Adult measures of general health and health-related quality of life

AUTHOR(S)

Ljoudmila Busija, E Pausenberger, T Haines, S Haymes, R Buchbinder, Richard Osborne

PUBLICATION DATE

01-11-2011

HANDLE

10536/DRO/DU:30047108

Arthritis Care & Research
Vol. 63, No. S11, November 2011, pp S383–S412
DOI 10.1002/acr.20541
© 2011, American College of Rheumatology

Adult Measures of General Health and Health-Related Quality of Life

Medical Outcomes Study Short Form 36-Item (SF-36) and Short Form 12-Item (SF-12) Health Surveys, Nottingham Health Profile (NHP), Sickness Impact Profile (SIP), Medical Outcomes Study Short Form 6D (SF-6D), Health Utilities Index Mark 3 (HUI3), Quality of Well-Being Scale (QWB), and Assessment of Quality of Life (AQoL)

LUCY BUSIJA,[1] EVA PAUSENBERGER,[2] TERRY P. HAINES,[2] SHARON HAYMES,[1]
RACHELLE BUCHBINDER,[3] AND RICHARD H. OSBORNE[4]

INTRODUCTION

The aim of this review is to provide a summary of adult measures of general health and health-related quality of life (HRQOL) commonly used in rheumatology research studies. Currently, there is no single generally agreed upon definition or conceptual model of health or HRQOL, and developing a comprehensive definition of these complex concepts was beyond the scope of the review. For the purposes of this review, we define measures of general health and HRQOL as multi-item questionnaires that assess perceived health status and overall physical and emotional well-being that is not specific to any disease. The health measures included in this review were further subdivided into generic health profiles (questionnaires that provide assessment of more than 1 dimension of health status) and health utility measures that provide an overall measure of health status rated between perfect health (1.0) and death (0.0).

Relevant measures were identified (using medical subject headings [MeSH]) through a systematic search of medical publications indexed to PubMed database. The following search queries were used: [Quality of life (title) AND Outcomes assessment (MeSH terms) AND Rheumatic diseases (MeSH terms)] and [Quality of life (abstract) AND Outcomes assessment (MeSH terms) AND Rheumatic diseases (MeSH terms)]. In MeSH, “rheumatic diseases” are defined as “disorders of connective tissue, especially the joints and related structures, characterized by inflammation, degeneration, or metabolic derangement,” and include rheumatoid arthritis, Caplan’s syndrome, Sjögren’s syndrome, Still’s disease, fibromyalgia, gout, hyperostosis, osteoarthritis, and polymyalgia rheumatica among others.

Inclusion criteria were 1) the study was concerned with a rheumatology condition and 2) participants were human adults. The first query returned 129 items and the second query returned 494 items, with 623 abstracts in total. After removal of 77 duplicates, 38 pediatric studies, and 2 animal studies, 2 reviewers (LB, EP) screened abstracts independently of the remaining 506 publications to identify relevant multi-item questionnaires (i.e., those generic questionnaires that were identified by the study authors as being used for the purpose of assessing general health or HRQOL). Where abstracts contained insufficient information to determine the type of measures used, full publications were obtained.

The reviewers, working independently, identified 10 generic health utility measures and 5 generic health profiles (Table 1). Agreement about the type and number of occurrences of relevant measures in the sample of reviewed

Drs. Busija and Haymes’ work was supported by internal funding from the Centre for Eye Research Australia. Ms Pausenberger’s work was supported by internal funding from the Southern Health and Monash University Allied Health Research Unit. Dr. Haines’s work was supported by a National Health and Medical Research Council Career Development Award. Dr. Buchbinder’s work was supported in part by a National Health and Medical Research Council Practitioner Fellowship. Dr. Osborne’s work was supported in part by a National Health and Medical Research Council Population Health Career Development Award.

[1]Lucy Busija, PhD, Sharon Haymes, PhD: University of Melbourne, Melbourne, Victoria, Australia; [2]Eva Pausenberger, BPhysiother (Hons), Terry P. Haines, PhD: Southern Health, Cheltenham, Victoria, Australia, and Monash University, Melbourne, Victoria, Australia; [3]Rachelle Buchbinder, MBBS (Hons), FRACP, PhD: Cabrini Institute and Monash University, Melbourne, Victoria, Australia; [4]Richard H. Osborne, PhD: Deakin University, Melbourne, Victoria, Australia.

Dr. Buchbinder has received speaking fees (less than $10,000 each) from Pfizer, Wyeth, Abbott, and Roche, and (more than $10,000) from Abbott for participating as an international mentor for 3e Initiative.

Address correspondence to Lucy Busija, PhD, Center for Eye Research Australia, Royal Victorian Eye and Ear Hospital, 32 Gisborne Street, East Melbourne, VIC 3002, Australia. E-mail: lbusija@unimelb.edu.au.

Submitted for publication February 22, 2011; accepted in revised form June 20, 2011.

Table 1. Measures used for the assessment of general health and health-related quality of life in rheumatology literature

Questionnaire	Occurrences in reviewed abstracts, no. (n = 506)
Generic health profiles	
Short Form 36*	146
Nottingham Health Profile*	21
Short Form 12*	13
Sickness Impact Profile*	7
Duke Health Profile	1
Generic health utility measures	
EuroQol 5-domain†	32
Short Form 6D*	6
Health Utilities Index 3*	6
Quality of Wellbeing Scale (self-administered)*	5
Assessment of Quality of Life Scale*	4
15D	3
Quality of Life Scale	2
World Health Organization Quality of Life/Bref	2
Perceived Quality of Life Scale	1
Profile of Quality of Life in the Chronically Ill	1

* Review included in this article.
† Review of this measure is included in Measures of Disability article in this issue.

abstracts was very high (intraclass correlation coefficient 0.996). Any disagreements were resolved through a discussion between all authors. Given the large number of potentially relevant measures identified, only those measures that were used at least 4 times in the screened abstracts were selected for this review. Consequently, this report provides reviews of 4 generic health profiles: the Medical Outcomes Study Short Form 36 and Short Form 12 Health Surveys, the Nottingham Health Profile, and the Sickness Impact Profile, and 4 health utility measures: the Medical Outcomes Study Short Form 6D, the Health Utilities Index Mark 3, Quality of Well-Being Scale (self-administered), and the Assessment of Quality of Life Scale. Although the EuroQol 5-domain is also frequently used in rheumatology, the review of this measure is included in Measures of Disability article in this issue.

MEDICAL OUTCOMES STUDY SHORT FORM 36-ITEM (SF-36) AND SHORT FORM 12-ITEM (SF-12) HEALTH SURVEYS

Description

Purpose. The SF-36 and SF-12 are multi-item generic health surveys intended to measure "general health concepts not specific to any age, disease, or treatment group" (1). The SF-12 is a shorter version of the SF-36 and uses only 12 questions to measure functional health and wellbeing from the patient's perspective. The original objective was to develop a short, generic health-status measure that reproduces the 2 summary scores of the SF-36, i.e., the physical component summary (PCS) score and the mental component summary (MCS) score (2).

The SF-36 and SF-12 are suitable for use in general, as well as in clinical populations and, as such, can be used to compare health between populations and between diseases. The SF-36 and the SF-12 health surveys are available in original and revised versions. The SF-36 and SF-12 were first published in 1992 and 1996, respectively, with the revised versions of both questionnaires published in 2000. The revised versions are very similar to their original forms, with major differences involving changes in item wording, revision of the response scale to incorporate a greater number of response options, and norm-based scoring (3).

Content. Both the SF-36 and SF-12 measure 8 health domains: physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional, and mental health. Physical functioning covers limitations in daily life due to health problems. The role physical scale measures role limitations due to physical health problems. The bodily pain scale assesses pain frequency and pain interference with usual roles. The general health scale measures individual perceptions of general health. The vitality scale assesses energy levels and fatigue. The social functioning scale measures the extent to which ill health interferes with social activities. The role emotional scale assesses role limitations due to emotional problems, and the mental health scale measures psychological distress.

The SF-36 and SF-12 can also be used to derive 2 aggregate summary measures: the PCS and the MCS. Summary scores are calculated by summing factor-weighted scores across all 8 subscales, with factor weights derived from a US-based general population sample (4). Country-specific weights are also available for Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, the UK (5), and Australia (6). In the calculation of the PCS summary score, highest weights are given to the physical functioning, role physical, bodily pain, and general health scales, whereas for the MCS summary score, higher weights are given to the vitality, social functioning, role emotional, and mental health scales.

Number of items. The SF-36 consists of 36 items, 35 of which are used in the calculation of 8 separate scale scores. The physical functioning scale (10 items) is the longest scale. The general health and mental health scales have 5 items each, and the vitality and role physical scales have 4 items each. The role emotional scale has 3 items, and the bodily pain and social functioning scales have 2 items each. The remaining item of the SF-36 is a health transition question that asks about a change in general health over the past 12 months.

