Published in final edited form as:
Int Urogynecol J Pelvic Floor Dysfunct. 2007 July ; 18(7): 773–778. doi:10.1007/s00192-006-0224-5.

Interrater reliability of assessing levator ani muscle defects with magnetic resonance images

Daniel M. Morgan
Pelvic Floor Research Group, Department of Obstetrics and Gynecology, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109−0276, USA e-mail: morgand@umich.edu

Wolfgang Umek
Pelvic Floor Research Group, Department of Obstetrics and Gynecology, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109−0276, USA

Department of Obstetrics and Gynecology, Medical University Vienna, Vienna, Austria

Tamara Stein
Pelvic Floor Research Group, Department of Obstetrics and Gynecology, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109−0276, USA

Division of Anatomical Sciences, University of Michigan, Ann Arbor, MI, USA

Yvonne Hsu,
Pelvic Floor Research Group, Department of Obstetrics and Gynecology, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109−0276, USA

Kenneth Guire, and
Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

John O. L. DeLancey
Pelvic Floor Research Group, Department of Obstetrics and Gynecology, University of Michigan, 1500 E. Medical Center Drive, Ann Arbor, MI 48109−0276, USA

Abstract

The objective of this study is to determine interrater reliability of assessing pubovisceral levator ani muscle defects with magnetic resonance images. Normal pubovisceral muscle was assigned a grade of 0; PVM defects were graded as mild=1 (less than half missing), moderate=2 (more than half missing), and severe=3 (total or near total loss). Among six pairs of examiners, percent agreement and weighted kappa coefficients were calculated to determine agreement between pairs of examiners and among all examiners (i.e., “overall”). For unilateral scoring, exact agreement was found in 83.7%, and differences of one, two, and three grades were found in 14.7, 1.5, and 0.1%, respectively. For bilateral scoring, exact agreement and differences of one, two and three grades were found in 75.4, 15.9, 6.9, and 1.6%, respectively. Thus, exact agreement or a one-point difference was reached in 91.3% of cases. When defect status was categorized as none/normal, minor, and major, the overall weighted kappa coefficient was 0.86 (95% CI 0.83, 0.89). There was variation among examiner pairs with unilateral (p=0.002) and bilateral (p=0.02) scoring, but not when defect status was categorized as none/normal, minor, and major (p=0.59). There was agreement to within one point in 91% of cases when six examiner pairs scored levator ani defects on a seven-point scale. Examiner pairs discriminated injury similarly when defect status was categorized as normal/none, minor, or major.

Correspondence to: Daniel M. Morgan.

Keywords

Levator ani; Pelvic organ prolapse; Urinary incontinence; Magnetic resonance imaging; Muscle defects; Interrater reliability

Introduction

Women with pelvic floor dysfunction can have abnormal appearing levator muscles with a loss of muscle thickness and gaps between muscle and the public bone [1]. Visible defects in the levator ani muscles have been found after vaginal birth [2,3]. A study of recently delivered primiparas and nulliparas revealed that the pubovisceral muscle (PVM) was the most often injured portion of the levator ani muscle group [3]. A defect scoring system was developed to describe these injuries seen with magnetic resonance (MR) imaging, and a relationship was subsequently found between obstetrical dystocia and the severity of PVM defects [4].

This scoring system focuses specifically on the severity of a defect, not the amount of muscle present. This is important because the thickness of normal muscles varies among nulliparas [5], and it is likely that vaginal birth leads to even greater variation among parous women. For instance, a woman with a very large PVM could develop a defect after delivery, yet still have more muscle than a woman with a small PVM and no defect. This system, developed to categorize those abnormalities found in primiparas, but not in nulliparas, assesses birth damage separately from variation in muscle bulk.

For any scoring system, it is necessary to know how consistently different raters can assess scores. The purpose of this report is to evaluate interrater reliability among five different examiners in assessing the integrity and severity of levator ani PVM defects.

