GYNECOLOGY

Responsiveness and minimally important difference of SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse

Heidi S. Harvie, MD; Amanda A. Honeycutt, PhD; Simon J. Neuwahl, MSPH; Matthew D. Barber, MD; Holly E. Richter, MD; Anthony G. Visco, MD; Vivian W. Sung, MD; Jonathan P. Shepherd, MD; Rebecca G. Rogers, MD; Sharon Jakus-Waldman, MD; Donna Mazloomdoost, MD; for the NICHD Pelvic Floor Disorders Network

BACKGROUND: Utility preference scores are standardized, generic, health-related quality of life (HRQOL) measures that quantify disease severity and burden and summarize morbidity on a scale from 0 (death) to 1 (optimal health). Utility scores are widely used to measure HRQOL and in cost-effectiveness research.

OBJECTIVE: To determine the responsiveness, validity properties, and minimal important difference (MID) of utility scores, as measured by the Short Form 6D (SF-6D) and EuroQol (EQ-5D), in women undergoing surgery for pelvic organ prolapse (POP).

MATERIALS AND METHODS: This study combined data from 4 large, U.S., multicenter surgical trials enrolling 1321 women with pelvic organ prolapse. We collected condition-specific quality of life data using the Pelvic Floor Distress Inventory (PFDI) and Pelvic Floor Impact Questionnaire (PFIQ). A subset of women completed the SF6D; women in 2 trials also completed the EQ5D. Mean utility scores were compared from baseline to 12 months after surgery. Responsiveness was assessed using effect size (ES) and standardized response mean (SRM). Validity properties were assessed by (1) comparing changes in utility scores at 12 months between surgical successes and failures as defined in each study, and (2) correlating changes in utility scores with changes in the PFDI and PFIQ. MID was estimated using both anchor-based (SF-36 general health global rating scale "somewhat better" vs "no change") and distribution-based methods.

RESULTS: The mean SF-6D score improved 0.050, from 0.705 ± 0.126 at baseline to 0.761 ± 0.131 at 12 months ($P < .01$). The mean EQ-5D score improved 0.060, from 0.810 ± 0.15 at baseline to 0.868 ± 0.15 at 12 months ($P < .01$). The ES (0.13—0.61) and SRM (0.13—0.57) were in the small-to-moderate range, demonstrating the responsiveness of the SF-6D and EQ-5D similar to other conditions. SF-6D and EQ-5D scores improved more for prolapse reconstructive surgical successes than for failures. The SF-6D and EQ-5D scores correlated with each other ($r = 0.41$; $n = 645$) and with condition-specific instruments. Correlations with the PFDI and PFIQ and their prolapse subscales were in the low to moderate range ($r = 0.09$—0.38), similar to other studies. Using the anchor-based method, the MID was 0.026 for SF-6D and 0.025 for EQ-5D, within the range of MIDs reported in other populations and for other conditions. These findings were supported by distribution-based estimates.

CONCLUSION: The SF-6D and EQ-5D have good validity properties and are responsive, preference-based, utility and general HRQOL measures for women undergoing surgical treatment for prolapse. The MIDs for SF-6D and EQ-5D are similar and within the range found for other medical conditions.

Key words: EuroQol, health-related quality of life, minimal important difference, pelvic floor disorders, pelvic organ prolapse, Short Form 6D, utility score

Pelvic organ prolapse (POP) is common, with more than 188,659 inpatient procedures performed annually in the United States.[1] Treatment is recommended if symptoms are bothersome and affect quality of life.[2] Understanding the impact of POP and its treatment on health-related quality of life (HRQOL) is important both clinically and for cost-effectiveness research. Utility preference scores are generic HRQOL measures that quantify disease severity/burden and treatment impact; they summarize morbidity on a scale from 0 (death) to 1 (optimal health).[3] Utility scores allow comparison across a wide range of disease states, populations, and treatment modalities, and serve as an integral component of the quality-adjusted life years (QALYs) annualized measure of HRQOL. QALYs are commonly used when quantifying the benefits of a medical intervention for cost—utility analysis, the most common health economics evaluation.[4] Evaluating the psychometric properties of utility scores in women with POP will allow researchers and health care providers to measure HRQOL, to assess the effect of treatment on women's quality of life, to perform health economic evaluations, and to compare the cost-effectiveness of treatments for pelvic organ prolapse to treatments for other medical conditions.

General scales have been developed to measure utility preference scores for a wide variety of disease conditions and populations. These include the widely used multi-item, multi-attribute EuroQol (EQ-5D) and Short Form 6D (SF-6D).[5,6] Use of these indices with varied medical conditions facilitates the interpretation of results and comparison of disease and treatment outcomes. Although the SF-6D and EQ-5D have been evaluated in women with pelvic floor disorders, including POP,[7] the responsiveness and minimally important difference (MID) of these instruments in women undergoing

Cite this article as: Harvie HS, Honeycutt AA, Neuwahl SJ, et al. Responsiveness and minimally important difference of SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019;220:265.e1-11.

0002-9378/$36.00

https://doi.org/10.1016/j.ajog.2018.11.1094

AJOG at a Glance

Why was this study conducted?

This study was conducted to determine the responsiveness, validity properties, and minimal important difference (MID) of utility scores, as measured by the Short Form 6D (SF-6D) and EuroQol (EQ-5D), in women undergoing surgery for pelvic organ prolapse.

Key findings

The SF-6D and EQ-5D have good validity properties and are responsive preference-based utility and general health−related quality of life (HRQOL) measures for women undergoing surgical treatment for prolapse. The MIDs for SF-6D and EQ-5D are similar and within the range found for other medical conditions

What does this add to what is known?

The SF-6D and EQ-5D provide valid measures of HRQOL and utility scores in women with pelvic organ prolapse, and will allow comparison of the impact of these conditions to other disease states and provide essential data for cost-effectiveness research

surgery for POP have not been established. It is unknown whether these generic instruments that do not contain POP- or pelvic floor−related components and the general scores produced are sensitive to change following surgical treatment of POP.

Our objective was to determine the responsiveness, validity properties, and MID of utility scores as measured by the SF-6D and EQ-5D in women undergoing surgical repair for POP.

