

# A comparison of six outcome measures across the recovery period after distal radius fixation—Which to use and when?

Journal of Orthopaedic Surgery  
29(1) 1–10

© The Author(s) 2021

Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/2309499020971866  
journals.sagepub.com/home/osj



Christian Fang<sup>1</sup> , Evan Fang<sup>1</sup>, Dennis KH Yee<sup>1</sup>, Kenny Kwan<sup>1</sup> , Gladys Leung<sup>2</sup> and Frankie Leung<sup>1</sup>

## Abstract

**Purpose:** Many standardized outcome measures exist to measure recovery after surgical fixation of distal radius fractures, however, choosing the optimal instrument is difficult. We evaluated responsiveness, ceiling/floor effects, and criterion validity over multiple time intervals across a 2-year follow-up period for six commonly used instruments. **Methods:** A total of 259 patients who received open reduction and internal fixation for distal radius fractures between 2012 and 2015 were recruited. Patients were administered the Patient-Rated Wrist Evaluation (PRWE), Shortened Disabilities of the Arm, Shoulder and Hand questionnaire (QuickDASH), Green and O'Brien score (Cooney modification) (CGNO), Gartland and Werley score (Sarmiento modification) (SGNW), flexion-extension arc (FEArc), and grip fraction test (GripFrac) at 1.5, 3, 6, 12, and 24 months postoperatively. Responsiveness was evaluated by calculating standardized response means (SRM) and Cohen's *d* effect sizes (ES), and by correlating each instrument's change scores against those of QuickDASH and PRWE, which were also used as external comparators to assess criterion validity. Ceiling/floor effects were calculated for all measures at each time point. **Results:** SRM (1.5–24 months) were 1.81, 1.77, 1.43, 1.16, 2.23, 2.45 and ES (1.5–24 months) were 1.81, 1.82, 1.95, 1.31, 1.99 and 2.90 for QuickDASH, PRWE, CGNO, SGNW, FEArc, and GripFrac respectively. Spearman correlation coefficients against QuickDASH at 24 months were: 0.809, 0.248, 0.563, 0.285, and 0.318 for PRWE, CGNO, SGNW, FEArc, and GripFrac respectively. Significant (>15% of patients reaching maximum score) ceiling effects were observed before 6 months for PRWE and SGNW. **Conclusions:** Our evidence supports the use of QuickDASH, PRWE, FEArc and GripFrac up to 6 months postsurgery, and QuickDASH and PRWE after 6 months.

**Level of evidence:** Level II.

## Keywords

Gartland and Werley score, Green and O'Brien score, patient-reported outcome measure, PRWE, QuickDASH, responsiveness

Date received: 15 July 2020; Received revised 6 October 2020; accepted: 15 October 2020

## Introduction

Standardized outcome measurement in orthopedics is essential for distinguishing the effects of different treatment methods and aiding research to produce better ones. Distal radius fractures are one of the most common fractures, occurring from a variety of mechanisms in people of all ages, from simple falls to high-energy sports injuries.<sup>1</sup> A wide range of standardized instruments are available to evaluate patient outcomes for these and other upper limb

<sup>1</sup> Department of Orthopedics and Traumatology, Queen Mary Hospital, The University of Hong Kong, Hong Kong, China

<sup>2</sup> Occupational Therapy Unit, David Trench Rehabilitation Centre, Hong Kong, China

### Corresponding author:

Dennis KH Yee, Department of Orthopedics and Traumatology, Queen Mary Hospital, The University of Hong Kong, 102 Pok Fu Lam Road, Hong Kong, China.

Email: yeedns@gmail.com



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

injuries, however choosing the optimal tool to do so can be challenging.

Different instruments measure different dimensions of clinically relevant outcome variables. Commonly measured outcomes of distal radius fractures include joint alignment, range of motion, strength, pain, task-specific functioning, perceptions of daily living, and emotional and mental health. These variables comprise five levels of quality of life: (1) biological and physiological status, (2) symptoms, (3) function, (4) general health perceptions, and (5) overall quality of life.<sup>2</sup> Because each patient is unique in his or her values, concerns, and expectations, it is difficult to strictly define a “best measure.” A questionnaire that measures the ability to perform heavy chores may be of little relevance to a low-demand elderly patient, and a measure that reveals a perfect range of motion may overestimate the patient’s actual wellbeing. Due to the subjective nature of quality of life, standardized measurement becomes increasingly complex at each additional level,<sup>2</sup> and yet increasingly relevant to the patient.

Choosing the ideal instrument is further complicated by the fact that the importance of certain outcome variables changes across the rehabilitation period. In the early stages following surgery, pain tends to be the most salient outcome, whereas the ability to perform tasks becomes of greater concern as healing progresses and pain subsides.<sup>3</sup> Return to work could be a priority immediately following treatment, or not at all, depending on the patient. Thus, we hypothesize that certain scoring systems are better suited than others for distinguishing good and bad outcomes across different phases of recovery.

We assessed the performance of six commonly used outcome instruments in terms of their responsiveness, ceiling/floor effects and criterion validity over a 2-year period following surgical treatment of distal radius fractures. Based on their popularity in the literature and relevance to our daily practice, the instruments selected were: (1) the Shortened Disabilities of the Arm, Shoulder and Hand questionnaire (*QuickDASH*),<sup>4</sup> (2) the Patient-Rated Wrist Evaluation (*PRWE*),<sup>5</sup> (3) wrist flexion and extension range of motion arc (*FEArc*), (4) handgrip strength fraction (*GripFrac*), (5) the Cooney modification of the Green and O’Brien score (*CGNO*),<sup>6</sup> and (6) the Sarmiento modification of the Gartland and Werley score (*SGNW*).<sup>7,8</sup> We summarize the evidence for the use of each instrument across the rehabilitation period based on an adapted set of predefined criteria.

