CrossMark

ORIGINAL ARTICLE

Discrepancies between patient-reported outcome measures when assessing urinary incontinence or pelvic- prolapse surgery

Michael Due Larsen[1] · Gunnar Lose[2] · Rikke Guldberg[1,3] · Kim Oren Gradel[1,4]

Abstract

Introduction and hypothesis In order to assess the outcome following surgery for urinary incontinence (UI) and pelvic organ prolapse (POP) the importance of patient-reported outcome measures, in addition to the clinical objective measures, has been recognised. The International Consultation on Incontinence has initiated the development and evaluation of disease-specific questionnaires (ICIQ) to compare the patient's degree of improvement. Alternatively, the Patient's Global Impression of Improvement (PGI-I score) with an inherent before–after assessment has been widely accepted in recent studies. The aim of this study was to compare the PGI-I versus the ICIQ score for women undergoing UI or POP surgery.

Methods This study is based on self-administered pre- and postoperative questionnaires, completed by women undergoing surgery for UI or POP in Denmark in 2013. Weighted Kappa statistics and 95 % limits of agreement method were used when comparing the PGI-I and ICIQ scores.

Results Among the 3,310 women included the PGI-I score showed a higher improvement than the IQIC score, for UI

0.83 (CI 95 %: 0.80–0.85) vs 0.62 (0.60–0.64) and for POP 0.77 (0.75–0.78) vs 0.66 (0.65–0.67).

Conclusions The PGI-I score renders higher satisfaction than the ICIQ score and the PGI-I score overestimates the improvement following UI and POP surgery.

Keywords Urinary incontinence · Pelvic organ prolapse · Patient-reported outcome measures · Patient's Global Impression of Improvement · Surgery · Questionnaires

Introduction

To assess the outcome following surgery for urinary incontinence (UI) and pelvic organ prolapse (POP) the importance of patient reported outcome measures (PROMs), in addition to the clinical objective measures, has been recognised [1–4]. PROMs are used to evaluate the effectiveness and quality of treatment in routine practice, to improve quality, benchmarking, and decision-making [4, 5]. However, critics argue that PROMs do not provide unambiguous answers about whether an intervention succeeds and thereby can be used as an evaluation tool for clinical interventions [6–8]. Different types of PROMs have been used in questionnaires following surgery for UI and POP. The International Consultation on Incontinence (ICI) initiated the development and evaluation of disease-specific questionnaires (ICIQ) [1, 9–12]. This scoring system is based on visual analogue scales (VAS) and the scoring of patients' subjective symptoms in defined categories (Likert scales), and the ICIQ is the sum of a combination of these (Fig. 1). When asking the patients before and after an intervention, PROMs are used to compare the patient's degree of satisfaction and improvement following the intervention. Alternatively, simpler

✉ Michael Due Larsen
   michael.due.larsen@rsyd.dk

1 Center for Clinical Epidemiology, Odense University Hospital, Sønderboulevard 29, Opg. 101, Odense 5000, Denmark

2 Gynaecology and Obstetrics Department, Herlev Hospital & University of Copenhagen, Copenhagen, Denmark

3 Gynaecology and Obstetrics Department, Hospital Lillebaelt, Kolding, Denmark

4 Research Unit of Clinical Epidemiology, Institute of Clinical Research, University of Southern Denmark, Odense, Denmark

 Springer

Incontinence

A How often do you leak urine?
(Tick one box)

never	0
about once a week or less often	1
two or three times a week	2
about once a day	3
several times a day	4
all the time	5

B We would like to know how much urine you think leaks.
How much urine do you usually leak (whenever you wear protection or not)?
(Tick one box)

none	0
a small amount	2
a moderate amount	4
a large amount	6

C Overall, how much does leaking urine interfere with your everyday life?
Please ring a number between 0 (not at all) and 10 (a great deal)

0 1 2 3 4 5 6 7 8 9 10
not at all   a great deal

ICIQ scores: sum A + B + C = _____

D How will you describe your condition now, compared with the condition before surgery" with the following response categories:
(Tick one box)

quite a lot better	0
much better	1
a little better	2
no change	3
a little worse	4
much worse	5
significantly worse	6

