difference between M Line rest and strain	11.84 (11.4)	-14, 54	1.83 (5.44)	-1.53, 6.2	0.74	0.51

There are two components to the variation; one component is the between-image variability and the second component is the reader measurement error (within-image variation). Poor reliability is due to the disagreement between the two readers, i.e. relatively large within-image variation comparing to the between image variation.

*Differences between the two readings provide an estimate of the reader measurement error. A positive difference means the site reading was on average higher than the QA reading and a negative difference means the opposite.

**The reported SDs for this mean difference represents the within-image variation, which are $\sqrt{2}$ ×SDs of the corresponding differences.

Table 3a. Difference between re-reads from three readers (bony tissue).

Variable (mm)		All	Reader 1	Reader 2	Reader 3	ICC	SD ratio
B11. Interspinous distance	**N**	60	20	20	20		
	Mean	108.73	0.07	1.67	-1.73	0.81	0.45
	Std	8.96	4.21	3.80	3.13		
B12. Intertuberous diameter	**N**	60	20	20	20		
	Mean	117.13	-0.18	-1.28	1.47	0.92	0.29
	Std	10.21	3.23	2.47	2.28		
B13. At same level as in B12, measure angle of pubic arch with symphosis as apex	**N**	60	20	20	20		
	Mean	85.10	-0.20	2.20	-2.00	0.72	0.54
	Std	7.54	4.29	3.36	2.99		
C7. Obstetric conjugate (measured from sacral promontory to superior aspect of symphysis	**N**	60	20	20	20		
	Mean	126.88	4.67	-2.83	-1.83	0.75	0.52
	Std	9.43	3.08	1.73	3.40		
C8. Anterior-posterior outlet (measured from last vertical joint of coccyx to inferior aspect of symphysis)	**N**	60	20	20	20		
	Mean	107.80	2.15	-3.65	1.50	0.62	0.64
	Std	7.90	4.37	3.84	3.75		
C9. Sacral length (measured from the sacral promontory to the tip of the coccyx)	**N**	60	20	20	20		
	Mean	122.25	0.60	0.45	-1.05	0.81	0.44
	Std	8.96	2.84	3.61	5.12		
C10. Depth of sacral hollow (measured perpendicular from the line in C9 to the anterior context of the deepest portion of the sacral curvature)	**N**	60	20	20	20		
	Mean	37.56	-1.16	-0.23	1.39	0.84	0.41
	Std	6.93	2.86	2.69	2.05		
D2. Transverse inlet (measured at the level of the fovea of femoral heads from inner aspect of ischial cortex on right to left)	**N**	60	20	20	20		
	Mean	105.37	0.68	-0.02	-0.67	0.82	0.43
	Std	5.35	2.11	2.57	2.10		
E1. Transverse diameter (measured as maximal transverse width of pelvis along plane of sacrum)	**N**	60	20	20	20		
	Mean	125.40	1.10	-1.00	-0.10	0.91	0.31
	Std	7.12	1.77	2.05	2.15		

Table 3b. Difference between re-reads from three readers (soft tissue).

Variable (mm)		All	Reader 1	Reader 2	Reader 3	ICC	SD ratio
B3. Width of levator hiatus at superior most aspect symphysis	**N**	60	20	20	20		
	Mean	35.15	-2.45	1.55	0.90	0.76	0.52
	Std	6.89	3.45	2.81	2.12		
B4. Width of right levator sling muscle	**N**	60	20	20	20		
	Mean	5.34	1.73	-0.46	-1.28	0.30	0.88
	Std	2.84	2.45	1.86	1.56		
B5. Width of left levator sling muscle	**N**	60	20	20	20		
	Mean	6.25	1.56	-0.74	-0.82	0.61	0.65
	Std	2.97	1.82	1.21	1.10		
C1. Distance from bladder neck to PCL -- Rest	**N**	60	20	20	20		
	Mean	21.61	0.30	0.84	-1.14	0.77	0.49
	Std	4.29	1.78	1.98	1.91		
C3. H Line (leviator hiatus AP) -- Rest	**N**	60	20	20	20		
	Mean	50.05	6.50	-4.00	-2.50	0.43	0.83
	Std	7.70	3.16	2.75	2.85		
C3. H Line (leviator hiatus AP) -- Strain	**N**	60	20	20	20		
	Mean	57.77	4.03	-3.82	-0.22	0.70	0.57
	Std	9.27	3.74	3.01	3.93		
Difference between H Line rest and strain	**N**	60	20	20	20		
	Mean	7.72	-2.47	0.18	2.28	0.78	0.49
	Std	8.09	3.70	3.31	2.50		
C4. M Line (PCL to H Line) – Rest	**N**	60	20	20	20		
	Mean	19.58	-5.93	-2.53	8.47	0.24	0.96
	Std	9.32	4.78	4.99	5.08		
C4. M Line (PCL to H Line) -- Strain	**N**	60	20	20	20		
	Mean	34.38	-1.53	0.72	0.96	0.49	0.76
	Std	12.44	5.59	5.69	6.23		
Difference between M Line rest and strain	**N**	60	20	20	20		
	Mean	14.80	0.00	0.00	0.00	0.68	0.58
	Std	10.61	6.83	6.22	5.68		