The SF-12 consists of 12 items: 2 items on physical functioning, 2 items on role physical, 1 item on bodily pain, 1 item on general health, 1 item on vitality, 1 item on social functioning, 2 items on role emotional, and 2 items on mental health. Since more items permit better representation of each domain, the domains are best represented by the SF-36. The most useful measures derived from the SF-12 are the 2 aggregate summary measures: the PCS and MCS.

Response options/scale. The response scales for the SF-36 and SF-12 items vary across and within the scales, with the number of response options ranging from 3 (physical functioning) to 6 (vitality and mental health). The health transition item is scored on a 5-point scale where 1 indicates much better than a year ago, and 5 indicates much worse than a year ago.

Recall period for items. The SF-36 and SF-12 are available in 2 forms: a standard form, which uses a 4-week recall period, and an acute form, which uses a 1-week recall. The standard 4-week recall form is appropriate when the instrument will be administered only once to the respondent, or when at least 4 weeks will pass between re-administration of the instrument. The acute 1-week recall form is appropriate when more frequent administration is required and changes are likely to occur rapidly.

Examples of use. The SF-36 is ubiquitous in rheumatology and has been used to capture health-related outcomes in a variety of rheumatic conditions, including knee osteoarthritis (7), Sjögren's syndrome (8), fibromyalgia (9), rheumatoid arthritis (10,11), ankylosing spondylitis (12), and gout (13). The SF-36 has been used to assess efficacy of a broad range of interventions in rheumatology, including orthopedic surgery (14–16), drug treatment (8,17), acupuncture (18), physiotherapy (19), electromagnetic field therapy (20), Tai Chi (21), and self-management education (22).

The SF-12 has been used in population-based studies to assess the impact of musculoskeletal diseases on general health (23,24). In addition, it has been used as an outcome measure to evaluate the efficacy of a broad range of interventions for rheumatic conditions, including pharmacologic treatment (25,26); hydrotherapy treatment for osteoarthritis (27) and fibromyalgia (28); Tai Chi (29); surgical procedures (e.g., total hip arthroplasty) (30), fore foot arthroplasty (31); total knee arthroplasty (32); and medication adherence programs (33).

Practical Application

How to obtain. The original version of the SF-36 (Research and Development [RAND] 36-Item Health Survey 1.0 Questionnaire) can be obtained free of charge from the RAND Corporation (http://www.rand.org/health/surveys_tools/mos/mos_core_36item_survey.html). English and Arabic language versions are available. The revised SF-36 and the SF-12 can be obtained from QualityMetric (http://www.qualitymetric.com/). Annual license fee applies. License fees are available on application and depend on whether the survey is used in a commercial or nonprofit setting. Manuals can also be purchased.

Method of administration. The SF-36 and SF-12 can be self-administered or interviewer-administered. Multiple modes are available, such as static (paper), online, e-form, personal digital assistant, tablet, and interactive voice response (IVR) via telephone. Several studies reported a consistent bias for lower SF-36 and SF-12 scores (indicating worse health) when self-completed as compared with interviewer administration (34–38). For SF-36, data quality also tends to be better for interviewer administration with a lower proportion of missing data, lower ceiling effects, and better internal consistency estimates (35,39). Data collection costs, on the other hand, are lower (up to 77%) for self-administration (35,39). IVR and live telephone methods for administering the SF-12 have been compared in a study of back pain patients, with similar results obtained for PCS scores but not MCS scores (mean MCS 44.22 and 48.50 for IVR and live telephone methods, respectively; $P < 0.01$), and the greatest discrepancy occurring for the item about feeling "downhearted and blue" (40).

The SF-36 can also be administered by proxy, but concordance between self and proxy ratings varies across proxy types. Generally, professional proxies (e.g., occupational therapists, nurses) provide a more accurate description of an individual's health state compared with lay proxies, who tend to overestimate the level of impairment (41,42).

Scoring. The SF-36 and SF-12 contain a mixture of positively- (higher scores indicate better health) and negatively-worded response scales, so some items need to be recoded prior to scoring. The scale scores are calculated by summing responses across scale items and then transforming these raw scores to a 0–100 scale. Computerized scoring algorithms are available and can be used to produce norm-based T scores for each scale (with a mean of 50 and SD of 10) as well as the PCS and MCS summary scores (4). If using the IVR mode, data can be loaded directly into the QualityMetric database for scoring, interpretation, and reporting in real time.

In computing scale scores for the SF-36 and SF-12, missing values have traditionally been calculated only for those respondents who provided data on at least half the scale items (4). More recently, pattern matching and regression methods of missing data imputation for these questionnaires have been developed (43). These new algorithms can be implemented using QualityMetric's purpose-developed software.

Score interpretation. Scores on the SF-36 and SF-12 scales range from 0–100, with higher scores indicating better health. On the physical functioning scale, low scores are typical of someone who experiences many limitations in physical activities, including bathing or dressing, while high scores represent someone who is able to perform these types of activities without limitations. Low scores on the role physical scale represent someone who experiences many limitations in work or other daily activities, and high scores characterize someone who has no difficulties with these activities. Low scores on the social functioning typify a person who experiences a great deal of difficulties in normal social activities due to physical and emotional health problems, and high scores represent someone who is able to perform normal social activities without interference due to physical or emotional health. Low scores on the bodily pain scale are typical of a person who has very severe and extremely limiting pain, and high scores represent individuals who have no pain or pain-related limitations. On the mental health scale, low scores represent high levels of nervousness and depression, while high scores characterize someone who feels peaceful, happy, and calm. Low scores on the role emotional scale represent someone who experiences many problems with work or other daily activities as a result of emotional ill health, and high scores represent those who have no problems with work or other daily activities as a result of emotional health. On the vitality scale, low scores are typical of someone who feels tired and worn out all of the time, while high scores characterize those who feel full of pep and energy. Low scores on the general health scale represent a person who believes their health to be poor and likely to get worse, and high scores represent someone who sees their health as excellent (1).

Age- and sex-based norms for the SF-36 are available for several countries, including the US (4,44), the UK (34,45), Australia (6,46,47), Sweden (48), China (49), and New Zealand (50). Normative data for MCS and PCS summary scores are also available for Denmark, France, Germany, Italy, the Netherlands, Norway, and Spain (5). Notable cross-country differences in normative SF-36 scores have been reported (6), which may reflect cultural differences in health perceptions. Contextual factors, such as survey methodology, mode of administration, and item order have also been reported to affect normative scores on the SF-36 (34). Age- and sex-based norms for the SF-12 are also available for several countries, in particular the US (2,51,52). Unlike for the SF-36, SF-12 data from general population surveys in 9 European countries suggest there is little difference between standard US-derived scoring algorithms and country-specific algorithms, and standard scoring algorithms are recommended (53).

Respondent burden. The data on the respondent burden of the SF-36 are mixed. The self-reported version takes only 7–10 minutes to complete (54), although the presence of cognitive or physical impairment and depressed mood are associated with substantially longer completion time (55). The SF-12 takes only 2–3 minutes to complete (in a small pilot test, 81% completed the SF-12 in <2 minutes), less than one-third the time required to complete the SF-36 (2).

Generally, although the SF-36 and SF-12 use plain, easy-to-understand language, some of their items contain more than 1 concept (e.g., moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf), which could make it difficult for the participants to select the most appropriate answer. Evidence of item or response misinterpretation on the SF-36 has been reported in at least 2 studies (56,57).

Administrative burden. The SF-36 and SF-12 have relatively low administration burden. Interviewer administration of the SF-36 by telephone takes 16 to 17 minutes (58). No specific training for administration of the SF-36 or SF-12 is required, and completion instructions are self-explanatory. Computerized scoring algorithms for the revised versions of the SF-36 and SF-12 are available for purchase from QualityMetric and require basic knowledge of statistical software.

Translations/adaptations. The original versions of the SF-36 and SF-12 are available in English and Arabic. The revised versions are available in 121 languages. A list of translated versions is available at http://www.qualitymetric.com/WhatWeDo/LanguageTranslations/SurveysandTranslationsAvailable/tabid/215/Default.aspx, and further information can be obtained from the International Quality of Life Assessment web site (http://www.iqola.org). QualityMetric offers a translation service if required. Cultural adaptations of the original US version to other English-speaking countries are also available (59).

Psychometric Information

Method of development. The SF-36 was developed out of the RAND Corporation Health Insurance Experiment (60). The initial study measured 40 health concepts, 8 of which were selected for the inclusion into the new questionnaire. These 8 concepts were chosen to represent issues that were frequently used in health surveys and were affected by disease and treatment (3). Items for the questionnaire were generated following review of the content of various instruments that, at the time of SF-36 construction, were used to measure mental health and general health perceptions as well as limitations in physical, social, and emotional role functioning (1). No patient groups or representatives of the general population were involved in questionnaire construction. The SF-12 was developed using regression analysis methods to select and score 12 items from the SF-36 to reproduce the PCS and MCS scales in the general US population (2).

Acceptability. Data on acceptability of the SF-36 are mixed. The proportion of missing data varies from 0% for interviewer administration (61) to 26% for mailed versions (62) in nonhospitalized rheumatoid arthritis patients to 47% in hospitalized patients (including musculoskeletal patients) (63). Higher proportion of missing data is significantly associated with increasing age and disability (55,63).