Materials and Methods

Study design and participants

This study is a retrospective analysis that combined data from 4 large, U.S., multicenter POP surgical trials conducted by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD)−sponsored Pelvic Floor Disorders Network (PFDN). The 4 trials were Outcomes following vaginal prolapse repair and mid urethral sling (OPUS),[8] Operations and pelvic muscle training in the management of apical support loss (OPTIMAL),[9] Colpopexy and Urinary Reduction Effort (CARE),[10] and the Colpocleisis Trial (COLPO).[11] All sites had institutional review board approval, and all women provided written informed consent. All participants underwent surgical correction of stage II−IV prolapse by experienced pelvic surgeons at 17 sites throughout the United States.

Two common, preference-based, multi-attribute health-status classification system instruments were used to estimate utility preference scores: EuroQol (EQ-5D) (EuroQol Group, http://www.euroqol.org), and Short Form 6D (SF-6D) (QualityMetric Incorporated, http://www.qualitymetric.com). The EQ-5D is scored on a scale of −0.59 to 1.00 and has 5 dimensions (mobility, self-care, usual activities, pain/discomfort, and anxiety/depression); each dimension has 3 levels for 243 possible unique health states.[5,12] The SF-6D is scored on a scale of 0.29−1.00 and has 6 dimensions (physical functioning, role limitation, social functioning, pain, mental health, vitality); each dimension has 2 to 6 levels for 18,000 possible unique health states.[6,12] Higher scores indicate better quality of life.

Pelvic floor symptoms were assessed by the Pelvic Floor Distress Inventory (PFDI), a validated, condition-specific questionnaire with 46 items and 3 scales, designed to evaluate distress caused by bowel, urinary, and POP complaints.[13] Pelvic floor−related quality of life was measured by the Pelvic Floor Impact Questionnaire (PFIQ), a validated, condition-specific HRQOL questionnaire with 93 items including bladder, bowel, and POP domains, each with 4 subscales.[13] Higher scores on the PFDI and PFIQ indicate worse symptoms and quality of life.

Surgical success was previously defined in the 4 individual trials. The 3 reconstructive studies (OPUS, OPTIMAL, and CARE) defined success using the following criteria: (1) absence of bothersome bulge symptoms as measured by the PFDI; (2) no prolapse beyond the hymen on Pelvic Organ Prolapse Quantification (POPQ) examination[14]; and (3) no subsequent retreatment for prolapse. OPTIMAL had an additional criterion: no descent of the vaginal apex more than one-third into the vaginal canal. Women not meeting all criteria were considered surgical failures. The obliterative study (COLPO) defined success as no prolapse beyond 1 cm inside the hymen on the POPQ. POPQ

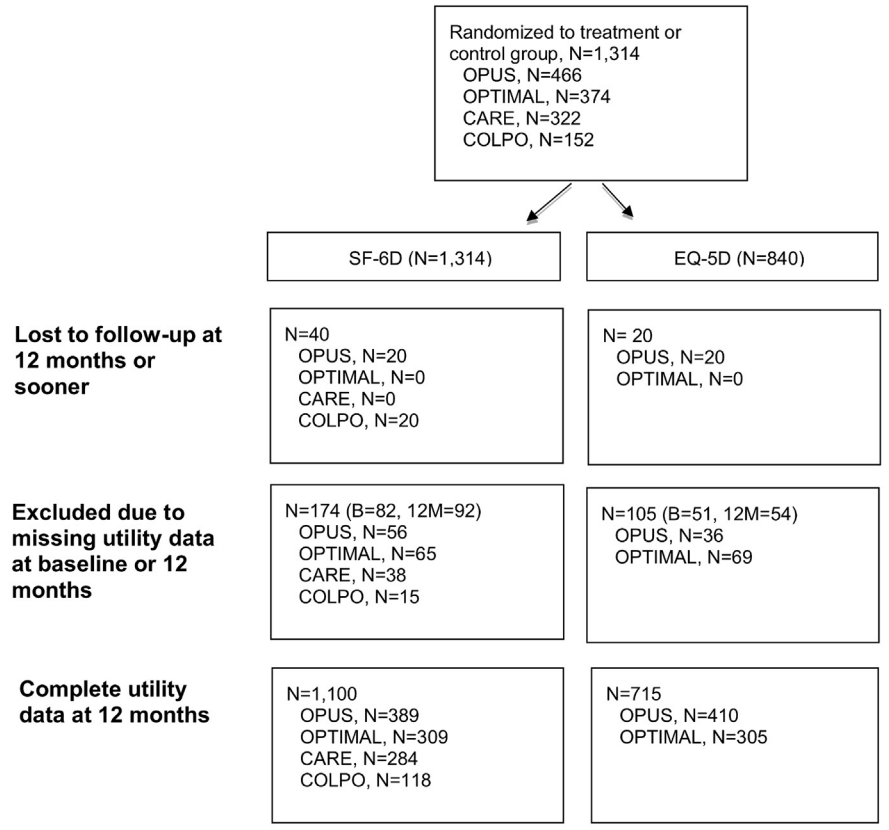
TABLE 1
Summary of studies

Study	Study type	Study n	Prolapse surgery type	SF-6D, n at 12 mo	EQ-5D. n at 12 Months
OPUS	RCT	337 RCT, 129 preference arm	Reconstructive	389	410
OPTIMAL	RCT	374	Reconstructive	309	305
CARE	RCT	322	Reconstructive	284	n/a
COLPO	Cohort	152	Obliterative	118	n/a
Total		1314		1,100	715

CARE, Colpopexy and Urinary Reduction Effort; *COLPO*, Colpocleisis Trial; *EQ-5D*, EuroQol; *OPTIMAL*, Operations and pelvic muscle training in the management of apical support loss; *OPUS*, Outcomes following vaginal prolapse repair and mid urethral sling; *RCT*, randomized controlled trial; *SF-6D*, Short Form 6D.

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

FIGURE 1
STROBE diagram



Note: B = Excluded due to missing baseline data, 12M = Excluded due to missing 12-month data

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

examinations, EQ-5D, SF-6D, PFDI, and PFIQ measures were administered at baseline and at 12-month follow-up visits. Although the 4 individual trials had different lengths of follow-up, ranging from 1 to 2 years, this analysis uses data from the common 12-month visit.