Table 4a. Summary of categorical variables in initial trials.

Parameter			Reader		
			All	Site	QA
B6. Is there a gap or tear of the levator sling present?	Yes	**N**	35	27	8
		%	9.00	13.57	4.21
B7.Vaginal Shape	Normal H or butterfly	**N**	292	143	149
		%	75.06	71.86	78.42
	Flattened	**N**	51	19	32
		%	13.11	9.55	16.84
	U shape, concave anterior	**N**	29	23	6
		%	7.46	11.56	3.16
	Asymmetric within sling	**N**	17	14	3
		%	1.75	5.56	0
B8. Is there a gap at the sling insertion to symphysis present?	Yes	**N**	38	23	15
		%	9.77	11.56	7.89
B9. Is there an internal anal sphincter tear	Yes	**N**	12	9	3
		%	3.02	4.50	1.52
	No	**N**	377	186	191
		%	94.72	93.00	96.46
	Cannot visualize	**N**	9	5	4
		%	2.26	2.50	2.02
B9. Is there an internal anal sphincter tear	Missing	**N**	2	0	2
		%	0.50	0	1.01
	Yes	**N**	20	18	2
		%	5.03	9.00	1.01
	No	**N**	369	179	190
		%	92.71	89.50	95.96
	Cannot visualize	**N**	7	3	4
		%	1.76	1.50	2.02
C5. Rectocele	Yes	**N**	78	38	40
		%	20.05	19.10	21.05
C6. Enterocele	Yes	**N**	2	1	1
		%	0.51	0.50	0.53
D1. Contour of	Bowed superiorly	**N**	348	173	175

iliococcygeus		**%**	89.46	86.93	92.11
	Flat	**N**	40	26	14
		%	10.28	13.07	7.37
	Bowed inferiorly	**N**	1	0	1
		%	0.26	0	0.53

The SDs listed under the mean differences are calculated as $\sqrt{3/2}$ ×SDs of the corresponding differences, which represents the within-image variation.

Table 4b. Summary of categorical variables in initial trials.

Parameters		NN+	NY+	YY+	Kappa
B6. Is there a gap or tear of the levator sling present?	**N**	165	29	3	0.12
	%	83.8	14.7	1.5	
B7.Vaginal Shape*	**N**	18	58	122	0.31
	%	9.1	29.3	61.6	
B8. Is there a gap at the sling insertion to symphysis present?	**N**	163	26	6.000	0.25
	%	83.6	13.3	3.1	
B9. Is there an internal anal sphincter tear#	**N**	186	12	0	-0.023
	%	93.9	6.1	0	
B10. Is there an external anal sphincter tear#	**N**	177	19	0	-0.019
	%	90.3	9.7	0	
C5. Rectocele	**N**	142	29	25	0.54
	%	72.4	14.8	12.8	
C6. Enterocele	**N**	193	2	0	NA
	%	99.0	1.0	0	
D1. Contour of iliococcygeus&	**N**	11	17	166	0.52
	%	5.7	8.8	85.6	

* B7: "Normal H or butterfly" is considered to be Y, "Flattened", "U shape, concave anterior", and "Asymmetric within sling" are considered to be N.
# B9 and B10: "Cannot visualize" is considered to be N.
&D1: "Bowed superiorly" is considered to be Y, "Flat" and "Bowed inferiorly" is consider to be N.
+ Potential combined answers to “Yes” or “No” questions from 2 different readers