Pelvic Organ Prolapse

A Are you aware of a lump or bulge coming down in your vagina?
(Tick one box)

never	0
occasionally	1
sometimes	2
most of the time	3
always	4

B How much does this bother you?
(Tick one box)

not at all	0
a little	1
somewhat	2
a lot	3

C Overall, how much does do vaginal symptoms interfere with your everyday life?
Please ring a number between 0 (not at all) and 10 (a great deal)

0 1 2 3 4 5 6 7 8 9 10
not at all   a great deal

ICIQ scores: sum A + B + C = _____

D How will you describe your condition now, compared with the condition before surgery" with the following response categories:
(Tick one box)

quite a lot better	0
much better	1
a little better	2
no change	3
a little worse	4
much worse	5
significantly worse	6

Fig. 1 Pre- and post-questionnaires. The International Consultation on Incontinence Questionnaire (ICIQ) score is based on the sum of questions A–C and the PGI-I score is question D [1, 12]

PROMs with an inherent before–after assessment have been suggested, such as asking the Patient's Global Impression of Improvement (PGI-I score). Such scoring systems have been validated for the evaluation of IU and POP [10, 11, 13, 14], and are widely accepted in the recent literature. To the best of our knowledge, no study has evaluated the scoring systems in a large body of data, and there is concern relating to possible recall bias for the PGI-I score compared with the traditional pre- and postoperative questionnaires.

The Danish Urogynaecological Database (DugaBase) was established to monitor, ensure and improve the quality of urogynaecological surgery for all UI and POP surgeries in public and private hospitals in Denmark [15, 16]. Since its establishment in 2006, pre- and postoperative questionnaires regarding POP and UI surgery have been systematically collected. In 2013 the DugaBase was supplemented with a postoperative PGI-I score. The DugaBase is a national register with a high completeness, 95.0 % for UI and 91.3 % for POP surgeries [15].

The aim of this study was to compare the concordance of patients' evaluation of surgery using the single postoperative PGI-I score versus the use of a pre- and postoperative ICIQ score system for women undergoing UI or POP surgery.

Materials and methods

This study is based on pre- and postoperative questionnaires completed by women aged 18 years or older undergoing surgery for UI or POP in Denmark in 2013. Definitions conform to the international joint report on terminology for female pelvic floor dysfunction and urinary incontinence [17]. Only those who completed both the pre- and postoperative questionnaires were included in the analyses.

Data sources

For all Danish hospital departments and private hospitals/ clinics performing POP and UI surgery it is mandatory by Danish law to report data to the DugaBase and the data collection is based on a national web-based input module.

The DugaBase contains information on five areas: referrals; a pre-operative self-administered patient questionnaire based on the ICIQ scoring system; a pre-operative questionnaire completed by the gynaecologists including information on preoperative examination; information on surgical procedures; and finally, a post-surgery questionnaire consisting of the same self-administered questionnaires as those used before surgery, supplemented with a PGI-I score.

The ICIQ scoring system is based on visual analogue scales (VAS) from 0 to 10, and two symptom-specific Likert scales, and the ICIQ is the sum of a combination of scores (Fig. 1). Satisfaction of improvement is the difference between the pre- and post-surgery ICQI scores. The pre-surgery questionnaire is completed in connection with the preoperative examination. The post-surgery questionnaire is either sent to the patient 3 months after surgery or filled in by a nurse conducting a telephone interview with the patient. Question scores relevant for this study are presented in Fig. 1, and a detailed description of the database is available elsewhere [16].

Statistical analysis

All results are reported using descriptive statistics in numbers and means with 95 % confidence intervals. We computed the ceiling effect, the percentage of respondents who achieved the highest possible score, and we determined a cut-off point at 15 % as an acceptable ceiling for an operational scale [18]. The ceiling effect tells if a score only uses the top end of a scale; thus, changes in improvement would not be recognised.

In order to estimate the agreement between two methods of measurements (ICIQ and PGI-I), which express the same clinical intervention, a traditional correlation analysis was not appropriate. We therefore analysed the agreement of the PGI-I and ICIQ scores by converting these to a comparable scale ranging from −1 to 1, where 1 is the highest possible improvement and −1 is the lowest, and we further analysed the agreement as categorical variables and as if they were continuous variables. As categorical variables we calculated the inter-rater strength of agreement between the ICIQ and the PGI-I score by weighted Kappa statistics, using Altman's definitions: poor (kappa value <0.21), fair (0.21–0.40), moderate (0.41–0.60), good (0.61–0.80), and very good (0.81–1.00) [19]. Considering the scores as continuous variables we used the 95 % limits of agreement method, which is also called the Bland–Altman plot [19, 20]. All calculations were performed using STATA Release 13.0.