Responsiveness of the utility scores, an instrument's ability to detect change that occurs as the result of therapy (ie, POP surgery), was assessed in 2 ways via the effect size (ES) and the SRM. ES is the mean change in utility score from baseline to 12 months divided by the standard deviation (SD) of the baseline score. SRM is the mean change in score from baseline to 12 months divided by the SD of the change. ES and SRM were classified as small (0.2−0.49), moderate (0.5−0.79), and large (>0.8).[15]

Validity properties of the utility scores, that is, whether an instrument measures what it is intended to measure, was assessed in 2 ways. *Convergent and discriminant validity*, the degree to which 2 measures that should be related or unrelated are in fact related or unrelated, was analyzed by comparing changes in scores of the EQ-5D and SF-6D at 12 months between surgical successes and failures and assessing whether surgical successes had larger utility gains than failures. *Concurrent validity*, the relationship of an instrument with other measures of the same or similar construct that are measured at the same time, was analyzed by correlating changes in scores of the EQ-5D and SF-6D with each other and with the condition-specific instruments PFDI and PFIQ. Correlations were classified as low (<0.3), moderate (0.3−0.5), and high (>0.5).[16] Potential biases in the utility score measures from the SF-6D and EQ-5D were explored using a Bland−Altman plot to examine whether differences between the 2 measures depended on the initial health status of a patient.[17]

The minimal important difference (MID), the smallest change that can be regarded as clinically meaningful, was estimated for the utility scores by applying both anchor-based and distribution-based methods per current recommendations.[18−21] Anchor-based methods examine the relationship between a HRQOL measure and an independent external measure (or anchor) to elucidate the meaning of a particular change in the health construct.[18,20,21] The anchor-based MID approach used the difference in utility score corresponding to a self-reported small, but important, change on question 2 of the SF-36, 12 months postoperatively. Question 2, which is not part of the SF-6D, asks whether general health is much better, somewhat better, the same, somewhat worse, or much worse compared with before surgery. The MID was defined as the difference between the mean change in utility scores for patients whose global rating score was "somewhat better" and the mean change in utility for patients who reported they were "the same." Distribution-based MID approaches relate utility changes to either variability (eg, standard deviation) or reliability (eg, Cronbach's α).[22] Our approach focused on variability and was based on the baseline standard deviation of utility scores. The distribution-based MID was defined as 0.5 × baseline SD (ie, medium effect size) and 0.2 × baseline SD (ie, small effect size).[16] The MID estimates of the anchor- and distribution-based approaches were compared, and recommendations for the MID of the SF-6D and EQ-5D were made by consensus, consistent with the recommendations of Revicki et al.[19]

Statistical methods

Demographic data are presented as percentages or means. Categorical data were compared between studies using the Pearson χ^2; continuous variables were compared using analysis of variance (ANOVA). Ordinal variables such as pelvic organ prolapse stage were compared using a Kruskal−Wallis test.

TABLE 2
Baseline characteristics for SF-6D and EQ-5D samples by study

Baseline characteristics	SF-6D Sample						EQ-5D Sample			
	Overall	OPUS	OPTIMAL	CARE	COLPO	Within SF-6D *P* value[a]	Overall	OPUS	OPTIMAL	Within EQ-5D *P* value[a]
n	1100	389	309	284	118		715	410	305	
Age (mean ± SD)	62.9 (11.5)	63.6 (10.0)	57.2 (10.9)	61.6 (10.3)	78.1 (5.4)	<.01	61.2 (10.7)	63.8 (9.9)	57.7 (10.8)	<.01
Race						<.01				.94
White	89%	88%	87%	94%	91%		87%	87%	87%	
Black	6%	6%	5%	5%	8%		6%	6%	6%	
Other Race	5%	6%	7%	1%	1%		7%	7%	7%	
Ethnicity (%)						<.01				.03
Hispanic	11%	12%	20%	3%	1%		15%	12%	18%	
Non-Hispanic	89%	88%	80%	97%	99%		85%	88%	82%	
Baseline prolapse stage (%)						<.01				<.01
Stage 2	25%	28%	38%	13%	2%		32%	28%	38%	
Stage 3	63%	63%	57%	69%	64%		61%	63%	57%	
Stage 4	13%	9%	5%	18%	34%		7%	9%	5%	
Prior surgery for prolapse (%)	19%	14%	6%	38%	24%	<.01	11%	14%	8%	<.01
Prior surgery for urinary incontinence (%)	5%	3%	4%	7%	14%	<.01	3%	2%	4%	.18
Prior hysterectomy (%)	46%	39%	26%	71%	61%	<.01	34%	38%	28%	<.01

CARE, Colpopexy and Urinary Reduction Effort; *COLPO*, Colpocleisis Trial; *EQ-5D*, EuroQol; *OPTIMAL*, Operations and pelvic muscle training in the management of apical support loss; *OPUS*, Outcomes following vaginal prolapse repair and mid urethral sling; *SD*, standard deviation; *SF-6D*, Short Form 6D.

[a] *P* value comparing studies.

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

Responsiveness

Utility score change from baseline to 12 months was estimated and compared to 0 for each study using a *t* test and for the combined studies using a *t* test from an individual patient data random effects meta-analysis mixed effects linear model using the Kenward–Roger method to estimate degrees of freedom.[23,24] The ES was calculated as the mean change divided by the SD of the baseline score. The SRM was calculated as the mean change divided by the SD of the change. For the combined studies, the within-study SD estimate was obtained using ANOVA.

Validity

Change in utility scores for those with surgical success were compared with surgical failures for each study using ANOVA and for the combined studies using a meta-analysis mixed effects model as described above. Pearson correlations were calculated for changes in scores with each other and with condition-specific instruments.

