NIH Public Access
Author Manuscript
Clin Plast Surg. Author manuscript; available in PMC 2013 January 14.

Published in final edited form as:
Clin Plast Surg. 2008 April ; 35(2): 239–250. doi:10.1016/j.cps.2007.10.001.

Measuring Outcomes in Hand Surgery

Aviram M. Giladi, MD[1] and **Kevin C. Chung, MD, MS**[2]

[1] Resident, Department of Surgery, Section of Plastic Surgery, The University of Michigan Health System

[2]Professor of Surgery, Section of Plastic Surgery, Assistant Dean for Faculty Affairs, The University of Michigan Medical School

Synopsis

Methods for measuring outcomes after hand and upper extremity surgery continue to evolve, yet remain inconsistent in quality. In this article we review the use of patient-reported outcomes measures in upper extremity surgery patients, and provides a practical guide to questionnaire selection, assessment, and utilization. We also present the future direction of health services research, and how it will drive changes in measuring outcomes in hand surgery.

Keywords

Patient-reported outcomes; hand surgery outcomes; patient-reported measures of hand function; COSMIN; CAT; PROMIS; Health services research

OVERVIEW

The upper extremity is a highly specialized functional, sensory, and aesthetic unit. The upper extremity can also suffer a unique range of insults. In 2010, the United States Bureau of Labor Statistics reported the annual incidence rate of hand injuries at 25.1 per 10,000 workers, and the most frequently injured population were young, active workers.(1) When the costs of medical care, rehabilitation, and productivity loss are computed, for this younger population of trauma patients as well as the often-older population with arthritis, neuropathies, and other sources of pain and functional loss, the burden of hand pathology is massive.(2) How we as providers evaluate and manage hand pathology is critical to individuals and society as a whole.

On the national level, as health care delivery and reimbursement in the United States undergoes rapid and substantial change, the focus on quality and value of care continues to increase. A shift towards value-based insurance models has begun.(3, 4) These programs aim to reduce patient costs and increase access and utilization of high-value treatments, while discouraging low-value treatments. “Choosing Wisely” and other similar campaigns

© 2012 Elsevier Inc. All rights reserved.

Corresponding Author: Kevin C. Chung, MD, MS, Section of Plastic Surgery, University of Michigan Health System, 2130 Taubman Center, SPC 5340, 1500 E. Medical Center Drive, Ann Arbor, MI, 48109-5340, kecchung@umich.edu, Phone 734-936-5885, Fax 734-763-5354.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosures: None

$watermark-text
$watermark-text
$watermark-text

$watermark-text

$watermark-text

$watermark-text

are also emphasizing appropriate and evidence-based surgical interventions.(5) Fee for service reimbursement is changing, and quality of care will play an increasing role in provider compensation.(6, 7) This all has resulted in a renewed focus on the need for high-quality evidence to support provider decision-making and delivery of care.

In the US, Canada, the UK, and many other nations, health services research has been a substantial area of focus for over 15 years.(8, 9) This field analyzes how patients access health care, what care costs, and what outcomes the patients experience as a result of this care. As work in this area continues to increase, the volume of literature aimed at addressing challenging issues in treatment quality, value, effectiveness, and appropriateness is growing. Unfortunately, the quality of this literature is inconsistent. Interpreting the various results and their potential impact continues to be a challenge as well. Considering the volume of hand pathology in the US and worldwide, providing high-quality, sustainable, effective, and cost-conscious care is paramount. Especially when preparing for the changing landscape of health care, an awareness of the various factors affecting outcomes after hand surgery is critical for continued improvement and success.

Patient-Reported Outcomes in Hand Surgery

Traditionally, when evaluating the quality of care in hand surgery, standard functional metrics have been measured -- fracture healing, range of motion, strength, sensation, and others.(10–12) In many cases, however, what we as providers consider substantial improvement does not align with the perceptions and experiences of our patients.(13, 14) That is not to say that traditional objective metrics cannot show significant differences in outcomes; rather, what is measured by these functional tests often does not translate to the outcomes desired by the patient, provider, or society. For example, fracture union on x-ray does not equate with a patient having high satisfaction with their outcome or with returning to activities of daily living (ADL). A growing appreciation of this dichotomy has led the drive to using patient reported outcome (PRO) metrics in the assessment of upper extremity disease. PRO questionnaires allow providers to assess function, health-related quality-of-life (HRQL), and satisfaction from the patient's perspective.(15, 16)

Understanding of a patient's HRQL requires an appreciation of physical, mental, and social well-being.(17) How satisfaction, function, pain control, and other components can affect HRQL has substantial impact on treatment decisions and outcomes. In addition, the degree to which expenditures can be justified is guided by the expected improvement in HRQL. Improving how these components are measured has formed the basis for design, application, and evaluation of numerous PRO instruments. The design and refinement of a PRO instrument is a weighty task. It requires a mix of qualitative and quantitative assessments. It must be tested with pilot patient cohorts, and relatively complex statistical analysis is needed to determine reliability and consistency. The instrument must then be evaluated for validity and responsiveness for the disease-state in question. This requires that each new metric be examined for each specific subset of patients the investigators aim to evaluate. The details of this process, and the various statistical measurements that are used, have been described by numerous authors and will not be covered in detail here.(15, 18–21) Table 1 contains a list of key quality domains and definitions.