On the SF-12, missing data in rheumatology settings occur less frequently, with just 15% of Danish respondents with arthritis missing ≥1 items of the PCS and 16% missing ≥1 items of the MCS (64). There is also a low individual item missing rate (<2.30%) and high percentage score computability (>90%) (25).

In arthritis studies, ceiling effects (>10% of participants obtaining the lowest possible score) are commonly reported for SF-36 role physical (21–76%), role emotional (49–60%), social functioning (23–64%), and bodily pain (20–40%) scales (14,61,65). Ceiling effects have also been observed on the mental health (20–28%) (14,65) and vitality (18%) scales (65). Floor effects (>10% of participants obtaining the highest possible score) are frequently found on role physical (29–80%) and role emotional scales (27–48%) (14,61,62,66,67,68), while at least 1 study has found there to be no notable floor effects (65). There do not appear to be ceiling and floor effects for the SF-12 among patients with rheumatic conditions (25,64).

Reliability. In musculoskeletal settings, results for reliability of the SF-36 are mixed. In several studies, all of the SF-36 scales were reported to have good internal consistency, with Cronbach's $\alpha \geq 0.70$ (61,62,65,67,69). In addition, internal consistency estimates were in excess of 0.90 for physical functioning and bodily pain scales in at least 2 studies (61,62) and for general health in at least 1 study (70), indicating suitability of these scales for use at the level of individual. Results for test–retest reliability of the SF-36 are less encouraging with the intraclass correlation coefficient (ICC) below the recommended standard of 0.70 on mental health (ICC 0.55) (68), role emotional (ICC 0.66) (68,71), as well as role physical (ICC 0.44), and vitality (ICC 0.03) scales (71).

Evidence also indicates a high proportion of measurement error on the SF-36 questionnaire in rheumatology, with large SDs for one-time administration (up to or exceeding the mean score) (72) and large variations (change of up to 200% from the initial score) in the SF-36 scores over a one-week test–retest period (70–72). In orthopedic surgery, minimal detectable change at an individual level ranged from 22% (general health) to 97% (role physical) of the total score range (14,70).

Internal consistency of the SF-12 component summary scores is generally high (Cronbach's $\alpha \geq 0.82$ and 0.75, for SF-12 PCS scale and MCS scale, respectively) (73–76). Test–retest reliability of the SF-12 administered 2 weeks apart is adequate in the US and the UK general populations: $r = 0.89$ for PCS and $r = 0.76$ for MCS (2), and others (74,77,78). For both the PCS and MCS scales, changes in scores between test and retest averaged less than 1 point, and at the second administration, 85% scored within the 95% confidence interval (95% CI) of the score at the first administration (2).

Validity. *SF-36.* Given the uncertainty about the reliability of the SF-36, findings on the validity of this measure need to be interpreted with caution. The SF-36 appears to have good face validity, with all items referring to health-related issues. Although content validity of an outcome measure is largely determined by the concept being measured, and may vary from one setting to another, the SF-36 captures a broad range of health states. However, the presence of severe floor and/or ceiling effects on the number of the SF-36 scales in rheumatic conditions indicates that this questionnaire does not adequately capture the full range of health experiences in this setting. Empirical studies of the construct validity of the SF-36 have shown mixed results for its validity.

The dimensional structure of the SF-36 (8 first-order and 2 higher-order factors) has been questioned in several studies. For example, higher-order factor analysis of the scale scores have confirmed separation of the scale scores into mental and physical health summary scores in some rheumatic studies (62,69) but not others (79). First-order factor analysis has also failed to confirm the 8-dimensional structure of the SF-36 in either exploratory (61,69) or confirmatory factor analysis (79,80).

Results for the convergent validity of the SF-36 are generally favorable. In several studies (62,65,70,72), the SF-36 scales had higher correlations with measures of similar constructs (such as the Western Ontario and McMaster Universities Osteoarthritis Index, the Nottingham Health Profile, the Health Assessment Questionnaire [HAQ], and rheumatoid arthritis disease activity measures) and lower correlations with dissimilar domains. In at least 2 other studies, the SF-36 had expected strong correlations ($r > 0.60$) with measures of similar concepts, including the Arthritis Impact Measurement Scales (AIMS), the HAQ, and EuroQol 5-domain instrument (EQ-5D) (67,68). However, at least one of these studies reported that the SF-36 scales did not correlate as well as expected with disease-specific measures of rheumatoid arthritis (61). However, the discriminant validity of the SF-36 has received less support. In a study of 200 patients with rheumatoid arthritis, physical functioning, role physical, general health, and bodily pain scales were expected to have correlations <0.3 with questionnaires hypothesized to measure dissimilar concepts (including fatigue and rheumatoid arthritis disease activity) (68). The discriminant correlations were much higher than expected, ranging between 0.51 (correlation between physical functioning and visual analog scale of fatigue) and 0.78 (correlations between bodily pain and rheumatoid arthritis activity scale). Higher than expected correlations between SF-36 scales and conceptually dissimilar measures were reported in at least one other rheumatoid arthritis study (67).

Overall, the evidence supports the known-groups validity of the SF-36 in rheumatology. With the exception of the mental health scale, the SF-36 had been able to differentiate between levels of osteoarthritis severity (72). The difference between those who had moderate and severe osteoarthritis, assessed using standardized effect sizes (ES), were in the small to moderate range, varying from 0.35 for general health to 0.75 for physical functioning. In the same study, all scales but role emotional and pain were able to differentiate among rheumatology patients with and without comorbid conditions, also with small to moderate ES, ranging from 0.49 (physical functioning and mental health) to 0.78 (general health). In another study, the SF-36 physical functioning and bodily pain scales discriminated well between patients receiving the disability pension versus those who did not, with medium ES values (0.69 and 0.50, respectively) (68). The SF-36 has also been shown to be able to differentiate between people with and without lower extremity osteoarthritis (81).

SF-12. Given that the primary purpose of the SF-12 was to reproduce the PCS and MCS scores of the SF-36, how well it does so is the important criterion, and there is strong evidence for the criterion-related validity of the SF-12. The SF-12 PCS and MCS scores correlate 0.95 and 0.96 with the SF-36 PCS and MCS scores, respectively (2,53). These findings have been replicated in the general populations of 9 European countries (Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, and the UK) (53), with very high correlations between SF-12 PCS and SF-36 PCS scores ($r = 0.94-0.96$) and SF-12 MCS and SF-36 MCS scores ($r = 0.94-0.97$). Clinical trials data from patients with osteoarthritis and rheumatoid arthritis also indicate good criterion validity of the SF-12 in rheumatology, with strong correlations between the SF-12 and SF-36 PCS scores and the SF-12 and SF-36 MCS scores ($r = 0.92–0.96$) (25).

The 2-factor conceptual structure of the SF-12 (PCS and MCS) has been confirmed in several population-based (82,83) and clinical studies (25,84). However, a recent study has challenged the 2-factor structure of the SF-12 (73). The standard orthogonally-weighted SF-12 scoring algorithm has been cautioned against, with oblique scoring algorithms appearing preferable (73,85).

Convergent and discriminant validity of the SF-12 in a general population is supported by relationships found with the EQ-5D (82). Comparable summary scores and dimensions correlate better, e.g., PCS with mobility (r = −0.69), usual activities (r = −0.71), and pain discomfort (r = −0.61) and MCS with anxiety/depression (r = −0.47), indicating good convergent validity. Less comparable summary scores and dimensions correlate weakly, e.g., PCS and anxiety/depression (r = −0.28) and MCS and mobility (r = −0.34), supporting discriminant validity of the SF-12.

Results for the convergent and discriminant validity of the SF-12 in rheumatic diseases are somewhat variable. In Danish patients with rheumatoid arthritis (64), the SF-12 PCS and MCS have been found to have unexpectedly weak correlations with measures of similar constructs, such as the HAQ (r = −0.15 for PCS and r = −0.25 for MCS) and lower correlations with dissimilar domains. In spinal clinic patients, back pain and disability have been found to be significantly moderately correlated with SF-12 PCS (r = −0.41 and −0.63, respectively, $P < 0.0001$) and MCS (r = −0.33 and −0.55, respectively, $P < 0.0001$) (76), as hypothesized. In addition, as expected in these patients, stress has been found to be weakly correlated with SF-12 PCS (r = −0.07, $P = 0.001$), but moderately correlated with SF-12 MCS (r = −0.33, $P < 0.0001$) (76).

The ability of the SF-12 to differentiate between groups based on severity of health impairment is generally good and is similar to that of the SF-36. PCS-12 and MCS-12 reach the same statistical conclusions about group differences as PCS-36 and MCS-36; they do so with relative validity coefficients that are typically 10% below those observed for the SF-36 (2). More specifically, among the Greek general population, the SF-12 PCS score has been found to be significantly worse among those reporting hip and knee problems compared with those not reporting such problems ($P < 0.01$) (82). However, among Danish patients with rheumatoid arthritis, the SF-12 did not seem to be sensitive to variation between patient groups with different disease severity (64).

