Published in final edited form as:

J Hand Surg Am. 2013 February ; 38(2): 401–406. doi:10.1016/j.jhsa.2012.11.028.

Measurement Scales in Clinical Research of the Upper Extremity, Part 1: General Principles, Measures of General Health, Pain, and Patient Satisfaction

Marie Badalamente, PhD, **Laureen Coffelt, BScOT**, **John Elfar, MD**, **Glenn Gaston, MD**, **Warren Hammert, MD**, **Jerry Huang, MD**, **Lisa Lattanza, MD**, **Joy Macdermid, PhD**, **Greg Merrell, MD**, **David Netscher, MD**, **Zubin Panthaki, MD**, **Greg Rafijah, MD**, **Douglas Trczinski, MD**, and **Brent Graham, MD [Chair] for the American Society for Surgery of the Hand Clinical Trials and Outcomes Committee**

American Society for Surgery of the Hand, Chicago, IL

Abstract

Measurement is a fundamental cornerstone in all aspects of scientific discovery, including clinical research. To be useful, measurement instruments must meet several key criteria, the most important of which are satisfactory reliability, validity, and responsiveness. Part 1 of this article reviews the general concepts of measurement instruments and describes the measurement of general health, pain, and patient satisfaction.

Keywords

Measurement outcomes; upper extremity

Measurement is a fundamental cornerstone in all aspects of scientific discovery, including clinical research. To be useful, measurement instruments must meet several key criteria, the most important of which are satisfactory reliability, validity, and responsiveness.[1] The instrument should also have sensibility, an attribute that Feinstein[1] defines as “enlightened common sense.” Some of the features that confer sensibility are those that make the instrument user friendly, such as its format and output. Sensibility is also defined by features that interface with the idea of validity, especially the concepts of content and face validity.

The concepts of reliability and validity are more familiar to clinicians and are the attributes on which most clinical researchers focus when choosing a measurement instrument. Excellent reliability is an absolutely necessary feature for any measurement instrument. Poor reliability greatly reduces the usefulness of any scale, regardless of how valid it might be considered. Simply stated, if the instrument cannot measure the phenomenon in a reproducible manner, it has little value. Nonetheless, reliability by itself, although a necessary condition, is not sufficient to qualify an instrument as useful, because it must also

Corresponding author: Brent Graham, MD, Toronto Western Hospital, 399 Bathurst Street, East Wing 2-425, Toronto, Ontario M5T 2S8; Brent.Graham@uhn.on.ca.

No benefits in any form have been received or will be received related directly or indirectly to the subject of this article.

be valid. In other words, it must measure the phenomenon of interest with accuracy. Usually, accuracy denotes comparison with a reference standard; however, there are many clinical phenomena for which an external gold standard will be lacking. For this reason, validation of a measurement instrument should be seen as an ongoing process that takes place as more experience is gained with the instrument through time.[2]

Finally, the concept of responsiveness should be considered a key feature of any measurement instrument. This reflects the sensitivity of the instrument to detect change in the phenomenon under study. Often, responsiveness appears to be poor because the instrument is not suited to the task. A common example of this in clinical research in hand surgery is the use of generic health measures such as the SF-36. This scale is intended to measure various facets of general health, many of which would not be materially affected by relatively minor conditions that affect the hand; for example, the sensory disturbance associated with carpal tunnel syndrome. The idea of selecting or developing responsive instruments has been captured in the concept of the minimally clinically important difference,[3,4] which has now been established for some of the instruments used by clinical researchers in hand and upper extremity care.[5] This identifies the smallest change in the output of an instrument that can be considered to be clinically meaningful. Knowledge of this can allow the most efficient planning possible of costly investigations such as randomized clinical trials.

Responsiveness can also be linked to floor or ceiling effects. These characteristics refer to the ability of the instrument to identify changes at the extremes of the measurement scale. For example, if increments of improvement exist above what the scale can measure, the instrument has a ceiling effect because these changes are not identified. Conversely, a scale that has a floor effect might not be able to measure below a certain level so that all individuals falling below this level will be classified as being equivalent. Ceiling or floor effects might be important in judging the responsiveness of a scale if measurements at the extremes of the domain of interest are likely to be important to the goals of the study.

If a number of different instruments could be chosen to evaluate a particular outcome in a research setting, the investigator should select those instruments that are best suited to the goal of the study. Although this would seem to be self-evident, selection of an inappropriate instrument is a common mistake made by clinical researchers. It should be clear that there is no one instrument that is optimal for all needs. The objective of this project was to provide an overview of the characteristics of measurement instruments that have been developed for use in clinical research in the upper extremity and report on their performance as reflected in the literature. The goal is to help guide researchers to instruments that will best meet their needs.