## Methods

### Study protocol

This prospective cohort study was carried out in a publicly funded university healthcare institute in a high-income region (GDP per capita USD\$48,915).<sup>9</sup> Patients who received open reduction and internal fixation (ORIF) for

a distal radius fracture between July 2012 and June 2015 were screened for enrollment. Inclusion criteria were: (1) AO/OTA classification type 23.A, B, or C<sup>10</sup> distal radius fracture, (2) treatment within 3 weeks of injury, (3) treatment by volar or dorsal plate fixation, (4) willingness to participate in a protocol-driven rehabilitation schedule, and (5) expected capacity to complete multiple consecutive questionnaires for up to 1 year. Exclusion criteria were: (1) treatment delayed beyond 3 weeks of injury, (2) pathological fracture, (3) polytrauma (Injury Severity Score >16), (4) concomitant upper limb trauma, and (5) compromised cognitive state. Ethical approval was waived for non-interventional studies on routinely collected clinical data by our institutional review board at the time of the study. Verbal informed consent was obtained from all patients prior to study inclusion. All patient-rated outcome instruments were administered under supervision by trained research personnel, and all observer-rated items were graded by physicians or occupational therapists.

Patients followed a standardized rehabilitation regime at the same designated outpatient rehabilitation center under the supervision of physiotherapists and occupational therapists. Rehabilitation protocol dictated early gentle active range of motion training before 1.5 months, progressing to passive range of motion and strengthening exercises after 1.5 months if fracture union was evident on AP and lateral radiographs.

All patients were seen in the outpatient follow-up clinic at 1.5, 3, 6, 12 and 24 months post-surgery with all six outcome instruments administered at each visit. The study period of 24 months was chosen based on our hypothesis that the change in outcome measures might not plateau before 2 years.

### Outcome measures

*The Shortened Disabilities of the Arm, Shoulder and Hand questionnaire (QuickDASH).* *QuickDASH* is a patient-rated measure of function, symptoms, and quality of life pertaining to the upper limb.<sup>4</sup> It was developed as a more convenient abbreviation of the original Disabilities of the Arm, Shoulder and Hand (*DASH*) questionnaire, a highly validated measure of upper extremity functional status.<sup>11</sup> *QuickDASH* consists of 11 self-administered items with a final score ranging from 0 (least disability) to 100 (most severe disability). It has been extensively validated in the literature and has demonstrated good concurrent validity and responsiveness compared to the original *DASH* in the context of distal radius fractures.<sup>12</sup> A language- and culture-validated version of the *QuickDASH* that matched the study population<sup>13–15</sup> was used in our study.

*Patient-Rated Wrist Evaluation (PRWE).* *PRWE* is a patient-rated, wrist-specific instrument developed specifically to measure pain and disability in patients after distal radius fracture.<sup>5</sup> It consists of two subscales: Pain, which contains

5 items rated from 0–10, and Function, which consists of 10 items rated from 0–10. The Function score is divided by 2 and added to the Pain score to give a total score out of a maximum of 100 points, with higher scores indicating poorer results. *PRWE* is commonly used and extensively validated in the literature,<sup>16</sup> although in this study we used a local language- and culture-matched version that has been validated to a lesser extent.<sup>17</sup>

**Flexion-extension arc range (FEArc) and grip strength fraction (GripFrac).** Range of motion and grip strength are clinician-rated tests that have been shown to be sensitive to change and to significantly predict *DASH* scores.<sup>18</sup> For the *FEArc* test, wrist joint range of motion was manually assessed via a goniometer. Grip strength was assessed via hydraulic hand dynamometer (JAMAR, Bolingbrook, IL) with the patient seated, elbow flexed to 90 degrees, and forearm in the neutral position. *GripFrac* was calculated as the grip strength fraction of the injured side to the contralateral side, expressed as a percentage. Hand dominance was not considered in the *GripFrac* measurement as we aimed to evaluate the measurement properties of the instrument as a whole, and due to a lack of reliable pre-injury data for grip strength. Both *FEArc* and *GripFrac* were measured and documented by an occupational therapist with an average of three trials recorded.

**Cooney modification of the Green and O'Brien score (CGNO).** *CGNO* is a clinician-rated, wrist-specific assessment measuring pain (25 points), functional status (25 points) range of motion (25 points), and grip strength (25 points) (Online Appendix A).<sup>6</sup> This measure has not been extensively validated although it is credited for its simplicity and ease of use.<sup>19</sup> The final score is graded as Excellent (90–100 points), Good (80–89 points), Fair (65–79 points) or Poor (<65 points).

**Sarmiento modification of the Gartland and Werley score (SGNW).** *SGNW* is a mixed clinician- and patient-rated, wrist-specific assessment system (Online Appendix B).<sup>7</sup> The tool consists of clinician-rated items including residual deformity (3 points), range of motion and grip strength (5 points), nerve compression (3 points), finger stiffness (2 points), and arthritis change (5 points), as well as a patient-rated subjective evaluation (5 points). Several different methods of scoring have been reported in the literature.<sup>20</sup> Our method allowed for a maximum score of 24 points, with 0–2 points being graded as “Excellent,” 3–8 points as “Good,” 9–20 points as “Fair” and 21 or more points as “Poor.” This measure has commonly been reported in the literature despite a lack of validation.<sup>21</sup>

### Statistical analysis

**Standardization of scores.** To enable direct comparison of mean scores between outcome measures, each score was standardized to a score of 0–100, with 0 representing lowest

function and most severe symptoms, and 100 representing best function and least severe symptoms. For the purposes of this study, the maximum value obtained from the sample (140 degrees) was considered the ceiling for the standardized *FEArc* scale. For *GripFrac*, values greater than 100 were truncated to a maximum score of 100. The conversions were calculated as follows:

$$\begin{aligned} \text{Standardized QuickDASH} &= 100 - \text{QuickDASH} \\ \text{Standardized FEArc} &= (\text{FEArc}/140) \times 100 \\ \text{Standardized GripFrac} &= \text{GripFrac (truncated to 100)} \\ \text{Standardized PRWE} &= 100 - \text{PRWE} \\ \text{Standardized CGNO} &= \text{CGNO (no standardization required)} \\ \text{Standardized SGNW} &= [1 - (\text{SGNW}/24)] \times 100 \end{aligned}$$

