REPORT

Science of assessment

N Bellamy

Ann Rheum Dis 2005;**64**(Suppl II):ii42–ii45. doi: 10.1136/ard.2004.031567

Psoriatic arthritis is a multisystem disorder which, from a measurement standpoint, demands consideration of its cutaneous manifestations and both axial and peripheral musculoskeletal involvement. Measurements of various aspects of impairment, ability/disability, and participation/handicap are feasible using existing measurement techniques, which are for the most part valid, reliable, and responsive. Nevertheless, there remain opportunities for the further development of consensus around core set measures and responder criteria, as well as for instrument development and refinement, standardised assessor training, cross-cultural adaptation of health status questionnaires, electronic data capture, and the introduction of standardised quantitative measurement into routine clinical care.

The psoriatic diathesis presents a significant challenge in clinical metrology, given the propensity for varying patterns of involvement, encompassing different tissues and anatomical regions. Musculoskeletal involvement can affect the peripheral and axial skeleton and can result in not only arthritis or spondylitis but also enthesitis or dactylitis. Extra-articular involvement can manifest as psoriatic skin lesions, nychodystrophy, or ocular inflammation (conjunctivitis, iritis). Occasionally psoriatic arthritis may be complicated by oral ulceration, urethritis, or aortic valve disease.[1] The challenge of developing measurement tools for one aspect of the disorder is relatively small compared with that of developing a multidimensional index, which becomes particularly severe when attempting to weight and aggregate information from different measures into a single composite score.[2] Nevertheless, there is precedent for the successful development of all three types of measurement tool,[2 3] and the definition of basic principles that underpin instrument development and selection for clinical research and clinical practice applications.[4]

The measurement of patient outcomes, that is, the consequence of disease and health management decisions, is an essential component of health care. For the policy maker and epidemiologist, it provides opportunity for estimating the burden of disease. In clinical research, standardised outcome measurement procedures facilitate protocol harmonisation and facilitate benchmarking activities and systematic reviews. In health economic analyses, measures of patient health status are necessary for the conduct of cost effectiveness and cost utility analyses. For the practitioner, clinical measurement can become an integral part of decision making, shared goal setting, and monitoring health status, goal attainment, and response to treatment. Structured health status assessment can also play a key role in case management and adjudication in compensation/litigation environments. Clinical assessments, therefore, are an integral part of healthcare delivery systems.

CONCEPTUAL FRAMEWORK

In chronic diseases such as psoriatic arthritis, an as yet undefined aetiological event triggers a pathological response, resulting in a number of clinical manifestations and outcomes. Two types of measures may be used to assess events in that sequence: process measures and outcome measures. Process measures include histological analyses, the measurement of biological markers, and various imaging procedures. It is important to note that scores derived from process measures are often poorly correlated with outcome measurement scores and generally have little or no value as surrogate measures for those clinical outcomes. For this reason direct measurement of clinical outcomes, using tools that are valid, reliable, and responsive is a key requirement in clinical practice and clinical research environments. It is convenient, for operational reasons, to place measurement procedures within a framework, such as the paradigm proposed by Fries *et al*, the World Health Organization's International Classification of Impairments, Activities and Participation (ICIDH-2), or the International Classification of Function (ICF).[5–7] For the most part, the outcomes encompassed by these frameworks are all relevant to patients because they are discernible at the individual patient level and represent the consequence of disease and the ultimate outcome of the disease process and its clinical manifestations.

SCIENTIFIC PRINCIPLES

For evaluative applications, outcome measurement procedures should meet each of four major criteria: ethics, validity, reliability, and responsiveness. The first three are important in all measurement procedures, but responsiveness (sensitivity to change) is the quintessential requirement of a measurement procedure for use in evaluating change following effective treatment.

Ethics

The measurement process must be ethical. Processes that are potentially hazardous to patients require disclosure. Where possible, less invasive procedures should be employed. Furthermore, the importance and necessity of acquiring new information should be weighed against any attendant risks.

Validity

Validity is a measure of the extent to which an instrument specifically measures the phenomenon of interest.[8] More specifically, validity is concerned with sources of non-random error. Such systemic error or bias may prevent an instrument from truly measuring what is intended, which results in inaccuracy. There are four types of validity: face, content, construct, and criterion.

Abbreviations: HRQOL, health related quality of life; HSQ, health status questionnaire

Face validity
A measure has face validity if relevant experts judge that it measures at least part of the defined phenomenon. In many instances this is self-evident, whereas in others, particularly in measures of functional status, it may not be entirely obvious whether the measurement reflects physical, social, or emotional function, or some combination.