GENERIC HEALTH MEASURES

By their nature, generic health instruments seek to quantify general health and well-being. The most familiar[6,7] are scales such as the SF-36 or its abbreviated version the SF-12; however, there are many other similar scales in use.[8] These scales measure general health by probing various facets of functioning, often by using multiple subscales. In some instances,

the various aspects of physical, social, emotional, and mental functioning can be summed into an overall measure of general health status. Other instruments are not intended to have the various subscales combined in any way, but rather to survey these as individual characteristics.

Many of these scales have been used extensively in a wide range of clinical disciplines, and their reliability and validity have been well established. Nonetheless, they have meaning only when the phenomenon of interest is likely to have impact on general health and well-being. In other words, they are responsive only to conditions that affect general health.[9,10] With a few exceptions, conditions that affect the hand are unlikely to have a meaningful impact on general health. However, there are diseases that affect the hand and also have substantial systemic effects, such as rheumatoid arthritis and other conditions, such as injuries to the brachial plexus or severe mutilating hand injuries that impair upper extremity so substantially that there are systemic ramifications that might be measured by generic health measures. Emotional functioning might be affected by some of these conditions, and instruments with subscales that measure this aspect of general health might provide useful information. There might also be instances in which severe, acute pain from a relatively small problem, such as de Quervain's tenosynovitis, for example, might have an impact on a patient's functioning that is sufficient to be reflected in changes on a generic health measure. To a certain extent, this reflects the subjective nature of patient reports that comprise generic health measures, in general. It seems obvious that there is also a relationship between general health and quality of life; however, this is complex and not always clear.[11] In fact health-related quality of life is really a subset of quality of life in general, insofar as it attempts to measure the extent to which physical, emotional, and social well-being is affected by disease and treatment.[12]

In general, generic health measures play a minor, supporting role in the evaluation of most conditions that affect the upper extremity. They should be used to evaluate ideas, such as the burden of a specific disease. Instances in which they would represent the primary outcome measure will be rarely encountered; however, together with disease-specific scales, they might provide important insight into the overall health-related quality of life of an individual.[13] The main value of generic health measures might be in the comparison of different conditions; however, this might not be germane to many conditions affecting the upper extremity.

DISEASE-SPECIFIC MEASURES

There have been a number of disease-specific measures that are relevant to clinical research in the upper extremity. Many of these evaluate disease status and measure changes in disease activity. As a result, they might only indirectly measure an outcome related to a surgical intervention. In other words, insofar as the disease is somehow changed by an intervention, this might be reflected in a change in disease status as measured by a particular outcome instrument. Some examples include the Arthritis Impact Measurement Scale[14] and the Health Assessment Questionnaire,[15] both of which were developed for the evaluation of rheumatoid arthritis; the Systemic Lupus Erythematosis Disease Activity Index for lupus[16]; and the Western Ontario and McMaster Universities Osteoarthritis Index for osteoarthritis.[17]

Disease-specific measures should be chosen when responsiveness is paramount in importance, assuming that reliability and validity are acceptable.

These instruments derive their responsiveness from a focus on important aspects of a particular condition. Therefore, they are usually well suited to measure changes in the status of these features. In general, the performance of these measures is related to the disease for which they were developed. As a consequence, they will not necessarily function well when used to evaluate conditions for which they were not necessarily created, although there are a few circumstances in which measures developed for one condition might function adequately in the evaluation of other similar conditions. One example is the Carpal Tunnel Syndrome Severity Score,[18] which, with minor changes, could conceivably be validated for use in the evaluation of a condition such as cubital tunnel syndrome.[19]