Approvals

The DugaBase operates under the Danish law on data protection, with a license granted by the Danish Data Protection Agency and the Danish Health and Medicines Authority. This specific study has been approved by the Danish Data Protection Agency (Region Syddanmark: 2008-58-0035/sagnr. 14/15130). According to Danish law, ethical approval is not required for purely registry-based studies.

Results

Of the 5,476 women registered in the DugaBase in 2013, 3310 (60.4 %) were included in this study, 738 after surgery for UI and 2,581 after POP (9 women underwent both POP and UI concomitantly). Among the 2,166 excluded, 525 had not filled in the preoperative questionnaire, 1,141 the postoperative questionnaire, and 499 neither of them.

Overall, the PGI-I score showed higher improvement than the IQIC score on a converted comparable scale, PGI-I 0.83 (95 % confidence interval [CI]: 0.80–0.85) vs ICIQ 0.62 (CI 0.60–0.64) for UI, and 0.77 (CI 0.75–0.78) vs 0.66 (CI 0.65–0.67) for POP (Table 1). Among the subgroups, the elderly (>70 years) had a lower degree of improvement following UI surgery than other age groups at both scores, and the youngest (18–39 years) had the lowest degree of improvement after POP surgery at the ICIQ score (Table 1). The subgroup of women with a previous UI surgery intervention had a lower degree of improvement following UI surgery, which supports other reported studies of repeat IU surgery [21, 22].

The scatterplot of PGI-I versus ICIQ illustrates the relatively higher score for PGI-I (Fig. 2). Moreover, only few women reported a negative improvement after surgery, regardless of score. The dotted line in Fig. 2 illustrates where the two scores were identical. There was a higher concordance for UI than for POP, although the regression line did not coincide with the line of equality for both UI and POP. Figure 2 also illustrates the ceiling, which is especially high for PGI-I. For UI, the ceiling was 3.8 % for ICIQ and 69.9 % for PGI-I, whereas for POP it was 14.1 and 53.2 % (Table 2).

Using Kappa statistics the agreements between the PGI-I and the ICIQ score were fair for both POP and UI surgery interpreted by using Altman's definition of strength of agreement (Table 3). We computed a Bland–Altman plot to show the differences in ICIQ and PGI-I against the mean for the same ICIQ and PGI-I scores (Fig. 3). The interpretation of the plot tells us that the scores were equivalent from −0.29 to 0.71 for UI and −0.57 to 0.79 for POP. The histograms showed that the difference in the scores was 0.21 for IU and 0.11 for POP, on a comparable −1 to 1 scale. They further showed an almost normal distributed difference, thus fulfilling the assumption for the 95 % limits of agreement method.

Discussion

Main findings

In general, women who undergo surgery for UI and POP express a high improvement of their disease-related symptoms in addition to improved quality of their everyday life [12, 23, 24]. In this study we also found a high degree of improved patient satisfaction following UI and POP surgery, using both

Table 1 Pre- and post-surgery improvement at Patient's Global Impression of Improvement (PGI-I) and International Consultation on Incontinence Questionnaire (ICIQ) scores (scores converted to comparable −1 to 1 scales)