Materials and methods

Study subjects

Women were recruited under IRB-approved protocols for case control studies of pelvic organ prolapse between November 2000 and October 2004 and stress urinary incontinence (SUI) between February 2003 and November 2005. The prolapse study had interpretable MR scans for 137 women with prolapse and 134 women with normal support. Nine of the 280 scans (3.3%) in the prolapse study were not scored due to either motion artifact or the axial sequence was inadvertently omitted by the technician. The urinary incontinence cohort had interpretable MR scans for 78 women with daily SUI and 77 continent women. Three of the 158 scans (1.9%) in the incontinence study were not scored—all due to the inadvertent omission of the axial study sequence. Patients were recruited from the Urogynecology Clinic at our institution, and controls were recruited by advertisement in the local communities. The controls were recruited to be of similar age, race, and hysterectomy status. All data were prospectively collected, including a Pelvic Organ Prolapse-Quantitation [6].

Inclusion criteria for the studies were as follows. To be considered as a prolapse case, a portion of the vaginal wall or the cervix had to be at least 1 cm or more below hymen. To be a normal support control, all areas of vaginal support had to be at least 1 cm above the hymen and subjects had to demonstrate continence during a prestudy full bladder stress test. The criteria for the stress urinary incontinence study differed from the prolapse study. The SUI patients had to report daily symptoms of stress urinary incontinence and demonstrate SUI in the standing position with a full bladder. The continent controls could not report symptoms of urinary incontinence and had to demonstrate continence during a prestudy standing full bladder stress

test. For both SUI patients and continent controls, no area of vaginal support could be more than 1 cm below the hymen.

Exclusion criteria for the studies were similar. Women who had undergone hysterectomy were eligible if the surgery had been done at least 1 year before enrollment and was not performed for symptoms of pelvic floor dysfunction (e.g., pelvic organ prolapse, urinary incontinence, or fecal incontinence). Women who had had surgery for pelvic floor dysfunction and who were diagnosed with a cancer within a year could not participate. Controls were also excluded if they reported symptoms of stress urinary incontinence, demonstrated SUI during a prestudy standing, full bladder stress test, had a history of radiation therapy, or had a history of genital anomalies.

Imaging

All subjects underwent a MR imaging scan using our established protocol [6]. This included axial, sagittal, and coronal two-dimensional fast spin proton density MR scans (echo time: 15 ms, repetition time 4,000 ms) obtained in the supine position with a 1.5-T superconducting magnet (Signa; General Electric Medical Systems, Milwaukee, WI). The slice thickness was 4 mm with a slice gap of 1 mm, yielding an image spacing of 5 mm using a 160×160 mm field of view and an imaging matrix of 256×256. Frequency encoding of the scan sequence was set in an anterior to posterior direction in all scans to avoid lateral asymmetric chemical shift.

Grading levator ani muscles

The grading system developed in our earlier work was used [4]. This system assesses the integrity of the PVM, which includes both the pubococcygeus and puborectalis muscles. Different grades of unilateral defects at the level of the midurethral 1 cm above the arcuate pubic ligament [7] are shown in the axial images in Fig. 1. This image location was chosen to demonstrate the integrity of the PVMs because it is the slice where they are often best evaluated. However, it is important to remember that this system requires that multiple images in both the axial and coronal planes be evaluated to assign a score.

The left and right PVMs were graded separately as follows:

Grade 0 normal PVM

Grade 1 less than half the PVM missing

Grade 2 more than half the PVM missing

Grade 3 total or near total loss of PVM

The scores were then analyzed as a unilateral score (0−3), a four-point scale, and as a bilateral score (0−6), a seven-point scale, in which the score of the two PVMs are summed. Using the seven-point bilateral scores, PVM defect status was then further categorized as follows:

1. *None*/*Normal* No defects observed and bilateral score equal to zero.
2. *Minor* Partial injuries to one or both sides with a bilateral score 1−3 (except for a complete grade 3 unilateral injury that is considered “major”).
3. *Major* Bilateral scores 4−6 or complete grade 3 unilateral injury.