REGION-SPECIFIC MEASURES

Region-specific scales, although not exclusively limited to the evaluation of extremity conditions, have been popular with musculoskeletal researchers, and a number of these have been developed and are in wide use.[20,21] Their attractiveness is linked to the global nature of the assessment they provide. They have also allowed some comparison of different conditions that affect the entire upper extremity, or a subset of the extremity, by providing a common tool for measurement.[22] The main characteristic of this kind of scale is the implicit summing of the overall status of the limb as a unit. As a result, the output of the scale might reflect the influence that one part of the limb has on the extremity as a whole. Although this information might provide some useful insights, it also might have the effect of lowering the responsiveness of the scale if the impact of disease or treatment in one area of the limb is diluted by the status of other parts of the extremity. To take the Disabilities of the Arm, Shoulder, and Hand (DASH) questionnaire[21] as an example, if shoulder pain and movement are improved by an arthroplasty, but the ability to carry out the activities evaluated by the instrument is not changed because of poor hand/wrist or elbow function as might be the case in a condition like rheumatoid arthritis, the impact of the shoulder treatment as measured by the scale might not be as clear as expected. The converse might also occur if disability in the limb is minimal because the main problems are in the shoulder. Even if the shoulder problems are fully remedied, the result of the DASH might not change much. It is also important to recognize that the construct measured by the DASH is that of disability, which is patient reported. It is clear that for a given degree of physically measureable impairment, patient-reported disability can vary widely. This has to be kept in mind when interpreting the meaning of the DASH.

The Michigan Hand Outcomes scale[20] also approaches the construct of disability by exploring a variety of concepts linked to function, including the ability to perform activities of daily living and work, as well as a report of pain and aesthetics. There is also a scale that evaluates satisfaction with hand function.

PAIN

Feinstein states “Unfortunately the idea of accuracy cannot always be applied to … clinical indexes because an unequivocal reference standard does not exist or cannot be obtained. We have no unique reference standard against which to compare ratings … of pain.”[1] The evaluation of pain will always be a self-report by patients, in much the same way that other sensory experiences pertinent to the upper extremity, such as abnormalities of sensation, are described. An understanding of this and the recognition that these self-reports can be influenced by other factors external to the fundamental physiologic phenomenon of pain is essential to interpreting the results of any pain assessment.

Visual analog scales

Visual analog scales (VAS) have gained considerable traction as a method of quantifying pain in both clinical and research settings. A recent review of pain scales used for clinical trials in general medicine and musculoskeletal disease found that more than 60% of the studies reviewed had used a VAS as a pain outcome measure. In 34% of the trials reviewed, the VAS was the only measure of pain.[23] As a status measure, this approach has some merit, at least in reflecting change in the symptom of pain over time in an individual. The VAS has been shown to have good intra-rater reliability,[24] and this is one characteristic that has led to the widespread use of this and similar scales, such as numerical rating systems. However, there are numerous limitations to this approach to the measurement of pain that often go unaddressed by clinical researchers.[25]

The first is the assumption that pain is a linear phenomenon. This suggests that, for example, a score of 8 represents pain that is exactly twice as bad as pain assigned a score of 4. This seems extremely unlikely, and yet this is the assumption that is being made when VAS scores are averaged across patients and used as an aggregate measure of the effect of a therapeutic intervention, for example. The assumption that pain measured with a VAS is linear, continuous, and normally distributed is probably not tenable in most instances, although this is how these data are usually analyzed in clinical research.

The second major deficiency in the VAS is the tacit assumption that all patients use essentially the same scaling when they evaluate their pain. Insofar as they make a comparison to an earlier state, an individual person probably does use a reasonably stable form of scaling, and this is why the VAS has acceptable intra-rater reliability. However, most clinicians clearly recognize that the pain experience is highly variable among patients, and so it should not be expected that they would respond to a VAS in a uniform manner.[26] Even when standard anchors such as “no pain whatsoever” and “the worst pain imaginable” are used to suggest a scaling approach for the patients, the responses should be expected to be variable. Although most patients would probably come close to agreement on what would constitute “no pain whatsoever,” there would likely be enormous variability on the upper end, “the worst pain imaginable.” Clearly this relates to a patient’s past pain experiences, as well as psychosocial factors[25] such as catastrophization and depression. In fact, Litcher-Kelly et al[23] found that, in over half of the studies that used VAS as a primary or sole measure of pain, anchors were not even described.

Multidimensional pain scales

Although the use of multidimensional pain scales places a larger burden on patients because of their greater complexity, they also provide much more information on dimensions of pain beyond the simple factor of intensity. Behavioral and affective components of pain experience might be substantially influenced by therapeutic interventions and yet not be measured by an approach that focuses solely on pain intensity, such as a VAS. It is also true, however, that single-item measures like VAS might be more responsive than multidimensional scales, and for some applications in clinical research, this might trump the more detailed information obtained by multidimensional instruments.[27] Once again, a clear understanding of the goal of the research question might allow a determination of whether a VAS is appropriate of whether a more detailed analysis of pain is required.