Minimal important difference

Anchor-based MID and 95% confidence interval (CI) in utility score mean changes for those with a global rating score of "somewhat better" minus the mean change for those with a global rating score of "the same" were estimated for each study using ANOVA and for the combined studies using the meta-analysis mixed effects model. Homogeneity of MID across studies was assessed by a test of interaction between the study and global rating score category in a supportive ANOVA model. The percentage of all patients with a change equal to or better than the MID was calculated. Distribution-based MID was calculated as 0.2 and 0.5 times the baseline standard deviation of utility scores. For the combined studies, the within-study SD estimate was obtained using ANOVA. The 95% CI of the distribution-based MID estimates were calculated using a bootstrapping approach.

All reported *P* values were 2-sided. All statistical analyses were performed using Stata Statistical Software, Release 15 (StataCorp, College Station, TX) or SAS, version 9.4 (SAS Institute, Cary, NC).

Results

The methods and results of OPUS, OPTIMAL, CARE, and COLPO have been previously published.[8–11] These 4

TABLE 3
Responsiveness—changes in SF-6D and EQ-5D scores at 12 months after POP surgery

	SF-6D					EQ-5D				
	N	Baseline, mean (SD)	12 mo, mean (SD)	Change,[a] mean (SD/SE)	*P* value[b]	n	Baseline, mean (SD)	12 mo, mean (SD)	Change,[a] mean (SD/SE)	*P* value[b]
Overall	1,100	0.705 (0.126)	0.761 (0.131)	0.050 (SE 0.004)	<.0001	715	0.810 (0.149)	0.868 (0.154)	0.060 (SE 0.006)	<.0001
OPUS	389	0.719 (0.132)	0.779 (0.132)	0.060 (0.124)	<.01	410	0.806 (0.154)	0.854 (0.158)	0.047 (0.146)	<.01
OPTIMAL	309	0.677 (0.127)	0.754 (0.137)	0.077 (0.136)	<.01	305	0.815 (0.143)	0.887 (0.146)	0.072 (0.151)	<.01
CARE	284	0.716 (0.117)	0.763 (0.120)	0.047 (0.115)	<.01	0	N/A	N/A	N/A	N/A
COLPO	118	0.703 (0.114)	0.717 (0.122)	0.014 (0.108)	.15	0	N/A	N/A	N/A	N/A

CARE, Colpopexy and Urinary Reduction Effort; *COLPO*, Colpocleisis Trial; *EQ-5D*, EuroQol; *OPTIMAL*, Operations and pelvic muscle training in the management of apical support loss; *OPUS*, Outcomes following vaginal prolapse repair and mid urethral sling; *SD*, standard deviation; *SE*, standard error; *SF-6D*, Short Form 6D.

[a] Change in utility score of 0.03 is generally considered clinically significant; [b] *P* values in this table represent a paired *t* test of utility at baseline and follow-up. Standard errors in the overall results are adjusted for clustering by trial.

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

studies enrolled a total of 1314 women (Table 1). The SF-6D was included in all 4 studies; the EQ-5D was included in CARE and COLPO. Women who had utility data at 12 months for the SF-6D (n = 1100) and EQ-5D (n = 715) were included in the current analysis. We included 410 of 466 women (88%) from OPUS, 309 of 374 (83%) from OPTIMAL, 284 of 322 (88%) from CARE and 118 of 152 (78%) from COLPO in the current analysis; the number of included women from each trial reflects the maximal sample from the SF-6D or the EQ-5D responses (Figure 1).

Baseline characteristics for women completing the SF-6D and the EQ-5D from each of the 4 studies are shown in Table 2. Notable differences between study groups include older average age for COLPO at 78.1 ± 5. 4 years, whereas the mean age in the other studies ranged from 57.2 ± 10.9 to 63.8 ± 9.9 years. There were higher stages of POP in CARE and COLPO and higher rates of prior surgeries at baseline in these cohorts. We found no difference in baseline utility scores, and found there were few and small differences in baseline characteristics of subjects who were excluded from our analysis because of missing data and those who were included (data not shown).

Responsiveness

The mean SF-6D and EQ-5D scores improved at 12 months for each of the reconstructive studies ($P < .01$) but not for the obliterative study ($P = 0.15$). The overall SF-6D and EQ-5D scores also showed improvement ($P < .01$) (Table 3). Effect sizes for the SF-6D and EQ-5D were in the small (0.2−0.49) to moderate (0.5−0.79) ranges for the reconstructive studies but were <0.2 for the obliterative study (Table 4).

Validity

The SF-6D and EQ-5D scores improved more for POP surgical successes than failures, both overall and for the reconstructive studies. This was not seen for the obliterative study. These particular analyses were performed among the subset of women who had complete surgical success outcomes data at 12

TABLE 4
Responsiveness—effect size and standardized response mean for the SF-6D and EQ-5D

	SF-6D			EQ-5D		
	n	Effect size[b]	SRM[c]	n	Effect size[b]	SRM[c]
Overall[a]	1100	0.399	0.404	715	0.390	0.393
OPUS	389	0.459	0.487	410	0.308	0.324
OPTIMAL	309	0.608	0.571	305	0.502	0.477
CARE	284	0.400	0.408	N/A	N/A	N/A
COLPO	118	0.126	0.133	N/A	N/A	N/A

CARE, Colpopexy and Urinary Reduction Effort; *COLPO,* Colpocleisis Trial; *EQ-5D,* EuroQol; *N/A,* not available because the underlying EQ-5D data were not collected; *OPTIMAL,* Operations and pelvic muscle training in the management of apical support loss; *OPUS,* Outcomes following vaginal prolapse repair and mid urethral sling; *SF-6D,* Short Form 6D; *SRM,* standardized response mean.

[a] Overall effect size and SRM use standard deviations that account for within-trial clustering; [b] Effect size is the mean change in score divided by the standard deviation of the baseline score; [c] Standardized response mean (SRM) is the mean change in score divided by the SD of the change.

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

months for the SF-6D (n = 991) and EQ-5D (n = 666) (Table 5). The SF-6D and EQ-5D scores moderately correlated with each other (r = 0.41) and with the PFDI, PFIQ, and all their subscales. The correlations were low (r = 0.0−0.30) or moderate (r = 0.30−0.50) (Table 6). The Bland−Altman plot to assess agreement of the SF-6D and EQ-5D in measurement of baseline utility scores suggested that the differences between the 2 measurements was somewhat dependent on the individual's baseline health status. Similar to other studies' findings, women with low baseline quality of life (average utility scores <0.6) had lower scores on the EQ-5D than on the SF-6D, whereas those with high baseline quality of life (average utility scores >0.8) had higher scores on the EQ-5D. Most women had mid-range utility values where the EQ-5D and SF-6D were more aligned[25] (Figure 2).