Ability to detect change. Systematic examinations of minimum clinically important differences (MCID) for the SF-36 in rheumatic conditions are rare. More generally, a minimal detectable change of 5 points on a 100-point scale was previously reported for the SF-36 and is based on 95% CIs from a normative sample (4). Other measures of responsiveness (standardized response mean [SRM] and ES) support the ability of the SF-36 to detect change in this setting. The SF-36 has been demonstrated to be able to detect large improvements in health status at 3 and 6 months following joint replacement surgery (81). The most responsive scales were physical functioning, role physical, bodily pain, and social functioning, with SRMs of 1.04 or higher at 3 months. The least responsive scales were general health (SRM 0.20) and role emotional (SRM 0.37). Similar results have been obtained in another joint replacement study, with large improvements (ES >0.80) recorded for physical functioning, role physical, and bodily pain, moderate improvements (ES 0.50–0.80) for role emotional, vitality, and social functioning, and small improvements (ES <0.50) for mental health and general health at 6-months followup (14).

Veehof compared the responsiveness of the SF-36 with the responsiveness of disease specific scales (including the AIMS2 and the HAQ) in 168 patients with rheumatoid arthritis (86). The study showed that the responsiveness of the SF-36 is comparable to that of disease-specific measures. The bodily pain, vitality, physical functioning, and role physical scales have also been shown to have good ability (assessed by SRM) to identify rheumatoid arthritis patients who were classified as improved based on self-rating of disease activity (61,68). Responsiveness of the SF-36 to deterioration is more limited, with no scale able to capture self-reported deterioration (68).

The MCID of the SF-12 in rheumatology are also not known and there is limited information on its responsiveness. There is some evidence among those with back pain attending a spinal clinic (76), with a large ES for SF-12 PCS (0.82) and a small to moderate ES for SF-12 MCS (0.37) observed in patients whose self-reported back pain became much better after 3–6 months of followup. Similarly, moderate ES for SF-12 PCS and MCS (−0.46 and −0.21, respectively) were observed in patients whose self-reported back pain became much worse. Among workers with neck or upper extremity musculoskeletal disorders, SF-12 PCS scores have been shown to be responsive to clinically confirmed incident cases with a decrease in general physical function observed (ES −0.9, SRM −0.6). SF-12 MCS has been shown to be not responsive to such change, and neither PCS nor MCS scores were responsive to self-reported symptomatic incident cases, self-reported symptomatic recovered cases, or clinically-confirmed recovered cases (ES or SRM <0.2 or changes not in the expected direction) (87). In addition, SF-12 was reported to be responsive to a wide range of treatments and programs for musculoskeletal diseases (26–29,33).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SF-36 and SF-12 can be used when the assessment of a broad range of health aspects is needed. The SF-12 is brief, appears to adequately reproduce the 2 summary scores of the SF-36, and generally has comparable psychometric properties to the SF-36. Since the concepts represented in these questionnaires are not disease specific, the SF-36 and SF-12 are especially suited when comparisons between disease groups or with the general population are required. Availability of population norms also provides context for score interpretation. The SF-36 and SF-12 also appear to differentiate between levels of disease severity in rheumatic conditions and between people with and without rheumatic conditions, as well as respond to treatment-related changes in health status of people with rheumatic conditions.

Caveats and cautions. Given equivocal evidence of psychometric robustness of the SF-36 in rheumatic conditions, its use in this setting needs to be approached with caution. The role physical, role emotional, and social functioning scales are frequently reported to have low reliability, which puts their validity into question. Large test–

retest variations in the SF-36 scores at the individual level make the SF-36 unsuitable for individual assessments. Floor and ceiling effects in rheumatic conditions also indicate that the SF-36 does not adequately target the full range of health experiences of this population.

In contrast to the extensive SF-36 literature, the SF-12 has been less well-studied. Findings related to the SF-36 may not be transferable to the SF-12. There is a small loss (10%) in the ability of the SF-12 to distinguish between different disease groups compared with the SF-36. Use of the SF-12 for assessing and/or monitoring individuals is discouraged. There is limited evidence of its responsiveness to treatment-related changes in the health status of people with musculoskeletal conditions.

Clinical/research usability. In clinical settings, large intra-individual variations in the SF-36 scale scores and its low ability to detect deterioration make it unsuitable for use with individual patients, although the scale appears to have satisfactory ability to detect treatment-related improvements in health at a group level. In research settings, the SF-36 can be used to compare different disease groups or disease groups with the general population. Low measurement precision reported for the SF-36 scales in cross-sectional and test–retest studies can also dramatically increase sample size required to detect the desired ES for either between-group differences or within-group change over time. Ease of administration, availability of an online version, and availability of a computerized scoring algorithm support the usability of the SF-36 and SF-12 in research settings. However, financial costs can limit the use in low-budget studies, although the original version of the SF-36 is available at no cost. The SF-12 is a suitable measure where information on the SF-36 PCS and MCS scores is required. Psychometric evaluation does not support interpretation of scores to make decisions for individuals and, therefore, limits its clinical use.

NOTTINGHAM HEALTH PROFILE (NHP)

Description

Purpose. The NHP was developed in the 1970s (88) for measuring the impact of illness on patients and the assessment of changes in health status over time (89). As a generic health status questionnaire, it provides a brief indication of a patient's perceived emotional, social, and physical health and is intended for use in the general population (90).

Content. There are 2 parts of the NHP. The domains covered in part 1 are related to the health status of the individual (89) and include energy levels, pain, emotional reactions, sleep, social isolation, and physical abilities. Part 2 addresses the impact of ill health on daily life (89) and covers paid employment, home duties, social life, home life (relationships), sex life, interests and hobbies, and vacations. The 2 parts of the NHP can be used together or separately, with part 1 frequently used on its own.

Number of items. Part 1 consists of 38 items. The energy levels domain has 3 items, pain has 8 items, emotional reactions consists of 9 items, sleep and social isolation domains have 5 items each, and physical abilities domain has 8 items. Part 2 consists of 7 items that cover the 7 life areas listed in the above section (91).

Response options/scale. Responses are measured on a dichotomous scale, with respondents asked to check a yes box or a no box, according to whether a statement applies to them. If unsure, the instructions are to select an answer that is more applicable at the time of answering the questionnaire.

Recall period for items. Respondents are asked to identify whether each statement applies to them "at the moment."

Examples of use. In rheumatology, the NHP has been used in several randomized controlled trials, including evaluation of outcomes of exercise programs in rheumatoid arthritis (92), manual lymph drainage therapy and connective tissue massage in primary fibromyalgia (93), balneotherapy and tap water (94), and balneotherapy and mud-pack therapy (95) in patients with knee osteoarthritis. The NHP has also been used in several observational studies for the purpose of evaluating health status of people with osteoarthritis after knee arthroplasty (96) and after hip revision surgery (97), assessment of the efficacy of disease-modifying antirheumatic drugs in rheumatoid arthritis (98), and to evaluate outcomes of multidimensional rehabilitation program in chronic myofascial pain and/or fibromyalgia (99).

Practical Application

How to obtain. A copy of the NHP can be viewed at www.cebp.nl/media/m83.pdf. The official web site (www.galen-research.com) was under development at the time of writing; however, a copy of the NHP can also be obtained by contacting Galen Research (gr@galen-research.com). The noncommercial license fee for one language version of the NHP is approximately $192 (£120) per study. The scoring manual is included in this cost. The minimum cost for commercial studies is approximately $8,000 (£5,000) for 1 language version and increases to approximately $24,000 (£15,000) for 2 languages with an additional $8,000 (£5,000) for each subsequent language.

Method of administration. The NHP is designed to be self-administered.

Scoring. A scoring algorithm is available with the purchase of the questionnaire. Scoring instructions can also be downloaded from https://www.cebp.nl/media/m83.pdf. Scores for each of the 6 domains in part 1 are computed by summing weighted values given to each positive response. The weights for the NHP were derived using Thurstone's method of paired comparisons from a sample of 215 members of the general public. The sum total of the weighted scores is 100, with weights intended to reflect the perceived severity of a health state represented by the item from the point of view of the general public, rather than a specific patient population (89). Only domain scores are calculated, with no overall score.

There appear to be no specific instructions for handling missing values. Developers of the NHP recommend scoring responses to missing items as "no" since the respondents did not answer "yes" to these questions. However, Kersten et al (100) caution against using this approach routinely, since it could substantially underestimate the level of disability, particularly for severely disabled people, such as those using wheelchairs who are unable to walk at all.

Score interpretation. The scores on NHP domains range from 0 (best health state) to 100 (worst health state) (91). No normative data for the NHP are available.

Respondent burden. The NHP appears to have low respondent burden, taking 5–10 minutes to complete (91). Developers of the NHP have described this questionnaire as being a simple instrument that is acceptable and understood by a majority of people (91). In general, statements in the NHP are simple and easy to understand; for example, "I feel lonely" or "I have pain at night." However, some statements describing negative health states (e.g., "I feel that life is not worth living") may distress some respondents.