An example of a multidimensional pain scale is the McGill Pain Questionnaire.[28] This instrument attempts to incorporate aspects of the pain experience that might have an impact on how this is reported. Simply stated, if it were possible to quantitate pain into units, the impact and reporting by patients of a given unit of pain might vary substantially with the context in which the pain is experienced. This context might be defined by any number of behavioral, affective, or cognitive factors. There might be a large role played by the previous experience of pain, as well. The use of a multidimensional scale might improve the understanding of pain substantially by evaluating the factors that modify the way in which patients report the experience. For a clinical researcher measuring pain as an outcome, this might have critical importance, especially in instances in which the intervention of interest might seem to have improved pain but to have had little or no impact on overall functioning. Alternatively, there also might be settings in which an intervention appears likely to have been successful, and yet patients still report pain similar to before the intervention. In general, the use of multi-dimensional scales such as the McGill Pain Questionnaire is preferable to more basic approaches, such as the use of visual analog scales, simply because of the richer information they provide about this important outcome.

PATIENT SATISFACTION

The highly complex issue of patient satisfaction is one of increasing interest among clinical researchers, as well as for hospital administrators and payers.[29] An important consideration for all measures of satisfaction is identifying the scope of the construct.[30] There is an important distinction between satisfaction as it relates to the outcome of care and as it might relate to the process of care. Satisfaction with the process of care can be directly related to the outcome, partly related to outcome, or completely unrelated to outcome. For example, patients who have been carefully cared for but who ultimately have an undesirable outcome might express satisfaction with the process of care but still be unhappy or unsatisfied with the outcome. This might be frequently observed in cancer studies in which patients receive excellent care but eventually die anyway. Making clear what is assessed by asking questions about satisfaction is critical to understanding the results of such an evaluation; however, this is frequently neglected in studies of musculoskeletal outcomes.

Given the contextual sensitivity of the construct of satisfaction, there is no one scale that fits all needs in clinical research settings. Frequently, what is required is to evaluate components

of care, including the interaction between the patients and the providers, to gain an overall understanding of whether there is satisfaction with the process of care and with the outcome of care, regardless of how it might be evaluated by the clinicians. However, it is clear that attempts to summarize patient satisfaction by a single direct question or by determination of a willingness to have the same procedure again are probably not valid.[31]

In conclusion, the choice of an outcome measurement scale should always be dictated by the needs of the research question. In many instances, the use of a series of scales is appropriate. For conditions that have an impact on overall health, generic health measures are informative as a reflection of general health status. Among these, the SF-36 has been the most widely tested. In most instances, the measurement of overall health status will be of secondary importance in research on the upper extremity. Pain can be measured using simple instruments such as visual analog scales; however, the ability to combine data from these scales across individual patients might be limited. The best attribute of visual analog scales is their intra-rater reliability. Multidimensional pain scales, although more time-consuming for patients, might allow investigators a broader understanding of the pain experienced by patients. Patient satisfaction is a highly complex construct that should be evaluated from within the context of the specific research question. Simply questioning patients as to their satisfaction or willingness to hypothetically have treatment a second time is probably not adequate.

References

1. Feinstein, AR. Clinimetrics. New Haven: Yale University Press; 1987.
2. Streiner, DL., Norman, GR. Health Measurement Scales, A Practical Gude to Their Development and Use. New York: Oxford University Press; 1989.
3. Beaton DE, Boers M, Wells GA. Many faces of the minimal clinically important difference (MCID): a literature review and directions for future research. Curr Opin Rheumatol. 2002; 14(2):109–114. [PubMed: 11845014]
4. Beaton DE, Bombardier C, Katz JN, et al. Looking for important change/differences in studies of responsiveness. OMERACT MCID Working Group. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. J Rheumatol. 2001; 28(2):400– 405. [PubMed: 11246687]
5. Ozyurekoglu T, McCabe SJ, Goldsmith LJ, et al. The minimal clinically important difference of the Carpal Tunnel Syndrome Symptom Severity Scale. J Hand Surg Am. 2006; 31(5):733–738. [PubMed: 16713833]
6. Stewart AL, Hays RD, Ware JE Jr. The MOS short-form general health survey. Reliability and validity in a patient population. Med Care. 1988; 26(7):724–735. [PubMed: 3393032]
7. 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(3):220–233. [PubMed: 8628042]
8. Bowling, A. Measuring Health: a Review of Quality of Life Measurement Scales. Philadelphia: Open University Press; 1991.
9. Guyatt G. Understanding the fundamentals of quality of life measurement. Evid Based Cardiovasc Med. 1998; 2(2):35–36. [PubMed: 16379798]
10. Guyatt G, Feeny D, Patrick D. Issues in quality-of-life measurement in clinical trials. Control Clin Trials. 1991; 12(4 Suppl):81S–90S. [PubMed: 1663862]
11. Moi AL, Wentzel-Larsen T, Salemark L, et al. Impaired generic health status but perception of good quality of life in survivors of burn injury. J Trauma. 2006; 61(4):961–968. [PubMed: 17033569]
12. Khanna D, Tsevat J. Health-related quality of life—an introduction. Am J Manag Care. 2007; 13(Suppl 9):S218–223. [PubMed: 18095785]