Minimal important difference

Using the anchor-based method, the MID of the SF-6D ranged from 0.017 to 0.031 with a mean of 0.026; the MID of the EQ-5D ranged from 0.013 to 0.042 with a mean of 0.025. There was no statistical evidence of lack of homogeneity in the anchor-based MID estimates across the 4 studies that collected SF-6D utility scores ($P = .95$, F = 0.13, df = 3) or the 2 studies that collected EQ-5D sores ($P = 0.36$, F = 0.84 df = 1). MID estimates were therefore combined to produce overall weighted total mean MID estimates for the SF-6D and the EQ-5D (Table 7). The wide confidence intervals for the MID estimates, including negative values for the individual studies and EQ-5D overall, reflect both the uncertainty in the estimates and the relatively small study sizes.

The MID estimates using the distribution-based method indicated that the ES with 0.2 SD corresponded to an improvement in SF-6D score that ranged from 0.023 to 0.026, with a mean of 0.025; the EQ-5D improvement ranged from 0.029 to 0.031, with a mean of 0.030. The ES with 0.5 SD corresponded to an improvement in SF-6D score that ranged from 0.057 to 0.066, with a mean of 0.063. The corresponding improvement in the EQ-5D score ranged from 0.072 to 0.077, with a mean of 0.075. Anchor-based MID estimates were similar to distribution-based MIDs of 0.2 SD and substantially smaller than distribution-based MIDs of 0.5 SD. Figure 3 shows a forest plot of the SF-6D and EQ-5D anchor-based MID estimates for each trial and overall, along with their associated confidence limits. Distribution-based MIDs with small ES (0.2 SD) were similar to the anchor-based MID point estimates; distribution-based MIDs with medium ES (0.5 SD) corresponded closely to the upper confidence limits of the anchor-based MIDs.

Comment

This study shows that the SF-6D and EQ-5D have good validity properties and are responsive instruments for measuring the effect of reconstructive surgical treatment for POP. We observed moderate correlations of the SF-6D and EQ-5D scores with each other and condition-specific measures of POP symptoms and QOL, the PFDI and PFIQ, similar to other studies, demonstrating concurrent validity.[7] For women undergoing reconstructive pelvic surgery for POP, scores for both instruments improved at 12 months, with greater improvement for surgical successes than for failures, demonstrating responsiveness and convergent and discriminant validity. Finally, the MID of both instruments for the surgical treatment of POP were similar to values that have previously been reported for other medical conditions. These findings suggest that SF-6D and EQ-5D are valid preference-based utility and general HRQOL measures that could be used to evaluate the cost-effectiveness of reconstructive surgeries for POP.

The values of 2 measures of responsiveness, ES and SRM, were in the small-to-moderate range for reconstructive surgical procedures, ES (0.31−0.61) and SRM (0.32−0.57), comparable to the reported responsiveness of the SF-6D and EQ-5D for a variety of other medical conditions. In an analysis of 7 studies that included irritable bowel syndrome, leg ulcers, knee osteoarthritis, orthopedic limb reconstruction, early rheumatoid arthritis, chronic obstructive

TABLE 5
Validity—change in SF-6D and EQ-5D scores for surgical successes vs failures at 12 months[a]

	Subsample n[a]	% With surgical success[b]	SF-6D				EQ-5D		
			Surgical success utility score change[c] (SD/SE)	Surgical failure utility score change[c] (SD/SE)	*P* value	Subsample n[a]	Surgical success utility score change[c] (SD/SE)	Surgical failure utility score change[c] (SD/SE)	*P* value
Overall	991	80.8%	0.052 (0.005)	0.029 (0.009)	.023	666	0.066 (0.008)	0.038 (0.014)	.034
OPUS	357	75.1%	0.064 (0.125)	0.046 (0.113)	.211	381	0.052 (0.142)	0.037 (0.149)	.391
OPTIMAL	287	76.3%	0.087 (0.138)	0.045 (0.131)	.024	285	0.081 (0.148)	0.035 (0.153)	.035
CARE	260	92.7%	0.044 (0.107)	0.023 (0.139)	.538	N/A	N/A	N/A	N/A
COLPO	87	83.9%	0.014 (0.110)	0.016 (0.105)	.967	N/A	N/A	N/A	N/A

CARE, Colpopexy and Urinary Reduction Effort; *COLPO*, Colpoleisis Trial; *EQ-5D*, EuroQol; *OPTIMAL*, Operations and pelvic muscle training in the management of apical support loss; *OPUS*, Outcomes following vaginal prolapse repair and mid urethral sling; *SD*, standard deviation; *SE*, standard error; *SF-6D*, Short Form 6D.

[a] Uses only the subsample that had complete surgical success outcomes data at 12 months; [b] Percentage of the subsample that had surgical success outcome data available whose outcomes reflected successful surgery based on the definition of success used in the trial; [c] Change in utility score of 0.03 is generally considered clinically significant. SD shown for individual study results, and SE shown for overall study results. SE in the overall results are adjusted for clustering by trial.

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

pulmonary disease, and adults >65 years, Walters et al reported that the SRM for SF-6D was 0.30 (range, 0.11−0.48).[26] Similarly, another analysis of 11 studies, also with a variety of medical conditions, showed that the SRM of SF-6D was 0.39 (range, 0.12−0.87) and 0.24 for EQ-5D (range, −0.05 to 0.43).[12] Our findings suggest that SF-6D and EQ-5D measure the effect of surgical interventions for POP in a manner similar to treatments for other conditions.

The SF-6D was not responsive to the effect of obliterative surgery; scores did not significantly improve at 12 months (Table 3) and were not higher for surgical successes than for failures (Table 5). This may have been due to a relatively small sample size from a single study. Alternatively, the SF-6D may not be sensitive to the effect of obliterative surgical treatment of POP in older women who may have additional comorbidities that prevent them from undergoing reconstructive surgery, greatly reduce their general quality of life, or affect recovery after surgery. Our findings suggest that for women undergoing obliterative surgery, condition-specific instruments might be better than generic HRQOL instruments for measuring the efficacy of treatment, even though they do not allow calculation of utilities or comparison to other disease conditions.

