Feasibility, inter- and intra-rater reliability of physiotherapists measuring prolapse using the pelvic organ prolapse quantification system

Stark, Diane; Dall, Philippa M.; Abdel-Fattah, Mohamed; Hagen, Suzanne

Published in:
International Urogynecology Journal

DOI:
[10.1007/s00192-009-1089-1](https://doi.org/10.1007/s00192-009-1089-1)

Publication date:
2010

Document Version
Author accepted manuscript

[Link to publication in ResearchOnline](#)

Citation for published version (Harvard):
Stark, D, Dall, PM, Abdel-Fattah, M & Hagen, S 2010, 'Feasibility, inter- and intra-rater reliability of physiotherapists measuring prolapse using the pelvic organ prolapse quantification system', *International Urogynecology Journal*, vol. 21, no. 6, pp. 651-656. <https://doi.org/10.1007/s00192-009-1089-1>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

If you believe that this document breaches copyright please view our takedown policy at <https://edshare.gcu.ac.uk/id/eprint/5179> for details of how to contact us.

Title page

Feasibility, inter- and intra-rater reliability of physiotherapists measuring prolapse using the Pelvic Organ Prolapse Quantification System.

Authors

Diane Stark¹, Philippa Dall², Mohamed Abdel-Fattah³ and Suzanne Hagen⁴

¹ Physiotherapy Department, Obstetrics and Gynaecology, Southern General Hospital, Glasgow

² School of Health, Glasgow Caledonian University

³ Division of Applied Health Sciences, School of Medicine & Dentistry, University of Aberdeen.

⁴ Nursing, Midwifery and Allied Health Professions Research Unit, Glasgow Caledonian University

Corresponding Author:

Diane Stark

Conflicts of interest

The authors declare the following for research or advisory work. Mohamed Abdel-Fattah: Astellas and Coloplast.

Abstract

Introduction

There is no agreed assessment tool for physiotherapists treating pelvic organ prolapse. This study hypothesised that POP-Q assessment was a feasible measure for use by physiotherapists and tested inter- and intra-rater agreement.

Methods

Six physiotherapists and two gynaecologists participated. Women were recruited from uro/gynaecology clinics. Two POP-Q examinations were performed at the first clinic (Gynaecologist, Physiotherapist1), and one week later (Physiotherapist1, Physiotherapist2). The examination was timed and women completed a short questionnaire. Using weighted kappa, agreement of POP-Q stage was assessed.

Results

Forty-five women were recruited (median age 59, range 32-87 years). Agreement between gynaecologist and physiotherapist was substantial (weighted kappa=0.63). Weighted kappa was 0.67 for inter-rater agreement between two different physiotherapists; and 0.71 for intra-rater reliability for the same physiotherapist. Examination time was significantly shorter [difference 53±73s, $p<0.001$] for gynaecologists. Participants found the examination acceptable.

Conclusions

POP-Q is a feasible and reliable outcome measure for physiotherapists to use.

Keywords

Physiotherapists, POP-Q, prolapse, reliability

Brief summary

Forty five women had two POP-Q measurements completed on two occasions to assess inter- and intra-examiner reliability of physiotherapists using the POP-Q.

Introduction

Physiotherapists routinely use pelvic floor muscle training to treat women with pelvic organ prolapse (POP), however no common outcome measure is used to assess the effect of this intervention [1]. The pelvic organ prolapse quantification (POP-Q) system is an objective, standardised and validated measure of POP recognised by the International Continence Society [2]. The POP-Q measures nine individual points from which a categorical stage is derived. It is used within research and in clinical practice to assess the extent of the prolapse and the outcome of treatment. The regular and routine use of such a measure by physiotherapists treating women with prolapse would allow the effect of physiotherapy for prolapse to be clearly demonstrated.