Choosing an Outcomes Instrument

Even when properly vetted, validated PRO metrics do not all perform at the same level. For example, the Michigan Hand Questionnaire (MHQ)(22) and the Disability of Arm, Shoulder, and Hand questionnaire (DASH)(23) have both been validated for patients with carpal tunnel syndrome (CTS)(24); however, with additional sub-domains geared towards more than just functional aspects of disease, the MHQ is better able to evaluate the

$watermark-text

$watermark-text

$watermark-text

symptomatic components of CTS.(24) The Short Form-36 (SF-36)(25) is validated for rheumatoid arthritis (RA).(26, 27) For patients with RA, DASH scores were highly correlated with SF-36 for pain, but DASH was only moderately correlated for physical and mental function.(28) In contrast, for patients after distal radius fracture fixation, MHQ and DASH are significantly more responsive than SF-36.(29, 30) The challenge in hand surgery is deciding which metrics should be used for each patient population. Although the number of available PROs continues to grow, the number of valid and robust outcomes measures remains few and inconsistently utilized.(10, 31) Appropriate selection of PRO metrics will govern the value of any study results.

PRO Instruments

PRO questionnaires are classified as general, system-specific, and disease-specific.(21) General PRO measures evaluate qualitative and quantitative aspects of the patient's life without focusing on any specific disease or organ system. They ascertain general well-being, including components of pain, vitality, emotional and mental health, and self-assessment of ability to perform daily functions and activities. The SF-36 and Arthritis Impact Measurement Scales 2 (AIMS2)(32) are frequently used general PRO measures in hand surgery outcomes research.

System-specific, or domain-specific, instruments focus on an organ system or functional unit. These PRO metrics are geared towards better understanding of how the specific system of interest is affected by a diseased-state, what effects this has on the patient, and how these problems improve after intervention. This makes domain-specific instruments more valuable in intervention trials, but less likely to detect broader features of health states.(15) The most commonly used instruments in upper extremity studies are the MHQ, DASH, and Patient-Rated Wrist Evaluation outcomes questionnaire (PRWE).(19, 33)

Disease-specific instruments are geared towards a population grouped by a particular pathology. These metrics are utilized in evaluating treatment of the specific disease. The focused nature of the questionnaire often results in high responsiveness when used in the appropriate patient population.(15) However, the design often limits use in evaluating other diseases even within the same system, which restricts how the results from a disease-specific instrument are used. The Carpal Tunnel Questionnaire (CTQ) is a commonly used disease-specific instrument.(34)

For PRO metrics of all types, it is important to consider cross-cultural applications as well. Validity and responsiveness are population dependent, and this is an even greater issue when the different populations of interest do not speak the same language or live with similar cultural norms. The process of translating and subsequently validating quantitative and PRO instruments is challenging. It not only requires language conversion, but also ensuring that subtle nuances and organizational aspects of the translated questionnaire do not adversely affect the way patients understand and answer questions.(35–37) This can be something as clear as Korean patients showing limited understanding of questions related to self-feeding with a spoon rather than using chop-sticks.(38) It can also be far more complex, such as loss of idiomatic quality in translation from English to Spanish resulting in patients perceiving the questionnaire as less "serious."(39) The details of these concepts are beyond the scope of this paper. However, as health care delivery and research is increasingly more global, instruments with adequate cross-cultural equivalence will have broader usability in patient care and health services research.

$watermark-text

$watermark-text

$watermark-text

Understanding the Literature on PRO Metric Quality

Understanding the above classification scheme is only a small part of the decision tree in selecting outcomes tools. Adequate consistency, reliability, validity, and responsiveness of the instrument are a large component of this decision process as well. Although the volume of literature evaluating these quality measures of the different PRO metrics continues to increase, understanding these studies and the quality of their results remains challenging for most. Making this even more problematic, definitions and utilization of terms are inconsistent across various studies. This results in difficult decision-making in planning for a PRO-focused study, and limits quality of methodology and content of systematic reviews.(40)

A common concern when using PRO measures is how to interpret the scores. For example, what does a 10-point difference in the MHQ after treatment really mean -- although it is statistically significant, is it clinically significant? Interpretability provides an indication as to how well the quantitative data can be translated into qualitatively (clinically) relevant results.(41) This is most often done by determining the Minimally Clinically Important Difference (MCID).(42) In patients with CTS, the MCID of the MHQ pain sub-domain is 23, whereas the MCID for the function sub-domain is 13.(43) For RA patients, the MHQ sub-domain MCID for pain is 11 and for function is 13.(43) Although useful when available, the applicability is somewhat limited because meaningful clinical change varies between patient groups. However, having the MCID for a questionnaire in the population being evaluated gives an indication as to the clinical relevance of study results.

