OPEN

PRS GLOBAL OPEN

ORIGINAL ARTICLE
Hand/Peripheral Nerve

Site-specific Patient-reported Outcome Measures for Hand Conditions: Systematic Review of Development and Psychometric Properties

Justin C.R. Wormald, MBBS, MRCS*†
Luke Geoghegan, BSc*
Kyra Sierakowski, MD‡
Andrew Price, MBB(Ch), FRCS, PhD*
Michele Peters, PhD§
Abhilash Jain, MBBS, FRCS, PhD*¶
Jeremy N. Rodrigues, MBChB, FRCS, PhD*

Background: There are a number of site-specific patient-reported outcome measures (PROMs) for hand conditions used in clinical practice and research for assessing the efficacy of surgical and nonsurgical interventions. The most commonly used hand-relevant PROMs are as follows: Disabilities of the Arm, Shoulder and Hand (DASH), QuickDASH (qDASH), Michigan Hand Questionnaire (MHQ), Patient Evaluation Measure (PEM), Upper Extremity Functional Index (UEFI), and Duruoz Hand Index (DHI). There has been no systematic evaluation of the published psychometric properties of these PROMs.
Methods: A PRISMA-compliant systematic review of the development and validation studies of these hand PROMs was prospectively registered in PROSPERO and conducted to assess their psychometric properties. A search strategy was applied to Medline, Embase, PsycINFO, and CINAHL. Abstract screening was performed in duplicate. Assessment of psychometric properties was performed.
Results: The search retrieved 943 articles, of which 54 articles met predefined inclusion criteria. There were 19 studies evaluating DASH, 8 studies evaluating qDASH, 13 studies evaluating MHQ, 5 studies evaluating UEFI, 4 studies evaluating PEM, and 5 studies evaluating DHI. Assessment of content validity, internal consistency, construct validity, reproducibility, responsiveness, floor/ceiling effect, and interpretability for each PROM is described.
Conclusions: The psychometric properties of the most commonly used PROMs in hand research are not adequately described in the published literature. DASH, qDASH, and MHQ have the best-published psychometric properties, though they have either some poor psychometric performance or incompletely studied psychometric properties. There are more limited published data describing the psychometric properties of the UEFI, PEM, and DHI. *(Plast Reconstr Surg Glob Open 2019;7:e2256; doi: 10.1097/GOX.0000000000002256; Published online 21 May 2019.)*

INTRODUCTION

Patient-reported outcome measures (PROMs) are instruments that provide information of the patient's health status without external interpretation.[1] PROMs can be generic, domain specific, site specific, or disease specific.[2] A previous review by our team identified common PROMs used electively managed hand condition research [accepted, in press, *The Journal of Hand Surgery* (Asian-Pacific Volume)]. PROMs demand well-performing psychometric properties in relevant patient populations.[3] Psychometrics involves the design, deployment, and interpretation of tools to quantify psychological variables, such as questionnaires used to measure hand function or health-related quality of life. Analyses of psychometric properties can be performed after a PROM's implementation to verify the analyses conducted during development, to investigate unstudied psychometric properties, or to assess its performance under other clinical conditions.

*From the *Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Science (NDORMS), University of Oxford, Oxford, United Kingdom; †Department of Plastic and Reconstructive Surgery, Stoke Mandeville Hospital, Buckinghamshire Healthcare NHS Trust, Aylesbury United Kingdom; ‡Flinders University, Adelaide, SA, Australia; §Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom; ¶Department of Plastic and Reconstructive Surgery, Imperial Healthcare NHS Foundation Trust, London, United Kingdom.*

Received for publication January 20, 2019; accepted March 18, 2019.

DOI: 10.1097/GOX.0000000000002256

Disclosure: The authors have no financial interest to declare in relation to the content of this article. JR is an NIHR Postdoctoral Fellow (PDF-2017-10-075). JCRW is an NIHR Academic Clinical Fellow.

A recent systematic review has assessed the validity and psychometric properties of PROMs used in traumatic hand and wrist conditions but predated the most recent consensus-based psychometric assessment framework.[4,5] Furthermore, the evidence supporting the validity of the most commonly used PROMs in elective hand conditions has not been assessed against this framework. Prinsen et al[5] recently published criteria for assessing the quality of published psychometrics. Despite the recency of this publication, the standards by which psychometrics are assessed are well established in the literature. For example, the concept of Rasch analysis has been in circulation since the 1960s.[6] Evaluation of the published psychometric properties of commonly used hand PROMs, in both elective and traumatic hand conditions, is vital to determine their utility.

The objective of this systematic review is to appraise the psychometric properties of site-specific PROMs commonly used in clinical studies of electively managed and traumatic conditions of the hand against current standards—the Prinsen et al[5] criteria. The results of this analysis can then be used to assess the suitability of hand-specific PROMs for ongoing use.

METHODS

The protocol for this systematic review was developed in accordance with the PRISMA statement and prospectively registered in the international prospective register of systematic reviews (PROSPERO) on January 26, 2018 (CRD42018081508). Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) were applied to this systematic review, in accordance with guidance by Prinsen et al.[5]

Search Strategy

A bespoke search strategy was developed (index and free terms) to identify developmental studies and published psychometric properties of the 6 most commonly used hand site-specific PROMs (Appendix 1). The search strategy was applied in parallel to Medline (1946-February 2018*), Embase (1974-February 2018*), and PsycINFO (1806-February 2018*) provided by the Ovid interface and CINAHL (1981-January 2018*) provided by the NICE Healthcare Databases Advanced Search interface. The search was run on February 15, 2018, and was limited to human studies. The reference list of included articles was hand searched for further relevant publications, and gray literature was searched using Google Scholar at the time of the original literature search. Electronic deduplication was used.

Eligibility Criteria

The PROMs evaluated were those identified in our recent systematic review of elective hand surgery and those identified in the equivalent trauma-based systematic review by Dacombe et al[4] (excluding wrist-specific PROMs). These were the Disabilities of the Arm, Shoulder and Hand (DASH),[7] QuickDASH (qDASH),[8] Michigan Hand Questionnaire (MHQ),[9] Patient Evaluation Measure (PEM),[10] Upper Extremity Functional Index (UEFI),[11] and Duruoz Hand Index (DHI).[12] All studies of these PROMs in an adult cohort with a condition affecting the hand (distal to the carpal bones) (P) with any applicable intervention and/or comparator (I and C) presenting data related to the psychometric properties of the PROM(s) according to the criteria defined by Prinsen et al[5] were included. Clinical studies not evaluating one of these PROMs, pediatric studies, and cross-cultural validation studies were excluded. Two authors (JCRW and LG) independently screened all abstracts against a prespecified checklist of criteria for inclusion (Fig. 1). Any disagreement was resolved by consensus discussion and consultation with the third author (JNR) if required.