In practice it is currently gynaecologists who use the POP-Q, most commonly in research situations. Studies have looked at the inter-rater reliability of the POP-Q [3,4] and at the effects on the POP-Q measurement of various examination techniques [5]. Although physiotherapists treating women by assessment and re-education of pelvic floor muscles will usually perform a vaginal examination and pelvic floor muscle assessment, the POP-Q measurement is not usually carried out [1].

Hall et al. [3] assessed 48 women twice with different examiners to test inter-rater reliability of the POP-Q, and 25 women were examined twice by the same examiner (three weeks apart) for intra-rater reliability. They found moderate to almost perfect correlation for the nine site-specific points (range of correlation coefficients: 0.488 to 0.913). There were seven different examiners (physicians in an obstetrics and gynaecology department) with a range of experience. Experienced examiners took a mean of 2.05 minutes to perform a POP-Q examination, and inexperienced examiners 3.73 minutes. It was concluded that the reliability of the POP-Q was independent of examiner experience and that the system was easy to learn and teach.

Kobak et al. [4] compared two methods of measuring prolapse, the vaginal profile and the POP-Q. Two examiners, a physician and a nurse clinician, examined 49 consecutive women referred for evaluation of urinary incontinence or prolapse, using first the vaginal profile and then the POP-Q system. A high degree of inter-rater

agreement for both systems was reported (vaginal profile kappa = 0.68 and POP-Q kappa = 0.79). It was concluded that both measuring systems were reliable and that close attention to examination technique may be the most important factor in measurement of prolapse.

There is currently no published work documenting the reliability of the POP-Q system when used by physiotherapists. This study aimed to determine the feasibility, inter- and intra-rater reliability of physiotherapists using the POP-Q.

Materials and methods

Setting

The study took place in two hospitals (Southern General Hospital and Victoria Infirmary) in Glasgow, UK. Participating women were recruited on attendance at outpatient clinics, and all examinations took place on-site at the hospitals.

Participants: women

Women were recruited from one urogynaecology and two gynaecology outpatient clinics. Participants were attending for various reasons, including some with symptoms of POP, and all participants would have had a vaginal examination as part of their routine care on attending the clinic. Women were excluded from the study if they were pregnant, had had pelvic surgery in the last 12 weeks, were attending the clinic for psycho-sexual problems, had vaginal infection or inflammation of the vulva, or were unable to provide informed consent. Women were sent information regarding the study by post prior to attending the clinic. On attending the clinic, women had an opportunity to talk with the researcher (a physiotherapist with experience in women's health). Informed consent was obtained from all participants. The study was approved by the South Glasgow Research Ethics Committee on 1st March 2006 (REC reference number 06/S0702/9).

Participants: clinical staff

Six physiotherapists, with a range of experience in women's health, and two consultant gynaecologists, took part in the study. The physiotherapists consisted of: two staff grade physiotherapists (undertaking a single 4 month rotation period in women's health); two senior physiotherapists (with 6 and 8 years' experience working in women's health); one superintendent physiotherapist (16 years' experience); and

one clinical specialist physiotherapist in women's health (16 years' experience). The two consultant gynaecologists ran the clinics at which women were recruited, and used the POP-Q examination as part of their clinical practice. For the purposes of this study, the consultant gynaecologists were regarded as the gold standard. A single nurse chaperone participated in the study, and attended all the examinations.

Training

All clinical staff participating as POP-Q assessors completed a standardised training programme in the use of the POP-Q based on the original article describing the method [2] and the American Urogynecological Society POP-Q DVD. At the 1.5 hour long training session the DVD was viewed, followed by a discussion amongst the attendees with the opportunity to ask questions of a gynaecologist who used the POP-Q in regular practice (not one of the participating gynaecologists), and finally the DVD was viewed for a second time. The physiotherapists also completed practical training sessions consisting of clinical observation of the POP-Q being undertaken and the opportunity to perform the POP-Q under supervision. The physiotherapists continued POP-Q training informally in the eight weeks between the training session and the first recruitment clinic, including: further observation of the DVD; construction of teaching props; informal discussion; and reading the original POP-Q article [2].