An additional approach to addressing the challenges in PRO metric evaluation has been to set guidelines and quality standards. Terwee et al published quality criteria for measurement properties (see Table 2), and provided guidelines as to how readers can critically evaluate published results.(44) The COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) study group has presented results of a four-round Delphi study, releasing additional guidelines on evaluating the methodological quality of studies on health status measurement instruments.(45) These guidelines include taxonomy of relationships of measured properties (see Figure 1), and a thorough analysis of what properties and methods must be employed and reported for the study to be of adequate quality.(41, 45) The COSMIN group has challenged some of the traditional tools and methods used in vetting these studies, and developed a series of checklists that guide thorough analysis of published results.(45, 46) These sets of standards and checklists are not intended to rate the specific instruments; rather, they provide a systematic approach to evaluating the studies that report on instrument quality, regardless of the study's conclusion.(46) Based on these standards, numerous systematic reviews have assessed the measurement properties and clinimetrics of available PRO metrics.(33, 47, 48) One such study evaluated the clinimetric properties of instruments used to assess patients with hand injuries.(49) They concluded that a significant majority of functional and patient-reported measures have been inadequately evaluated. MHQ, DASH, and CTQ are three of only five questionnaires to receive strong ratings, in that well-executed studies properly report reliability, validity, and responsiveness of these metrics.

In addition, the COSMIN group published consensus definitions for the extensive terminology used in determining PRO metric quality.(41) Consistency in the language used to report results will serve to improve the value and reliability of primary analyses, systematic reviews, and meta-analyses. The guidelines released by Terwee, followed by the COSMIN group's reports, have been large strides towards improving the quality of outcomes measurement. However, these tools are still limited, as use of COSMIN guidelines has shown lower inter-rater reliability than desired.(33) This is in part due to inexperience in

using the guidelines, as well as inconsistent use of terminology by the raters. When implemented by experienced authors, COSMIN guidelines are a useful tool.

Practical Decisions in Instrument Selection

Utilizing these guidelines and definitions can remove some of the challenge in reviewing this complex subset of hand surgery literature. However, even before looking through the literature, a more practical decision must be made – does the metric I plan to use contain the proper components to thoroughly evaluate the disease, treatment, and patient population in question? For this there are no developed guidelines, and the investigator must use empirical analysis and questionnaire evaluation.

Consider CTS, which has symptomatic and functional disease manifestations. Understanding that functional metrics would not capture all aspects of the patient's experience and post-operative recovery, the CTQ was designed for use with CTS patients, and is more responsive than traditional functional metrics.(34) The MHQ and DASH are also both responsive instruments for CTS.(24) The pain subdomain of MHQ has a large effect size, and overall MHQ score and DASH score have a moderate effect. Both out-perform the SF-36 for CTS.(24, 50)

When preparing to do a study on CTS, it is important to consider what each questionnaire can provide. A well-validated disease-specific metric is available; however, this would limit the ability to compare results across other diseases, as the CTQ is not validated for use in most upper extremity conditions. Using the MHQ or DASH would provide valid outcomes evaluation that could then be further analyzed and even compared to other patient populations for utility, cost-effectiveness, or other health-related outcomes.(51) Considering that MHQ has separate sub-scales for symptom and functional scores and evaluates the right and left hand separately, one must decide whether these additional elements provide desirable benefits when compared to other available questionnaires.

Similarly, with distal radius fracture patients, functional metrics have traditionally been measured. Range of motion and grip strength are important indicators of patient recovery.(52) Unlike with carpal tunnel syndrome, these measures perform as well as PROs in indicating treatment outcomes after distal radius fracture.(30) The MHQ and DASH have been validated for use in these patients.(29, 30) These instruments adequately evaluate function, ADLs, and pain, and DASH is shown to be highly responsive in the first 3 months after injury when functional metrics are more difficult to evaluate.(29) A third questionnaire, the PRWE, has also been found to have a large effect size and slightly greater responsiveness than DASH.(29) This instrument, although robust for wrist pathology, has not been validated for use with as many other upper extremity disorders. The questionnaires also evaluate patient satisfaction, which functional metrics do not explore.