Data Extraction

Data extraction was performed in duplicate using a bespoke proforma comprising source, study design, criteria used to assess content validity, internal consistency, criterion validity, construct validity, reproducibility, responsiveness, floor and ceiling effects, and interpretability. Content validity was assessed by the examination of development studies for each PROM. Development methodology was scrutinized for comprehensiveness and relevance based on how the items were generated and selected. Further data regarding study populations in which the PROM was utilized, including patient demographics and disease characteristics, were extracted to determine the generalizability of the study population.

Data Analysis

A narrative synthesis was utilized to evaluate the published psychometric properties of included PROMs using quality assessment criteria outlined by Prinsen et al.[5] Specifically, the following properties were assessed for each PROM:

- Content validity: the appropriateness to the targeted patient group based on comprehensiveness and relevance.
- Structural validity: how well it measures the relevant underlying construct (including Classical Test Theory and Item Response Theory and Rasch measurement theory analyses).
- Internal consistency: the extent to which the items within the PROM correlate with each other.
- Measurement invariance: whether there are important differences found between group factors (age, sex, and language).
- Reliability: the reproducibility of the answers given.
- Measurement error: whether smallest detectable change (SDC) in the PROM is smaller than the minimal important change (MIC) and/or minimal important difference.
- Construct validity/hypothesis testing: whether it correlates with instruments measuring similar constructs and does not correlate with instruments measuring unrelated constructs.
- Criterion validity: whether it correlate with a true "gold standard", where one exists (typically, this is only ap-

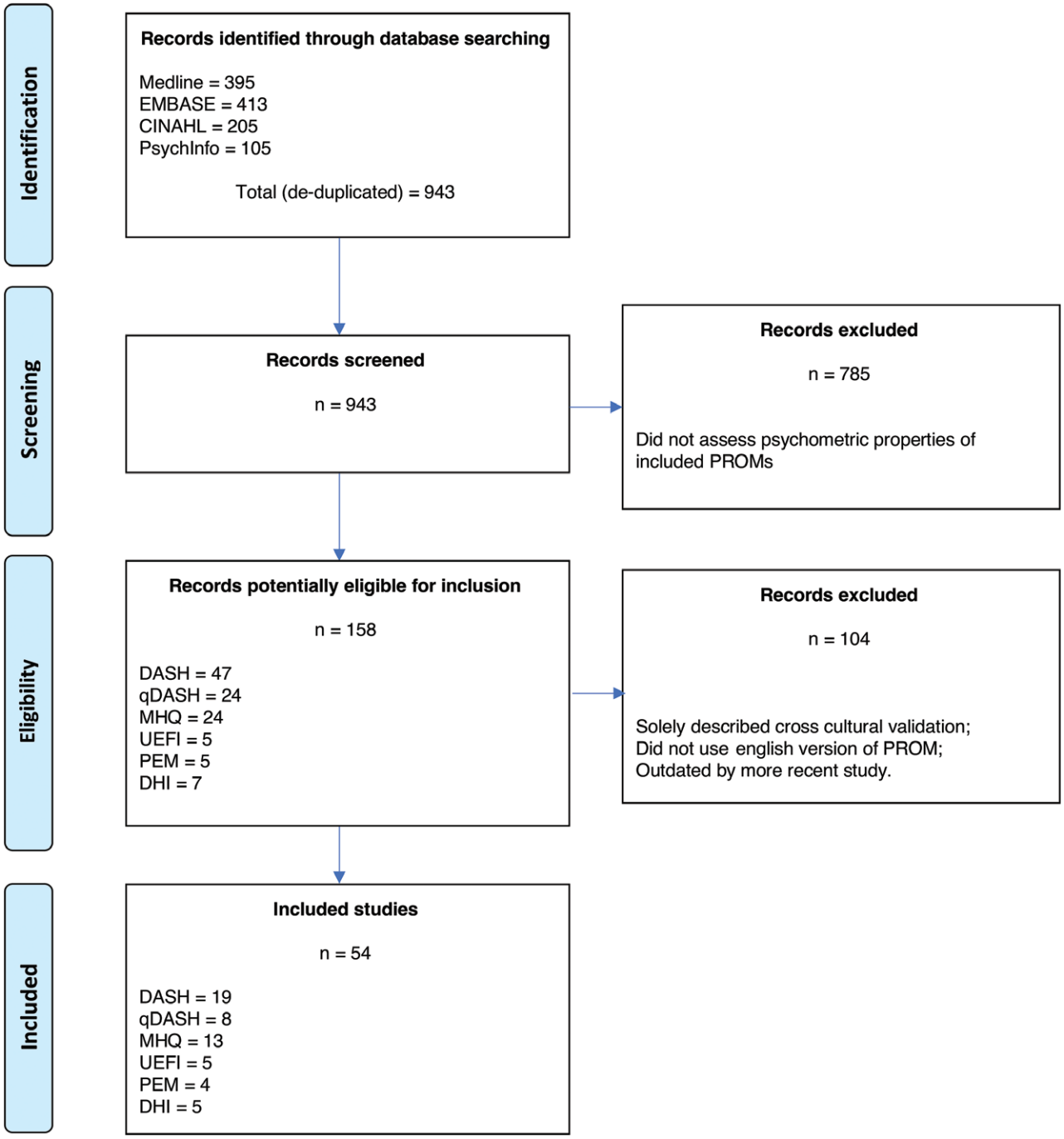

Fig. 1. PRISMA flowchart. Adapted with permission from *PLoS Med* 2009;6:e1000097. For more information, visit www.prisma-statement.org.

plicable when comparing a shortened PROM with its original long form).

Responsiveness: whether the PROM responds to change.