Content validity
An instrument can have face validity but still fail to capture, in its entirety, the dimension of interest. Content validity, therefore, is a measure of comprehensiveness—the extent to which the measure encompasses all relevant aspects of the defined attribute. Content validity is generally determined by group consensus (that is, nominal or Delphi techniques) and can be decided either by patients, who rate the importance of their symptoms, or by clinical assessors, whose decision is based on their perception of the patient's symptoms. Decisions regarding which combination of items should be included in an instrument are critical because they define the nature of the instrument and its future applicability.

Construct validity
Construct validity is of two types: convergent and discriminant. Both are tested by demonstrating relations between measurement scores and a theoretical manifestation (that is, construct or consequences of an attribute) of the disease. Convergent construct validity testing is assessed by the statistical correlation between scores on a single health component, as measured by two different instruments. If the correlation coefficient is positive and appreciably above zero, the new measure is said to have convergent construct validity. In contrast, discriminant construct validity testing compares correlation coefficients between scores on the same health component, as measured by two different instruments (such as separate measures of physical function), and between scores on that health component and each of several other health components (such as measures of social and emotional function). A measure has discriminant construct validity if the proposed measure correlates better with a second measure, accepted as more closely related to the construct, than it does with a third, more distantly related measure.

Criterion validity
Criterion validity is assessed by statistically testing a new measurement technique against an independent criterion or standard (concurrent validity) or against a future standard (predictive validity). Criterion validity is an estimate of the extent to which a measure agrees with a ''gold standard'' (an external criterion of the phenomenon being measured). The major problem in criterion validity testing is the general paucity of gold standards. Indeed, even some purported gold standards may not provide completely accurate estimates of the true value of a phenomenon.

Reliability
Repeatability, consistency, and reproducibility are synonyms for reliability.[8] Reliability is the extent to which a measurement procedure yields the same result on repeated determinations. This determination may be the result of either different measurements performed at the same time (internal consistency) or the same measurements performed at different times (stability). Repeated measurements are rarely exactly the same, since there is almost always some random error (noise) or degree of inconsistency. Defined sources of measurement error include the subject, the assessor, and the measuring instrument. There are various methods of calculating reliability, each method reflecting a different aspect of instrument performance such that different coefficients are derived using different methods.

Responsiveness (sensitivity to change)
In order to detect change, a measurement technique needs to be targeted on aspects of the disease amenable to change, using format and scaling methods that allow detection of change, and it needs to be applied at a point in time when change might have occurred. It is important to note that an assessment technique may fail to record clinical improvement for a number of reasons (for example, patient lacks response potential, malcompliance with the treatment programme, inefficacious treatment, insensitivity of the outcome measure).

OPERATIONAL CONSIDERATIONS
The efficient and effective conduct of outcome measurement procedures requires skilled personnel and standardised techniques, employing measures that are valid, reliable, and responsive.

Administration
Before assessors conduct observer dependent measurements, such as counting the number of tender and swollen joints or quantifying the magnitude of axial skeletal movement, they should be trained to acceptable levels of reliability. Training can be provided individually or in groups and may involve the use of procedure manuals, experienced trainers, or the use of standardised audiovisual aids. Where patients self complete measurements, adequate instructions (and, where necessary, supervision), should be provided, particularly to verify the completeness of data collection. It is important in the case of health status questionnaires (HSQ) to obtain authentic versions, usually by directly contacting the originator, and to review the most recent version of any available user guide.

Cross-cultural adaptation
The issue of cross-cultural adaptation is particularly relevant to the conduct of multicentre studies and the use of HSQs outside their country of origination. Many HSQs have originated in Europe and North America but subsequently been applied on a global basis. Although many countries are multicultural, cross-cultural adaptation of HSQs should be conducted with a standard protocol meeting the requirements of tandem forward translation, followed by tandem backward translation using experienced, fluently bilingual translators. The development should also involve on-site linguistic validation in the relevant culture in a group of representative individuals, followed, where necessary, by further refinement of the questionnaire.