		PGI-I	ICIQ		
	n	Improved	Pre	Post	Improved
Incontinence					
All	738	0.83 (0.80–0.85)	0.78	0.16	0.62 (0.60–0.64)
18–39	79	0.92 (0.88–0.97)	0.73	0.09	0.64 (0.59–0.69)
40–69	548	0.86 (0.83–0.88)	0.77	0.13	0.64 (0.62–0.66)
>70	111	0.62 (0.53–0.70)	0.83	0.33	0.50 (0.43–0.57)
BMI [a]	430	0.81 (0.78–0.84)	0.80	0.18	0.61 (0.58–0.65)
Smoking	122	0.85 (0.80–0.90)	0.81	0.14	0.67 (0.62–0.73)
Previous					
Hysterectomy	124	0.76 (0.70–0.83)	0.82	0.22	0.60 (0.54–0.66)
POP surgery	76	0.78 (0.70–0.87)	0.78	0.20	0.58 (0.50–0.65)
UI surgery	48	0.55 (0.41–0.69)	0.81	0.35	0.46 (0.35–0.58)
Prolapse					
All	2,581	0.77 (0.75–0.78)	0.73	0.08	0.66 (0.65–0.67)
18–39	82	0.74 (0.66–0.81)	0.70	0.14	0.56 (0.47–0.66)
40–69	1,613	0.77 (0.75–0.78)	0.73	0.08	0.65 (0.63–0.67)
>70	886	0.76 (0.74–0.78)	0.75	0.07	0.68 (0.66–0.70)
BMI [a]	1,429	0.76 (0.74–0.77)	0.74	0.08	0.66 (0.64–0.68)
Smoking	308	0.75 (0.72–0.79)	0.77	0.08	0.68 (0.64–0.72)
Previous					
Hysterectomy	541	0.76 (0.73–0.79)	0.76	0.09	0.67 (0.64–0.70)
POP surgery	546	0.76 (0.73–0.78)	0.77	0.09	0.69 (0.66–0.71)
UI surgery	121	0.70 (0.64–0.77)	0.73	0.07	0.66 (0.59–0.72)

[a] BMI (body mass index)>25

Table 2 Ceiling for PGI-I and ICIQ scores [18]

Scale	Pre-surgery	Post-surgery	Improvement
Incontinence		Ceiling (%)	
ICIQ	0.7	62.3	3.8
PGII	–	–	69.9
Prolapse			
ICIQ	4.8	79.2	14.1
PGII	–	–	53.2

the ICIQ score and the PGI-I score. Nevertheless, we found that the PGI-I score was higher than the ICIQ score, and graphically we observed a bad correlation with the equality line and a fair inter-rater agreement using kappa statistics.

Strengths and limitations

The DugaBase is a national clinical database containing 92.2 % of all Danish POP and IU surgeries carried out in 12 private clinics and 23 public hospitals reporting in the same web-based data-entering system [15, 16]. The response rate being 60.4 % for answering both the pre- and the post-questionnaire, we found our body of data valid for the purpose of this study.

When comparing a new measurement with an established scoring system, it is necessary to test whether they agree adequately. Without a gold standard, analytical correlation models would have been misleading because the scores would have been measuring the same clinical intervention. Instead, we used a graphical approach to describe the relation between
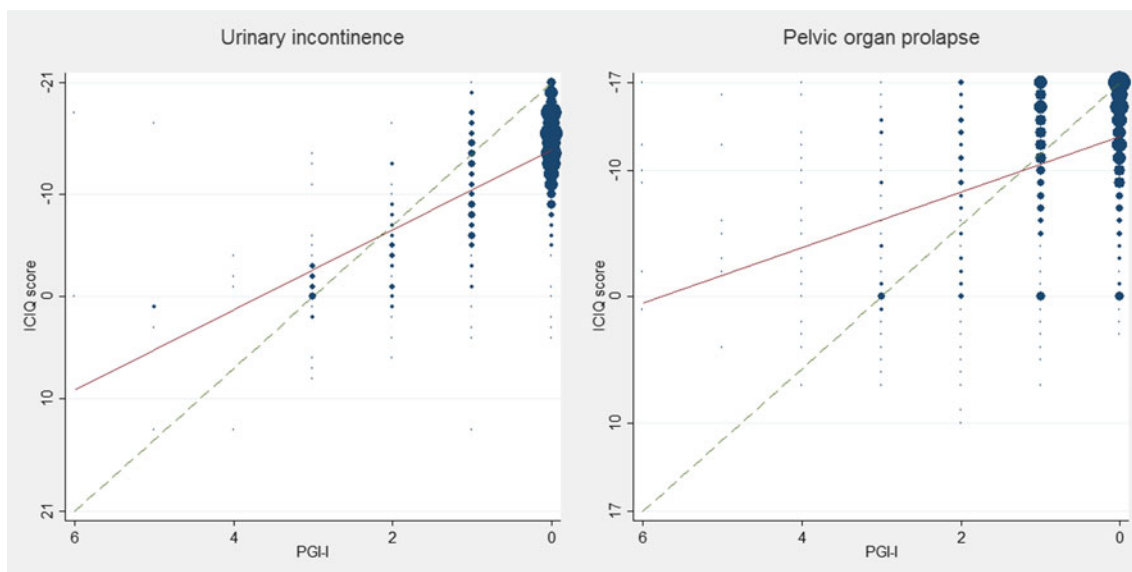