Evidence for each psychometric property was appraised as good performance (+), poor performance (–), indeterminate (?), inappropriate (!), or no evidence (0) to support each psychometric criterion. This was determined by the reviewers and was based on the information across all available studies. "Good performance" ratings denoted that the performance of the PROM across all identified studies of that psychometric measurement property was predominantly good, whereas "poor performance" ratings denoted predominantly poor performance in analyses of that psychometric measurement property in the published data included. "Indeterminate" ratings signified a lack of, or conflicting evidence of, performance, and "inappropri-

ate" ratings were assigned when the measurement property had not been studied using the methods outlined by Prinsen et al.[5]

RESULTS

Database searching yielded 943 articles, of which 889 studies were excluded. This left 54 studies, which met predefined inclusion criteria and underwent full analysis (Fig. 1): 19 studies evaluated DASH, 8 studies evaluated qDASH, 13 studies evaluated MHQ, 4 studies evaluated PEM, 5 studies evaluated UEFI, and 5 studies evaluated DHI. Table 1 provides an overview of the reviewed PROMs and populations in which they were validated. Figure 2 demonstrates the cumulative frequency of psychometric evaluation studies for each constituent PROM over time.

All PROMs were validated in at least one domain. Carpal tunnel syndrome was the most common pathology in which the psychometric properties had been assessed, but most studies evaluated psychometric properties across mixed cohorts (Table 2). None were validated across all domains. The DASH and the qDASH had the most good performance evidence, determined by the total amount of evidence available across all properties evaluated.

Development studies were available for analysis for the DASH, qDASH, MHQ, PEM, and UEFI. Only the MHQ explicitly sought information from a group of patients in item generation. DASH, MHQ, and UEFI were based on literature reviews of existing upper limb PROMs, whereas the PEM was developed by expert opinion only. Item selection was performed using a combination of expert opinion and clinimetric/psychometric item reduction for the DASH, MHQ, and UEFI and for the conversion of the DASH to the qDASH. The development studies for the DHI were not available despite email request.

The DASH had the most published research assessing structural validity. These studies demonstrated poor performance for factor analysis and item response theory (IST) analyses. The qDASH used Rasch modeling in its development, along with 2 other methods of item reduction but has not undergone formal structural validity assessment. The original 20-item UEFI failed Rasch measurement model and so was refined into a 15-item questionnaire, which fitted a Rasch model. All other PROMs had no published evidence for or against their structural validity.

All included PROMs had supporting evidence of internal consistency with values for Cronbach's α ranging from 0.87 to 0.98, with many showing evidence of redundancy of items (Cronbach's α >0.90).

Almost all of the included PROMs had evidence of reliability with intraclass correlation coefficients (ICCs) >0.80 apart from the PEM, where no reliability studies were

Table 1. Overview of Included PROMs and Studies

PROM	PROM Scope	Broad Domains Covered by PROM Items	Specific Populations Demonstrating Positive Evidence
DASH	Site specific to entire upper limb	1.Activities of daily living 2.Social activities 3.Work activities 4.Symptoms 5.Sleeping 6.Confidence 7.Sports/hobbies* 8.Work*	Carpal tunnel syndrome: Greenslade et al[16], Bakhsh et al[20], Gay et al[25], and Hobby et al[26] Rheumatoid arthritis Hammond et al[22] Trapeziometacarpal arthritis: Angst et al[19] Mixed cohorts (elective): Dias et al[18] and Gummesson et al[57] Mixed cohorts (trauma and elective): Sorensen et al (2013), Whalley and Adams[27] and Gummesson et al[23] Not specified: Gabel et al[17]
qDASH	Site specific to entire upper limb	1.Activities of daily living 2.Social and work activities 3.Recreation 4.Symptom severity 5.Sleeping 6.Sports/hobbies* 7.Work*	Dupuytren's disease: Budd et al[34] Carpal tunnel syndrome: Lyrén and Atroshi[31] Mixed cohorts (elective): Beaton et al[8] and Gummesson et al[57] Mixed cohorts (trauma): Polson et al[33] Mixed cohorts (trauma and elective): Gabel et al[32] and Whalley and Adams[27]
MHQ	Site specific to hand	1.Function 2.Activities of daily living a.One-handed activities of daily living b.Two-handed activities of daily living 3.Work 4.Pain 5.Aesthetics 6.Satisfaction	Rheumatoid arthritis: Massy-Westropp et al[35], Waljee et al[36], and Dritsaki et al[44] Dupuytren's disease: Thoma et al[37] Carpal tunnel syndrome: Kotsis and Chung[41] Systemic sclerosis: Schouffoer et al[43] Not specified: Chung et al[9] Mixed cohort (elective): Dias et al[18] and London et al[38] Mixed cohort (trauma): Horng et al[39] and Weinstock et al[42]
UEFI	Site specific to entire upper limb	20 individual items (no individual domains)	Not specified: Gabel et al[17], Stratford[11], and Lehman et al[49]
PEM	Site specific to hand	1.Treatment 2.Subjective hand function 3.Overall assessment	Carpal tunnel syndrome: Hobby et al[26] Scaphoid fracture: Dias et al[45] Mixed cohort (elective): Dias et al[18]
Duruoz Hand Index	Site specific to hand	1.Kitchen work: 8 items 2.Dressing: 2 items 3.Hygienic practices: 2 items 4.Office work: 2 items 5.Other: 4 items	Systemic sclerosis: Brower and Poole[51], Duarte et al (2014), Gheorghiu et al[50] Flexor tendon injury: Ercalik et al (2011)

*Optional modules.