Scaling options for HSQs
Pain and disability are two of the commonest consequences of most musculoskeletal conditions. The patient's pain and physical function are often quantified through patient self-report using a disease specific HSQ or a generic health related quality of life (HRQOL) questionnaire. Experience in patients with rheumatoid arthritis and osteoarthritis suggests that 100 mm visual analogue pain scales tend to be slightly more responsive than five point Likert scales and that 11 point numerical rating scales are intermediate in their responsiveness.[9 10] Furthermore, users place priority not only on validity, reliability, and responsiveness but also on brevity, speed of completion, and ease of scoring.[11 12] As a consequence, it is important to weigh up the trade-offs between different types of scaling format. Some HSQs have been developed in multiple formats, but probably the majority exist in only a single format.

Flexible data capture

Traditionally, HSQs have been administered either in paper format for self completion, or by interviewers in face to face settings. More recent innovations have included computer assisted telephone interviews (CATI) or self completion using a mouse driven cursor or touch screen, so called electronic data capture. The availability of HSQs in different administration formats provides considerable flexibility in data capture.

PATIENT GLOBAL ASSESSMENT

An alternative, or additional, approach to conducting symptom based measurement on a dimension by dimension basis, is to incorporate a patient global assessment question into the measurement battery. It is extremely important to specify for the patient, in the wording of the global question, which aspects of the condition are to be considered (such as symptom severity, disease activity, anatomical area, or overall health). The patient global assessment question can be phrased to assess current status or change in status and be focused on a particular anatomical area, the condition in general, or the patient as a whole. The timeframe over which the patient should consider his or her status should be defined. At the present time, there is no international consensus on the exact working of the global assessment question or the preferred scaling format.

PHYSICIAN GLOBAL ASSESSMENT

Whether physician global assessment adds significant information to the measurement process, over and above the patient global assessment, is debatable. The physician (or other assessor) can consider additional aspects of the condition which are not assessable by the patient (such as radiographic change) and may have insight into whether the patient tends to amplify or minimise reported symptoms. Physicians require clear specification as to which aspects of the condition should be considered when making their global assessment. The timeframe for the physician global assessment usually should be specified as ''today'' since the assessor generally has no knowledge of the patient's interval status, other than that described by the patient and captured by the patient global assessment. There is no international consensus on the exact working of the physician global assessment question or the preferred scaling format.

In selecting outcome measures for clinical research and clinical practice applications, each of the aforementioned issues should be given due consideration and the most appropriate measurement battery constructed. The necessity for including measures of disease activity, symptom severity, limitation in range of movement, and multisystem involvement, as well as including disease specific and generic HRQOL measures should be considered in the context of the prevailing research question(s) or clinical management scenario.

INTERNATIONAL STANDARDISATION

International consensus in outcome measurement has generally required multinational, multi-stakeholder collaboration and has been established through a combination of statistical and judgmental processes.

Core sets

The existence of valid, reliable, and responsive measures is a prerequisite for the establishment of core sets, since some measures having these attributes will be selected for inclusion in the core set. International consensus has been established on core set outcome measures for several musculoskeletal diseases.[13–16] Once agreement has been reached on the core domains, appropriate measures can be selected for inclusion in the measurement battery. If the application will be multinational, the availability of the necessary alternative language forms should be established or plans put in place for their development and validation.

Responder criteria

Responder criteria are threshold values used to assign individual patients as responders or non-responders to treatment. They can be based on one or more variables, and can be defined by relative (percentage) or absolute (normalised units) change or a combination of both relative and absolute change.[17–21] Implicit in the establishment of responder criteria is the exact specification of the magnitude of change that is considered important. There are separate definitions of several types of detectible difference, each associated with acronyms such as, MCPD (minimum change potentially detectable), MPCPD (minimum percentage change potentially detectable), MPCI (minimum perceptible clinical improvement), and MCID (minimum clinically important difference). Beaton *et al* have recently proposed a tridimensional framework for categorising detectable differences.[22]

OUTCOME MEASUREMENT IN PSORIATIC ARTHRITIS

The measurement challenge in psoriatic arthritis is in part related to requirements to evaluate the axial as well as the peripheral skeleton. In addition, studies assessing any positive or negative effects on cutaneous lesions need to incorporate measures of psoriatic skin involvement. As a consequence, outcome measurement in studies of psoriasis and psoriatic arthritis often employs a battery of outcome measures, rather than just a single measure.