The MID places the magnitude of change in a context to help clinicians assess whether surgeries result in meaningful improvement in patient HRQOL. The present study used 1 anchor-based approach and a conservative distribution-based approach to establish the MID for the SF-6D and EQ-5D. The application of multiple methods to determine the MID in a specific patient population generally results in a range of values, as seen in this study. Clinically, a single point estimate or narrow range of the MID is most helpful. Using an integrated approach, we report MIDs of 0.026 (0.017−0.031) for the SF-6D and 0.025 (0.013−0.042) for the EQ-5D for women undergoing surgery for POP. Approximately half of women across

TABLE 6
Validity—SF-6D and EQ-5D correlations[a] with each other and with indices of condition severity

	Total		OPUS		OPTIMAL		CARE		COLPO	
	SF-6D n = 1100	EQ-5D n = 715	SF-6D n = 389	EQ-5D n = 410	SF-6D n = 309	EQ-5D n = 305	SF-6D n = 260	EQ-5D N/A	SF-6D n = 87	EQ-5D N/A
SF-6D	—	0.41 (n = 645)	—	0.43 (n = 382)	—	0.37 (n = 263)	—	—	—	—
PFDI	0.29 (n = 1047)	0.24 (n = 711)	0.24 (n = 384)	0.15 (n = 406)	0.31 (n = 269)	0.31 (n = 305)	0.34 (n = 278)	—	0.21 (n = 116)	—
Prolapse subscore	0.29 (n = 1056)	0.25 (n = 715)	0.29 (n = 389)	0.19 (n = 410)	0.28 (n = 269)	0.31 (n = 305)	0.35 (n = 281)	—	0.27 (n = 117)	—
Bladder subscore	0.26 (n = 1055)	0.24 (n = 713)	0.23 (n = 386)	0.13 (n = 408)	0.30 (n = 269)	0.32 (n = 305)	0.30 (n = 283)	—	0.16 (n = 117)	—
Bowel subscore	0.20 (n = 1050)	0.14 (n = 713)	0.14 (n = 387)	0.05 (n = 408)	0.23 (n = 269)	0.21 (n = 305)	0.22 (n = 278)	—	0.13 (n = 116)	—
PFIQ	0.31 (n = 1043)	0.27 (n = 703)	0.32 (n = 380)	0.21 (n = 398)	0.31 (n = 269)	0.29 (n = 305)	0.38 (n = 279)	—	0.13 (n = 115)	—
Prolapse subscore	0.30 (n = 1043)	0.17 (n = 703)	0.33 (n = 380)	0.09 (n = 400)	0.27 (n = 269)	0.22 (n = 305)	0.37 (n = 279)	—	0.13 (n = 115)	—
Bladder subscore	0.31 (n = 1043)	0.23 (n = 703)	0.32 (n = 380)	0.17 (n = 398)	0.31 (n = 269)	0.26 (n = 305)	0.33 (n = 279)	—	0.19 (n = 115)	—
Bowel subscore	0.19 (n = 1043)	0.29 (n = 705)	0.15 (n = 380)	0.26 (n = 398)	0.24 (n = 269)	0.29 (n = 305)	0.20 (n = 279)	—	0.05 (n = 115)	—

EQ-5D, EuroQol; *N/A*, not available because the underlying EQ-5D data were not collected; *PFDI*, Pelvic Floor Distress Inventory; *PFIQ*, Pelvic Floor Impact Questionnaire; *SF-6D*, Short Form 6D.

[a] Pearson correlation coefficient.

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

all studies had a utility change greater than or equal to the MID. Our reported MIDs are comparable to those for other treatments and are within the range previously reported for other medical conditions. Results from preference-based measures in other populations suggest that MID = 0.03.[27] One analysis with 7 studies reported a mean SF-6D MID of 0.033 (range, 0.010−0.048)[26]; another with 11 studies reported a mean SF-6D MID of 0.041 (range, 0.011−0.097) and mean EQ-5D MID of 0.074 (range −0.011 to 0.140).[12]

Although both the SF-6D and EQ-5D weigh health states on a scale of 0 (dead) to 1 (optimal health) and the MID values for both instruments are similar, they are not directly comparable. In the current study, the Bland−Altman plot demonstrated differences in baseline SF-6D and EQ-5D scores. The pattern is similar to other studies that have demonstrated that the SF-6D does not appear to describe health states at the lower end of the scale as well as the EQ-5D but is better able to describe health states and to detect improvements toward the top of the utility scale.[28] These findings suggest that the SF-6D and EQ-5D scores are similar in their detection of changes in utility but different in the absolute amount of HRQOL measured. Therefore, although both the SF-6D and EQ-5D are able to measure the effect of POP, scores are not directly comparable. The baseline health state of women might inform the choice between SF-6D and EQ-5D utility instruments. EQ-5D may be preferred for women with lower baseline health and SF-6D preferred for women with higher baseline health.

The strengths of this study are the use of 4 multicenter studies of women undergoing various POP surgical procedures, the use of multiple approaches to establish MID estimates, and the use of validated and widely accepted patient-reported utility- and condition-specific outcome measures. Our data allow evaluation of the reliability and validity of generic HRQOL scales to assess the impact of surgical interventions for POP

TABLE 7
Minimal important difference (MID) of the SF-6D and EQ-5D

	SF-6D			EQ-5D		
	n	Participants with change ≥MID[a]	MID[a] (95% CI)	n	Participants with change ≥MID[a]	MID[a] (95% CI)
Overall	548	57.8%	0.026 (0.007, 0.045) n = 548	444	47.0%	0.025 (−0.005, 0.054) N = 350
OPUS	198	60.9%	0.031 (0.000, 0.062)	305	47.3%	0.013 (−0.026, 0.051)
OPTIMAL	154	64.1%	0.030 (−0.009, 0.068)	139	47.5%	0.042 (−0.005, 0.088)
CARE	132	60.2%	0.017 (−0.019, 0.053)	N/A	N/A	N/A
COLPO	64	43.2%	0.025 (−0.031, 0.080)	N/A	N/A	N/A

A total of 548 subjects were included in the SF-6D MID estimate (267 reporting "better" and 281 reporting "same"). A total of 350 subjects were included in the EQ-5D MID estimate (176 reporting "better" and 174 reporting "worse").