Administrative burden. Scoring of part 1 produces 6 domain scores plus a further 7 scores are produced if part 2 is used, therefore scoring may be cumbersome if done by hand (91). However, scoring and administration instructions are self-explanatory and require no specific training.

Translations/adaptations. The NHP is available in numerous languages including English, Greek (101), French (102), Swedish (103), Dutch (104), and Spanish (105). For an extended list of available translations, see http://www.proqolid.org/instruments/nottingham_health_profile_nhp.

Psychometric Information

Method of development. Information on the development of the NHP part 1 is generally scant and lacking in detail. In the development of part 1, statements describing the typical effects of ill health (social, psychological, behavioral, and physical) were collected from more than 700 people (91). This initial stage produced 2,200 statements, with 138 statements left after the removal of redundant and ambiguous items. The properties of these 138 statements were evaluated in a number of studies using diverse patient populations, after which the number of statements was reduced to 82 (91). No further information on characteristics of study participants, or types of tests or criteria used in item refinement and selection, is provided in the original publication describing the development of the NHP.

Part 2 was subsequently developed for the purpose of assessing how perceived health problems may affect daily living (106). The original statements collected during the development of the NHP were reviewed to identify areas of "task performance" most often affected by health problems. The areas of job, housework, social life, family life, sex, spare time activities, holidays, and travel were identified. Interviews were conducted with patients attending a hospital outpatient clinic. Difficulties in wording and presentation were identified, and further interviews were conducted with outpatients and a range of university employees. In total, 114 interviews were conducted. The wording of the items was revised by the developers with the intent of making them more understandable and acceptable for the average person with no university background and possibly limited education (106).

Acceptability. Missing data may be an issue when the NHP is administered to people who are severely disabled. In a study of 92 people with a range of disabilities (including 7 with rheumatoid arthritis), 46 people were unable to complete the NHP due to questions referring to activities that they were unable to perform (100). Missing data were present on 14 of 38 (37%) questions. Questions relating to pain, standing, walking, and other physical activities such as climbing stairs were particularly problematic. The pain domain was not completed by 48% of participants, and the physical functioning domain was not completed by 49%.

The NHP appears to be better able to capture states of ill health rather than states of good health. More than 50% of respondents in a study comparing the NHP sores for consulters of a general practice and nonconsulters scored 0 (best health) on each of the NHP domains (90). In a more recent study of 111 people using wheelchairs who live independently (including 30 who had rheumatic conditions), the emotional reactions, social isolation, and sleep scales of the NHP all had median scores of 0 (107).

Reliability. A limited number of studies have examined the internal consistency of the NHP in rheumatic conditions and have reported mixed results. The internal consistency of the NHP pain subscale was found to be acceptable in a sample of 160 people with rheumatoid arthritis (Cronbach's α = 0.83) (108). In another study conducted with a sample of 111 wheelchair-using people with a range of chronic conditions (including rheumatic diseases), the internal consistency of the pain and emotional reactions subscales were similar (Cronbach's α = 0.82), although the internal consistency of the mobility subscale was very poor (Cronbach's α = 0.34), and no internal consistency estimates were reported for the remaining subscales (107). In a sample of 1,063 individuals drawn from the general population, the internal consistency of the social isolation subscale was slightly below the acceptable lower limit of 0.70 (Cronbach's α = 0.65) while the internal consistency of the remaining subscales ranged from 0.71 (energy) to 0.88 (pain) (109).

Information on test–retest reliability of the NHP in rheumatology settings is very limited. In a sample of 73 patients with osteoarthritis who had no other comorbidities, 4-week test–retest reliability of the NHP (assessed using Pearson's correlation coefficient) ranged from 0.77 (energy) to 0.85 (sleep and physical mobility) on part 1 and 0.44 (hobbies/interests) to 0.86 (paid employment) on part 2 (110). However, it is well recognized that Pearson's correlation coefficient is a poor measure of temporal stability, since it is unable to capture systematic changes in scores over time. Hence, the "real" test–retest reliability of the NHP might be even lower. In a sample of 49 individuals with musculoskeletal disorders, test–retest reliability (assessed using intraclass correlation coefficient [ICC]) was within the acceptable range for pain (ICC 0.87) and physical ability (ICC 0.76) scales, with no information provided for the remaining NHP scales (111). Test–retest reliability of the NHP subscales in a French study, conducted with 111 individuals with rheumatoid arthritis, of the NHP subscales ranged from 0.57–0.73 (112).

Validity. The NHP also appears to have good face validity, with all items referring to an aspect of health. Although content validity of a measure is largely dependent on the concept being measured, the NHP covers a broad range of health-related functions (physical abilities, pain, sleep) that could be expected to be affected in rheumatic as well as in many other chronic health conditions, and therefore appears to have good content validity as a measure of general health status. In the development of the NHP, patient consultation was combined with expert consultation, therefore enhancing the relevance of the questionnaire to patients and clinicians. However, there has been limited investigation of the factor structure of the NHP in rheumatology-specific populations and more generally, so very little is currently known about the factorial validity of this questionnaire.

Convergent and discriminant validity of the NHP in rheumatology settings appears to be supported. A 3-year followup study of people with rheumatoid arthritis (n = 160 at baseline, n = 124 at 3-year followup) found that correlations between the pain subscale of the NHP and the General Health Questionnaire-28 ranged from 0.45 and 0.64, and from 0.25 and 0.41 between the pain subscale of the NHP and the Ritchie Articular Index (108). Less pain was also significantly correlated with greater psychological well-being. This profile of associations was in line with author hypotheses. Similarly, in a sample of 72 individuals with rheumatoid arthritis, the NHP showed an expected pattern of correlations with the Arthritis Impact Measurement Scales (AIMS), Beck Depression Inventory, and Health Assessment Questionnaire (HAQ) (113). In another study, all 6 NHP domain scores were significantly ($P < 0.0005$) related to disease activity measured by the Modified Disease Activity Score in rheumatoid arthritis, ranging from 0.25 for social isolation to 0.55 for pain (114).

Known-groups validity of the NHP was assessed in several studies and also received robust support. The NHP was able to differentiate between people with rheumatoid arthritis and a sample of well, community-dwelling people ages 40–59 years, with significantly lower scores for the rheumatoid arthritis group on the domains of energy, pain, physical mobility, and sleep (113). Scores for emotion and social subscales of the NHP were more similar between these groups, although mean scores were poorer in the rheumatoid arthritis sample for each domain. In another study, 200 outpatients with rheumatoid arthritis had higher NHP scores than both a random population sample and a second sample of patients with a variety of common diseases (114). However, neither of the above 2 studies provided standardized measures of differences, therefore information on magnitude of the differences in NHP scores between people with and without rheumatic conditions awaits further research.

The NHP also appears to be able to differentiate people with rheumatic conditions from those with other types of chronic illness. In a study of 82 individuals with rheumatoid arthritis or migraine (89), the authors hypothesized that rheumatoid arthritis would have greater impact on individual health than migraine, which would be reflected in higher NHP domain scores for individuals with rheumatoid arthritis. This hypothesis was partially supported. People with rheumatoid arthritis did have significantly worse health than the migraine group, but only on 3 out of the 6 NHP domains, including energy (migraine 43.6, rheumatoid arthritis 74.3), pain (migraine 14.9, rheumatoid arthritis 67.3) and physical mobility (migraine 2.0, rheumatoid arthritis 64.6), with no significant differences between the groups on domains of sleep, emotional reactions, and social support (89).

Ability to detect change. Information about the ability of NHP to detect change in rheumatic conditions is somewhat inconsistent, although it generally indicates that the NHP may not be as sensitive to change as other instruments that measure similar concepts. In one study, the ability of the NHP to detect self-reported improvements in health status in rheumatoid arthritis was compared with that of the AIMS, the HAQ and the Functional Limitations Profile (FLP) (115). Not one instrument outperformed the others across all domains. Compared with other questionnaires, the NHP had the lowest ability to detect self-reported change in mobility (effect size [ES] 0.27), pain (ES 0.38), and emotion (ES 0.59) domains, with only small to moderate ES recorded. At the same time, ES for other questionnaires measuring similar concepts were in moderate to high range, ranging from 0.69 to 0.83. In the social domain, NHP (ES 0.24) was worse at detecting change than FLP (ES 0.60) but better than the AIMS (ES 0.06). In another study involving 276 people with unilateral osteoarthritis of the hip waiting for joint replacement surgery, NHP was able to detect change in health status, with all NHP domain scores showing significant improvements 1 year following the surgery (116), although no information on the magnitude of change had been reported.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The NHP encompasses several domains that are of relevance to rheumatology, including energy, pain, physical mobility, emotions, sleep, and social and holds several areas in common with other disease specific instruments in this field. Being a generic measure, the main advantage of the NHP compared with disease-specific measures is that it can be used to compare the impact of rheumatic conditions with that of other illnesses or with the general population.