Table 5a. Summary of categorical variables in rereads

Parameter			Reader			
			All	1	2	3
B6. Is there a gap or tear of the levator sling present?	**Yes**	**N**	6	2	3	1
		%	10.53	11.11	15.00	5.26
B7.Vaginal Shape	**Normal H or butterfly**	**N**	45	14	19	12
		%	78.95	77.78	95.00	63.16
	Flattened	**N**	5	2	0	3
		%	8.77	11.11	0	15.79
	U shape, concave anterior	**N**	6	1	1	4
		%	10.53	5.56	5.00	21.05
	Asymmetric within sling	**N**	1	1	0	0
		%	1.75	5.56	0	0
B8. Is there a gap at the sling insertion to symphysis present?	**Yes**	**N**	42	18	20	4
		%	73.68	100.00	100.00	21.05
C5. Rectocele	**Yes**	**N**	13	4	7	2
		%	22.81	22.22	35.00	10.53
C6. Enterocele	**Yes**	**N**	0	0	0	0
		%	0	0	0	0
D1. Contour of iliococcygeus	**Bowed superiorly**	**N**	48	15	14	19
		%	84.21	83.33	70.00	100.00
	Flat	**N**	9	3	6	0
		%	15.79	16.67	30.00	0
	Bowed inferiorly	**N**	0	0	0	0
		%	0	0	0	0

Table 5b. Summary of categorical variables in rereads

Parameters		NNN	NNY	YYN	YYY	kappa
B6. Is there a gap or tear of the levator sling present?	**N**	14	1	5	0	0.030
	%	70.0	5.0	25.0	0	
B7.Vaginal Shape*	**N**	0	6	2	11	0.21
	%	0	31.6	10.5	57.9	
B8. Is there a gap at the sling insertion to symphysis present?	**N**	0	13	0	4	-0.34
	%	0	76.5	0	23.5	
C5. Rectocele	**N**	11	1	6	2	0.35
	%	55.0	5.0	30.0	10.0	
C6. Enterocele	**N**	20	0	0	0	NA
	%	100.0	0	0	0	
D1. Contour of iliococcygeus[&]	**N**	0	3	3	14	0.22
	%	0	15.0	15.0	70.0	

* B7: "Normal H or butterfly" is considered to be Y, "Flattened", "U shape, concave anterior", and "Asymmetric within sling" are considered to be N.
[&]D1: "Bowed superiorly" is considered to be Y, "Flat" and "Bowed inferiorly" is consider to be N.

CAPTIONS FOR ILLUSTRATIONS

Figure 1. Transverse T2-weighted MRI shows disruption of the right levator muscle (arrow) with lack of normal continuity to the levator symphysis.

Figure 2. Cystocele. Dynamic sagittal T2-weighted image (a) at rest and (b) during valsalva show the abnormal descent of the urinary bladder neck (arrow) 2 cm below the pubococcygeal line (PCL), consistent with cystocele. On 2a, the H-line (H) and M-line (M) are included for illustrative purposes.

Figure 3. Rectocele. Sagittal T2-weighted image demonstrates convex bowing of the anterior rectal wall > 2 cm (arrows).

Figure 4. Levator muscle thickness and signal intensity. Transverse T2-weighted images show normal thickness of the puborectalis muscles (arrows) with normal signal intensity (ROI circle).

Figure 5. Coronal T2-weighted image through the rectum shows abnormal straightening of the right puborectalis muscle (arrow).

Figure 6. Graphical Summary of Comparison of Initial Readings and Rereads. Intraclass Correlations and Ratio of reader standard deviations are shown for each parameter. Description for each parameter numbered in the figure is listed in tables 2-5.

Figure 1. Transverse T2-weighted MRI shows disruption of the right levator muscle (arrow) with lack of normal continuity to the levator symphysis.

Figure 2. Cystocele. Dynamic sagittal T2-weighted image (a) at rest and (b) during valsalva show the abnormal descent of the urinary bladder neck (arrow) 2 cm below the pubococcygeal line (PCL), consistent with cystocele. On 2a, the H-line (H) and M-line (M) are included for illustrative purposes.

Figure 3. Rectocele. Sagittal T2-weighted image demonstrates convex bowing of the anterior rectal wall > 2 cm (arrows).

Figure 4. Levator muscle thickness and signal intensity. Transverse T2-weighted images show normal thickness of the puborectalis muscles (arrows) with normal signal intensity (ROI circle).

Figure 5. Coronal T2-weighted image through the rectum shows abnormal straightening of the right puborectalis muscle (arrow).

Figure 6. Graphical Summary of Comparison of Initial Readings and Rereads. Intraclass Correlations and Ratio of reader standard deviations are shown for each parameter. Description for each parameter numbered in the figure is listed in tables 2-5.