CARE, Colpopexy and Urinary Reduction Effort; *CI*, confidence interval; *COLPO*, Colpocleisis Trial; *EQ-5D*, EuroQol; N/A, not available because the underlying EQ-5D data were not collected; *OPTIMAL*, Operations and pelvic muscle training in the management of apical support loss; *OPUS*, Outcomes following vaginal prolapse repair and mid urethral sling; *SF-6D*, Short Form 6D.

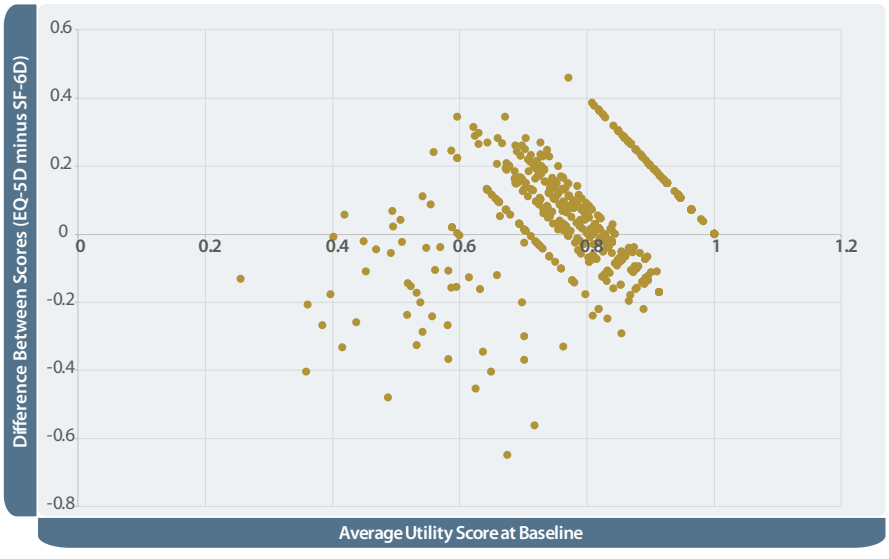
[a] MID anchor-based method uses mean difference between the subsample of subjects reporting "somewhat better" on the global measure of change question from the SF-6D and the subsample of subjects reporting "the same."

Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

and allow evaluation of utility measures for each study and in aggregate across multiple studies.

The limitations of this study include a smaller amount of EQ-5D data from 2 of the 4 studies and a smaller number of subjects undergoing obliterative procedures from only 1 study. This retrospective study was limited to the SF-6D and EQ-5D instruments administered as part of the 4 trials; other possible utility instrument options include the 15-D[29] and Health Utilities Index Mark 3 (HUI-3).[30] The use of multiple utility instruments in future studies for women undergoing surgery for the treatment of POP can further help inform the choice between these instruments in this population. Confidence in the MID values of SF-6D and EQ-5D from this study should evolve over time through additional research on different populations and contextual characteristics. As with other aspects of construct validity, responsiveness and MID values are confirmed based on accumulating evidence from multiple studies.

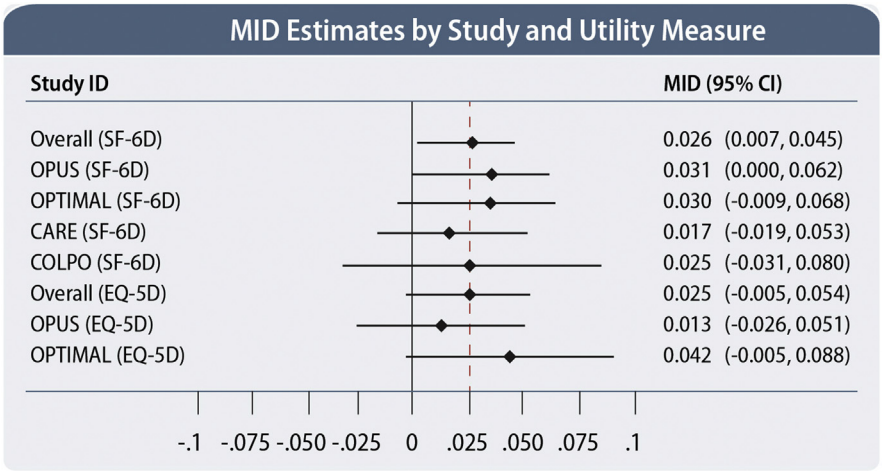
FIGURE 2
Bland Altman plot for SF-6D and EQ-5D scores at baseline



Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

In conclusion, the SF-6D and EQ-5D allow valid measurement of the impact of POP on HRQOL, facilitate comparison of HRQOL changes to other diseases and general population norms, and provide utility scores for calculating quality-adjusted life years for health economic evaluation. Our data suggest that reasonable estimates of MID in women undergoing surgical treatment for POP are approximately 0.026 for the SF-6D and 0.025 for the EQ-5D. Concern that these generic instruments with non−pelvic floor−specific items may lack sensitivity to the unique aspects of POP, and the impact on patients' lives has limited their use. Our findings support the use of these general utility preference instruments in women undergoing reconstructive surgical treatment for

FIGURE 3
Minimally important differences for the SF-6D and EQ-5D with anchor- and distribution-based methods



Harvie et al. SF-6D and EQ-5D utility scores for the treatment of pelvic organ prolapse. Am J Obstet Gynecol 2019.

POP; future intervention trials should include these measures to provide HRQOL outcome data and allow for cost-effectiveness analysis. ■

References

1. Bradley SL, Weidner AC, Siddiqui NY, Gandhi MP, Wu JM. Shifts in national rates of inpatient prolapse surgery emphasize current coding inadequacies. Female Pelvic Med Reconstr Surg 2011;17:204–8.