Fig. 2 PGI-I in relation to ICIQ score (scores not converted to a −1 to 1 scale) flowing urinary incontinence and pelvic organ prolapse (*unbroken line* regression line, *dotted line* line of equality)

Table 3 Inter-rater agreement using weighted Kappa statistics for the improvement on the PGI-I and ICIQ score (scores converted to comparable −1 to 1 scales)

	Kappa	Strength of agreement[a]
Incontinence $n=738$		
PGI-I ICIQ	0.37 (0.34–0.41)	Fair
Prolapse $n=2,581$		
PGI-I ICIQ	0.27 (0.25–0.29)	Fair

[a] Strength of agreement interpreted by using Altman's definition of kappa statistic

the established and the alternative measure [19, 20, 25]. The advantage of the 95 % limits of agreement method is that it does not give an answer to whether the scores are equivalent or not. Instead, the method shows that an interval between the scales is equivalent and measures a mean difference, and the researcher has to interpret whether the results are meaningful in a clinical content. We did not find that differences between the converted scores of 0.11 and 0.21, (corresponding to a 5 and 10 % relative difference) were acceptable from a clinical point of view. Even more critically, we found the 95 % limits of agreement, ranging from −0.57 to 0.79 for POP, showing that the interval of agreement between the scores ranged most of the scales.

Interpretation

There may be more reasons for these findings. First, it may have been be a matter of a simple recall bias when the women had to compare their symptoms and inconvenience before surgery indirectly when reporting the PGI-I measure. Second, it seems that the women's answers reflect their current status and not a change in improvement; thus, the inter-rater agreement of the PGI-I matches their post-surgery ICIQ scores better than the scores of improvement (results not shown). Therefore, are they actually answering our questions? Apparently, they answer whether they were satisfied with the operation in general and do not compare their original symptoms and inconvenience.

Third, we cannot rule out that the differences may, at least partly, be due to different phrasing of the questions (Fig. 1). However, we find it unlikely that this covers the entire explanation because results for separate comparisons between PGI-I on the one hand and the separate Likert and VAS scales on the other (data not shown) were very similar to the overall results, which corroborates that the findings were more likely related to measurement issues. In this study we only focused on PROMs and did not implicate objective clinical measures, even if these measures were available and did not necessarily correlate with the PROMs [4, 26, 27], and objective clinical measures could therefore not have been a gold standard in comparison.
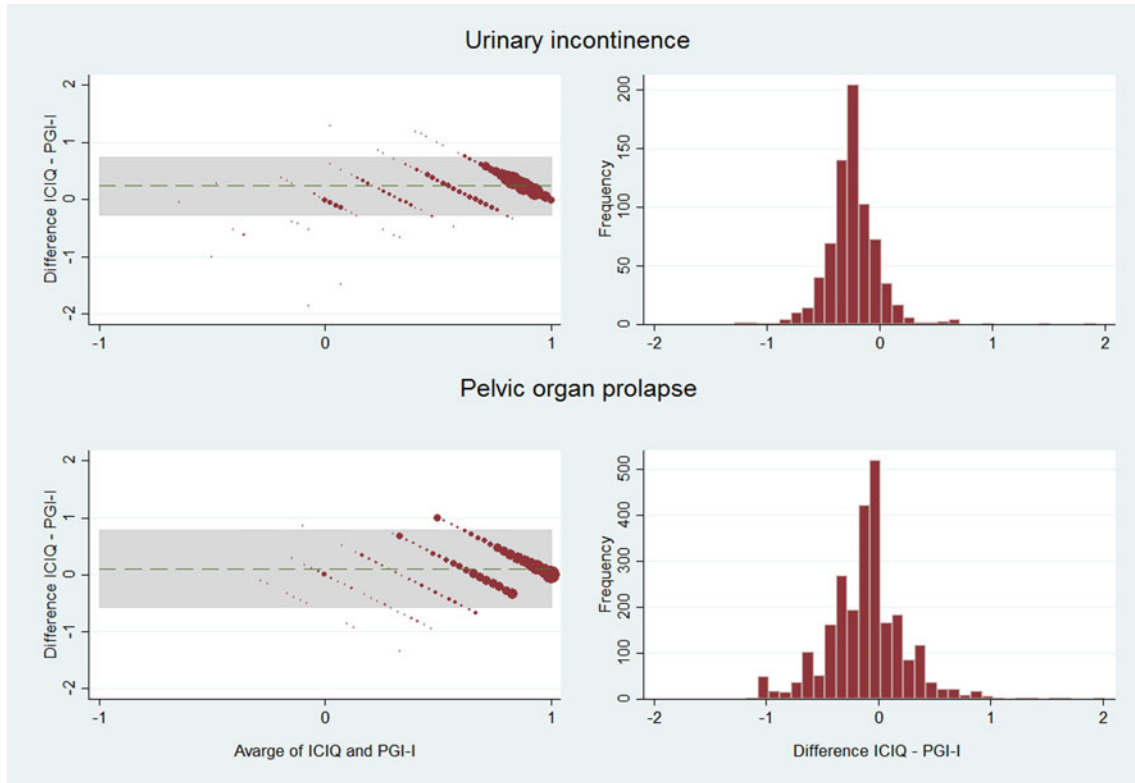