Caveats and cautions. Although studies assessing construct validity of the NHP produced favorable results, low test–retest reliability of some domains (namely emotion and social) raise doubts about psychometric robustness of this measure in rheumatology. It would also appear that the NHP is not appropriate for use in people with minor disability due to severe ceiling effects. The presence of ceiling effects could also pose problems in pre- and post-intervention studies, since improvement in condition for those who score zero at baseline cannot be demonstrated. Furthermore, the NHP also appears to be less sensitive to change than other health status measures used in rheumatology. The use of the NHP with severely disabled people might also present problems due to large amounts of missing data (100), and there appears to be no adequate methods for handling missing data on this questionnaire.

Clinical/research usability. The NHP is easy to use and score. Part 1 contains several items within each domain that combine to form a moderately detailed picture of the patient's current health. The areas of life affected by health listed in part 2 could serve to flag areas for further assessment in a clinical context. However, the high cost of obtaining the questionnaire could limit its usefulness in clinical settings. Sensitivity to change of the NHP is lower than that of other instruments measuring similar aspects of health, therefore its use in clinical trials where ES largely dictates the size and cost of a trial is less assured. As a measure of general health and health-related quality of life, the NHP may be of greater interest in epidemiologic rather than clinical research, although pronounced ceiling effects in well populations may severely limit its usefulness in population-based studies.

SICKNESS IMPACT PROFILE (SIP)

Description

Purpose. The SIP is a generic measure of health-related functional status (117), designed to be broadly applicable across types and severities of illness and across demographically and culturally diverse groups (118). The purpose of the scale is to provide a descriptive profile of changes in a person's behavior due to sickness (119). The SIP is intended for use in health surveys, program planning, policy formation, and monitoring patients' progress (120), and was initially published in 1977. Due to the length and respondent burden of the original 136-item version of the SIP, a shorter version, the SIP68, with 68 items, was developed in 1994 (121). A number of disease-specific short-form adaptations of the SIP were also developed, including back pain (122) and rheumatoid arthritis (123) versions.

Content/number of items. This SIP136 has 12 domains addressing the impact of health on a range of day-to-day behaviors, including sleep and rest (7 items related to sleep quality and daytime tiredness), emotional behavior (9 items addressing emotional well-being), body care and movement (23 items related to self-care, balance, and body movement), household management (10 items related to activities of daily life), mobility (10 items related to the ability to move within and outside the home), social interaction (20 items addressing relationships with others), ambulation (12 items related to walking), alertness behavior (10 items describing alertness and ability to concentrate), communication (9 items related to spoken and written communication), work (9 items related to work productivity and relationships with coworkers), recreation and pastimes (8 items addressing frequency and type of recreational activities performed), and eating (9 items addressing quantity and type of food intake) (120).

The SIP68 has 68 items across 6 domains: somatic autonomy (17 items related to basic somatic functions, such as ability to move independently and self-care), mobility control (12 items related to walking and hand use), psychic autonomy and communication (11 items describing concentration and spoken and written communication), social behavior (12 items related to social activities and recreation), emotional stability (6 items related to emotional self-control), and mobility range (10 items related to tasks of daily life) (118). The dimension names of the SIP68 differ from those of the SIP136, since during the construction of the SIP68, the questionnaire items formed a configuration of factor loadings that was different from that originally reported for the SIP136, with somewhat different dimensions emerging.

Response options/scale. When answering the SIP, respondents are asked to check all the statements that apply to them. Statements that do not apply are left blank.

Recall period for items. The recall period for all items is "today."

Examples of use. In rheumatology, the SIP has been previously used to assess changes in health-related function status following total hip replacement surgery (124), to determine the effects of an exercise program in osteoarthritis of the hip (125), and to measure the impact of multidisciplinary team care versus regular outpatient clinic care on overall health in people with rheumatoid arthritis (126).

Practical Application

How to obtain. A copy of the SIP136 is available under a limited use agreement from MAPI Research Trust at http://www.mapi-trust.org/questionnaires/53. The SIP costs ~$677 (€500) per study for funded academic research and ~$1,354 (€1,000) per study for commercial studies. There are no distribution fees for nonfunded academic research and individual clinical practice. Distribution fees are ~$400 (€300) per study, plus $68 (€50) per language version in funded academic research, and ~$677 (€500) per study plus ~$203 (€150) per language version in commercial studies (127). The SIP68 is available at no cost in de Bruin et al (121).

Method of administration. The scale can be self-administered or interviewer-administered.

Scoring. SIP scores can be calculated manually or using a scoring algorithm, which is available with the purchase of the SIP (128). The 12 categories of the SIP136 can be scored separately to provide a health profile. Alternatively, the SIP items can be combined to obtain 2 summary dimension scores, including physical dimension (ambulation, mobility, and body care/movement) and psychosocial dimension (emotional behavior, alertness behavior, communication, and social interaction) scores. An overall score based on all 136 items can also be obtained (120).

The category scores are calculated by adding the weights assigned to each item checked within the category. The sum total is then divided by the value of the highest weight for the category and multiplied by 100 to obtain the category score. The 2 dimension scores and the overall score are calculated in a similar manner. The item severity weights for the SIP have been derived using equal-appearing interval scaling method from a sample of more than 100 judges, including patients and health professionals in Seattle, Washington (120).

The SIP68 can be used to calculate an overall total score, 2 dimension scores (physical and psychosocial), or 6 subscale scores. The physical dimension score includes somatic autonomy, mobility control, and mobility range scales, and the psychosocial dimension consists of psychological attention and communication, social behavior, and emotional stability scales. The SIP68 is scored by adding the number of items that were checked for each category, dimension, or overall to obtain category, dimension, and total score, respectively. In scoring of SIP instruments, all unchecked items are given a score of 0.

Score interpretation. The score range for the SIP136 category, dimension, and total scores is 0 (best health) to 100 (worst health) (128). The score range for the SIP68 is 0 (best health) to 68 (worst health), with the score range for the category and dimension scores varying according to the number of items that make up a given category/dimension (118).

Respondent burden. The SIP136 may have moderate respondent burden, with an average completion time of 20–30 minutes (129). One study reported that 84% of respondents self-completed the SIP in <40 minutes (130). In a study of 168 male veterans residing in nursing homes in the US, the interviewer-administered SIP136 completion time ranged from 20–65 minutes, with a mean of 35 minutes. Longer completion times were associated with impaired verbal functioning. Interviewer assessment indicated that, in general, the instructions were well understood and items were not considered to be unduly sensitive. The SIP68 has been reported to take 15–20 minutes to complete (132).

Administrative burden. The SIP questionnaires have minimal administrative burden. The manual scoring procedure has been reported to take 5–10 minutes to complete for the SIP136. No special training is needed to either administer the questionnaire or interpret the results (129). Administration and scoring instructions are self-explanatory and are easy to follow for the SIP68.

Translations/adaptations. The original language of the SIP136 is US English. Existing translations (which may not have undergone a full linguistic validation process) are available in Arabic, Chinese for Hong-Kong, Danish, Dutch, Dutch for Belgium, English for Mexico, English for the UK, Finnish, French, French for Belgium, German, Italian, Norwegian, Portuguese, Romanian, Russian, Spanish, Spanish for Mexico, Spanish for the US, Swedish, Tamil, and Thai (133). A UK adaptation of the SIP, the Functional Limitations Profile, is also available (134). Since the SIP68 is a shortened version of SIP136, this questionnaire can also be made readily available in multiple language versions.

Psychometric Information

Method of development. Statements describing sickness-related changes in behavior were elicited from general practice patients, health care professionals, significant others, and apparently healthy individuals (135). A total of 1,100 responses to the survey were collected. These statements, together with a review of function assessment instruments designed for the evaluation of circumscribed patient groups, resulted in 1,250 specific statements of behavioral change. These statements were subjected to standard grouping techniques according to a set of criteria, which yielded 312 unique statements, each describing a behavior or activity and specifying a dysfunction. A standard sorting procedure yielded 14 groups of statements, each of which appears to describe dysfunction in an area of living or a type of activity (119). The 14-item groups were further refined to produce the current scale with 12 domains.

The SIP68 was developed using principal components analysis of the data obtained from studies in 10 different diagnostic groups with a total of 2,527 respondents to the Dutch translation of the original SIP (121). Of the 2,527 respondents, data from 835 individuals were used in the construction of the SIP68, with a maximum of 100 from any of the 10 diagnostic groups (n = 100 for rheumatoid arthritis, n = 100 for ankylosing spondylitis, n = 41 for spinal cord injury, n = 53 for stroke, n = 100 for cancer, n = 100 for neuromuscular disease, n = 100 for back/neck pain, n = 100 for head injury, n = 99 for hemodialysis, and n = 42 for Crohn's disease). Items which applied to <10% or >90% of any diagnostic subpopulation were removed, as were items that did not contribute substantially (using an a priori definition of substantial loading as >0.40) to the scales or the total score.

