Int Urogynecol J (2010) 21:1005–1009
DOI 10.1007/s00192-010-1139-8

ORIGINAL ARTICLE

"Eyeball" POP-Q examination: shortcut or valid assessment tool?

Deborah R. Karp · Thais V. Peterson · Marjorie Jean-Michel · Roger Lefevre · G. Willy Davila · Vivian C. Aguilar

Received: 27 January 2010 / Accepted: 5 March 2010 / Published online: 4 May 2010

Abstract
Introduction and hypothesis The objective of this study was to compare the results of the Pelvic Organ Prolapse Quantification (POP-Q) examination by visual estimation to measurement.
Methods Women with pelvic organ prolapse underwent both "eyeball"/estimated and measured POP-Q examinations by two trained examiners in a randomized order. POP-Q points and stage were analyzed using the paired *t* test, chi-square, Pearson's correlation, and kappa statistics.
Results Fifty subjects had a mean age of 60, mean BMI 27.8, and median parity of 2. The POP-Q stages by the measured technique were 18% (9/50) stage 1, 38% (19/50) stage 2, 44% (22/50) stage 3, and 0% (0/50) stage 4. The POP-Q stages based on estimation and measurement were highly associated ($p<0.05$). Individual points did not differ significantly between the techniques and did not differ significantly between examiners (all $p > 0.05$).
Conclusion Among examiners who routinely perform POP-Q examinations, there is no significant difference between "eyeball"/estimated and measured POP-Q values and stage.

Keywords Pelvic organ prolapse · Pelvic Organ Prolapse Quantification system · Evaluation

Presented at the American Urogynecologic Society Annual Meeting, Hollywood, Florida, September 24–26, 2009.

D. R. Karp · T. V. Peterson · M. Jean-Michel · R. Lefevre · G. W. Davila · V. C. Aguilar
Department of Gynecology, Section of Urogynecology and Reconstructive Pelvic Surgery, Cleveland Clinic Florida, Weston, FL, USA

V. C. Aguilar (✉)
2950 Cleveland Clinic Boulevard,
Weston, FL 33331, USA
e-mail: aguilav@ccf.org

Introduction

The Pelvic Organ Prolapse Quantification (POP-Q) was introduced in 1996 as an objective and precise system for the description of pelvic floor anatomy in women with pelvic floor disorders [1]. Since its introduction, it has been adopted by multiple professional societies including the American Urogynecologic Society, the Society for Gynecologic Surgeons, and the International Continence Society. It has become a useful clinical and research tool for the longitudinal evaluation of women with prolapse and their treatment outcomes and allows accurate communication of pelvic floor anatomy among pelvic floor surgeons.

There have been several studies published in the gynecologic literature on the validation of the POP-Q system [2, 3]. These studies have demonstrated excellent intra-examiner and inter-examiner reliability and reproducibility. Yet, as originally described, the POP-Q system makes no specific recommendations for variables such as patient position, type of exam table or chair, degree and type of strain, and method by which quantitative measures should be made. Subsequent research has shown that such variables influence POP-Q results [4–6].

Despite its precision, reproducibility, and reliability, the POP-Q system is underutilized in clinical practice. It has been estimated to be used by 40% of pelvic floor surgeons in clinical practice and 60% for research purposes [7]. Reasons given for this low utilization rate include perceptions that the POP-Q is too time-consuming, complicated and confusing, difficult to learn, and, by some, to have no clinical relevance [7]. Thus, there has been an interest in simplifying the POP-Q system to make it more practical and efficient for clinical and research applications [8, 9].

Because of the suggestion that in an effort to save time clinicians modify the exam by estimating POP-Q points as

opposed to measuring them, this study was performed with the primary objective of comparing the POP-Q stage and its individual points obtained between a standard measured POP-Q examination and an estimated technique.

Materials and methods

Between April and November 2008, 50 consecutive women presenting to the Section of Urogynecology and Pelvic Reconstructive Surgery at Cleveland Clinic Florida with a primary complaint of pelvic organ prolapse were consecutively recruited to participate in this Institutional Review Board-approved study. Inclusion criteria were a primary complaint of genital prolapse and exclusion criteria consisted of inability to consent to study enrollment or tolerate two consecutive genital examinations. After consent was obtained, subjects underwent a standard POP-Q examination with a rigid marked measuring stick (POPStix^{TM}, Auckland, New Zealand) and a POP-Q by estimation (the "eyeball" POP-Q) in a randomized order by two successive examiners.