Fig. 3 Difference against average and histogram of difference of PIG-I and ICIQ score, Bland–Altman plot; scores converted to comparable −1 to 1 scales)

When designing and choosing PROMs of clinical interventions a number of criteria have to be fulfilled: reliability, responsiveness, interpretability, and response burden [18, 28]. Regarding the responsiveness of PROMs used in questionnaires, their applicability for evaluating changes is relevant. Only the ICIQ score showed an acceptable ceiling effect under 15 %; hence, the ability of the PGI-I score to detect improvement in clinical quality over time will be limited, although a deterioration in quality could be detected.

Conclusion

Questionnaires including questions based on the validated ICIQ are developed by the ICI [1, 29], and two studies suggest using a PGI-I score as a supplemental or surrogate measure for women undergoing surgery for UI and POP [10, 13]. The PGI-I score has been widely accepted [24, 30] and used in a study of incontinence disorders in men [14], as well as other incontinence disorders [31]. The question is, how reliable is a global measure of improvement for measuring the clinical quality of an intervention? Because only a post-surgery questionnaire is needed, the PGI-I reduces the response burden [18]. It is therefore tempting to use the PGI-I as a surrogate for more complicated pre- and post-questionnaires, but this study demonstrates that the PGI-I has to be used carefully or in addition to other PROMs to evaluate improvement following UI or POP surgery, because this score does not take recall bias into consideration.