**Responsiveness.** Responsiveness is the ability of an outcome measure to detect a change in the construct of interest.<sup>22</sup> In addition to validity and reliability, responsiveness is an essential property of repeated outcome measures when a change in the construct of interest is expected to have occurred.<sup>23</sup> Responsiveness was evaluated using a mixed distribution- and criterion-based approach. In a distribution-based approach, the distribution and change scores of the sample are analyzed.<sup>24</sup> In a criterion-based approach, as defined by the COSMIN panel, change scores are correlated against those of a comparator measure that is presumed to be a “gold standard.”<sup>25</sup>

For the distribution-based approach, Cohen’s *d* effect size (ES) and standardized response mean (SRM) were calculated for the six outcome measures for the 1.5- to 3-month, 3- to 6-month, 6- to 12-month, 12- to 24-month and 1.5- to 24-month time intervals. Both indices were calculated since there is no consensus on which index is superior.<sup>26</sup> ES was calculated as:

$$d = \frac{\bar{x}_f - \bar{x}_i}{SD_{pooled}}$$

where  $\bar{x}_f$  and  $\bar{x}_i$  are the mean scores for the final and initial time points, respectively, and  $SD_{pooled}$  is the pooled standard deviation between the two time points, which was calculated as:

$$SD_{pooled} = \sqrt{\frac{SD_i^2 + SD_f^2}{2}}$$

where  $SD_i$  and  $SD_f$  are the standard deviations for the initial and final time points, respectively. SRM was calculated as:

$$SRM = \frac{\bar{x}_f - \bar{x}_i}{SD_c}$$

where  $\bar{x}_f$  and  $\bar{x}_i$  are the mean scores for the final and initial time points, respectively, and  $SD_c$  is the standard deviation for the change scores between the two time points. Effect sizes were evaluated using Cohen’s thresholds, with 0.20, 0.50, and 0.80 considered low, moderate and high respectively.<sup>27</sup>

For the criterion-based approach, change scores for *FEArc*, *GripFrac*, *CGNO*, and *SGNW* were calculated for the 1.5- to the 24-month interval, and correlated against corresponding change scores for *QuickDASH* and *PRWE* using Spearman's rho correlation coefficient. Correlation coefficient values were taken to represent high (>0.7), moderate (0.5–0.7), low (0.3–0.5) or negligible (<0.3) correlations.<sup>28</sup> *QuickDASH* and *PRWE* were considered valid comparators based on their extensive validation and demonstrated responsiveness in the literature.<sup>11,12</sup> Higher correlation of change scores with *QuickDASH* and *PRWE* was considered better evidence for responsiveness to changes in pain and function. Furthermore, all change scores and effect sizes were expected to occur in the direction indicating improved function and/or symptoms, since the true change was assumed to occur in this direction for all outcome measures at all time intervals. Values in the opposite direction would be taken as evidence for poor responsiveness for a given outcome measure.

**Ceiling and floor effects.** Ceiling and floor effects occur when a substantial proportion of patients obtain the maximum or minimum score for a given scoring system. Instrument scales should be “in-range” in order to discriminate outcomes at patients' best or worst statuses. The proportion of patients reaching the maximum and minimum scores for each scoring system were calculated for each follow-up interval. Ceiling or floor effects were considered significant if 15% of patients or more reached the upper or lower bounds of the scale, as per the definition by McHorney.<sup>29</sup> Ceiling/floor effects greater than 15% at or prior to 6 months follow-up were considered unsatisfactory as full recovery was not expected to occur by this time in a

majority of cases. It should be noted that while ceiling effects are expressed for *GripFrac* for the purposes of this study, in practice there is no reliable maximum score since the relative strength of the contralateral wrist varies by patient.

**Criterion validity.** In order to “double-check” for any unexpected non-agreements, all outcome instruments were correlated against *PRWE* and *QuickDASH* scores for each follow-up point using Spearman's rho correlation coefficient, with coefficient values taken as high (>0.7), moderate (0.5–0.7), low (0.3–0.5) or negligible (<0.3) correlations.<sup>28</sup> *PRWE* and *QuickDASH* were assumed to be valid comparators due to their extensive validation in the literature.<sup>11,12</sup>

**Handling of missing data.** As there currently exists no consensus on the best method for handling missing data when assessing PROM measurement properties,<sup>30</sup> missing instrument responses were excluded pairwise to minimize data exclusion.

## Results

A total of 259 patients (147 female, 112 male) were recruited. Mean age was 55.6 years (range 16–86 years). The number of patients with complete data at 12 and 24 months was 209 and 111, respectively. The number of patients who completed a minimum of two, three, four or all five follow-ups was 251, 186, 130, and 52, respectively. Fracture characteristics are presented in Table 1.

**Patient outcomes.** All six outcome measures demonstrated improvements across all time intervals. Table 2 shows the non-standardized mean and SD of the six measures at each follow-up point. Figure 1 shows the mean trends of the six outcome measures on the standardized 1–100 scale with 95% confidence intervals.

**Responsiveness by distribution-based approach.** All outcome measures demonstrated mean differences in the direction indicating improved function and/or symptoms, as hypothesized. All outcome measures had moderate to high (0.5 or higher) ES and SRM between 1.5 and 3 months. Between 3 and 6 months, *FEArc*, *GripFrac* and *CGNO* had

**Table 1.** Fracture characteristics of 259 patients.

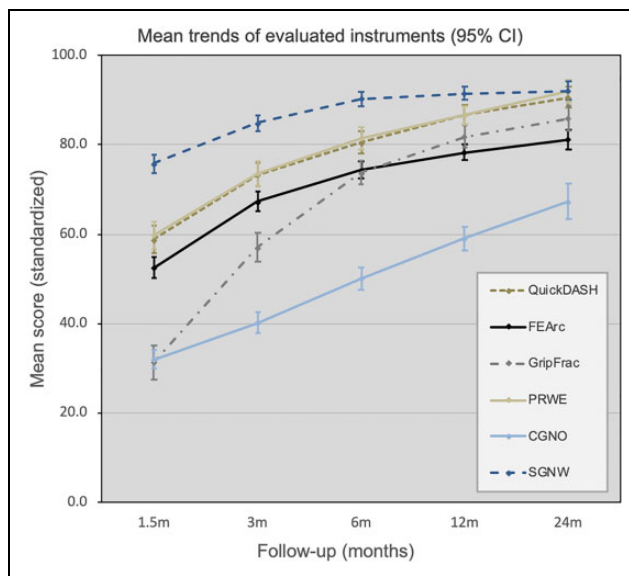
AO 23. Fracture Class	Frequency	Percent
A2	25	9.7%
A3	34	13.1%
B2	4	1.5%
B3	21	8.1%
C1	38	14.7%
C2	80	30.9%
C3	57	22.0%

**Table 2.** Mean, SD (raw scores) and number of patients with complete data at each follow-up point for the six outcome measures.