Acceptability. The evidence for acceptability of the SIP instruments is generally not favorable. In a pilot study of the interviewer-administered SIP136 to 246 general practice enrollees (inpatients, home care patients, walk-in patients, outpatients, and nonpatients) in the US, all subjects completed the interview, with 9% of participants not finding at least 1 item on the questionnaire that applied to them (119). Similarly, in a study of 85 people with rheumatic conditions, who consented to participate in an evaluation study of the SIP68 in the Netherlands, 9% were unable to complete the instrument due to physical limitations or difficulty in understanding the instructions (118).

The proportion of missing data on the SIP is difficult to estimate, due to respondents instructed to leave items that do not apply to them unchecked. In a study of 301 people age ≥65 years, the question asking about sexual activity was left unchecked most frequently (12% of respondents) (136). In another study of 329 poststroke patients who participated in the interviewer-administered SIP136, responses from only 10 people (3%) could not be used for data analysis due to high proportion of missing data (137). While this study was not conducted in a rheumatology specific population, results may be indicative of a broader acceptability among frailer populations (121).

The SIP appears to have good range of functioning at the levels of very good health, with no floor effects generally reported for the total scale or dimension and category scores (136,138). However, substantial ceiling effects were found for category scores on the SIP136 in a study of 301 people, age ≥65 years, ranging from 31% for social interaction scale to 87% for the work scale. It should also be noted that persons who do not work at all (e.g., retired individuals) are given the maximum score for this category, therefore potentially inflating the ceiling effect for this scale. The physical and psychosocial dimension scores also had ceiling effects, with 27% and 22% of respondents recording best possible health state, respectively (136). These results indicate that the SIP may not be suitable to use with people who have low to moderate levels of ill health.

Ceiling effects were also reported for the SIP68. In a study of 329 people with disabilities (138), ceiling effects were found on emotional stability (54%), mobility range (24%), psychic autonomy and communication (24%), and somatic autonomy (17%) categories, as well as psychological dimension (19%). De Bruin et al (121) also found mild ceiling effect for the SIP68 total score in a sample of 83 outpatients with rheumatoid arthritis (12%). These individuals judged their health as good to very good, which was matched by a rheumatologist's rating of their functional status.

Reliability. In a sample of 299 patients with musculoskeletal disorders recruited from the Hospital for Rheumatic Diseases (Bad Wurzach, Germany), internal consistency (Cronbach's alpha) of the SIP136 was very low for sleep and rest (0.28), eating (0.33), communication (0.41), and emotional behavior (0.59); marginal for home management (0.66), work (0.64), recreation and hobbies (0.67), and mobility (0.69); and was within the acceptable range for social interaction (0.71), ambulation (0.76), alertness behavior (0.76), and body care and movement (0.80) (139). Internal consistency of the overall score was 0.83.

Test–retest information on the SIP136 in rheumatology and more generally is scant, but indicates good temporal stability of dimension and overall scores. In a sample of 49 individuals with musculoskeletal disorders who completed a second SIP 3 weeks after the initial administration, intraclass correlation coefficient (ICC) was 0.94 for the physical function dimension and 0.93 for the overall SIP136 score, indicating excellent reliability; no information was provided about test–retest reliability of the remaining scores (111). The SIP overall score also showed good temporal stability in another study, involving 130 patients with chronic low back pain. The ICC for the SIP136 total score was 0.70 over a 2-week test–retest interval (140).

Temporal stability of subscale scores seemingly had been assessed only using Pearson's correlation coefficient, with results indicating below optimal reliability for some of the scale. Correlation coefficients ranged between 0.49 (eating) and 0.86 (mobility) over a 4-week interval in a study of 299 musculoskeletal patients (139) and between 0.62 (household management) and 0.85 (ambulation) in a study involving 119 individuals with a range of chronic conditions (141). However, Pearson's correlation coefficient is unable to capture any systematic changes in scores over time, so that the actual stability of SIP136 subscale scores might be even lower.

Internal consistency reliability of the SIP68 in rheumatic conditions is not well studied, with the available evidence indicating suboptimal internal consistency, at least for some subscale scores. Internal consistency estimates (Cronbach's alpha) for the SIP68 in a study of 51 outpatients with rheumatic conditions (118) ranged from 0.49–0.87 (coefficients for specific domains not specified). However, this sample size was relatively small and generalizability of this finding is not clear. More broadly in the field of disability, for a study conducted with 111 independently living wheelchair users, Cronbach's alpha of the total SIP68 score was $\alpha = 0.88$, with scores for the individual scales ranging from $\alpha = 0.53$ (for mobility control) to $\alpha = 0.85$ (somatic autonomy) (107). However, given the much higher levels of physical disability in this sample than would be expected in rheumatic diseases, it is not known whether these results can be extrapolated to rheumatology.

Test–retest reliability of the SIP68 was assessed using a 48 hour test–retest interval in a study of 51 outpatients with rheumatic health problems using self-completed questionnaires (118). The ICCs for different categories ranged from 0.90 (mobility range) to 0.97 (somatic autonomy) and was 0.97 for the overall SIP68 score, indicating excellent test–retest reliability of this questionnaire. In another study, involving 401 people with disabilities (including arthritis), the ICCs for test–retest reliability of SIP68 were above 0.75 for all subscales and dimensions, except the physical dimension (0.61) score (142).

Validity. The items of SIP instruments appear to have good face validity as a health measure, reflecting aspects of everyday life that are likely to be affected by illness. However, no specific a priori conceptual model was used in the SIP construction, which makes it difficult to comment on its content validity. Furthermore, content validity of an instrument varies from context to context, depending on the nature of the concept being studied. Nonetheless, the SIP covers a wide range of health-related behaviors, many of which are likely to be relevant in rheumatology. Development of the SIP136 involved both patient and expert consultation, during both the item construction and selection phases of development; hence, the questionnaire is likely to be relevant to patients and clinicians. However, content validity of the SIP instruments is undermined by the presence of ceiling effects, which indicates that these questionnaires do not adequately capture the full range of health problems at a less severe end of the ill health continuum. This could potentially pose problems when assessing interventions in populations that are not severely affected by illness.

Although results for the construct validity of SIP instruments are generally favorable, given the unsatisfactory reliability of some subscales, findings about their construct validity should be viewed with caution. Factorial validity of the SIP68 and SIP136 is not well supported, with different pattern of factor loadings to that reported in the original publications generally emerging (138,142). Construct validity of the SIP136 in rheumatic conditions was supported by the expected pattern of correlations with the Arthritis Impact Measurement Scales (AIMS), a multidimensional questionnaire designed to measure ill health in arthritis. Over 12 months of followup in a study of 115 patients with knee or hip osteoarthritis (143), SIP subscales had moderate to strong correlations ($r = 0.37-0.76$) with the corresponding subscale of the AIMS ($P < 0.001$). Correlations for the total scores of the SIP136 and AIMS ranged between 0.70 and 0.73 (143). Further support for the construct validity of the SIP was found in another study, involving 299 patients with musculoskeletal conditions, where the hypothesized pattern of correlations was found between the SIP total score and a range of functional and psychosocial measures, including the Measurement of Patient Outcome Scale (arthritis-specific questionnaire in

German) (r = 0.72) and the Keitel Index of Functional Status (r = 0.60) (139).

The SIP68 also received support for its construct validity, with a correlation coefficient of 0.94 for its total score with the SIP136 in a study of 401 people with mobility disabilities, including spinal cord injury, multiple sclerosis, and arthritis (142). In another study with people with physical disabilities (n = 398), physical and psychosocial dimensions of the SIP68 had high correlations with corresponding dimensions of SIP136 (r = 0.91 and r = 0.92, respectively) (138). In the same study, construct validity of the SIP68 was further supported by the expected pattern of correlations with other generic health measures, including the SF-36 and the Katz Index of Activities of Daily Living.

Known-groups validity of the SIP instruments in rheumatology appears to be well supported. In one study (n = 172) (139), the SIP136 overall score was able to differentiate people with musculoskeletal disorders from healthy controls, although the difference was much larger for women (standardized effect size [ES] 1.07) than for men (ES 0.66). Across individual dimensions, work, eating, mobility, alertness behavior, sleep and rest, and communication were unable to differentiate between patients and controls for men, with no significant differences in the scores of these groups ($P > 0.05$). On the remaining subscales, male patients scored higher than controls, with ES ranging from 0.21 (communication) to 1.07 (body care). For women, only work was unable to differentiate between patients and controls ($P > 0.05$), with ES ranging from 0.42 (mobility) to 1.53 (home management). Deyo et al (144) have also published mean scores and SDs for the SIP overall and the physical and psychosocial dimensions that correspond with functional impairment levels of the American Rheumatism Association Functional Classification.

The SIP68 was also able to differentiate between people with spinal cord injury and those who had rheumatic diseases, with significantly worse health status scores on emotional stability, social behavior, mobility range, and psychic autonomy and communication for the rheumatic conditions groups (Z scores 2.10, 5.10, 5.71, and 2.01, respectively). Somatic autonomy and mobility control scores did not differ significantly between the study groups (145). Ability of the SIP68 to differentiate between spinal cord injury and rheumatic conditions was comparable to that of the Nottingham Health Profile.