2. Committee on Practice Bulletins—Gynecology, American Urogynecologic Society. Practice bulletin no. 185: Pelvic organ prolapse. Obstet Gynecol 2017;130:e234–50.

3. Drummond MF, Schulpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the economic evaluation of health care programmes, 3rd ed. New York: Oxford University Press; 2005.

4. Mehrez A, Gafni A. Quality-adjusted life years, utility theory, and healthy-years equivalents. Med Decis Mak 1989;9:142–9.

5. Brooks R. EuroQol: the current state of play. Health Policy 1996;37:53–72.

6. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.

7. Harvie HS, Lee DD, Andy UU, Shea JA, Arya LA. Validity of utility measures for women with pelvic organ prolapse. Am J Obstet Gynecol 2018;218:119.

8. Wei JT, Nygaard I, Richter H, et al. A midurethral sling to reduce incontinence after vaginal prolapse repair. N Engl J Med 2012;366:2358–67.

9. Barber MD, Brubaker L, Burgio KL, et al. Comparison of 2 transvaginal surgical approaches and perioperative behavioral therapy for apical vaginal prolapse: the OPTIMAL randomized trial. JAMA 2014;311:1023–34.

10. Brubaker L, Cundiff GW, Fine P, et al. Abdominal sacrocolpopexy with Burch colposuspension to reduce urinary stress incontinence. N Engl J Med 2006;354:1557–66.

11. Fitzgerald MP, Richter HE, Bradley CS, et al. Pelvic support, pelvic symptoms, and patient satisfaction after colpocleisis. Int Urogynecol J Pelvic Floor Dysfunct 2008;19:1603–9.

12. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523–32.

13. Barber MD, Kuchibhatla MN, Pieper CF, Bump RC. Psychometric evaluation of 2 comprehensive condition-specific quality of life instruments for women with pelvic floor disorders. Am J Obstet Gynecol 2001;185:1388–95.

14. Bump RC, Mattiasson A, Bo K, et al. The standardization of terminology of female pelvic organ prolapse and pelvic floor dysfunction. Am J Obstet Gynecol 1996;175:10–7.

15. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63:737–45.

16. Cohen J. Statistical power analysis for the behavioral sciences. Mahwah, NJ: Lawrence Erlbaum Associates; 1988.

17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.

18. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol 2003;56:395–407.

19. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 2008;61:102–9.

20. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc 2002;77:371–83.

21. Wyrwich KW, Bullinger M, Aaronson N, Hays RD, Patrick DL, Symonds T. Estimating clinically significant differences in quality of life outcomes. Qual Life Res 2005;14:285–95.

22. Jelovsek JE, Chen Z, Markland AD, et al. Minimum important differences for scales assessing symptom severity and quality of life in patients with fecal incontinence. Female Pelvic Med Reconstr Surg 2014;20:342–8.

23. Higgins JP, Whitehead A, Turner RM, Omar R, Thompson SG. Meta-analysis of continuous outcome data from individual patients. Stat Med 2001;20:2219–41.

24. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 1997;53:983–97.

25. Obradovic M, Lal A, Liedgens H. Validity and responsiveness of EuroQol-5 dimension (EQ-5D) versus Short Form-6 dimension (SF-6D) questionnaire in chronic pain. Health Qual Life Outcomes 2013;11:110.

26. Walters SJ, Brazier JE. What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual Life Outcomes 2003;1:4.

27. Drummond MF. Introducing economic and quality of life measures into clinical studies. Ann Med 2001;33:344–9.

28. Longworth L, Stirling B. An empirical comparison of EQ-5D and SF-6D in liver transplant patients. Health Econ 2003;12:1061–7.

29. Sintonen H. The 15D instrument of health-related quality of life: properties and applications. Ann Med 2001;33:328–36.

30. Feeny D, Furlong W, Boyle M, Torrance GW. Multi-attribute health status classification systems. Health Utilities Index. PharmacoEconomics 1995;7:490–502.

Author and article information

From the Department of Obstetrics and Gynecology (Dr Harvie), University of Pennsylvania, Philadelphia, PA; Clinical Research Network Coordination (Dr Honeycutt and Mr Neuwahl), RTI, Research Triangle Park, NC; Department of Obstetrics and Gynecology (Dr Barber), Cleveland, Cleveland, OH; Department of Obstetrics and Gynecology (Dr Richter), University of Alabama at Birmingham, Birmingham, AL; Department of Obstetrics and Gynecology (Dr Visco), Duke University, Durham, NC; Department of Obstetrics and Gynecology (Dr Sung), Brown, Providence, RI; Department of Obstetrics and Gynecology (Dr Shepherd), University of Pittsburgh, Pittsburgh, PA; Department of Obstetrics and Gynecology (Dr Rogers), University of New Mexico and Department of Women's Health, Dell Medical School, University of

Texas, Austin, TX; Department of Obstetrics and Gynecology (Dr Jakus-Waldman), Kaiser Downey, Downey, CA; *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (Dr Mazloomdoost), National Institutes of Health, Bethesda, MD.

Received May 4, 2018; revised Nov. 3, 2018; accepted Nov. 15, 2018.

M.D.B reports Royalties from Elsevier, UpToDate; H.E.R. reports the following: Pelvalon, Consultant and grant support, Renovia, Consultant; Royalties from UptoDate; Travel funds IUGA; and Travel funds, ICS; NICHD, NIA, PCORI. A.G. reports Ninomed stock ownership. J.P.S. reports research support from Site PI for Myrbetriq trial supported by Astellas. R.G.R is DSMB chair for the TRANSFORM trial sponsored by American Medical Systems, and reports Stipend and travel from ABOG and IUGA and Royalties from Uptodate. The other authors report no conflict of interest.

Financial support for this project was provided by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development and the National Institutes of Health Office of Research on Women's Health (5U24HD069031-07).

Presented in part at PFD Week 2017, Providence, RI, Oct. 3, 2017.

Corresponding author: Heidi Harvie, MD, MBS, MSCE, Division of Urogynecology and Reconstructive Pelvic Surgery, 3400 Spruce Street, Fifth Floor, Dulles Building, Hospital of the University of Pennsylvania, Philadelphia, PA 19104. hharvie@obgyn.upenn.edu