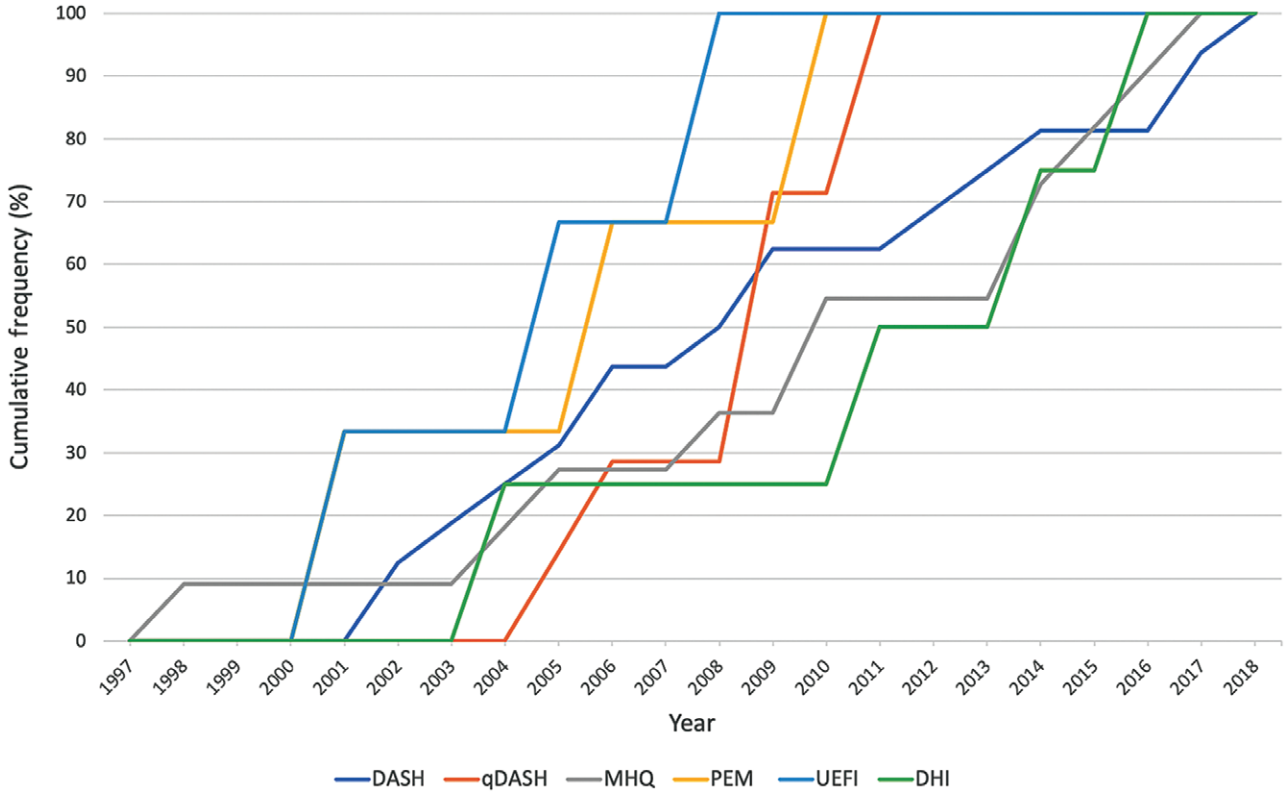

Fig. 2. Cumulative frequency of published psychometric evidence over time for included PROMs.

found. The DASH, qDASH, and MHQ had published MICs that were larger than the SDC of the PROM. The PEM and DHI did not have published MICs that could be identified. The UEFI has 2 studies published that show an SDC greater than the MIC, indicating potential for measurement error.

All PROMs in the analysis had evidence of construct validity when they were compared with similar instruments.

The DASH, UEFI, and PEM underwent formal testing of measurement invariance. Differential item functioning (DIF) found sex-related issues in the DASH and UEFI for task-specific items. Multiple group factor analysis for PEM showed no differences between age and sex for its items.

Criterion validity was often inappropriately assessed. The only appropriate analysis possible in this review's scope was to assess the qDASH against the DASH, which had been done, and demonstrated good performance. Studies evaluating against inappropriate gold standards were present for MHQ, PEM, and UEFI.

Almost all of the PROMs in the analysis had published evidence of responsiveness to change in the tested populations. There were some discrepancies in the published evidence depending on the populations. For example, DASH was not responsive to change in patients with Dupuytren's disease but was in populations with other hand conditions.

DISCUSSION

This systematic review has evaluated the psychometric properties of commonly used PROMs in hand conditions using internationally accepted, consensus-based criteria. The psychometric properties of the 6 included PROMs were largely only studied incompletely. In some cases, poor psychometric performance was identified, with no single outcome measure demonstrating sufficient psychometric robustness to meet the Prinsen et al[5] criteria.

Content Validity

Modern development of a PROM requires extensive, multicenter, and multinational data collection exercises to ensure a comprehensive list of items is generated, applicable to different populations and cultures. Alongside primary data collection, systematic reviews of existing PROMs are required to ensure comprehensive inclusion of all relevant items. Following this, item reduction should be performed through field testing and psychometric statistical techniques, including the use of item response theory and Rasch modeling. For some PROMs (DASH, qDASH, MHQ, and UEFI), psychometric analyses were used to reduce the items to the current questionnaires, whereas the PEM employed expert opinion alone. The most commonly used hand PROMs, therefore, do not fully meet contemporary standards for PROM development, particularly in the item generation process, raising questions about their comprehensiveness.

Structural Validity

Only 2 PROMs had evidence of structural validity testing: the DASH and the UEFI. Development of the qDASH involved Rasch modeling for item reduction, but we could

Table 2. Analysis and grading of published psychometric properties of hand PROMs accoring to the Prinsen criteria.