In studies on patients with distal radius fractures, the system-specific questionnaires provide additional insight into the domains of pain and satisfaction that functional measures do not. However, the MHQ, DASH, and PRWE all lack the ability to ascertain a greater sense of global well-being and overall health status that a general measure would provide. Adding an additional questionnaire that measures social and mental components of outcomes can give a more comprehensive evaluation of the whole patient experience. Rheumatoid arthritis (RA), with emotional and physical manifestations far beyond the upper extremity, sits on the opposite side of the spectrum from fractures. General PROs, including SF-36, AIMS2, and the Health Assessment Questionnaire (HAQ), have been validated for RA by numerous studies.(26, 27, 53, 54) These questionnaires have high responsiveness for RA patients. In evaluating rheumatoid hand function, a system-specific instrument can also be used. The MHQ is validated in this population, both in those who underwent metacarpophalangeal

$watermark-text $watermark-text $watermark-text

(MCP) arthroplasty and those who did not have surgical correction of MCP disease.(55) The sub-domains and overall scores correlated with aspects of the disease without surgical treatment as well as in post-operative recovery, and showed construct validity when compared to AIMS2. Even with these results, when considering the substantial psychosocial overtones of this condition, using any system-specific questionnaire alone for a study on hand function in RA would not capture general health status components of this disease. Therefore, it is important to consider adding a questionnaire that measures the psychosocial component of RA.

Minimizing Patient Burden

When attempting to capture the multidimensional aspects of a complex disease, longer questionnaires are known to have more incomplete data, lower response rates, and often lesser quality results because of responder fatigue and loss to follow-up.(56–58) Several outcomes tools have now been shortened. For example, the SF36 and the MHQ have been restructured to create the SF12 and the Brief-MHQ.(56, 59) These questionnaires use only one or two items to assess each domain. Although the shortened questionnaire may lose some precision, it can enhance responder compliance by making it less strenuous to complete the questionnaire. These shortened questionnaires have been developed through rigorous methodology, and thus-far have performed at or above the level of their more-comprehensive predecessors.(59–61)

Another difficulty to consider in upper extremity PRO metric design and utilization is ceiling and floor effects, in which too many patients score in the highest or lowest range due to inadequate discrimination between patients with different degrees of recovery.(62) With a ceiling effect, some patients with residual impairment are already scoring at the maximum level. A floor effect results in patients having the lowest possible score even when they are actually in worse condition than others with similar low scores. Upper extremity metrics more often risk ceiling effects.(15) The full-version questionnaires have optimized item quality to minimize a ceiling effect, which is one of the benefits of having the larger number of items. When reducing the number of questions, there is great potential for augmenting a ceiling effect. The ideal is to identify the fewest number of questions that provide precision and also allows for adequately stratifying patients.

Aiming to address these issues, investigators have used concepts guided by Item Response Theory (IRT) to design and refine Computerized Adaptive Testing (CAT). IRT uses each item as an indicator of ability or condition, modeling the answer by a respondent with a certain degree of function or ability to each of the items in the questionnaire.(57) This provides insight into the responder's abilities or skills based on how they answer each item. With IRT, item number can be decreased, metrics can be standardized, and results can be meaningfully compared. This is also true for questionnaires translated into different languages.(63, 64) Although initially used for educational and psychological testing, IRT has also provided unique tools for improving questionnaire design and utilization in health services research. For example, IRT was used to develop an alternate summary score for the 10-item Physical Functioning scale (PF-10) of the SF36.(65) The new summative scale had improved precision especially in patients well above or below median scores. Subsequently, the use of IRT has rapidly increased.

IRT has also guided the design and utilization of CAT, a model in which the instrument progressively adapts to the individual answering the questions.(57) Based on how one question is answered, the next question can be selected and geared to provide more discerning information about the patient's condition. This allows for the overall question bank to remain large – helping to minimize ceiling and floor effects – while asking questions

$watermark-text
$watermark-text
$watermark-text

that provide adequate stratification of patient conditions and keeping the overall number of questions low.

The Future of Outcomes Measurement

Use of IRT-based techniques, and the shift towards CAT, has changed how health services researchers approach PROs. CAT has been used to improve measurement precision over a wide range of health conditions while also having reduced testing burden.(57) Furthering these efforts, the NIH have put a large focus on developing and utilizing the Patient Reported Outcomes Measurement Information System (PROMIS).(66) PROMIS includes a growing bank of thoroughly vetted and tested questionnaire items, and divides them into key domains, for example, pain, fatigue, physical function, etc. Using IRT, different scales and questionnaires have been developed and geared towards specific patient populations, supplanting general PRO questionnaires. These methods have yet to substantially affect hand and upper extremity PRO evaluation. However, the physical function item banks include sub-sets for upper extremity function.(67, 68) In addition, an upper extremity function scale has been created for use with pediatric cancer patients.(66) Investigators have also shown improved responsiveness with reduction in floor and ceiling effects with IRT-based PROMIS instruments for patients with RA.(69) As these tools are developed and refined, it may usher us away from needing disease- or symptom-specific metrics.