Prior to study enrollment, we developed the "eyeball" POP-Q technique, a novel approach to assess the individual POP-Q points by both visual estimation and palpation. In this technique, the points along the anterior and posterior vaginal walls (Aa, Ba, Ap, and Bp) are visually estimated (not measured) in 0.5-cm increments with maximal Valsalva, as originally described in the standard POP-Q [1]. Visual estimation is performed on the perineum (GH & PB). Determination of vaginal depth (total vaginal length or TVL) and apical descent (points C and D) are assessed by both visual estimation and palpation with the examiner's dominant hand.

Four examiners including two attending physicians and two fellows were involved in this study at one clinical site. All study subjects underwent both exam techniques by two examiners (one attending and one fellow), blinded to each other's results, during the same clinic visit. The order for exam technique was randomized by computer generation for the first examiner. The second examiner performed both examinations, but in an opposite order. The study was designed such that upon study enrollment, half of the subjects underwent an estimated exam first followed by a measured exam by the first examiner and then a measured exam followed by an estimated exam by the second examiner. The other half of the subjects underwent a measured exam followed by an estimated exam by the first examiner and an estimated exam followed by a measured one by the second examiner (Fig. 1).

Other than estimation and measurement, exams were performed in a similar, standardized manner. This included performing them immediately following voiding, in the supine lithotomy position, and with maximal strain using Valsalva effort. For Valsalva, patients were asked to take a deep breath in, hold their breath, and bear down as if they were constipated and trying to have a bowel movement. Patients confirmed that their maximal prolapse was reproduced with each examination by either palpation or visual confirmation with a hand-held mirror. The bottom half of a bivalve speculum was used to assess all internal values (TVL, C, D, Aa, Ba, Ap, and Bp) for the measured and estimated examinations, with the exception of TVL, C, and D for the estimated technique where the examiner's dominant hand was used instead of a speculum.

Data was entered into an Excel spreadsheet following completion of data collection and then imported into Statistical Package for the Social Sciences (SPSS Inc., Chicago, IL, USA). Because this was considered a pilot study, an a priori power calculation was not performed. A post hoc power analysis based on the final sample size of 50 patients shows that assuming a 2-cm difference between measured and estimated exams as a clinically significant difference, a sample size of 50 provides 80% power (two-sided $\alpha = 0.05$). A 2-cm difference has been cited as a clinically important change within the POP-Q literature [4].

Individual POP-Q values and stages were compared between the estimated and measured techniques for the same examiners and between different examiners. In order to ensure that results were not skewed by examiner recall bias, estimated results obtained prior to measured values were analyzed primarily. Secondary analysis was performed to compare values obtained by estimation following measured ones and to assess the inter-examiner reliability of estimated examinations.

The two-tailed paired t test was used to compare individual POP-Q points and the chi-square test for stage. Correlation and agreement of POP-Q points and stage were performed by Pearson's correlation coefficient and kappa statistics, respectively. A p value of ≤ 0.05 was considered statistically significant.

Results

Patient demographics are displayed in Table 1. Forty-two percent (21/50) had prior hysterectomy. The POP-Q stages were 18% (9/50) stage 1, 38% (19/50) stage 2, 44% (22/50) stage 3, and 0% (0/50) stage 4 based on the measured technique.