	1.5 months			3 months			6 months			12 months			24 months		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
QuickDASH	41.6	20.4	178	28.9	19.4	194	22.7	17.9	196	18.7	16.6	222	14.4	14.2	136
FEArc	73.5	21.8	171	94.3	20.7	186	104.1	18.5	188	110.1	17.7	221	114.6	18.5	136
GripFrac	33.2	21.9	116	57.1	22.4	183	73.9	19.0	183	82.9	19.3	217	88.3	18.8	131
PRWE	40.6	20.8	176	28.4	19.1	192	21.8	17.8	186	18.0	16.0	214	12.5	13.7	114
CGNO	32.3	13.8	175	40.2	16.6	190	50.1	17.1	186	59.0	20.1	215	67.9	21.5	114
SGNW	6.2	3.4	169	4.7	3.2	187	3.6	2.6	178	3.2	2.6	209	3.0	2.6	111

moderate to high ES and SRM, while *QuickDASH*, *PRWE* and *SGNW* had one or both effect indices fall into the low range. By 6 months, all measures had low ES and SRM except *GripFrac*, which had an SRM of 0.59. *SGNW* had the lowest effect sizes overall (1.5–24 months). These results alone suggest that *SGNW* is not responsive to change in the context of distal radius fractures. ES and SRM values for each follow-up interval are presented in Table 3. Trends in ES and SRM across time intervals are displayed in Figure 2.

**Responsiveness by criterion-based approach.** *PRWE* and *QuickDASH* change scores correlated highly with each other overall ( $r_s = 0.754$ ) (Table 4). *FEArc* and *GripFrac* change scores showed moderate correlations with *QuickDASH* and low correlations with *PRWE*. *CGNO* and *SGNW* change scores showed low to negligible correlations with both *QuickDASH* and *PRWE*, suggesting a lack of responsiveness in these measures.



**Figure 1.** Converted (0–100) mean trends across follow-up with error bars representing 95% confidence intervals.

**Ceiling and floor effects.** Ceiling effects generally increased at each subsequent time interval. Some patients who reached the ceiling for a given score were subsequently lost to follow-up, particularly after 12 months. *SGNW* had the greatest ceiling effects at all time intervals. *PRWE* and *SGNW* demonstrated significant ceiling effects at or prior to 6 months follow-up. No significant floor effects were observed for any outcome measure. Proportions of ceiling and floor effects are presented in Figure 3.

**Criterion validity.** All outcome measures displayed highly statistically significant correlations ( $p < 0.01$ ) with *QuickDASH* and *PRWE* at all follow-up points, except *FEArc* with *QuickDASH* at 24 months ( $p < 0.05$ ). Correlations with *QuickDASH* and *PRWE* generally decreased with time. *PRWE* correlated highly with *QuickDASH* at all follow-up points. *FEArc* showed low correlations with *QuickDASH* and *PRWE* at all time points. *GripFrac*, *CGNO*, and *SGNW* showed low to moderate correlations with *QuickDASH* and *PRWE*. Values for Spearman correlations with *QuickDASH* and *PRWE* are presented in Table 5. Scatter plots of relationships between criterion comparators and other outcome measures are presented in Figure 4. A summary of evidence for use of each outcome measures is provided in Table 6.