Scoring of the MR scans was conducted by five observers including the principal investigator. All observers were blinded to case or control status and to one another's scores. Before participating as a rater, an observer was involved in group discussions about defect severity and had reviewed at least 50 scans before being included in the study to eliminate training effects. Six pairs of examiners had evaluated more than 20 scans in common, and before data analysis, it was decided that this represented the critical number necessary for a pair of

examiners to be included. In the prolapse study, the principal investigator and four other observers were involved in grading MR scans. In the SUI study, the first author (DMM) and the principal investigator (JOLD) were the only two observers involved in grading MR scans.

Statistical analysis

In assessing interrater reliability, we report percentage of exact agreement, the percentage of agreement within one scale unit, and weighted kappa statistics that use the Cicchetti–Alison weights [8]. The weighted kappa statistic describes agreement between two raters with a range of scores between 0 and 1. Exact agreement results in a score of 1, complete disagreement results in a score of 0, and discrepancies between these extremes are weighted to reflect the amount of partial agreement. Weighted kappa statistics were calculated for the six rater pairs in the prolapse population and for the one rater pair involved in the SUI study. The overall weighted kappa statistics were calculated to summarize the results of raters in the prolapse and SUI cohorts and their respective control populations. The test of equal kappa coefficients among the rater pairs was used to determine if pairs differed in their ability to agree on the severity of muscle injury. This test was performed for unilateral and bilateral scoring and the categorization of muscle injury into none vs minor vs major injury. All computations were carried out using PROC FREQ in SAS.

Results

The demographics of the cohorts are shown in Table 1 with successful matching for age, race, and hysterectomy status. There was a small difference in parity among the prolapse cohort despite attempts to recruit a control population of similar parity because of the strong effect of parity on prolapse [9]. There was a difference in BMI among the SUI cohort and its control population, but the control population was not recruited to match with respect to BMI.

The results of the six examiner pairs for unilateral scoring are shown in Table 2. The range of exact agreement among examiner pairs was 76−89% and agreement within one unit or grade was present in 95−100% of PVM assessments. The weighted kappa statistics range from 0.85 to 0.97. The overall weighted kappa for all muscle assessments was 0.87 (95% CI 0.85−0.88). There was variation among examiner pairs in the ability to agree (test for equal kappa coefficients, p=0.002).

The results of the six examiner pairs for bilateral scoring are shown in Table 3. The range of exact agreement was 61−84% and agreement within one unit or grade was present in 87−94% of cases. Weighted kappas for the six examiner pairs range from 0.78 to 0.90. The overall weighted kappa for all scans was 0.86 (95% CI 0.83−0.88). There was variation among examiner pairs in the ability to agree (test for equal kappa coefficients, p=0.02).

The overall summary weighted kappa statistics for unilateral and bilateral scoring for the prolapse cohort, the normal support control group, the SUI cohort, and the continent control group are shown in Table 4.

The analysis of PVM defect status by none, minor, and major was performed by categorizing bilateral scores as previously described. The weighted kappas among the six examiner pairs ranged from 0.80 to 0.88, and the overall weighted kappa was 0.86 (95% CI, 0.83−0.89). Examiner pairs did not demonstrate variation in the ability to agree on PVM defect status (test for equal kappa coefficients, p= 0.59).

Discussion

We found high levels of agreement among several pairs of examiners in assessing MR scans for defects in the pubovisceral levator ani muscle. Agreement within one grade was present in 95−100% of cases with unilateral scoring and in 87−94% of cases with bilateral scoring. This was possible in a large number of scans with many different raters in each of two different types of pelvic floor dysfunction. The lack of complete unanimity is not unexpected. The degree of muscle loss present in a group of women is a continuous spectrum from none to total loss. For example, a woman may have muscle loss halfway between two grades, and each observer must make a decision in scoring them. Modifying the system to have more grades, however, exceeds the ability to assess these fine nuances. As with other clinical grading systems (e.g., Pap smear grading), a balance must be struck between too many and too few grades.