	Content Validity		Structural Validity	Internal Consistency	Reliability	Measurement Error	Hypotheses testing for construct validity	Cross-cultural validity/measurement invariance	Criterion validity*	Responsiveness
	Item generation (comprehensive)	Item selection (relevance)								
DASH	**Hudak 1996 (-)** Literature review of existing upper limb questionnaires 13 existing upper limb questionnaires pooled (821 items)	**Hudak 1996 (-)** Expert opinion for item relevance Expert opinion for questionnaire acceptability Pretesting on 20 doctors **Marx 1999 (+)** Clinimetric/ psychometric item reduction to 30-item scale	**Forget 2014 (–)** Failed Rasch analysis in Dupuytren's population **Braitmayer 2017 (–)** Failed Rasch analysis in hand injuries population	**Greenslade 2004 (+)** Cronbach's α 0.97 **Gabel 2006 (+)** Cronbach's α 0.96 **Dias 2008 (+)** Cronbach's α 0.98 **Angst 2009 (+)** Cronbach's α 0.88 **Bakhsh 2012 (+)** Cronbach's α 0.95 **Braitmayer 2017 (+)** Cronbach's α 0.97 **Rodrigues 2015 (+)** Cronbach's α 0.98 **Hammond 2018 (+)** Cronbach's α 0.98	**Gabel 2006 (+)** ICC = 0.98 for the DASH **Bakhsh 2012 (+)** ICC = 0.87 **Hammond 2018 (+)** ICC = 0.97	**Gummesson 2003 (+)** MIC = 10 (> SDC) **Sorensen 2013 (+)** MIC = 10 (> SDC)	**SooHoo 2002 (+)** The result is in accordance with the hypothesis **Gummesson 2003 (+)** The result is in accordance with the hypothesis	**Braitmayer 2017 (–)** No DIF was found for age. Item 11 ('carrying a heavy object') showed DIF for gender.	**Soohoo 2002 (!)** Moderate correlation to the SF-36	**Gay 2002 (+)** The result is in accordance with the hypothesis **Hobby 2005 (+)** The result is in accordance with the hypothesis **Whalley 2009 (+)** The result is in accordance with the hypothesis **McMillan 2009 (+)** The result is in accordance with the hypothesis **Chapman 2008 (+)** The result is in accordance with the hypothesis **Rodrigues 2017 (-)** The result is not in accordance with the hypothesis
QuickDASH	As per DASH (-)	As per DASH **Beaton 2005 (+)** Item reduction performed using concept-retention approach	**Lyren 2012 (?)** IRT performed but not all information for '+' reported	**Beaton 2005 (+)** Cronbach's α >0.90 **Gummesson 2006 (+)** Cronbach's α 0.92-0.95 **Gabel 2009 (+)** Cronbach's α 0.92 **Rodrigues 2015 (+)** Cronbach's α 0.93	**Beaton 2005 (+)** ICC = 0.94-0.97 **Gummesson 2006 (+)** ICC = 0.90-0.93 **Gabel 2009 (+)** ICC = 0.91	**Polson 2009 (+)** SDC = 11 < MIC = 19	**Budd 2011 (+)** The result is in accordance with the hypothesis	No multiple group factor analysis OR DIF analysis performed (0)	**Beaton 2005 (+)** Correlation with gold standard (DASH) ≥ 0.70 (r = 0.98) **Rodrigues 2015 (+)** Correlation with gold standard (DASH) ≥ 0.70 (r = 0.98)	**Beaton 2005 (+)** The result is in accordance with the hypothesis **Whalley 2009 (+)** The result is in accordance with the hypothesis **Budd 2011 (+)** The result is in accordance with the hypothesis **Lyren 2011 (+)** The result is in accordance with the hypothesis
MHQ	**Chung 1998 (+)** Literature review of existing questionnaires Relevant items pooled plus input from a group of patients	**Chung 1998 (?)** Expert opinion to categorise and reduce items **(-)** Factor analysis for item reduction **(+)**	No test of structural validity (0)	**Chung 1998 (+)** Cronbach's α 0.87-0.97 **Massy-Westropp 2004 (+)** Cronbach's α 0.88 **Dias 2008 (+)** Cronbach's α 0.93 **Waljee 2010 (+)** Cronbach's α 0.87	**Massy-Westropp 2004 (+)** ICC = 0.95 **Thoma 2014 (+)** ICC=0.79	**London 2014 (+)** SDC = 4.9 < MIC = 10.3 [5.5–15.0]	**Chung 1998 (+)** The result is in accordance with the hypothesis	No multiple group factor analysis OR DIF analysis performed (0)	**Dias 2008 (!)** PROM scores compared to Levine symptom score and Gartland and Werley score. **Horng 2010 (!)** Correlation with gold standard (DASH) ≥ 0.70 (r=0.89) **Van der ven-stevens 2015 (!)** Correlates with DASH	**Kotsis 2005 (+)** The result is in accordance with the hypothesis **McMillan 2009 (+)** The result is in accordance with the hypothesis **Thoma 2014 (+)** The result is in accordance with the hypothesis **Weinstock 2015 (+)** The result is in accordance with the hypothesis **Schouffoer 2016 (+)** The result is in accordance with the hypothesis **Dritsaki 2017 (+)** The result is in accordance with the hypothesis
PEM	**Macey and Burke 1995 (-)** Expert opinion only	**Macey and Burke 1995 (-)** Expert opinion only	No test of structural validity (0)	**Dias 2001 (+)** Cronbach's α 0.93 **Hobby 2005 (+)** Cronbach's α 0.94 **Dias 2008 (+)** Cronbach's α 0.94	ICC or weighted Kappa ≥ 0.70 not published (0)	MIC not defined (0)	**Dias 2001 (+)** The result is in accordance with the hypothesis **Hobby 2005 (+)** The result is in accordance with the hypothesis **Dias 2008 (+)** The result is in accordance with the hypothesis	**Dias 2001 (+)** No important differences found between group factors (age, gender) in multiple group factor analysis	**Dias 2008 (!)** Correlation with gold standard (clinical assessments of pain, tenderness, swelling, ROM and grip strength) ≥ 0.70 (r=0.82)	**Dias 2001 (+)** The result is in accordance with the hypothesis **Hobby 2005 (+)** The result is in accordance with the hypothesis
UEFI	**Stratford 2001 (-)** Literature review of existing PROM items Responses from Patient Specific Functional Scale (n=40 patients) Expert opinion	**Stratford 2001 (+)** Condensation of similar items by development team Clinimetric/ psychometric item reduction	**Hamilton 2013 (+)** Rasch analysis did not support the validity of the 20-item UEFI – UEFI 15 item developed	No test of internal consistency (0)	**Hefford 2012 (+)** ICC 0.85 **Chesworth 2014 (+)** ICC 0.94 (UEFI-15 and UEFI-20)	**Stratford 2001 (-)** SDC 9.1 **Hefford 2012 (-)** SDC 17.6 > MIC 8.50 **Chesworth 2014 (-)** SDC 9.4 > MIC 8 (UEFI-20) and SDC 8.8 > MIC 6.7 (UEFI-15)	**Stratford 2001 (+)** The result is in accordance with the hypothesis	**Hamilton 2013 (-)** 2 items in the Rasch-refined UEFI (15 item) had DIF by sex	**Lehman 2010 (!)** Correlation with gold standard (DASH) ≥ 0.70 (r=0.90)	**Lehman 2010 (+)** The result is in accordance with the hypothesis **Hefford 2012 (+)** The result is in accordance with the hypothesis (AUC 0.88)
DHI	Original development study unavailable for analysis (0)		No test of structural validity (0)	**Ercalik 2011 (+)** Cronbach's α >0.70 **Duarte 2014 (+)** Cronbach's α 0.98 **Gheorghiu 2016 (+)** Cronbach's α 0.89	**Brower 2004 (+)** ICC0.81-0.97 **Ercalik 2011 (+)** ICC > 0.99 **Gheorghiu 2016 (+)** ICC 0.98	MIC not defined (0)	**Ercalik 2011 (+)** The result is in accordance with the hypothesis	No multiple group factor analysis OR DIF analysis performed (0)	**Bower 2004 (?)** No correlation with gold standard or AUC reported **Duarte 2014 (?)** No correlation with gold standard or AUC reported	**Ercalik 2011 (+)** The result is in accordance with the hypothesis **Dotu 2013 (?)** No hypothesis defined **Gheorghiu 2016(+)** The result is in accordance with the hypothesis