Conclusion

In measuring hand surgery outcomes, there are unique challenges. Improving how we evaluate the physical, emotional, aesthetic, and psychological components of disease has resulted in substantial change. Shifting to PROs ushered in the current era of hand surgery-related health services research. However, inconsistent design and use of PRO metrics contributes to continued deficiencies in appropriately measuring outcomes.

It is important to choose the questionnaire(s) that will suit study needs. Making assumptions about how the disease affects patients will lead to better study design and outcomes tool selection. In addition, the investigator must be careful to avoid over-burdening subjects with numerous tests and questionnaires. Finding the right combination of outcomes metrics without compromising study quality can be mitigated in part by thoughtful selection of robust and appropriate instruments. As use of IRT and CAT continues to mold the future of health services and outcomes research, measuring outcomes in hand surgery will again require a shift in technique, metric design, and study execution. As it becomes increasingly important to make these changes, both for patients and for developing sustainable practice models in an evolving health-care climate, improving and maintaining efficiency, quality, and consistency must define hand surgery outcomes research.

Acknowledgments

Supported in part by grants from the National Institute on Aging and National Institute of Arthritis and Musculoskeletal and Skin Diseases (R01 AR062066) and from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (2R01 AR047328-06) and a Midcareer Investigator Award in Patient-Oriented Research (K24 AR053120) (to Dr. Kevin C. Chung).

Abbreviations: Measuring Outcomes in Hand Surgery

AIMS2	Arthritis Impact Measurement Scales 2
ADL	Activities of daily living

$watermark-text
$watermark-text
$watermark-text

$watermark-text

$watermark-text

$watermark-text

CAT	Computerized Adaptive Testing
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
CTQ	Carpal Tunnel Questionnaire
CTS	Carpal tunnel syndrome
DASH	Disability of Arm, Shoulder, and Hand questionnaire
HAQ	Health Assessment Questionnaire
HRQL	Health-related quality-of-life
IRT	Item Response Theory
MCID	Minimally Clinically Important Difference
MCP	Metacarpophalangeal
MHQ	Michigan Hand Questionnaire
PF-10	10-item Physical Functioning scale
PRO	Patient reported outcome
PROMIS	Patient Reported Outcomes Measurement Information System
PRWE	Patient-Rated Wrist Evaluation outcomes questionnaire
RA	Rheumatoid arthritis
SF-36	Short Form-36

References

1. Nonfatal Occupational Injuries and Illnesses Requiring Days Away From Work, 2010. Bureau of Labor Statistics, USDL-11-1612. 2011 Nov.
2. de Putter CE, Selles RW, Polinder S, et al. Economic impact of hand and wrist injuries: health-care costs and productivity costs in a population-based study. J Bone Joint Surg Am. 2012; 94:e56. [PubMed: 22552678]
3. Chernew ME, Rosen AB, Fendrick AM. Value-based insurance design. Health Aff (Millwood). 2007; 26:w195–w203. [PubMed: 17264100]
4. Choudhry NK, Rosenthal MB, Milstein A. Assessing the evidence for value-based insurance design. Health Aff (Millwood). 2010; 29:1988–1994. [PubMed: 21041737]
5. www.choosingwisely.org
6. Share DA, Mason MH. Michigan's Physician Group Incentive Program Offers A Regional Model For Incremental 'Fee For Value' Payment Reform. Health Aff (Millwood). 2012; 31:1993–2001. [PubMed: 22949448]
7. Landman JH. Recommendations for delivering value. Healthc Financ Manage. 2012; 66:98–100. [PubMed: 22788044]
8. Epstein AM. The outcomes movement--will it get us where we want to go? N Engl J Med. 1990; 323:266–270. [PubMed: 2366836]
9. Eisenberg JM. Putting research to work: reporting and enhancing the impact of health services research. Health Serv Res. 2001; 36:x–xvii.
10. Bindra RR, Dias JJ, Heras-Palau C, et al. Assessing outcome after hand surgery: the current state. J Hand Surg Br. 2003; 28:289–294. [PubMed: 12849936]
11. Crosby CA, Wehbe MA, Mawr B. Hand strength: normative values. J Hand Surg Am. 1994; 19:665–670. [PubMed: 7963331]