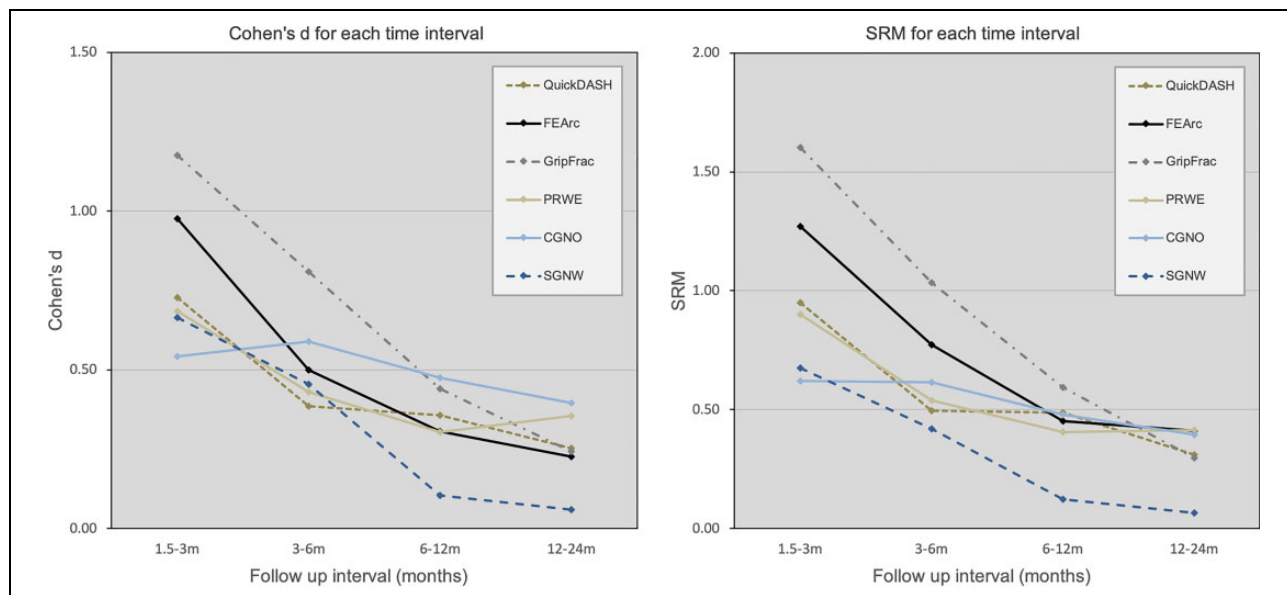
## Discussion

Based on our evaluation of responsiveness, ceiling/floor effects, and criterion validity, there is good evidence to support the use of *QuickDASH*, *PRWE*, *FEArc* and *GripFrac* up to 6 months postsurgery to evaluate recovery following distal radius fractures. After 6 months, our data support the use of *QuickDASH* and *PRWE* only, as there is negative evidence for responsiveness or criterion validity for the other measures. Our study provides evidence for just a few of the measurement properties that are of interest when selecting an outcome measure for clinical or research use. Recently, efforts have been made to reach consensus on which measurement properties of outcome measures are most important and how these properties should be measured.<sup>31,32</sup> Optimal selection of outcome measures ultimately relies on the accumulation of high-quality

**Table 3.** Cohen’s d (d) and standardized response mean (SRM) of outcome measures for each follow-up interval.<sup>a</sup>

	1.5–3 months		3–6 months		6–12 months		12–24 months		1.5–24 months	
	d	SRM	d	SRM	d	SRM	d	SRM	d	SRM
QuickDASH	0.73	0.95	0.39	0.49	0.36	0.49	0.25	0.31	<b>1.81</b>	<b>1.81</b>
FEArc	0.98	1.27	0.50	0.77	0.31	0.45	0.23	0.41	<b>1.99</b>	<b>2.23</b>
GripFrac	1.18	1.60	0.81	1.03	0.44	0.59	0.24	0.30	<b>2.90</b>	<b>2.45</b>
PRWE	0.69	0.90	0.43	0.54	0.30	0.41	0.35	0.41	<b>1.82</b>	<b>1.77</b>
CGNO	0.54	0.62	0.59	0.61	0.48	0.48	0.40	0.39	<b>1.95</b>	<b>1.43</b>
SGNW	0.66	0.68	0.46	0.42	0.10	0.12	0.06	0.07	<b>1.31</b>	<b>1.16</b>

<sup>a</sup>Directionality is not displayed as all effect sizes were in the expected direction indicating better functionality/decreased pain.



**Figure 2.** Cohen's d and SRM across follow-up intervals for the six outcome measures.

**Table 4.** Spearman correlation coefficients of 1.5- to 24-month change scores with *QuickDASH* and *PRWE* change scores.<sup>a</sup>

	PRWE	P	FEArc	P	GripFrac	P	CGNO	P	SGNW	P
QuickDASH	.754	.000	-.662	.000	-.526	.000	-.397	.001	.417	.001
PRWE			-.478	.000	-.428	.005	-.327	.008	.279	.031

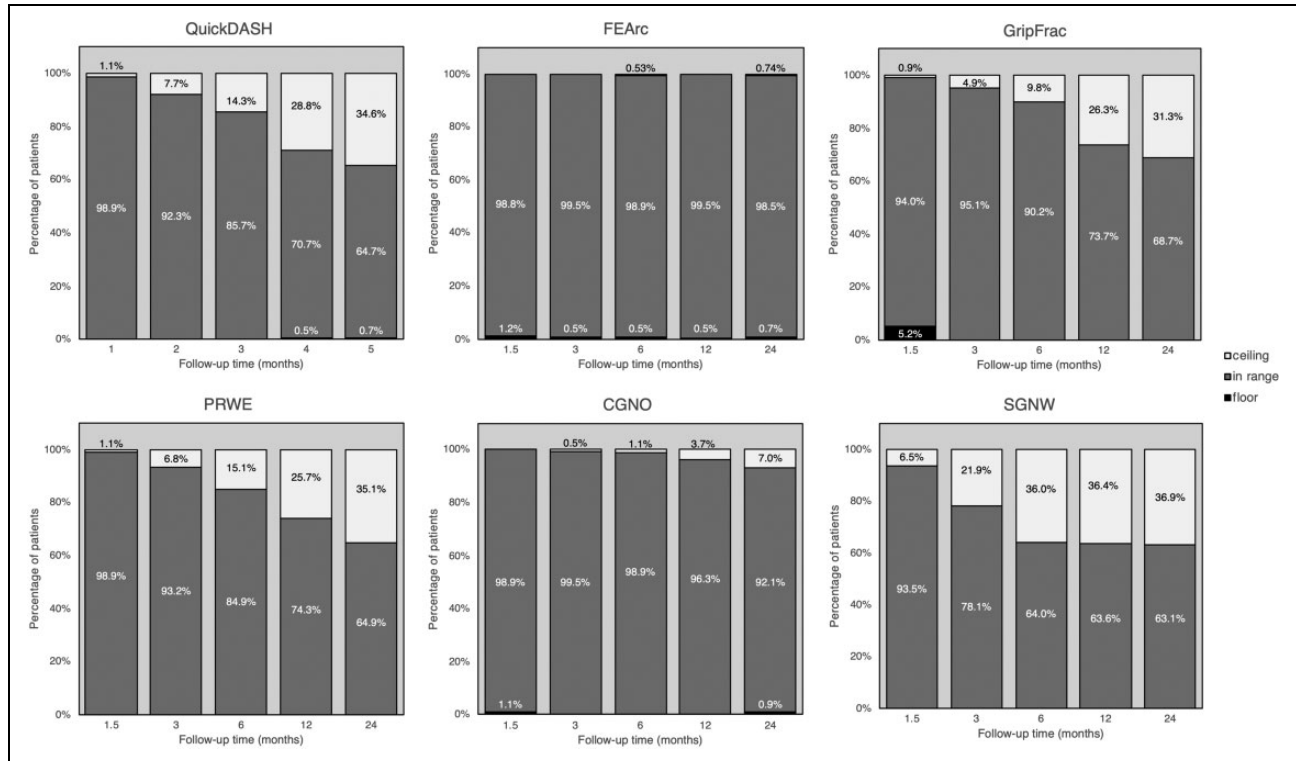
<sup>a</sup>All correlations were statistically significant.

evidence and critical consideration of available information. We demonstrate that responsiveness and ceiling/floor effects vary across different phases of the rehabilitation period. While many previous studies have followed up to 6 or 12 months, this period may not be of sufficient length to evaluate final patient outcomes after treatment of distal radius fractures as it is evident from our data that patients continue to improve beyond 12 months.