Ujfalussy and Koo recently assessed the validity of tender and swollen joint counts, disease activity score 4 (DAS 4), DAS 3, and DAS 28 in patients with psoriatic arthritis and compared response classification using the European League Against Rheumatism (EULAR) and Clegg criteria. They concluded that these measures are valid and useful.[23] The Clegg criteria are of particular interest since they were conceptualised for use in studies of psoriatic arthritis.[24] The criteria are based on change assessed on four measures: patient self-assessment, physician global assessment, joint pain/tenderness score, and joint swelling score. Treatment response in the Clegg criteria is defined as improvement in at least two of the four measures, one of which must be joint pain/tenderness or swelling, with no worsening on any of the four measures.[24 25] Gladman and colleagues have established the reliability of counts of actively inflamed and damaged joints.[26]

Experience with HSQs in patients with psoriatic arthritis has generally been positive. The validity and responsiveness of both arthritis specific[27–31] and generic HRQOL[32] instruments have been evaluated in patients with psoriatic arthritis, the validity data in particular showing complex relationships with disease activity and severity. For example, the Health Assessment Questionnaire (HAQ) and HAQ-S scores correlate with clinical measures of pain and function but not with disease severity,[28] and the Arthritis Impact Measurement Scales 2 (AIMS2) pain, function, and work scores, correlate with measures of function and disease activity but not with disease severity.[29] Furthermore, AIMS physical function scores are correlated with measures of clinical function, disease activity, and disease severity, and AIMS pain scores are correlated with measures of clinical function and disease activity but not disease severity.[27]

Outcome measurement in psoriatic arthritis continues to evolve. There are opportunities for the further development of consensus around core set measures and responder criteria,

as well as opportunities for instrument development and refinement, standardised assessor training, cross-cultural adaptation of HSQs, electronic data capture, and the introduction of standardised quantitative measurement into routine clinical care.

Correspondence to: Dr N Bellamy, CONROD, University of Queenland, Level 3, Mayne Medical School, Herston Road, Brisbane, Queensland, 4006, Australia; nbellamy@medicine.uq.edu.au