$watermark-text

$watermark-text

$watermark-text

12. LaStayo PC, Wheeler DL. Reliability of passive wrist flexion and extension goniometric measurements: a multicenter study. Phys Ther. 1994; 74:162–174. discussion 174-166. [PubMed: 8290621]
13. Berkanovic E, Hurwicz ML, Lachenbruch PA. Concordant and discrepant views of patients' physical functioning. Arthritis Care Res. 1995; 8:94–101. [PubMed: 7794992]
14. Hewlett SA. Patients and clinicians have different perspectives on outcomes in arthritis. J Rheumatol. 2003; 30:877–879. [PubMed: 12672220]
15. Szabo RM. Outcomes assessment in hand surgery: when are they meaningful? J Hand Surg Am. 2001; 26:993–1002. [PubMed: 11721242]
16. Alderman AK, Chung KC. Measuring outcomes in hand surgery. Clin Plast Surg. 2008; 35:239–250. [PubMed: 18298996]
17. Ware JE Jr. Standards for validating health measures: definition and content. J Chronic Dis. 1987; 40:473–480. [PubMed: 3298292]
18. Zlowodzki M, Bhandari M. Outcome measures and implications for sample-size calculations. J Bone Joint Surg Am. 2009; 91(Suppl 3):35–40. [PubMed: 19411498]
19. 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. [PubMed: 17534619]
20. Davis Sears E, Chung KC. A guide to interpreting a study of patient-reported outcomes. Plast Reconstr Surg. 2012; 129:1200–1207. [PubMed: 22544102]
21. Fitzpatrick R, Davey C, Buxton MJ, et al. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998; 2:i–iv. 1–74.
22. 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. [PubMed: 9708370]
23. Beaton DE, Katz JN, Fossel AH, et al. Measuring the whole or the parts? Validity, reliability, and responsiveness of the Disabilities of the Arm, Shoulder and Hand outcome measure in different regions of the upper extremity. J Hand Ther. 2001; 14:128–146. [PubMed: 11382253]
24. 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 Am. 2005; 30:81–86. [PubMed: 15680560]
25. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992; 30:473–483. [PubMed: 1593914]
26. Linde L, Sorensen J, Ostergaard M, et al. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D [corrected] RAQoL, and HAQ in patients with rheumatoid arthritis. J Rheumatol. 2008; 35:1528–1537. [PubMed: 18484697]
27. Oude Voshaar MA, ten Klooster PM, Taal E, et al. Measurement properties of physical function scales validated for use in patients with rheumatoid arthritis: a systematic review of the literature. Health Qual Life Outcomes. 2011; 9:99. [PubMed: 22059801]
28. Aktekin LA, Eser F, Baskan BM, et al. Disability of Arm Shoulder and Hand Questionnaire in rheumatoid arthritis patients: relationship with disease activity, HAQ, SF-36. Rheumatol Int. 2011; 31:823–826. [PubMed: 20680284]
29. MacDermid JC, Richards RS, Donner A, et al. Responsiveness of the short form-36, disability of the arm, shoulder, and hand questionnaire, patient-rated wrist evaluation, and physical impairment measurements in evaluating recovery after a distal radius fracture. J Hand Surg Am. 2000; 25:330–340. [PubMed: 10722826]
30. Kotsis SV, Lau FH, Chung KC. Responsiveness of the Michigan Hand Outcomes Questionnaire and physical measurements in outcome studies of distal radius fracture treatment. J Hand Surg Am. 2007; 32:84–90. [PubMed: 17218180]
31. Chung KC, Burns PB, Davis Sears E. Outcomes research in hand surgery: where have we been and where should we go? J Hand Surg Am. 2006; 31:1373–1379. [PubMed: 17027802]
32. Meenan RF, Mason JH, Anderson JJ, et al. AIMS2. The content and properties of a revised and expanded Arthritis Impact Measurement Scales Health Status Questionnaire. Arthritis Rheum. 1992; 35:1–10. [PubMed: 1731806]
33. Hoang-Kim A, Pegreffi F, Moroni A, et al. Measuring wrist and hand function: common scales and checklists. Injury. 2011; 42:253–258. [PubMed: 21159335]