The grading system was developed among the research team to describe the range of defects observed. Normal muscles and complete absence of muscles were initially appreciated [3]. With further experience, it became apparent that there were partial muscle injuries, leading to the differentiation of grade 1 and 2 muscle defects. The categorization of injuries based on bilateral scoring into normal, minor, and major defects was subsequently developed to characterize the muscle status of both sides for an individual. By grouping patients into one of three categories, the analysis of injury patterns is subject to less variability between pairs of examiners. It is interesting to note that the weighted kappas of none/minor/major defects analysis are the same as that for unilateral and bilateral scoring. This occurs because of the statistical properties of kappa coefficients. When less variation is possible (i.e., the scale is reduced from a seven- to a three-point scale), the likelihood of chance agreement increases and less partial credit is given for partial agreement in the weighted score.

Our experience with this system supports the validity of assessing levator ani muscle integrity. In a recent study, the severity of defects present was associated with obstetrical events which affect second-stage dystocia. Women with major degrees of defect were more likely than women without defects to have had a forceps delivery, a sphincter laceration, or a prolonged second stage. Findings in women with minor defects were intermediate between findings of normal women and those with major injuries [4]. Ongoing research has shown that the occurrence of defects is similar among patients with prolapse and provides further validation of this system [10].

This study suggests that the muscle grading technique is learnable and performs at least as well as other useful evaluation systems. Examiners ranged in experience from fellows in female pelvic medicine and reconstructive surgery to a senior investigator with extensive experience in anatomy and interpretation of pelvic imaging. Even with this range of experience, multiple pairs of examiners demonstrated reasonable levels of concordance in evaluation. These results compare favorably with other clinically useful evaluations. For example, in classification of Pap smears using the CIN I, II, and III system, there was exact agreement in only 35% cases and partial agreement in another 35% cases [11]. In grading differentiation of endometrial carcinomas, pathologists have reported agreement in 70% of cases [12].

As mentioned in the “Introduction,” this grading system differs from strategies intended to quantify muscle volume or thickness. Techniques are available to measure muscle thickness [1], muscle cross-sectional area [13], and the gap between the muscle and the pubic bone [14]. All of these approaches can be complimentary. For example, cross-sectional area and thickness could be expected to correlate with muscle force, but defect status might be better to assess the muscle injury patterns in pelvic floor dysfunction populations.

A current limitation of this system is that it was developed among a single research group. The examiners were all trained by the senior author and this could enhance agreement among

examiners. However, this system was developed with the study of a large image library of MR studies, and all possible scans from these two case control studies were eligible for inclusion in the study. The number of scans that were not interpretable was less than 3%. With electronic storage of the files, it is now possible to have others learn this system both as additional validation and as a means to expand research into pelvic floor dysfunction.

A simple levator ani grading system, which can readily be learned and can demonstrate consistency among examiners, is an important tool in the study of pelvic floor dysfunction. Although there are many competing hypotheses concerning the cause of pelvic organ prolapse, objective assessments that can be carried out by multiple examiners blinded to subject status have real value in testing these hypotheses. It is not a question of whether muscle or connective tissue assessment is important, but how often each type of impairment is found and how they interact to cause disease. Determining whether or not muscle damage is present will help to define the roles of connective tissue and muscle abnormalities by identifying and selecting women for study.

Acknowledgements

We gratefully acknowledge support for this research through the Office for Research on Women's Health and NICHD's SCOR on Sex and Gender Factors Affecting Women's Health P50 HD044406, NICHD grant R01 38665, and NICHD 2K12HD001438-06.