Table 2 illustrates the overall POP-Q stages obtained by different examiners for the estimated and measured techniques. In 90% of subjects, the stage was similar between the measured and estimated techniques. There was no trend toward a higher or lower stage: 4% (2/50) cases had a higher stage with estimation and 6% (3/50) cases had a

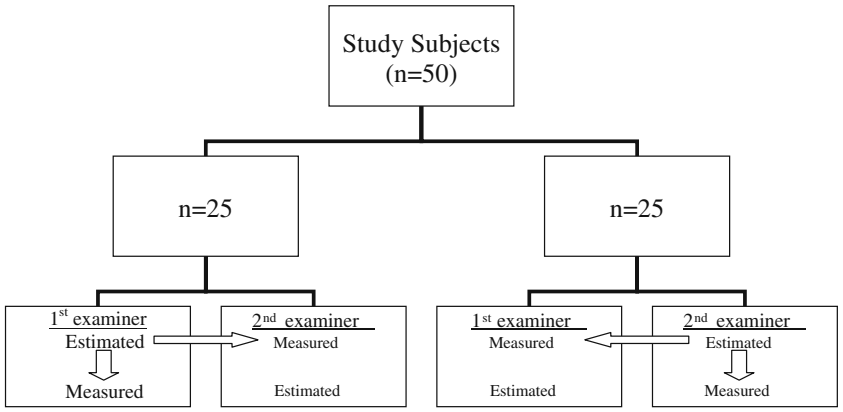

Fig. 1 Schematic of study design and analysis

lower stage using estimation. In no subject did the stage differ by greater than one. Overall, there was no difference in POP-Q stage between the measured and estimated techniques (p=0.83).

When comparing all individual POP-Q values, we found that 81% of values obtained by estimation and measurement were within 1 cm of each other, 2.6% of estimated values were 2 cm greater than measured values, 2.8% of estimated values were 2 cm less than measured ones, and 0.004% of estimated values had a 3-cm greater difference than measured points.

Because of the concern that measured exams could influence estimated ones, analyses of POP-Q points were also performed separately. For POP-Q exams in which estimated values were obtained prior to measured ones, there was no significant difference between measured and estimated POP-Q points for different examiners (all p > 0.05, Table 3). There was also no significant difference in estimated and measured values among exams in which measured values were obtained first (all p > 0.05).

Agreement between the techniques was assessed with Pearson's correlation coefficient (ρ) for integral data and kappa statistic (k) for categorical data. The value of each defines the strength of agreement. Coefficients between 0.81 and 1.00 are considered almost perfect, 0.61 and 0.80 substantial, and between 0.41 and 0.60 moderate. In this sample, there is substantial to almost perfect agreement between estimated and measured POP-Q stage and for the majority of individual POP-Q values (Table 3).

For assessment of overall POP-Q stage, good correlation was found between the two techniques with almost perfect intra-examiner (k=0.84, p<0.01) and substantial inter-examiner agreement (k=0.66, p<0.01). In terms of strength of agreement between the estimated and measured techniques, eight of nine intra-examiner and seven of nine inter-examiner correlations had "substantial or almost perfect" agreement, with the remainder (points D and TVL) having "moderate" agreement.

Secondary analysis to assess inter-examiner reliability between estimated exams showed no significant difference when comparing individual POP-Q values and substantial and almost perfect inter-examiner agreement for all points with the exception of points D and PB (Table 4). POP-Q stage obtained by estimation for different examiners did not differ significantly (p=0.70).

Table 1 Patient demographics

Demographic	Mean ± SD (range)
Age	60±11.6 (30–82)
BMI	27.8±4.5 (20.4–38.6)
Gravidity[a]	3 (1–9)
Parity[a]	2 (1–9)
Menopausal[b]	40 (80)
Previous hysterectomy[b]	21 (42)
Previous colporrhaphy[b]	14 (28)

[a] Data presented as median (range)

[b] Data presented as n (%)

Discussion

The POP-Q examination was developed to provide a precise and efficient model to allow pelvic floor clinicians and researchers to effectively communicate pelvic floor

Table 2 Overall POP-Q stage obtained by estimated and measured techniques by different examiners

		Measured exam			
Estimated exam	Stage	1	2	3	4
	1	8	0	0	0
	2	1	18	3	0
	3	0	1	19	0
	4	0	0	0	0

Table 3 Estimated and measured POP-Q points by different examiners (for estimation points performed before measurement)