REFERENCES

1 **Bruce IN**. Psoriatic arthritis: clinical features. In: Hochberg MC, Silman AJ, Smolen JS, Weinblatt ME, Weisman MH, eds. *Practical Rheumatology*, 3rd edn, vol 2. Edinburgh: Mosby, 2003:1241–52.
2 **Bellamy N**. *Musculoskeletal Clinical Metrology*. Lancaster: Kluwer Academic Publishers, 1993:135–46.
3 **Bellamy N**, Buchanan WW. Clinical outcome measurement. In: Dieppe P, Schumacher HR Jr, Wollheim FA, eds. *Classic Papers in Rheumatology*, 1st edn. London: Martin Dunitz Ltd, 2002:2–3.
4 **Bellamy N**. Principles of outcome assessment. In: Hochberg MC, Silman AJ, Smolen JS, Weinblatt ME, Weisman MH, eds. *Practical Rheumatology*, 3rd edn, vol 1. Edinburgh: Mosby, 2003:21–30.
5 **Fries JF**, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;**23**:137–45.
6 **World Health Organization**. *ICIDH-2, An international classification of impairments, activities and participation*. Geneva: WHO, 1997.
7 **World Health Organization**. *ICF. International Classification of Functioning, Disability and Health*. Geneva: WHO, 2001.
8 **Carmines EG**, Zeller RA. *Reliability and validity assessment*. Beverly Hills: Sage Publications, 1979:5–71.
9 **Bellamy N**, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in rheumatoid arthritis patients. *Curr Med Res Opin* 1999;**15**:121–7.
10 **Bellamy N**, Campbell J, Syrotuik J. Comparative study of self-rating pain scales in osteoarthritis patients. *Curr Med Res Opin* 1999;**15**:113–19.
11 **Bellamy N**, Muirden KD, Brooks PM, Barraclough D, Tellus MM, Campbell J. A survey of outcome measurement procedures in routine rheumatology outpatient practice in Australia. *J Rheumatol* 1999;**26**:1593–9.
12 **Bellamy N**, Kaloni S, Pope J, Coulter K, Campbell J. Quantitative rheumatology: a survey of outcome measurement procedures in routine rheumatology outpatient practice in Canada. *J Rheumatol* 1998;**25**:852–8.
13 **Altman R**, Brandt K, Hochberg M, Moskowitz R, Bellamy N, Bloch DA, *et al*. Design and conduct of clinical trials in patients with osteoarthritis: recommendations from a task force of the Osteoarthritis Research Society. Results from a workshop. *Osteoarthritis Cartilage* 1996;**4**:217–43.
14 **Bellamy N**, Kirwan J, Boers M, Brooks P, Strand V, Tugwell P, *et al*. Recommendations for a core set of outcome measures for future phase III clinical trials in knee, hip, and hand osteoarthritis. Consensus development at OMERACT III. *J Rheumatol* 1997;**24**:799–802.
15 **Felson DT**, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, *et al*. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. *Arthritis Rheum* 1993;**36**:729–40.
16 **van der Heijde D**, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. Assessments in Ankylosing Spondylitis Working Group. *J Rheumatol* 1997;**24**:2225–9.
17 **Felson DT**, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, *et al*. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;**38**:727–35.
18 **Felson DT**, Anderson JJ, Lange ML, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? *Arthritis Rheum* 1998;**41**:1564–70.
19 **Dougados M**, Leclaire P, van der Heijde D, Bloch DA, Bellamy N, Altman RD. Response criteria for clinical trials on osteoarthritis of the knee and hip: a report of the Osteoarthritis Research Society International Standing Committee for Clinical Trials response criteria initiative. *Osteoarthritis Cartilage* 2000;**8**:395–403.
20 **Pham T**, Van Der Heijde D, Lassere M, Altman RD, Anderson JJ, Bellamy N, *et al*. OMERACT-OARSI. Outcome variables for osteoarthritis clinical trials: The OMERACT-OARSI set of responder criteria, *J Rheumatol* 2003;**30**:1648–54.
21 **van Tubergen A**, van der Heijde D, Anderson J, Landewe R, Dougados M, Braun J, *et al*. Comparison of statistically derived ASAS improvement criteria for ankylosing spondylitis with clinically relevant improvement according to an expert panel. *Ann Rheum Dis* 2003;**62**:215–21.
22 **Beaton DE**, Bombardier C, Katz JN, Wright JG, Wells G, Boers M, *et al*. OMERACT MCID Working Group. Looking for important change/differences in studies of responsiveness. Outcome Measures in Rheumatology. Minimal Clinically Important Difference. *J Rheumatol* 2001;**28**:400–5.
23 **Ujfalussy I**, Koo E. Measurement of disease activity in psoriatic arthritis. Extended report. *Z Rheumatol* 2003;**62**:60–5.
24 **Clegg DO**, Reda DJ, Mejias E, Cannon GW, Weisman MH, Taylor T, *et al*. Comparison of sulfasalazine and placebo in the treatment of psoriatic arthritis. A Department of Veterans Affairs Cooperative Study. *Arthritis Rheum* 1996;**39**:2013–20.
25 **Clegg DO**, Reda DJ, Abdellatif M. Comparison of sulfasalazine and placebo for the treatment of axial and peripheral articular manifestations of the seronegative spondylarthropathies: a Department of Veterans Affairs cooperative study. *Arthritis Rheum* 1999;**42**:2325–9.
26 **Gladman DD**, Farewell V, Buskila D, Goodman R, Hamilton L, Langevitz P. Reliability of measurements of active and damaged joints in psoriatic arthritis. *J Rheumatol* 1990;**17**:62–4.
27 **Duffy CM**, Watanabe Duffy KN, Gladman DD, Brubacher BB, Buskila D. The utility of the arthritis impact measurement scales for patients with psoriatic arthritis. *J Rheumatol* 1992;**19**:1727–32.
28 **Blackmore MG**, Gladman DD, Husted J, Long JA, Farewell VT. Measuring health status in psoriatic arthritis: the Health Assessment Questionnaire and its modification. *J Rheumatol* 1995;**22**:886–93.
29 **Husted J**, Gladman DD, Farewell VT, Long JA. Validation of the revised and expanded version of the Arthritis Impact Measurement Scales for patients with psoriatic arthritis. *J Rheumatol* 1996;**23**:1015–19.
30 **Husted J**, Gladman DD, Long JA, Farewell VT. Relationship of the Arthritis Impact Measurement Scales to changes in articular status and functional performance in patients with psoriatic arthritis. *J Rheumatol* 1996;**23**:1932–7.
31 **Taccari E**, Spadaro A, Rinaldi T, Riccieri V, Sensi F. Comparison of the Health Assessment Questionnaire and Arthritis Impact Measurement Scale in patients with psoriatic arthritis. *Rev Rhum Engl Ed* 1998;**65**:751–8.
32 **Husted JA**, Gladman DD, Cook RJ, Farewell VT. Responsiveness of health status instruments to changes in articular status and perceived health in patients with psoriatic arthritis. *J Rheumatol* 1998;**25**:2146–55.