Responsiveness, also known as longitudinal validity, is the ability of an outcome measure to detect a meaningful change in the construct of interest.<sup>25</sup> Traditional distribution-based measures of responsiveness include effect sizes, such as Cohen's d (ES) and the standardized response mean (SRM).<sup>33</sup> These indices quantify the signal-to-noise ratio for the observed change in an outcome measure. More recently, it has been argued that effect sizes alone are inadequate for determining an instrument's responsiveness since they contain no information about whether the observed change in an instrument is due to a corresponding change in the construct of interest.<sup>34</sup> The COSMIN panel, therefore, proposes two valid methods for evaluating responsiveness: first, by comparing change scores of an instrument to those of a "gold standard" or external criterion (criterion approach), and second, by testing predefined hypotheses about the magnitude and direction of correlations of an instrument's

change scores against those of other measures which have been shown to be adequately responsive (hypothesis-based approach).<sup>22</sup> Neither approach provides a quantitative measure of an instrument's responsiveness; as the panel notes: "There is no criterion to decide whether an instrument is valid or responsive. Assessing validity or responsiveness is a continuous process of accumulating evidence."<sup>34</sup>

We employed a mixed distribution- and criterion-based approach to provide evidence for or against the responsiveness of the outcome measures. In our view, a responsive outcome measure should (1) have an adequately high signal-to-noise ratio in order to detect a change, and (2) change correspondingly to the construct of interest. Both conditions are necessary, as a large effect size may not necessarily correspond to a large change in the construct of interest, and a measure that detects the change in the desired construct may be subject to a high degree of variance that can lead to uncertainty in measurements. We, therefore, calculated both ES and SRM, two of the most commonly used effect sizes, as well as correlated the change scores of the instruments against those of the *QuickDASH* and *PRWE*. ES and SRM provide slightly different values due to differences in the calculation of the denominator for each ratio, however general trends over time were the same between the indices.



**Figure 3.** Ceiling and floor effects of the six outcome measures.

**Table 5.** Spearman correlation coefficients of raw instrument scores with *QuickDASH* and *PRWE* at each follow-up point.<sup>a</sup>

		PRWE	FEArc	GripFrac	CGNO	SGNW
QuickDASH	1.5 m	.849	.433	.548	.587	.493
	3 m	.850	.417	.513	.496	.582
	6 m	.883	.314	.438	.400	.489
	12 m	.875	.328	.427	.424	.524
	24 m	.809	.285	.318	.248	.563
PRWE	1.5 m		.450	.453	.580	.487
	3 m		.424	.512	.450	.568
	6 m		.303	.457	.450	.518
	12 m		.346	.447	.471	.542
	24 m		.218 <sup>b</sup>	.329	.394	.525

<sup>a</sup>Directionality is not displayed as all effect sizes were in the expected direction indicating better functionality/decreased pain. All correlations were significant at the 0.01 level except where indicated.

<sup>b</sup>Correlation is significant at the 0.05 level (two-tailed)

*QuickDASH* and *PRWE* are two of the most commonly used PROMs following distal radius fractures. Both measure what are generally considered to be important aspects of wrist health, and have been shown to have good validity, reliability and responsiveness in the context of distal radius fractures.<sup>11,12</sup> Correlation of the change scores of these two PROMs against each other confirmed a high level of agreement ( $r_s = .754$ ) and further justifies the use of either as a valid comparator for the other four measures.

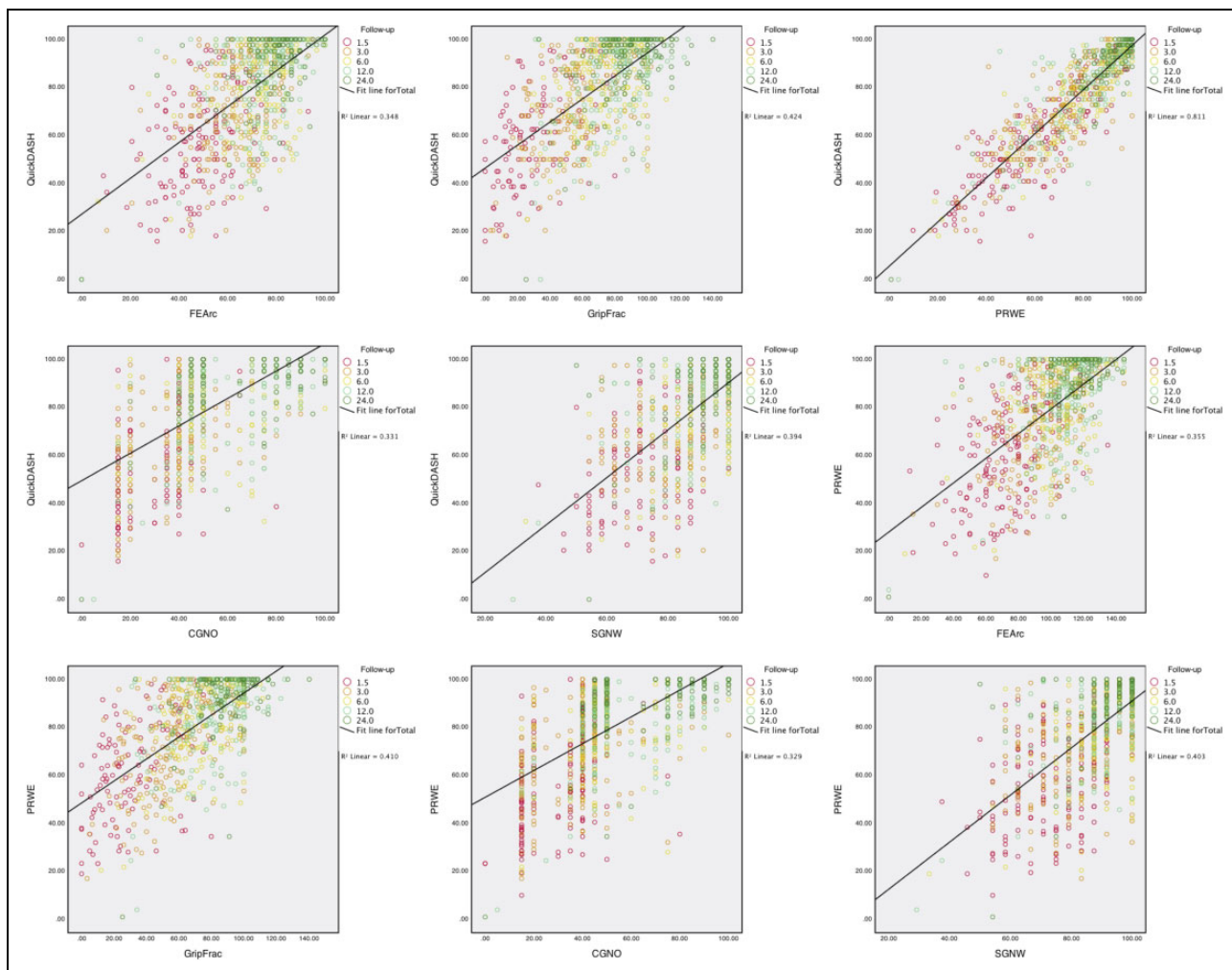
Both *FEArc* and *GripFrac* are indicated to be responsive by both the distribution- and criterion-based approaches,