Protocol

Two POP-Q examinations were performed for each participant at the clinic at which they were recruited (clinic 1); one by the consultant gynaecologist (this formed part of their routine care) and one by a study physiotherapist. Participants then attended the hospital one week later at the same time of day (clinic 2), and two further POP-Q examinations were performed; one by the same physiotherapist from the previous week (repeat physiotherapist); and one by another study physiotherapist (comparison physiotherapist). The order of the examiner was allocated randomly (using random numbers generated by Excel and placed in pre-sealed envelopes) at both clinic visits so that aspects such as discomfort or any tiredness experienced by the women from bearing down did not bias the results. All participants were blinded to the results of other POP-Q assessments, at both clinics. The attending chaperone scribed the nine POP-Q measurements and timed the duration of each examination using a stop watch. Participants were asked to complete a short questionnaire regarding their

experience of each examination. Women were offered travel expenses to attend the second hospital visit.

Instruments

The POP-Q assessment requires an internal vaginal examination, which records nine individual predefined measures [2]. Six measures are the location, relative to the hymen, of points on the anterior (Aa, Ba), posterior (Ap, Bp) and apical (C, D) vagina, which allow a description of the extent of any descent. Negative and positive values indicate locations above and below the hymen respectively. Additionally, the lengths of two external measures (perineal body; genital hiatus) and one internal measure (total vaginal length) are made. From these measures a stage of prolapse is calculated, based on the point of most prominent descent. The POP-Q stage is a five level category, ranging from a normal vaginal profile (stage 0) to complete vaginal eversion (stage IV). Measurements were made using a disposable wooden spatula which was marked in centimetres for the purpose of this study.

It is recommended that the POP-Q is recorded under the condition of maximum prolapse descent [2], although there is little agreement in the literature as to the standard position and conditions under which this is to be achieved. In this study, women were placed in a supine position, and asked to bear down by the examiner in order to see the prolapse at its maximum descent. Women were encouraged to empty their bladder prior to each examination.

A brief questionnaire was provided to the women after each pair of examinations, which they were asked to leave, once completed, in a sealed envelope at the clinic (this option was chosen by the majority of women) or to return by post (using a pre-paid envelope). The questionnaire asked three questions with categorical response options regarding the acceptability of each examination at the clinic (yes; no), the length of time that each examination took (acceptable; unacceptable) and the discomfort experienced during each examination (severe discomfort; mild discomfort; no discomfort). There was additional space for free comments to be added.

Data Analysis

The recorded values of the nine site-specific points and the assigned POP-Q stage were reviewed by a researcher. Commonly repeated errors in reporting positive and

negative values for the nine site specific points were highlighted and changed prior to data analysis, according to the following rules: all values of genital hiatus, perineal body and total vaginal length were corrected to be positive; for the six other points, where there was no written indication of sign, the value was assumed to be negative (as opposed to the usual convention that no written sign indicates a positive value).

Less than 5% of reported site-specific points (69 instances) were altered in this process. Changes were made to the data points from the gynaecologists (n=10) as well as from the physiotherapists (n=59). The weighted kappa statistic [6] was used to compare agreement in the POP-Q stage for three pairs of comparisons: inter-rater agreement between the gynaecologist and physiotherapist at clinic1; inter-rater agreement between the repeat physiotherapist and the comparison physiotherapist at clinic 2; and the intra-rater agreement of the repeat physiotherapist between clinics 1 and 2.

The absolute level of disagreement in the nine site-specific sites was calculated for each pair of comparisons. Mean and standard deviation of absolute differences and the percentage of comparisons where the absolute level of disagreement was under 2cm were reported. The duration of an examination was reported as mean and standard deviation. A paired t-test was performed to compare the duration of examination between gynaecologists and physiotherapists at clinic 1.