POP-Q point	Estimated exam mean ± SD	Measured exam mean ± SD	p value[a]	Pearson ρ[b]
Aa	0.13±1.9	0.15±1.9	0.91	0.77
Ba	0.38±2.4	0.43±2.3	0.81	0.80
Ap	−1.09±1.6	−1.03±1.5	0.7	0.62
Bp	−1.07±1.6	−0.93±1.7	0.48	0.65
C	−4.94±3.6	−4.90±3.7	0.89	0.86
D	−7.96±1.6	−7.67±1.7	0.36	0.46
TVL	9.13±1.2	9.08±1.3	0.75	0.58
GH	3.50±1.2	3.50±1.3	1.0	0.78
PB	3.51±0.9	3.58±0.9	0.49	0.66

[a] Paired t test

[b] Pearson's correlation coefficient, all $p < 0.05$

anatomy in a standardized manner. In order to avoid potential imprecision and subjectivity, the anatomic landmarks and quantitative points are explicitly described in the POP-Q system [10]. Though the POP-Q examination was developed with the objective to enhance uniformity and objectivity, modifications to the exam are routinely made in clinical practice.

In the original description of the POP-Q system, there are no specific recommendations as to how quantitative measurements should be made when performing a POP-Q exam. Thus, a variety of methods have been described for the quantitative measurements of POP-Q points including the use of marked measuring sticks, cotton swabs, ring forceps, and wooden or plastic spatulas. Technique modifications may reflect differences in the way the POP-Q is taught or modified in clinical practice to enhance convenience, use available tools and equipment, maximize time efficiency, and reduce repetitive steps during examination.

In this study, we compared POP-Q examination with a rigid marked measuring stick compared with visual estimation. While overall correlation was substantial, apical points (TVL and D) had the poorest correlation between measured and estimated exams. Our results are consistent with others who have shown that grading prolapse without measurement is least reliable in the apical segment [11]. Though differences did exist, a minority of values differed by >2 cm, the difference we and others believe to be clinically important [4, 12, 13].

The Pelvic Floor Disorders Network chose to investigate how a limited number of these technique modifications impact POP-Q measurements. Specifically, they assessed differences in POP-Q values obtained with and without a speculum, differences in perineal measurements at rest and with strain, and whether the leading edge of prolapse differed in lithotomy as opposed to standing [6]. Barber et al. [4] found a 2-cm increase in at least one POP-Q point in almost half of patients when comparing examinations in dorsal lithotomy to upright in a birthing chair. In addition to position, the effect of bladder volume has been examined [12]. This strongly suggests that modifications in exam technique alter exam results, are clinically relevant, and may affect management strategy (surgery vs. pessary) and potentially surgical approach (abdominal vs. vaginal and unaugmented vs. augmented repair).

Limitations of this study include its small sample size, limited patient population, and small number of examiners. Because the exams were performed at the same visit, there

Table 4 Comparison of estimated POP-Q points between examiners

POP-Q point	Estimated examiner 1 mean ± SD	Estimated examiner 2 mean ± SD	p value[a]	Pearson ρ[b]
Aa	0.31±1.8	−0.01±2.0	0.06	0.81
Ba	0.53±2.3	0.26±2.4	0.15	0.85
Ap	−0.95±1.5	−1.13±1.6	0.33	0.65
Bp	−0.81±1.8	−1.11±1.6	0.13	0.66
C	−5.04±3.7	−5.06±3.4	0.92	0.96
D	−7.93±1.6	−7.69±2.2	0.58	0.39
TVL	8.9±0.9	9.1±1.3	0.18	0.67
GH	3.51±1.2	3.64±1.2	0.31	0.72
PB	3.58±1.0	3.50±1.0	0.55	0.51

[a] Paired t test

[b] Pearson's correlation coefficient (all $p < 0.05$)

is a potential for examiner recall bias. In order to account for this, eyeball values obtained immediately preceded by a measured exam were initially not analyzed. It is also plausible that performing sequential POP-Q examinations during the same clinic visit could result in a "teaching effect," with patients straining more for the second examiner; however, these results did not demonstrate an upstaging for the second examination. In order to better assess intra-examiner reliability, patients could have returned for a second examination by the same examiner 1 or 2 weeks later. However, this would have significantly altered the standard practice in the clinic. In addition, it is a common belief that prolapse can vary in severity from day to day depending on physical activity level, time of day, and other unknown factors; thus test–retest reliability may not have been accurate for exams performed on different days [13]. Another possible fault may be poor generalizability of our results as they are only applicable to physicians experienced in the standard POP-Q exam. Finally, no time estimation or measurement for completion of each technique was performed in this study. It would have been prudent to time each examination in order to determine if estimating values indeed does save a significant amount of time during the clinical examination of a patient with genital prolapse.