particularly in the early recovery period, where they demonstrated large effect sizes in addition to moderate correlations with *QuickDASH* and *PRWE*. *CGNO* is also indicated to be responsive by both approaches, although with lower effect sizes in the 1.5- to 3-month interval and lower correlations with *QuickDASH* and *PRWE* than had *FEArc* and *GripFrac*. *SGNW* is not indicated to be responsive by either approach, particularly after 6 months, where it demonstrated the lowest effect sizes and correlations with *QuickDASH* and *PRWE* as compared to the other instruments.

It should be noted that multiple modifications and conflicting scoring methods for the Gartland and Werley system have been reported in the literature, which has led to confusion among authors.<sup>20</sup> Our scoring method allows for a maximum of 24 points, whereas other methods might have led to different results. However, based on the findings of ours and previous studies, there appears to be overall negative evidence for the responsiveness of the Gartland and Werley score.<sup>11</sup> In the presence of other instruments that have demonstrated good validity, reliability, and responsiveness, there appears to be little justification for its use when measuring recovery following distal radius fractures.

Ceiling and floor effects further complicate outcome assessment as they hinder the ability of outcome measures to detect improvement or deterioration once the maximum or minimum score has been reached.<sup>29</sup> Floor effects were not significant in any of the evaluated instruments. Ceiling





**Figure 4.** Scatter plots of correlations between criterion comparators (QuickDASH and PRWE) and other outcome measures. Standardized (0–100) scores for QuickDASH, PRWE, CGNO and SGNW were used, whereas raw scores were used for FEArc and GripFrac.

**Table 6.** Summary of evidence for use.

	QuickDASH		PRWE		GripFrac		FEArc		CGNO		SGNW	
	<6 months	>6 months	<6 months	>6 months	<6 months	>6 months	<6 months	>6 months	<6 months	>6 months	<6 months	>6 months
Responsiveness (distribution-based)	+	-	+	-	+	-	+	-	+	-	+	-
Responsiveness (criterion-based)	+	+	+	+	+	+	+	+	-	-	-	-
Ceiling effects	+		-		+		+		+		-	
Floor effects	+		+		+		+		+		+	
Criterion validity (QuickDASH)			+	+	-	-	+	-	∅	-	-	+
Criterion validity (PRWE)	+	+			-	-	∅	-	∅	-	∅	+
Evidence for use	Good	Good	Good	Good	Good	Poor	Good	Poor	Average	Poor	Poor	Poor

+ : evidence for use; - : evidence against use; ∅ : neutral evidence.



effects were observed, particularly at later time intervals; however, it can generally be expected that a larger proportion of patients will reach the upper bounds of an outcome measure at later stages of the recovery period.

Our study has several limitations: First, the results apply to the evaluated instruments as estimators only of the constructs measured by the comparators (i.e. function and symptoms as measured by *QuickDASH* and *PRWE*). Responsiveness and criterion validity depend on the construct of interest and evaluation of these properties yields different results depending on the chosen comparator measures.<sup>25</sup> Generalizability of our study is further limited to patients who have suffered distal radius fractures and been treated via ORIF. The applicability of these instruments to patients with different injuries or interventions requires evaluation through separate studies. Finally, our study is limited by the lack of an anchor measure, which typically involves a patient-rated measure of subjective change (i.e. “better,” “worse,” or “unchanged”) between time points. This provides an external criterion that more closely relates measured change to patient experience and can be used to estimate the minimal clinically important difference (MCID) of an instrument.<sup>33</sup>

In conclusion, our evidence supports the use of *QuickDASH*, *PRWE*, *FEArc* and *GripFrac* up to 6 months post-surgery, and *QuickDASH* and *PRWE* after 6 months. Other measurement properties of outcome measures which lay outside the scope of this study remain relevant and additional high-quality evidence should be considered to fully inform the clinician’s choice of instrument.

### Acknowledgments

The authors wish to thank the AO Trauma Asia-Pacific Research Grant for funding this study. They also wish to thank Margaret Ho, Elaine Tian, Lorraine Cheung, Grace Ho and Kathine Ching for data collection and keeping.

### Author contributions

CF: guarantor, principal investigator, manuscript preparation, data analysis, performing surgeries and follow-ups; EF: medical writer, manuscript preparation, data analysis; DY: monitoring visits, performing surgery and follow-ups, data analysis, manuscript preparation; KK: research grant application, manuscript review; GL: performing therapy and ensuring quality of scores, manuscript review; FL: study and grant application, manuscript preparation, study supervision.