Questionnaire responses from the women were summarised using proportions, reported for all examinations, and separately for those performed by gynaecologists and physiotherapists. Comments written on the questionnaires were transcribed, and used to clarify the responses to the questions.

Weighted kappa statistics were performed using a web application [<http://faculty.vassar.edu/lowry/kappa.html>]. All other analyses were performed using SPSS (version 15.0). A 5% level of statistical significance was used throughout.

Results

Forty-five women were recruited to the study [median age 59, range 32 to 87 years]. Their primary presenting complaint was prolapse (n=22), urinary incontinence (n=15), other conditions (n=7), or was not reported (n=1). Data analysis was based on 45 participants at the first clinic, and 39 of these participants who then attended the second clinic. Three women at the second clinic visit had only one POP-Q examination instead of two. The second POP-Q examination was not carried out due

to discomfort being reported at the first examination. In total, study physiotherapists performed 120 POP-Q examinations.

The agreement in POP-Q stage between the gynaecologist and physiotherapist at clinic 1 was substantial, with a weighted kappa statistic of 0.64 (table 1). Weighted kappa was 0.67 for inter-rater agreement between two different physiotherapists at clinic 2 (table 2); and 0.71 for intra-rater reliability for a single physiotherapist (clinic 1 versus clinic 2; table 3).

Tables 1, 2 and 3 inserted here

Mean absolute differences between examiners for the site specific points were between 0.45 and 2.18 cm (table 4). Most differences between examiners were less than 2cm, but differences could be substantial (figure 1). **In general, differences between examiners tended to be greater for the site-specific points C and D, and smaller for external points gh and pb.** The duration of examination was significantly shorter (mean difference $\pm$ standard deviation) for gynaecologists (171 $\pm$ 51s) compared with physiotherapists (224 $\pm$ 52s) for those same examinations (mean difference 53s; 95% CI 30 to 77s; $t=4.6$; $p<0.001$).

Table 4 inserted here

Figure 1 inserted here

All participants who expressed an opinion in feedback questionnaires ($n=44$) reported that the examination itself and the time taken to conduct the examination were acceptable. Participants predominantly rated the levels of discomfort as none or mild (table 5), with few differences between the rating given to gynaecologists and to physiotherapists. Two participants experienced severe pain during one of their examinations, in one instance this was caused by a cyst which had developed between clinics.

Table 5 inserted here

Discussion

The aim of this study was to determine the feasibility, inter- and intra-rater reliability of physiotherapists using the POP-Q. Six physiotherapists with a range of experience successfully completed 120 examinations.

The kappa statistics indicated a substantial agreement [7] between the examiners, for all of the comparisons tested in this study. It is difficult to directly compare kappa statistics between studies, however the agreement between examiners in this study was of a similar magnitude to that found between two examiners (a nurse and a gynaecologist; kappa=0.79) in Kobak et al. [4]. Use of correlation coefficients to assess agreement is recognised as incorrect, as an outcome could be perfectly correlated without agreeing. It is therefore not suitable to compare the outcome of an agreement study to those assessing correlation (such as in Hall et al. [3]).

Absolute differences in the nine individual POP-Q points were at times substantial, although mean absolute differences were under 2cm for all points. Hall et al. [3] reported mean differences in site specific points ranging from 0.04 and 0.40cm, however it was not reported whether those differences were absolute. The difference between stages can be quite sensitive to a small difference in measurement, if the leading edge of the prolapse is close to the cut-point between stages. However, a large discrepancy in actual measurement could be tolerated without changing the stage of prolapsed calculated from the POP-Q, if the measured point is not the one defining the POP-Q stage, or if it lies between stage cut-points.

As far as we are aware, this is the only study looking at the acceptability of the POP-Q examination to the woman. All participants reported that this was an acceptable examination to undergo at clinic. There was no difference in the reported experience of the participants during the examinations regarding whether the examination was performed by a gynaecologist or a physiotherapist.