The POP-Q system is underutilized in clinical practice and in the urogynecologic literature and has been criticized for being "time-consuming" and "confusing" [7, 14]. It has been suggested that in order to save time, practicing physicians estimate POP-Q values instead of measuring them. The results of this study suggest that estimating POP-Q values provides comparable results to measuring them in physicians well versed on the standard POP-Q.

Financial support None

Disclaimers Dr. Davila: Consultant for American Medical Systems (Minnetonka, MN), Astellas (Tokyo, Japan), and Watson (Corona, CA). Drs. Karp, Peterson, Jean-Michel, Lefevre, and Aguilar: No disclosures.

References

1. Bump RC, Mattiasson A, Bo K, Brubaker LP, DeLancey JO, Klarskov P et al (1996) The standardization of terminology of female pelvic organ prolapse and pelvic floor dysfunction. Am J Obstet Gynecol 175:10–17
2. Hall AF, Theofrastous JP, Cundiff GW, Harris RL, Hamilton LF, Swift SE et al (1996) Interobserver and intraoberver reliability of the proposed International Continence Society, Society of Gynecologic Surgeons, and American Urogynecologic Society pelvic organ prolapse classification system. Am J Obstet Gynecol 175 (6):1467–1471
3. Kobak WH, Rosenberger K, Walters MD (1996) Interobserver variation in the assessment of pelvic organ prolapse. Int Urogynecol J 7:121–124
4. Barber MD, Lambers AR, Visco AG, Bump R (2000) Effect of patient position on clinical examination of pelvic organ prolapse. Obstet Gynecol 96(1):18–22
5. Swift SE, Herring M (1998) Comparison of pelvic organ prolapse in the dorsal lithotomy compared with the standing position. Obstet Gynecol 91:961–964
6. Visco AG, Wei JT, McClure LA, Handa V, Nygaard I (2003) Effects of examination technique on pelvic organ prolapse quantification (POP-Q) results. Int Urogynecol J 14:136–140
7. Auwad W, Freeman RM, Swift SE (2004) Is the Pelvic Organ Prolapse Quantification system (POP-Q) being used? A survey of members of the International Continence Society (ICS) and the American Urogynecologic Society (AUGS). Int Urogynecol J 15:324–327
8. Swift SE, Morris S, McKinnie V, Freeman R, Petri E, Scotti R et al (2006) Validation of a simplified technique for using the POPQ pelvic organ prolapse classification system. Int Urogynecol J 17:615–620
9. Lemos N, Auge A, Lunardelli JL, Carramao SS, Faria A, Aoki T et al (2007) Validation of the Pelvic Organ Prolapse Quantification Index (POP-Q-I): a novel interpretation of the POP-Q system for optimization of POP research. Int Urogynecol J 18:609–611
10. Baden WF, Walker TA (1972) Genesis of the vaginal profile: a correlated classification of vaginal relaxation. Clin Obstet Gynecol 15:1048–1054
11. Prien-Larsen J, Mouritsen L (2001) Pelvic organ prolapse: is ICS grading without POP-Q measurement reliable? Int Urogynecol J12 (Sup 3):S45
12. Silva WA, Kleeman S, Segal J, Pauls R, Woods S, Karram M (2004) Effects of a full bladder and patient positioning on pelvic organ prolapse assessment. Obstet Gynecol 104(1):37–41
13. Pearce M, Swift SE, Goodnight W (2008) Pelvic organ prolapse: is there a difference in POP-Q exam results based on time of day, morning or afternoon? Am J Obstet Gynecol 199:200–205
14. Muir TW, Stepp KJ, Barber MD (2003) Adoption of pelvic organ prolapse quantification system in peer-review literature. Am J Obstet Gynecol 189(6):1632–1636