References

1. Hoyte L, Jakab M, Warfield SK, Shott S, Flesh G, Fielding JR. Levator ani thickness variations in symptomatic and asymptomatic women using magnetic resonance-based 3-dimensional color mapping. Am J Obstet Gynecol 2004;191:856–861. [PubMed: 15467553]
2. Dietz HP, Lanzarone V. Levator trauma after vaginal delivery. Obstet Gynecol 2005;106:707–712. [PubMed: 16199625]
3. DeLancey JO, Kearney R, Chou Q, Speights S, Binno S. The appearance of levator ani muscle abnormalities in magnetic resonance images after vaginal delivery. Obstet Gynecol 2003;101:46–53. [PubMed: 12517644]
4. Kearney R, Miller JM, Ashton-Miller JA, DeLancey JO. Obstetric factors associated with levator ani muscle injury after vaginal birth. Obstet Gynecol 2006;107:144–149. [PubMed: 16394052]
5. Tunn R, Delancey JO, Howard D, Ashton-Miller JA, Quint LE. Anatomic variations in the levator ani muscle, endopelvic fascia, and urethra in nulliparas evaluated by magnetic resonance imaging. Am J Obstet Gynecol 2003;188:116–121. [PubMed: 12548204]
6. Bump RC, Mattiasson A, Bo K, Brubaker LP, DeLancey JO, Klarskov P, Shull BL, Smith AR. The standardization of terminology of female pelvic organ prolapse and pelvic floor dysfunction. Am J Obstet Gynecol 1996;175:10–17. [PubMed: 8694033]
7. Chou Q, DeLancey JO. A structured system to evaluate urethral support anatomy in magnetic resonance images. Am J Obstet Gynecol 2001;185:44–50. [PubMed: 11483902]
8. SAS Institute Inc. SAS/STAT User's Guide. Cary, NC: 1999.
9. Mant J, Painter R, Vessey M. Epidemiology of genital prolapse: Observations from the oxford family planning association study. Br J Obstet Gynaecol 1997;104:579–585. [PubMed: 9166201]
10. DeLancey JOL, Kearney R, Umek WH, Ashton-Miller JA. Levator ani muscle structure and function in women with prolapse compared to women with normal support. J Pelvic Med Surg 2003;9(5):201.
11. Young NA, Naryshkin S, Atkinson BF, Ehya H, Gupta PK, Kline TS, Luff RD. Interobserver variability of cervical smears with squamous-cell abnormalities: a Philadelphia study. Diagn Cytopathol 1994;11:352–357. [PubMed: 7895574]
12. Scholten AN, Smit VT, Beerman H, van Putten WL, Creutzberg CL. Prognostic significance and interobserver variability of histologic grading systems for endometrial carcinoma. Cancer 2004;100:764–772. [PubMed: 14770433]

13. Chen L, Hsu Y, Ashton-Miller JA, DeLancey JO. Measurement of the pubic portion of the levator ani muscle in women with unilateral defects in 3-D models from MR images. Int J Gynaecol Obstet 2006;92:234–241. [PubMed: 16442111]

14. 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;188:910–915. [PubMed: 12712085]