The mean duration of examination physiotherapists (224s) was the same as that for inexperienced examiners reported by Hall et al. [3]. In both studies, more experienced examiners conducted the examinations more quickly; however the gynaecologists in this study (171s) took longer than the experienced examiners in Hall et al. (123s). It is reasonable to expect that the physiotherapists would become

quicker at conducting the examination with additional experience. In this study, the gynaecologists, on average, took approximately one minute less to conduct the examinations than the physiotherapists, but this made no difference to the acceptability of the assessment as reported by the participants.

Keeping the variable factors constant in repeated POP-Q measurements is an important factor in achieving repeatable measurements with this examination technique. The method of this study tried to control **a number of** variables, repeat clinical visits occurred at the same time of day, examination position and measuring instruments were standardised. However, differences in the women's activities prior to their clinic appointment, in **emptying the bladder or bowel prior to examination**, and measurement at different stages in the menstrual cycle may have altered the size of prolapse at examination. Similarly, some women examined had a ring pessary in situ which was removed at the first clinic visit and the women opted to wait to have the replacement ring pessary inserted after the second set of POP-Q measurements were completed. **These factors may have influenced comparisons between the first and second clinics, which were held a week apart. Examiners were asked to measure the POP-Q at maximum descent, however standardising a maximal valsalva between examiners is difficult to achieve, and could have had an effect on agreement in any of the reported comparisons. Differences in standardisation are likely to have worked to lower the reported level of agreement between examiners.**

Sample size was smaller than ideal, and this was especially true for the comparisons between physiotherapy examiners at the second clinic visit, where numbers were reduced through non-attendance and, in three cases, the inability of the second examination to be carried out. However the number of participants was comparable to other reliability studies on the POP-Q [3,4].

The POP-Q is an acceptable examination for women to have in the outpatient setting. The POP-Q is a feasible and reliable measure for use by physiotherapists. Its use both as a research tool and in clinical practice to assess physiotherapy interventions would prove to be a useful development for the profession, and would encourage multi-professional working via the application of a common standardised measurement system. All participating physiotherapists felt they would continue to use the POP-Q as an outcome measure in clinical practice.

There is little evidence to support the use of physiotherapy for the treatment of POP [8]. The use of a recognised, validated outcome measure is essential to the development of an evidence base for clinical practice. This study provides evidence of reliability and feasibility for physiotherapists using the POP-Q, and will contribute to the POP-Q being used with confidence in the future to evaluate both research and clinical outcomes.

Acknowledgements

Staff who provided the intervention: Angela Davies; Chris Hardwick; Julie Lang; Kate Lough; Karen McCulloch; Christine McFadden; June McGourty; Denise McKenzie; Stewart Pringle.

Funding

This study was funded by the Physiotherapy Research Foundation and the CSP Charitable Trust [PRF/05/3]. There is no conflict of interest.

References

- [1] Hagen S, Stark D, Cattermole D (2004) A United Kingdom-wide survey of physiotherapy practice in the treatment of pelvic organ prolapse. *Physiotherapy* 90:19-26
- [2] Bump R C, Mattiason A, Bo K, Brubaker LP, DeLancey JOL, Klarskov P, Shull BL, Smith ARB (1996) The standardization of terminology of female pelvic organ prolapse and pelvic floor dysfunction. *Am J Obstet Gynecol* 175:10-17
- [3] Hall AF, Theofrastus JP, Cundiff GW, Harris RL, Hamilton LF, Swift SE, Bump RC (1996) Interobserver and intraobserver reliability of the proposed International Continence Society, Society of Gynecologic Surgeons, and American Urogynecologic Society pelvic organ prolapse classification system. *Am J Obstet Gynecol* 175:1467-1471
- [4] Kobak WH, Rosenberger K, Walters MD (1996) Interobserver variation in the assessment of pelvic organ prolapse. *Int Urogynecol J* 7:121-124
- [5] Visco AG, Wei JT, McClure LA, Handa VL, Nygaard IE. (2003) Effects of examination technique modifications on pelvic organ prolapse quantification (POP-Q) results. *Int Urogynecol J* 14:136-140
- [6] Cohen J. (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70:213-220
- [7] Landis & Koch (1977) The measurement of observer agreement for categorical data. *Biometric* 33:159-174
- [8] Hagen S, Stark D, Maher C, Adams E. (2004) Conservative management of pelvic organ prolapse. *The Cochrane Library*, Issue 2. Chichester UK: John Wiley & Sons Ltd.