Legend:

(+) good performance
(–) poor performance
(?) indeterminate
(!) inappropriate
(0) no evidence

+, good performance; –, poor performance; ?, indeterminate; !, inappropriate; 0, no evidence.
ROM, range of movement; AUC, area under curve; SF, short form.
*Criterion validity assessed using comparison to DASH as 'gold-standard' for practicality of analysis.

not identify the evidence of structural validity testing specifically.[8] Forget et al[14] demonstrated that the DASH had large ceiling effects and failed to meet the assumptions of the Rasch measurement model in a population of patients with Dupuytren's disease. Additionally, the original UEFI 20-item questionnaire did not fit a Rasch model in a population of undefined upper extremity conditions. It is important to note that whereas the DASH and UEFI

performed poorly when subjected to Rasch modeling, it is unknown to what extent the other PROMs in this review would perform. There is, therefore, a risk of disproportionate criticism of the DASH and UEFI. Based on structural validity alone, the choice of a hand PROM is currently between those with known suboptimal performance and others with unknown performance, which may be the same, better, or worse.

This review only included PROMs that have been deployed in clinical studies, which was used as a benchmark of practical usefulness of the PROMs. There are other PROMs that have been recently developed, such as the PROM Information System Upper Extremity (PROMIS UE) tool.[53] The PROMIS UE is similar to the DASH in that it is site specific to the upper limb and correlates well with the qDASH.[54] It has been developed using item response theory and is a more complex, accurate, and dynamic tool.[53] Furthermore, it can be delivering using computer adaptive testing, which improves its deployment and utility across digital platforms.[55] This is an encouraging development, though it was not analyzed here as its application is yet to be confirmed in clinical studies. Based on this review, further study of the structural validity of existing PROMs is an important area for future research in this field.

Internal Consistency

Internal consistency was evaluated in all PROMs, each of which demonstrated positive ratings with Cronbach's α scores ranging from 0.87 to 0.98, suggesting high interrelatedness among constituent outcome measure items. Although an α of >0.70 is considered to be desirable in the framework used,[5] other sources suggest that very high values, >0.90, are indicative of redundancy of items within the PROM and are not desirable.[45]

Measurement Invariance

Only 3 studies reported multiple group factor analysis. Measurement invariance assesses the equivalence of items across specified groups, with only one study reporting no significant differences between group factors (sex, hand dominance, and side injured) and PEM score.[45] Braitmayer et al[15] conducted multifactor analysis for the DASH and reported DIF for sex. The Rasch-refined UEFI showed DIF for sex on 2 items "using tools/appliances" and "cleaning".[46] The PEM comprises symptom items (pain, stiffness, etc), whereas the DASH items with pronounced DIF were task based, specifically item 11 (carrying a heavy object) demonstrated important differences between sex,[15] which was mirrored by the task-based items with DIF in the UEFI.[46] The site-specific PROMs appraised here comprise some with mainly task-based items, some with symptom-based items, and some that are a mix of the 2.

Reliability

Reliability was frequently reported and rated positively across all PROMs except from the PEM, where the intraclass correlation coefficient or weighted Kappa was not reported. This was one of the better assessed aspects of psychometric performance across all PROMs.

Measurement Error

There was limited evidence of measurement error assessment, where the MIC is assessed relative to the SDC. This requires the MIC to be calculated. MIC is one element of the interpretability of a PROM, and interpretability is not central in the COSMIN system. MICs have been reported for 4 of the included PROMs (DASH, qDASH, MHQ, and UEFI),[21] and these were the PROMs with measurement error studies. For the UEFI, the SDC was reported as 8.8–17.6 with an MIC as 6.7–8.5 across 3 studies. This brings into question the ability of the UEFI to accurately measure the minimal clinically important change in the populations it has been tested in.

Criterion Validity

Criterion validity was inappropriately studied in the majority of PROMs, principally due to the lack of a gold standard for valid comparison to be made. Most studies reported correlation with the DASH as a gold standard instrument, which implies inappropriately that the DASH is a perfect outcome measure. Criterion validity is accepted in the comparison of shortened versions of PROMs to their longer constituents, and this was reported in the comparison of qDASH and DASH.[8]

Construct Validity

This property was widely studied, with all PROMs evaluated and demonstrating positive supporting evidence. The PROMs studied were frequently compared with generic quality of life measures such as the Short-Form 12 and Short-Form 36 to examine the behavior of the instrument in relation to the underlying concept it designed to measure. There are similarities between "construct validity" hypothesis testing the relationship between the PROM being studied and other measures, and the criterion validity studies criticized already. The distinction made here was that appropriate hypothesis testing looked for convergent or divergent validity by comparing with multiple similar measures, and/or seemingly unrelated measures.

Responsiveness

Several studies have reported that the DASH is a responsive instrument in cohorts of patients with carpal tunnel syndrome[25,26] and arthritis.[27] However, Rodrigues et al[30] examined the responsiveness of the DASH in patients with Dupuytren's disease following fasciectomy and dermofasciectomy and found that it could not distinguish patients who had experienced meaningful change in hand function following intervention. Consequently, the DASH was found to have moderate responsiveness and poor interpretability in patients with Dupuytren's disease, meaning that an MIC could not be estimated. Conflicting evidence regarding the responsiveness of the DASH may be due to the fact that clinically significant improvement is not always made following intervention. Moreover, as a site-specific PROM for the entire upper limb, the ability of the instrument to discern changes in specific functional aspects of the hand and digits is questionable. In contrast, the MHQ, a hand-specific PROM, was found to have a high effect size when used in a cohort of patients following fasciectomy for Dupuytren's

disease, indicating that the items in the MHQ are sensitive to change in hand function following intervention.[37]

Limitations

The findings of this review make it difficult to recommend the most suitable outcome measure at present. Indeed, the only core outcome set for hand surgery listed by the Core Outcome Measures in Effectiveness Trials initiative suggests combining the use of site-specific and disease-specific outcome measures in Dupuytren's disease but does not provide further detail.[56] Despite a robust search strategy, relevant publications may not have been identified from our original search. This review used only one framework for analysis.[5]

CONCLUSIONS

In summary, the results of this systematic review indicate that currently implemented PROMs have incomplete evidence to support their use in hand surgery research and practice, when compared with contemporary PROM standards (Appendix, http://links.lww.com/PRSGO/B116). The DASH, qDASH, and MHQ have the most published data evaluating their psychometric properties but have shortcomings and evidence gaps based on this review. There was more incomplete evidence to support the psychometric properties of the UEFI, PEM, and DHI for use in hand surgery practice and research. Future research in hand conditions could consider the role of contemporary PROMs, particularly those that have been developed using IRT, as such PROMs are more likely to have evidence that they meet modern psychometric standards.