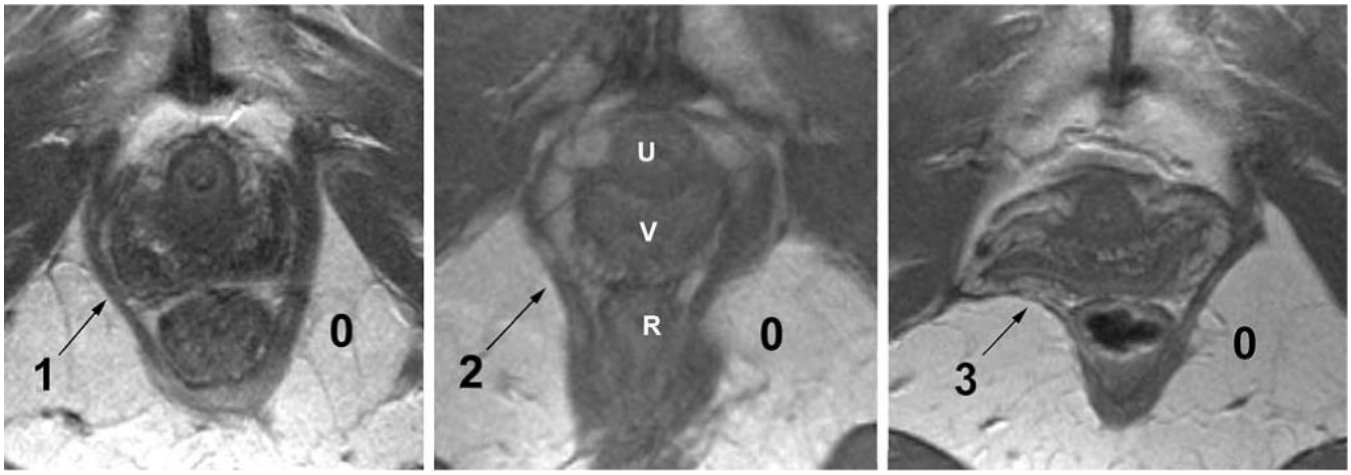

Fig. 1.

Examples of grade 1, 2, and 3 unilateral defects in axial images at the level of the midurethra. The urethra (*U*), vagina (*V*), and rectum (*R*) are labeled in the *middlepanel* of the figure. The *blackarrow* in each panel points to the left pubovisceral muscle in which there is loss of muscle bulk in comparison to the normal contralateral side

Table 1

Subject demographics

	Prolapse cases (*n*=137)	**Normal support controls (*n*=134)**	**SUI cases (*n*=78)**	**Continent controls (*n*=77)**
Age (years±SD)	56.5±12.7	56.5±13.0	47.3 (9.6)	47.8 (11.1)
Race (% white)	91.6%	93.0%	91.5%	96.5%
Hysterectomy (% yes)	19.9%	20.8%	9.8%	6.9%
BMI (kg/m^2)	26.1±4.7	26.5±4.8	31.1 (9.1)**	27.5 (5.4)**
Vaginal births (mean±SD)	2.9±3.2*	2.5±3.0*	2.3 (1.1)	2.3 (1.0)

* p=0.016

** p<0.01

Table 2

Interrater agreement for unilateral scoring

Rating pair	Number of sides	Exact agreement	Agreement within 1 unit	Weighted kappa
A vs B	496	82% (407)	97% (482)	0.85
A vs C	386	87% (337)	99% (384)	0.93
A vs D	138	84% (116)	984% (135)	0.92
A vs E	64	89% (57)	98% (63)	0.97
B vs C	46	78% (36)	95% (44)	0.91
B vs D	112	76% (85)	100% (112)	0.88
Overall	1,242	83.7% (1,038)	98.4% (182)	0.87

Table 3

Interrater agreement for bilateral scoring

Rating pair	Number of scans	Exact agreement (*n*)	Agreement within 1 unit (*n*)	Weighted kappa
A vs B	248	74.6% (185)	89.1% (221)	0.85
A vs C	193	79.8% (154)	94.8% (183)	0.93
A vs D	69	76.8% (53)	91.3% (63)	0.93
A vs E	32	84.4% (27)	90.7% (29)	0.98
B vs C	23	65.2% (15)	86.9% (20)	0.92
B vs D	56	60.7% (34)	91.1% (51)	0.88
Overall	621	75.3% (468)	91.2% (567)	0.86

Table 4

Weighted kappa statistic for defect grading among prolapse cases and controls and among SUI cases and controls

	Prolapse (*n*=137)	**Control (*n*=134)**	**SUI (*n*=78)**	**Control (*n*=77)**
Unilateral scoring	0.88	0.83	0.75	0.72
Bilateral scoring	0.88	0.82	0.77	0.70