$watermark-text

$watermark-text

$watermark-text

34. Levine DW, Simmons BP, Koris MJ, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. J Bone Joint Surg Am. 1993; 75:1585–1592. [PubMed: 8245050]
35. Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000; 25:3186–3191. [PubMed: 11124735]
36. Gonzalez-Calvo J, Gonzalez VM, Lorig K. Cultural diversity issues in the development of valid and reliable measures of health status. Arthritis Care Res. 1997; 10:448–456. [PubMed: 9481237]
37. Van Ommeren M. Validity issues in transcultural epidemiology. Br J Psychiatry. 2003; 182:376–378. [PubMed: 12724237]
38. Roh YH, Yang BK, Noh JH, et al. Cross-cultural adaptation and validation of the Korean version of the Michigan hand questionnaire. J Hand Surg Am. 2011; 36:1497–1503. [PubMed: 21783329]
39. Esposito N. From meaning to meaning: the influence of translation techniques on non-English focus group research. Qual Health Res. 2001; 11:568–579. [PubMed: 11521612]
40. Mokkink LB, Terwee CB, Stratford PW, et al. Evaluation of the methodological quality of systematic reviews of health status measurement instruments. Qual Life Res. 2009; 18:313–333. [PubMed: 19238586]
41. 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. [PubMed: 20494804]
42. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989; 10:407–415. [PubMed: 2691207]
43. Shauver MJ, Chung KC. The minimal clinically important difference of the Michigan hand outcomes questionnaire. J Hand Surg Am. 2009; 34:509–514. [PubMed: 19258150]
44. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007; 60:34–42. [PubMed: 17161752]
45. 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. [PubMed: 20169472]
46. Angst F. The new COSMIN guidelines confront traditional concepts of responsiveness. BMC Med Res Methodol. 2011; 11:152. author reply 152. [PubMed: 22099330]
47. Oftedal S, Bell KL, Mitchell LE, et al. A systematic review of the clinimetric properties of habitual physical activity measures in young children with a motor disability. Int J Pediatr. 2012; 2012:976425. [PubMed: 22927865]
48. Dobson F, Choi YM, Hall M, et al. Clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments: A systematic review. Arthritis Care Res (Hoboken). 2012
49. van de Ven-Stevens LA, Munneke M, Terwee CB, et al. Clinimetric properties of instruments to assess activities in patients with hand injury: a systematic review of the literature. Arch Phys Med Rehabil. 2009; 90:151–169. [PubMed: 19154842]
50. 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 Am. 2003; 28:250–254. [PubMed: 12671856]
51. Chatterjee JS, Price PE. Comparative responsiveness of the Michigan Hand Outcomes Questionnaire and the Carpal Tunnel Questionnaire after carpal tunnel release. J Hand Surg Am. 2009; 34:273–280. [PubMed: 19181227]
52. Chung KC, Haas A. Relationship between patient satisfaction and objective functional outcome after surgical treatment for distal radius fractures. J Hand Ther. 2009; 22:302–307. quiz 308. [PubMed: 19560317]
53. Tugwell P, Idzerda L, Wells GA. Generic quality-of-life assessment in rheumatoid arthritis. Am J Manag Care. 2008; 14:234. [PubMed: 18415966]
54. Russell AS. Quality-of-life assessment in rheumatoid arthritis. Pharmacoeconomics. 2008; 26:831–846. [PubMed: 18793031]

$watermark-text

55. 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 (Hoboken). 2010; 62:1569–1577. [PubMed: 20521331]
56. Ware J Jr, Kosinski M, Keller SD. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. Med Care. 1996; 34:220–233. [PubMed: 8628042]
57. Ware JE Jr. Improvements in short-form measures of health status: introduction to a series. J Clin Epidemiol. 2008; 61:1–5. [PubMed: 18083456]
58. Holland R, Smith RD, Harvey I, et al. Assessing quality of life in the elderly: a direct comparison of the EQ-5D and AQoL. Health Econ. 2004; 13:793–805. [PubMed: 15322991]
59. Waljee JF, Kim HM, Burns PB, et al. Development of a brief, 12-item version of the Michigan Hand Questionnaire. Plast Reconstr Surg. 2011; 128:208–220. [PubMed: 21701336]
60. Osthus TB, Preljevic VT, Sandvik L, et al. Mortality and health-related quality of life in prevalent dialysis patients: comparison between 12- items and 36-items short-form health survey. Health Qual Life Outcomes. 2012; 10:46. [PubMed: 22559816]
61. Gandhi SK, Salmon JW, Zhao SZ, et al. Psychometric evaluation of the 12-item short-form health survey (SF-12) in osteoarthritis and rheumatoid arthritis clinical trials. Clin Ther. 2001; 23:1080–1098. [PubMed: 11519772]
62. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. Ann Intern Med. 1993; 118:622–629. [PubMed: 8452328]
63. Mokkink LB, Knol DL, van Nispen RM, et al. Improving the quality and applicability of the Dutch scales of the Communication Profile for the Hearing Impaired using item response theory. J Speech Lang Hear Res. 2010; 53:556–571. [PubMed: 20530379]
64. Jafari P, Bagheri Z, Ayatollahi SM, et al. Using Rasch rating scale model to reassess the psychometric properties of the Persian version of the PedsQL 4.0 Generic Core Scales in school children. Health Qual Life Outcomes. 2012; 10:27. [PubMed: 22414135]
65. McHorney CA, Haley SM, Ware JE Jr. Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. J Clin Epidemiol. 1997; 50:451–461. [PubMed: 9179104]
66. www.nihpromis.org
67. Hung M, Clegg DO, Greene T, et al. Evaluation of the PROMIS physical function item bank in orthopaedic patients. J Orthop Res. 2011; 29:947–953. [PubMed: 21437962]
68. DeWitt EM, Stucky BD, Thissen D, et al. Construction of the eight-item patient-reported outcomes measurement information system pediatric physical function scales: built using item response theory. J Clin Epidemiol. 2011; 64:794–804. [PubMed: 21292444]
69. Fries J, Rose M, Krishnan E. The PROMIS of better outcome assessment: responsiveness, floor and ceiling effects, and Internet administration. J Rheumatol. 2011; 38:1759–1764. [PubMed: 21807798]