Jeremy N. Rodrigues, BSc, MBChB, MSc, PhD, PGDip, FRCS(Plast)
NIHR Postdoctoral Fellow in Plastic Surgery
Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDORMS)
University of Oxford
Windmill Road, Oxford OX3 7HE, United Kingdom
E-mail: j.n.rodrigues@doctors.org.uk

REFERENCES

1. Nelson EC, Eftimovska E, Lind C, et al. Patient reported outcome measures in practice. *BMJ*. 2015;350:g7818.
2. Wormald JCR, Rodrigues JN. Outcome measurement in plastic surgery. *J Plast Reconstr Aesthet Surg*. 2018;71:283–289.
3. 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.
4. Dacombe PJ, Amirfeyz R, Davis T. Patient-reported outcome measures for hand and wrist trauma: is there sufficient evidence of reliability, validity, and responsiveness? *Hand*. 2016;11:11–21.
5. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;27:1147–1157.
6. Rasch G. An item analysis which takes individual differences into account. *Br J Math Stat Psychol*. 1966;19:49–57.
7. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand) [corrected]. The Upper Extremity Collaborative Group (UECG). *Am J Ind Med*. 1996;29:602–608.
8. Beaton DE, Wright JG, Katz JN; Upper Extremity Collaborative Group. Development of the QuickDASH: comparison of three item-reduction approaches. *J Bone Joint Surg Am*. 2005;87:1038–1046.
9. Chung KC, Pillsbury MS, Walters MR, et al. Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. *J Hand Surg Am*. 1998;23:575–587.
10. Macey A, Burke F, Abbott K, et al. Outcomes of hand surgery. *J Hand Surg*. 1995;20:841–855.
11. Stratford PW. Development and initial validation of the upper extremity functional index. *Physiother Can*. 2001;52:259–267.
12. Duruöz MT, Poiraudeau S, Fermanian J, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. *J Rheumatol*. 1996;23:1167–1172.
13. Marx RG, Bombardier C, Hogg-Johnson S, et al. Clinimetric and psychometric strategies for development of a health measurement scale. *J Clin Epidemiol*. 1999;52:105–111.
14. Forget NJ, Jerosch-Herold C, Shepstone L, et al. Psychometric evaluation of the Disabilities of the Arm, Shoulder and Hand (DASH) with Dupuytren's contracture: validity evidence using Rasch modeling. *BMC Musculoskelet Disord*. 2014;15:361.
15. Braitmayer K, Dereskewitz C, Oberhauser C, et al. Examination of the applicability of the disabilities of the arm, shoulder and hand (DASH) questionnaire to patients with hand injuries and diseases using rasch analysis. *Patient*. 2017;10:367–376.
16. Greenslade JR, Mehta RL, Belward P, et al. Dash and Boston questionnaire assessment of carpal tunnel syndrome outcome: what is the responsiveness of an outcome questionnaire? *J Hand Surg*. 2004;29:159–164.
17. Gabel CP, Michener LA, Burkett B, et al. The Upper Limb Functional Index: development and determination of reliability, validity, and responsiveness. *J Hand Ther*. 2006;19:328–329.
18. Dias JJ, Rajan RA, Thompson JR. Which questionnaire is best? The reliability, validity and ease of use of the Patient Evaluation Measure, the Disabilities of the Arm, Shoulder and Hand and the Michigan Hand Outcome Measure. *J Hand Surg*. 2008;33:9–17.
19. Angst F, Goldhahn J, Drerup S, et al. How sharp is the short QuickDASH? A refined content and validity analysis of the short form of the disabilities of the shoulder, arm and hand questionnaire in the strata of symptoms and function and specific joint conditions. *Quality Life Res*. 2009;18:1043–1051.
20. Bakhsh H, Ibrahim I, Khan W, et al. Assessment of validity, reliability, responsiveness and bias of three commonly used patient-reported outcome measures in carpal tunnel syndrome. *Ortop Traumatol Rehabil*. 2012;14:335–340.
21. Rodrigues JN, Mabvuure NT, Nikkhah D, et al. Minimal important changes and differences in elective hand surgery. *J Hand Surg Eur Vol*. 2015;40:900–912.
22. Hammond A, Prior Y, Tyson S. Linguistic validation, validity and reliability of the British English versions of the Disabilities of the Arm, Shoulder and Hand (DASH) questionnaire and QuickDASH in people with rheumatoid arthritis. *BMC Musculoskelet Disord*. 2018;19:118.
23. Gummesson C, Atroshi I, Ekdahl C. The disabilities of the arm, shoulder and hand (DASH) outcome questionnaire: longitudinal construct validity and measuring self-rated health change after surgery. *BMC Musculoskelet Disord*. 2003;4:11.
24. SooHoo NF, McDonald AP, Seiler III JG, et al. Evaluation of the construct validity of the DASH questionnaire by correlation to the SF-36. *J Hand Surg*. 2002;27:537–541.
25. Gay RE, Amadio PC, Johnson JC. Comparative responsiveness of the disabilities of the arm, shoulder, and hand, the carpal tunnel questionnaire, and the SF-36 to clinical change after carpal tunnel release. *J Hand Surg*. 2003;28:250–254.
26. Hobby JL, Watts C, Elliot D. Validity and responsiveness of the patient evaluation measure as an outcome measure for carpal tunnel syndrome. *J Hand Surg Br*. 2005;30:350–354.