References

1. Abrams P, Avery K, Gardener N, Donovan J, Board IA (2006) The international consultation on incontinence modular questionnaire: www.iciq.net. J Urol 175:1063–1066
2. Freeman RM (2010) Do we really know the outcomes of prolapse surgery? Maturitas 65(1):11–14
3. Srikrishna S, Robinson D, Cardozo L (2009) Qualifying a quantitative approach to women's expectations of continence surgery. Int Urogynecol J Pelvic Floor Dysfunct 20(7):859–865
4. Marshall S, Haywood K, Fitzpatrick R (2006) Impact of patient-reported outcome measures on routine practice: a structured review. J Eval Clin Pract 12(5):559–568
5. Greenhalgh J (2009) The applications of PROs in clinical practice: what are they, do they work, and why? Qual Life Res 18(1):115–123
6. Jackowski D, Guyatt G (2003) A guide to health measurement. Clin Orthop Relat Res 413:80–89
7. Poolman RW, Swiontkowski MF, Fairbank JC, Schemitsch EH, Sprague S, de Vet HC (2009) Outcome instruments: rationale for their use. J Bone Joint Surg Am 91 [Suppl 3]:41–49
8. Srikrishna S, Robinson D, Cardozo L, Gonzalez J (2008) Is there a difference in patient and physician quality of life evaluation in pelvic organ prolapse? Int Urogynecol Pelvic Floor Dysfunct 19(4):517–520
9. Cartwright R, Srikrishna S, Cardozo L, Robinson D (2011) Validity and reliability of patient selected goals as an outcome measure in overactive bladder. Int Urogynecol J 22(7):841–847
10. Srikrishna S, Robinson D, Cardozo L (2010) Validation of the patient global impression of improvement (PGI-I) for urogenital prolapse. Int Urogynecol J 21(5):523–528
11. Reid FM, Smith AR, Dunn G (2007) Which questionnaire? A psychometric evaluation of three patient-based outcome measures used to assess surgery for stress urinary incontinence. Neurourol Urodyn 26(1):123–128
12. Price N, Jackson SR, Avery K, Brookes ST, Abrams P (2006) Development and psychometric evaluation of the ICIQ vaginal symptoms questionnaire: the ICIQ-VS. BJOG 113(6):700–712
13. Yalcin I, Bump RC (2003) Validation of two global impression questionnaires for incontinence. Am J Obstet Gynecol 189(1):98–101
14. Viktrup L, Hayes RP, Wang P, Shen W (2012) Construct validation of patient global impression of severity (PGI-S) and improvement (PGI-I) questionnaires in the treatment of men with lower urinary tract symptoms secondary to benign prostatic hyperplasia. BMC Urol 12:30
15. Danish Urogynaecological Database, Available via http://www.dugabase.dk/. Accessed 1 August 2014
16. Guldberg R, Brostrom S, Hansen JK, Kaerlev L, Gradel KO, Norgard BM et al (2013) The Danish urogynaecological database: establishment, completeness and validity. Int Urogynecol J 24(6):983–990
17. Haylen BT, de Ridder D, Freeman RM, Swift SE, Berghmans B, Lee J et al (2010) An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for female pelvic floor dysfunction. Int Urogynecol J 21(1):5–26
18. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J et al (2007) Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60(1):34–42
19. Altman DG (1999) Practical statistics for medical research. Chapman & Hall/CRC, Boca Raton, xii, 611 s. p
20. Bland JM, Altman DG (1986) Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1(8476):307–310
21. Jonsson Funk M, Siddiqui NY, Kawasaki A, Wu JM (2012) Long-term outcomes after stress urinary incontinence surgery. J Gynecol Oncol 120(1):83–90
22. Abdel-Fattah M, Familusi A, Fielding S, Ford J, Bhattacharya S (2011) Primary and repeat surgical treatment for female pelvic organ prolapse and incontinence in parous women in the UK: a register linkage study. BMJ Open 1(2):e000206
23. Altman D, Vayrynen T, Engh ME, Axelsen S, Falconer C (2011) Anterior colporrhaphy versus transvaginal mesh for pelvic-organ prolapse. N Engl J Med 364(19):1826–1836
24. Maher C, Feiner B, Baessler K, Schmid C (2013) Surgical management of pelvic organ prolapse in women. Cochrane Database of Systematic Reviews. (3):CD004014
25. Bland JM, Altman DG (2003) Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 22(1):85–93
26. Kingsley G, Scott IC, Scott DL (2011) Quality of life and the outcome of established rheumatoid arthritis. Best Pract Res Clin Rheumatol 25(4):585–606

27. Oh SJ, Ku JH, Hong SK, Kim SW, Paick JS, Son H (2005) Factors influencing self-perceived disease severity in women with stress urinary incontinence combined with or without urge incontinence. Neurourol Urodyn 24(4):341–347

28. Aaronson N, Alonso J, Burnam A, Lohr KN, Patrick DL, Perrin E et al (2002) Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res 11(3):193–205

29. Haylen BT, de Ridder D, Freeman RM, Swift SE, Berghmans B, Lee J et al (2010) An International Urogynecological Association (IUGA)/International Continence Society (ICS) joint report on the terminology for female pelvic floor dysfunction. Neurourol Urodyn 29(1):4–20

30. Ulrich D, Guzman Rojas R, Dietz HP, Mann K, Trutnovsky G (2014) Use of a visual analog scale for evaluation of bother from pelvic organ prolapse. Ultrasound Obstet Gynecol 43(6): 693–697

31. Tincello DG, Owen RK, Slack MC, Abrams KR (2013) Validation of the patient global impression scales for use in detrusor overactivity: secondary analysis of the RELAX study. BJOG 120(2):212–216