$watermark-text

$watermark-text

$watermark-text

Key Points

- A shift towards value-based insurance models has begun with programs that aim to reduce patient costs and increase access and utilization of high-value treatments, while discouraging low-value treatments.
- It is important to choose the questionnaire(s) that will suit study needs.
- In addition, the investigator must be careful to avoid over-burdening subjects with numerous tests and questionnaires.
- Finding the right combination of outcomes metrics without compromising study quality can be mitigated in part by thoughtful selection of robust and appropriate instruments.
- As use of IRT and CAT continues to mold the future of health services and outcomes research, measuring outcomes in hand surgery will again require a shift in technique, metric design, and study execution.

$watermark-text
$watermark-text
$watermark-text

Figure 1.
COSMIN taxonomy of relationships of measurement properties. COSMIN – COnsensus-based Standards for the selection of health Measurement Instruments, HR-PRO – health relatied-patient reported outcome. Duplicate from *Mokkink, L. B., Terwee, C. B., Patrick, D. L., 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 63: 737-745, 2010.* With Permission.

$watermark-text

$watermark-text

$watermark-text

Table 1

Definitions for key measurement properties used in evaluating the quality of patient reported outcomes instruments.

Measurement Property	Definition
Content Validity	The degree to which the content of an instrument is an adequate reflection of the construct to be measured
Criterion Validity	Strength of relationship between questionnaire scores and a measurable external criterion (the “gold standard”)
Construct Validity	The degree to which the scores of a questionnaire are consistent with the theoretical construct (hypothesis) that is being measured
Face Validity	The degree to which items in an instrument look as though they are an adequate reflection of the construct being measured
Internal Consistency	The extent to which the items are interrelated, and thus measure the same construct
Reliability	The extent to which patients can be distinguished from each other despite measurement errors
Test-Retest Reliability	The extent to which scores for patients who have not changed are the same in repeated measurements over time
Interrater Reliability	The extent to which scores for patients who have not changed are the same over repeated measurements by different examiners during the same visit
Responsiveness	The ability to detect clinically meaningful change over time in the construct being measured
Interpretability	The degree to which quantitative scores can be given qualitative meaning. Identifying clinically important differences in results.
Cross-Cultural Equivalence	The same measurement instrument used in different cultures measures the same construct without additional external cultural influences on results

$watermark-text

$watermark-text

$watermark-text

Table 2

Quality criteria for the key measurement properties used in evaluating patient reported outcomes instruments.

Measurement Property	Quality Criteria – Positive Rating
Content Validity	A clear description is provided of the measurement aim, the target population, the concepts that are being measured, and the item selection *AND* target population and investigators and/or experts were involved in item selection
Internal Consistency	Factor analyses performed on adequate sample size (7 × #items and ≥ 100) *AND* Cronbach's alpha(s) calculated per dimension AND Cronbach's alpha(s) between 0.70 and 0.95
Criterion Validity	Convincing argument that gold standard is "gold" *AND* correlation with gold standard ≥ 0.70
Construct Validity	Specific hypotheses were formulated *AND* at least 75% of the results are in accordance with these hypotheses
Reliability	Intraclass correlation coefficient (for continuous measures) or weighted Kappa (for ordinal measures) ≥0.70
Responsiveness	SDC *OR* SDC < MCID *OR* MCID outside the limits of agreement *OR* Guyatt's responsiveness ratio > 1.96 *OR* area under the receiver operating curve ≥ 0.70
Floor and Ceiling Effects	≤15% of the respondents achieved the highest or lowest possible scores
Interpretability	Mean and standard deviation scores presented for at least four relevant subgroups of patients *AND* MCID defined

SDC = Smallest Detectable Change; MCID = Minimal Clinically Important Difference. Adapted from *Terwee, C. B., Bot, S. D., de Boer, M. R., et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 60: 34–42, 2007.* With Permission