27. Whalley K, Adams J. The longitudinal validity of the quick and full version of the Disability of the Arm Shoulder and Hand questionnaire in musculoskeletal hand outpatients. *Hand Ther* 2009;14:22–25.
28. McMillan CR, Binhammer PA. Which outcome measure is the best? Evaluating responsiveness of the Disabilities of the Arm, Shoulder, and Hand Questionnaire, the Michigan Hand Questionnaire and the Patient-Specific Functional Scale following hand and wrist surgery. *Hand.* 2009;4:311–318.
29. Chapman TT, Richard RL, Hedman TL, et al. Combat casualty hand burns: evaluating impairment and disability during recovery. *J Hand Ther.* 2008;21:150–159.
30. Rodrigues JN, Zhang W, Scammell BE, et al. Recovery, responsiveness and interpretability of patient-reported outcome measures after surgery for Dupuytren's disease. *J Hand Surg Eur Vol.* 2017;42:301–309.
31. Lyrén PE, Atroshi I. Using item response theory improved responsiveness of patient-reported outcomes measures in carpal tunnel syndrome. *J Clin Epidemiol.* 2012;65:325–334.
32. Gabel CP, Yelland M, Melloh M, et al. A modified QuickDASH-9 provides a valid outcome instrument for upper limb function. *BMC Musculoskelet Disord.* 2009;10:161.
33. Polson K, Reid D, McNair PJ, et al. Responsiveness, minimal importance difference and minimal detectable change scores of the shortened disability arm shoulder hand (QuickDASH) questionnaire. *Manual Therapy.* 2010;15:404–407.
34. Budd HR, Larson D, Chojnowski A, et al. The QuickDASH score: a patient-reported outcome measure for Dupuytren's surgery. *J Hand Ther.* 2011;24:15–21.
35. Massy-Westropp N, Krishnan J, Ahern M. Comparing the AUSCAN Osteoarthritis Hand Index, Michigan Hand Outcomes Questionnaire, and Sequential Occupational Dexterity Assessment for patients with rheumatoid arthritis. *J Rheumatol.* 2004;31:1996–2001.
36. Waljee JF, Chung KC, Kim HM, et al. Validity and responsiveness of the Michigan Hand Questionnaire in patients with rheumatoid arthritis: a multicenter, international study. *Arthritis Care Res.* 2010;62:1569–1577.
37. Thoma A, Kaur MN, Ignacy TA, et al. Psychometric properties of health-related quality of life instruments in patients undergoing palmar fasciectomy for dupuytren's disease: a prospective study. *Hand (N Y).* 2014;9:166–174.
38. London DA, Stepan JG, Calfee RP. Determining the Michigan Hand Outcomes Questionnaire minimal clinically important difference by means of three methods. *Plastic Reconstr Surg.* 2014;133:616–625.
39. Horng YS, Lin MC, Feng CT, et al. Responsiveness of the Michigan Hand Outcomes Questionnaire and the Disabilities of the Arm, Shoulder, and Hand questionnaire in patients with hand injury. *J Hand Surg.* 2010;35:430–436.
40. van de Ven-Stevens LA, Graff MJ, Selles RW, et al. Instruments for assessment of impairments and activity limitations in patients with hand conditions: a European Delphi study. *J Rehabil Med.* 2015;47:948–956.
41. Kotsis SV, Chung KC. Responsiveness of the Michigan Hand Outcomes Questionnaire and the Disabilities of the Arm, Shoulder and Hand questionnaire in carpal tunnel surgery. *J Hand Surg.* 2005;30:81–86.
42. Weinstock-Zlotnick G, Page C, Ghomrawi HM, et al. Responsiveness of three patient report outcome (PRO) measures in patients with hand fractures: a preliminary cohort study. *J Hand Ther.* 2015;28(4):403–411.
43. Schouffoer AA, van der Giesen FJ, Beaart-van de Voorde LJ, et al. Validity and responsiveness of the Michigan Hand Questionnaire in patients with systemic sclerosis. *Rheumatology.* 2016;55:1386–1393.
44. Dritsaki M, Petrou S, Williams M, et al. An empirical evaluation of the SF-12, SF-6D, EQ-5D and Michigan Hand Outcome Questionnaire in patients with rheumatoid arthritis of the hand. *Health Qual Life Outcomes.* 2017;15:20.
45. Dias JJ, Bhowal B, Wildin CJ, et al. Assessing the outcome of disorders of the hand. Is the patient evaluation measure reliable, valid, responsive and without bias? *J Bone Joint Surg Br.* 2001;83:235–240.
46. Hamilton CB, Chesworth BM. A Rasch-validated version of the upper extremity functional index for interval-level measurement of upper extremity function. *Phys Ther.* 2013;93:1507–1519.
47. Hefford C, Abbott JH, Arnold R, et al. The patient-specific functional scale: validity, reliability, and responsiveness in patients with upper extremity musculoskeletal problems. *J Orthop Sports Phys Ther.* 2012;42:56–65.
48. Chesworth BM, Hamilton CB, Walton DM, et al. Reliability and validity of two versions of the upper extremity functional index. *Physiother Can.* 2014;66:243–253.
49. Lehman LA, Sindhu BS, Shechtman O, et al. A comparison of the ability of two upper extremity assessments to measure change in function. *J Hand Ther.* 2010;23:31–40.
50. Gheorghiu AM, Gyorfi H, Capota R, et al. Reliability, validity, and sensitivity to change of the simplified Duruoz Hand Index in systemic sclerosis. *Arthritis Rheumatol.* 2016;68.
51. Brower LM, Poole JL. Reliability and validity of the Duruöz Hand Index in persons with systemic sclerosis (scleroderma). *Arthritis Care Res.* 2004;51:805–809.
52. Dotu B, Usen A, Yilmaz F, et al. The assessment of the sensitivity to change in the Michigan Hand Outcome Questionnaire for patients with traumatic hand injuries. *Turk Fiz Tip Rehab D.* 2013;59:282.
53. Döring AC, Nota SP, Hageman MG, et al. Measurement of upper extremity disability using the Patient-Reported Outcomes Measurement Information System. *J Hand Surg Am.* 2014;39:1160–1165.
54. Overbeek CL, Nota SP, Jayakumar P, et al. The PROMIS physical function correlates with the QuickDASH in patients with upper extremity illness. *Clin Orthop Relat Res.* 2015;473:311–317.
55. Tyser AR, Beckmann J, Franklin JD, et al. Evaluation of the PROMIS physical function computer adaptive test in the upper extremity. *J Hand Surg Am.* 2014;39:2047.e4–2051.e4.
56. Prinsen CA, Vohra S, Rose MR, et al. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials.* 2014;15:247.
57. Gummesson C, Ward MM, Atroshi I. The shortened disabilities of the arm, shoulder and hand questionnaire (Quick DASH): validity and reliability based on responses within the full-length DASH. *BMC Musculoskelet Disord.* 2006;7:44.