Table 1

Inter-rater agreement in POP-Q stage between gynaecologist and physiotherapist.

Examiner	Stage	Physiotherapist					Total
		0	I	II	III	IV	
Gynaecologist	0	0	1	0	0	0	1
	I	2	7	3	0	0	12
	II	1	8	12	0	0	21
	III	0	0	4	6	1	11
	IV	0	0	0	0	0	0
Total		3	16	19	6	1	45

Table 2

Inter-rater agreement in POP-Q stage between repeat physiotherapist and comparison physiotherapist.

Examiner	Stage	Comparison Physiotherapist					Total
		0	I	II	III	IV	
Repeat Physiotherapist	0	0	2	0	0	0	2
	I	1	2	6	0	0	9
	II	0	5	14	1	0	20
	III	0	0	0	4	0	4
	IV	0	0	0	0	1	1
Total		1	9	20	5	1	36

Table 3

Intra-rater agreement in POP-Q stage for measurements taken by the same physiotherapist one week apart.

Clinic	Stage	Clinic 2					Total
		0	I	II	III	IV	
Clinic 1	0	1	1	1	0	0	3
	I	2	6	4	0	0	12
	II	0	3	15	1	0	19
	III	0	0	1	3	0	4
	IV	0	0	0	0	1	1
Total		3	10	21	4	1	39

Table 4. Agreement for nine site-specific POP-Q points

POP-Q Point	gynaecologist vs. physiotherapist			repeat physiotherapist vs. comparison physiotherapist			repeat physiotherapist clinic 1 vs. clinic 2		
	N	† Mean diff (SD)	‡ % <2cm	N	† Mean diff (SD)	‡ % <2cm	N	† Mean diff (SD)	‡ % <2cm
Gh	45	0.92 (0.90)	80	36	1.22 (0.99)	67	39	0.71 (0.59)	95
Pb	45	0.79 (0.52)	96	36	0.71 (0.57)	92	39	0.45 (0.48)	97
Aa	45	1.21 (1.36)	76	36	1.44 (1.09)	58	39	0.77 (1.32)	82
Ba	45	1.28 (1.30)	78	36	1.28 (1.31)	75	39	1.17 (1.73)	74
C	40	1.86 (1.98)	55	35	1.94 (1.53)	49	35	1.06 (1.46)	77
D	30	1.82 (2.35)	67	25	2.18 (1.74)	48	27	1.07 (1.46)	81
Ap	45	1.20 (1.15)	73	36	1.13 (1.33)	72	39	0.71 (1.15)	87
Bp	45	0.84 (0.76)	78	36	0.92 (1.21)	81	39	0.71 (1.19)	90
TVL	41	0.90 (0.83)	76	36	0.78 (0.81)	83	39	0.64 (0.71)	92

† mean absolute differences between examiner 1 and examiner 2, with standard deviation (SD), measured in cm

‡ percentage of absolute differences that were less than 2cm

Table 5

Summary of questionnaire responses regarding level of discomfort during the examination.

Discomfort level of examination	clinic 1 (n = 44)		clinic 2 (n = 34)	
	Doctor	Physiotherapist	Repeat Physiotherapist	Comparison Physiotherapist
none	27	27	18	16
mild	17	17	15	17
severe	0	0	1	1