### Declaration of conflicting interests


The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: CF and FL are speakers for DePuy Synthes.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the AO Trauma Asia Pacific Grant [AOTAP 12-09].

### ORCID iD

Christian Fang  <https://orcid.org/0000-0002-3827-0351>

Kenny Kwan  <https://orcid.org/0000-0002-4034-8525>

### Supplemental material

Supplemental material for this article is available online.

### References

1. Meena S, Sharma P, Sambharia AK, et al. Fractures of distal radius: an overview. *J Family Med Prim Care* 2014; 3: 325–332.
2. Wilson IB and Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995; 273: 59–65.
3. Bowyer AJ and Roysse CF. Postoperative recovery and outcomes—what are we measuring and for whom? *Anaesthesia* 2016; 71(1): 72–77.
4. Beaton DE, Wright JG, and Katz JN. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am* 2005; 87: 1038–1046.
5. MacDermid JC, Turgeon T, Richards RS, et al. Patient rating of wrist pain and disability: a reliable and valid measurement tool. *J Orthop Trauma* 1998; 12: 577–586.
6. Cooney WP, Bussey R, Dobyns JH, et al. Difficult wrist fractures. Perilunate fracture-dislocations of the wrist. *Clin Orthop Relat Res* 1987; (214): 136–147.
7. Souer JS, Lozano-Calderon SA, and Ring D. Predictors of wrist function and health status after operative treatment of fractures of the distal radius. *J Hand Surg Am* 2008; 33: 157–163.
8. Sarmiento A, Pratt GW, Berry NC, et al. Colles’ fractures. Functional bracing in supination. *J Bone Joint Surg Am* 1975; 57: 311–317.
9. Census and Statistics Department TGoHKSAR. Table 030: gross domestic product (GDP), implicit price deflator of GDP and per capita GDP, <https://www.censtatd.gov.hk/hkstat/sub/sp250.jsp?tableID=030&ID=0&productType=8> (2020, accessed 1 July 2020).
10. Meinberg EG, Agel J, Roberts CS, et al. Fracture and Dislocation Classification Compendium-2018. *J Orthop Trauma* 2018; 32(1): S1–S170.
11. Dacombe PJ, Amirfeyz R, and Davis T. Patient-reported outcome measures for hand and wrist trauma: is there sufficient evidence of reliability, validity, and responsiveness? *Hand (N Y)* 2016; 11: 11–21.
12. Tsang P, Walton D, Grewal R, et al. Validation of the QuickDASH and DASH in patients with distal radius fractures through agreement analysis. *Arch Phys Med Rehabil* 2017; 98: 1217–1222.
13. Kennedy CA, Beaton DE, Smith P, et al. Measurement properties of the QuickDASH (Disabilities of the Arm, Shoulder and Hand) outcome measure and cross-cultural adaptations of the QuickDASH: a systematic review. *Qual Life Res* 2013; 22: 2509–2547.
14. KY Chan R, Leung YC, KL Leung F, et al. Reliability and validity of the Chinese (Queen Mary Hospital, Hong Kong) version of the Disabilities of the Arm, Shoulder and Hand on

- patients with upper extremity musculoskeletal disorders in Hong Kong. *Hong Kong J Occup* 2019; 32(1): 62–68.
15. Wong JY, Fung BK, Chu MM, et al. The use of Disabilities of the Arm, Shoulder, and Hand Questionnaire in rehabilitation after acute traumatic hand injuries. *J Hand Ther* 2007; 20: 49–55.
  16. Mehta SP, MacDermid JC, Richardson J, et al. A systematic review of the measurement properties of the Patient-Rated Wrist Evaluation. *J Orthop Sports Phys Ther* 2015; 45: 289–298.
  17. Wah JW, Wang MK, and Ping CL. Construct validity of the Chinese version of the Patient-Rated Wrist Evaluation Questionnaire (PRWE-Hong Kong Version). *J Hand Ther* 2006; 19: 18–26.
  18. Kwok IH, Leung F, and Yuen G. Assessing results after distal radius fracture treatment: a comparison of objective and subjective tools. *Geriatr Orthop Surg Rehabil* 2011; 2: 155–160.
  19. Pynsent PB, Fairbank JCT, Carr A, et al. *Outcome measures in orthopaedics and orthopaedic trauma*. 2nd ed. London; New York, NY: Arnold; distributed in the United States by Oxford University Press, 2004, p. xiii, 381 pp.
  20. Davis TRC. Open reduction and internal fixation for distal radial fractures. *J Hand Surg* 1993; 18: 545.
  21. Changulani M, Okonkwo U, Keswani T, et al. Outcome evaluation measures for wrist and hand: which one to choose? *Int Orthop* 2008; 32: 1–6.
  22. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010; 19: 539–549.
  23. Frost MH, Reeve BB, Liepa AM, et al. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value Health* 2007; 10(2): S94–S105.
  24. Strand LI, Anderson B, Lygren H, et al. Responsiveness to change of 10 physical tests used for patients with back pain. *Phys Ther* 2011; 91: 404–415.
  25. Mokkink LB, Terwee CB, Knol DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010; 10: 22.
  26. Middel B and van Sonderen E. Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. *Int J Integr Care* 2002; 2: e15.
  27. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1988.
  28. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J* 2012; 24: 69–71.
  29. McHorney CA and Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Qual Life Res* 1995; 4: 293–307.
  30. Mokkink LB, Prinsen C, Patrick DL, et al. *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs). User Manual*. Amsterdam: COSMIN, [https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual\\_version-1\\_feb-2018-1.pdf](https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018-1.pdf) (2018, accessed 1 July 2020).
  31. Mason SJ, Catto JWF, Downing A, et al. Evaluating patient-reported outcome measures (PROMs) for bladder cancer: a systematic review using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist. *BJU Int* 2018; 122: 760–773.
  32. 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: 737–745.
  33. Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008; 61: 102–109.
  34. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. *BMC Med Res Methodol* 2011; 11: 152.