13. Netscher DT, Meade RA, Goodman CM, et al. Quality of life and disease-specific functional status following microvascular reconstruction for advanced (T3 and T4) oropharyngeal cancers. Plast Reconstr Surg. 2000; 105(5):1628–1634. [PubMed: 10809090]
14. Mason JH, Anderson JJ, Meenan RF. A model of health status for rheumatoid arthritis. A factor analysis of the Arthritis Impact Measurement Scales. Arthritis Rheum. 1988; 31(6):714–720. [PubMed: 3382446]
15. Pincus T, Summey JA, Soraci SA Jr, et al. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum. 1983; 26(11):1346–1353. [PubMed: 6639693]
16. Bombardier C, Gladman DD, Urowitz MB, et al. Derivation of the SLEDAI. A disease activity index for lupus patients. The Committee on Prognosis Studies in SLE. Arthritis Rheum. 1992; 35(6):630– 640. [PubMed: 1599520]
17. Hawker G, Melfi C, Paul J, Green R, Bombardier C. Comparison of a generic (SF-36) and a disease specific (WOMAC) (Western Ontario and McMaster Universities Osteoarthritis Index) instrument in the measurement of outcomes after knee replacement surgery. J Rheumatol. 1995; 22(6):1193–1196. [PubMed: 7674255]
18. 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(11):1585–1592. [PubMed: 8245050]
19. Zyluk A, Kosowiec L. The results of simple decompression of the ulnar nerve for cubital tunnel syndrome [in Polish]. Chi Narzadow Ruchu Ortop Pol. 2008; 73(4):248–251.
20. Chung KC, Pillsbury MS, Walters MR, et al. Reliability and validity testing of the Michigan Hand Outcomes Questionnaire. J Hand Surg Am. 1998; 23(4):575–587. [PubMed: 9708370]
21. Hudak PL, Amadio PC, Bombardier C. Development of an upper extremity outcome measure: the DASH (disabilities of the arm, shoulder and hand). The Upper Extremity Collaborative Group (UECG). Am J Ind Med. 1996; 29(6):602–608. [PubMed: 8773720]
22. 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(2):128–146. [PubMed: 11382253]
23. Litcher-Kelly L, Martino SA, Broderick JE, et al. A systematic review of measures used to assess chronic musculoskeletal pain in clinical and randomized controlled clinical trials. J Pain. 2007; 8(12):906–913. [PubMed: 17690014]
24. Clark P, Lavielle P, Martinez H. Learning from pain scales: patient perspective. J Rheumatol. 2003; 30(7):1584–1588. [PubMed: 12858463]
25. de Williams AC, Davies HT, Chadury Y. Simple pain rating scales hide complex idiosyncratic meanings. Pain. 2000; 85(3):457–463. [PubMed: 10781919]
26. Gagliese L, Katz J. Age differences in postoperative pain are scale dependent: a comparison of measures of pain intensity and quality in younger and older surgical patients. Pain. 2003; 103(1–2):11–20. [PubMed: 12749954]
27. Bellamy N, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in osteoarthritis patients. Curr Med Res Opin. 1999; 15(2):113–119. [PubMed: 10494494]
28. Melzack R. The McGill Pain Questionnaire: major properties and scoring methods. Pain. 1975; 1(3):277–299. [PubMed: 1235985]
29. Urden LD. Patient satisfaction measurement: current issues and implications. Lippincott Case Manag. 2002; 7(5):194–200.
30. Carr-Hill RA. The measurement of patient satisfaction. J Public Health Med. 1992; 14(3):236–249. [PubMed: 1419201]
31. Haverkamp D, Sierevelt IN, van den Bekerom MP, et al. The validity of patient satisfaction as single question in outcome measurement of total hip arthroplasty. J Long Term Eff Med Implants. 2008; 18(2):145–150. [PubMed: 19968623]