Ability to detect change. Sensitivity of SIP instruments in rheumatology has not been well studied. Although little is known about the sensitivity of the SIP68, results indicate that the SIP136 has high specificity to detect change in health status. In a study of 79 patients with rheumatoid arthritis, the SIP136 total score had high specificity to detect a 3-point change in self-rated function, with a specificity of 0.90 for detecting worsening and a specificity of 0.76 for detecting improvement (146). However, sensitivity to improvement was 0.25, sensitivity to worsening was 0.29, the predictive score of improvement was 0.50, and the predictive score of worsening was 0.31. This indicates that a cutoff threshold of 3 points on the SIP score change had only moderate accuracy in identifying self-perceived change in health functioning.

More favorable results were obtained in a study of 54 patients undergoing joint replacement surgery, where overall and psychosocial dimension scores of SIP136 were able to detect large improvement at 6-months postsurgery, with standardized response means (SRMs) of 0.94 and 0.88 (147). As might be expected with surgical intervention, smaller improvement (SRM 0.77) was recorded for psychosocial dimension. Similar results were obtained in another orthopedic surgery cohort at 1-year followup (148). SIP136 overall and physical dimension scores were also able to detect self-reported change in health status in a sample of 127 musculoskeletal patients, with an ES of 0.42 and 0.39, respectively (111). In another study involving 299 musculoskeletal patients, SIP overall, body care and movement, emotional behavior, and sleep and rest scores showed small improvements (ES 0.20–0.28) following 4 weeks of conservative treatment (139). Statistically significant improvements were also found on alertness behavior, ambulation, home management, social interaction, and mobility subscales, although these changes failed to reach practical importance (ES <0.20); communication, recreation and hobbies, eating, and work subscales showed no change over the study period.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SIP includes several items relevant in rheumatology settings including mobility, pain, and functional capacity. As generic measures of health status, SIP instruments would be useful when comparisons of the impact of rheumatic disease on individual health status with that of other illness is required. There is also good evidence to suggest that the SIP (especially the SIP136) is able to detect change in a range of interventions in rheumatology.

Caveats and cautions. Several studies have revealed considerable weaknesses of the SIP68 and SIP136, particularly in the area of reliability. Both versions of the SIP also exhibit severe ceiling effects, which suggest that these instruments may not be useful for low to moderate levels of health impairment.

Clinical/research usability. The low reliability of some subscales in the SIP indicates that the overall score, rather than subscale scores, is more likely to return more robust data. Administration and responder burden may be barriers to clinical use of the SIP136; however, this appears to have been rectified in the SIP68. The relatively high cost of the SIP136 may further limit its usability in clinical settings and research settings. Given the comparable psychometric properties of the 2 versions of the SIP and the greater administrative burden and cost of the SIP136, it appears there are no advantages of using the SIP136 over the SIP68.

INTRODUCTION

Health Utility Measures

The EuroQol 5-domain, Short Form 6D, Health Utility Index Mark 3, the Quality of Wellbeing Scale, and the

Assessment of Quality of Life Scale are health utility measures of generic health-related quality of life (HRQOL) originating from the field of health economics. These scales are also defined as multi-attribute utility instruments, which means that they consider multiple independent attributes of an individual to create an indication of overall HRQOL, ranging from perfect health (1.0) to death (0.0), and may even include states worse than death (<0.0). The individual attributes of HRQOL contained within the questionnaire (often described as the "descriptive system") are weighted by society's strength of preference for those health states. The strength of preference is termed the utility of a health state and is obtained by asking members of the community to rank desirability of a given health state relative to perfect health and death. The utility of health states represented in the descriptive systems is generally achieved through specialized interviews such as time trade-off or standard gamble.

Although all HRQOL instruments purport to measure the same thing, across the perfect health to death continuum, they often do not. The values obtained from each for the same health state (person- or community-tested) vary because each instrument has different content, and different weights are used to generate its overall utility score (149–153). Since each instrument can generate a different value, different change scores will be obtained across instruments (154–157). Given this, the choice of instrument included in a study has the potential to generate results suggesting a null or positive result (158–160), although Ruchlin et al suggest that there is no specific pattern emerging (161). The recent work of Seymour et al suggest that choosing an instrument is difficult without good prior information surrounding the expected magnitude and direction of health improvement related to a health care intervention (162).

MEDICAL OUTCOMES STUDY SHORT FORM 6D (SF-6D)

Description

Purpose. The SF-6D utility score is derived from items within the widely used SF-36 and SF-12. The purpose of the SF-6D is to provide ratings of an individual's health-related quality of life (HRQOL) across all health conditions. The ratings of HRQOL are also called "utilities" or preferences for health states that are used in health economic evaluation and to derive quality-adjusted life years (QALYs) for use in cost utility analysis. Brazier et al published an algorithm for estimating SF-6D utilities scores from the SF-36 in 2002 (163) and from the SF-12 in 2004 (164). The initial scoring algorithms were updated in 2008 (www.shef.ac.uk/scharr/sections/heds/mvh/sf-6d).

Content. The SF-6D covers 6 domains, including physical function, role limitation, social function, bodily pain, mental health, and vitality.

Number of items. The SF-6D utility score can be derived from 11 items of the SF-36 or from 7 items of the SF-12 (164).

Response options/scale. Items are scored on a Guttman scale, where the health states have increasing severity (disutility) expressed as limitations (in activities, the kind of work one can do, social activities), degree of pain interference with daily life, frequency of feeling down-hearted, or frequency of feeling fatigued. The number of response levels on the SF-6D items ranges between 3 and 5 (SF-12 derivation [164]) or 4 and 6 levels (SF-36 derivation [163]).

Recall period for items. The SF-6D is available in 4-week or 1-week recall periods.

Endorsements. The utilities derived from the SF-6D are being used in a wide range of health economic studies to provide estimates of cost per QALY, therefore enabling comparison of alternative treatments. These data inform policy makers of the relative value of new interventions. In several countries including the UK, Australia, and New Zealand, the calculation of QALYs is essential for economic evaluations of pharmaceuticals submitted to the government agencies (i.e., National Institute for Health and Clinical Excellence in the UK).

Examples of use. Several studies have used the SF-6D to estimate QALYs, therefore providing the economic dimension in treatment effectiveness studies, including studies of tumor necrosis factor–blocking agents (12,165), spa treatment for people with fibromyalgia (166), a physical exercise program for people with rheumatoid arthritis (RA) (167), and in a survey to express the burden of disease of people with RA (168).

Practical Application

How to obtain. The SF-6D can be obtained from SF-36 or SF-12 questionnaire scores, therefore it is necessary to obtain these questionnaires. For details, see the How To Obtain section for the SF-36 and SF-12 in this issue.

Method of administration. As with the SF-36 or SF-12, the SF-6D can be self- or interviewer-administered. For details, see the Method of Administration section for the SF-36 and SF-12 in this issue.

Scoring. The SF-6D utility score is calculated as a function of weighted scores across the items that comprise this tool. The algorithm to obtain SF-6D scores from the SF-36 and SF-12 questionnaire data can be obtained through 3 types of licenses, as described on the University of Sheffield web site (www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d): 1) a license is available free of charge for all noncommercial applications including work funded by research councils, government agencies and charities, 2) for commercial applications there is a per-study license (e.g., clinical trial), although an open license for a fixed period is available, and 3) the SF-6D can be calculated using purpose-developed software available from QualityMetric.

Score interpretation. The SF-6D produces an interval scale utility score, ranging from 0.30 (poor HRQOL) to 1.0 (perfect health). The SF-6D utility measure can also be used as an indicator of relative disease burden across diseases. This is dependent on reliable population norms being available, such as those proposed by Fryback et al for the US (169). Uhlig and colleagues used the SF-6D to compare the HRQOL of people on the Oslo Rheumatoid Arthritis Register with people from the general population and found that people with RA have 0.16 lower utility than the population. They were therefore able to present a case that RA contributes a substantial disease burden on individuals and society (168).

Respondent burden. See the Respondent Burden section for the SF-36 and SF-12 in this issue.

Administrative burden. See the Administrative Burden section for the SF-36 and SF-12 in this issue.

Translations/adaptations. The SF-36 is available in 121 languages, therefore the SF-6D is similarly available. Specific information can be obtained from the International Quality of Life Assessment web site, http://www.iqola.org.

Psychometric Information

Method of development. The SF-36 and SF-12 items were revised to only cover 6 dimensions of health while maintaining maximum coverage of the original breadth of the questionnaires. The challenge for the developers was to provide valuations of all the different combination of health states that could be represented across the 6 items, each with 3 or more levels. The total number of possible health state combinations is 18,000, which is far too many to value in practice. A common procedure in health economics is to select a minimum range of these using an orthogonal design, and therefore infer the valuations of the health states not directly valued. A total of 49 combinations of levels of the 6 items was valued by a representative community sample using a technique called standard gamble (SG) (170). In the SG interview, the respondent is asked to choose between the certain prospect (A) of living in an intermediate state defined by the SF-6D and the uncertain prospect (B) of 2 possible outcomes, the best state defined by the SF-6D or the worst state. The chances of the best outcome occurring is varied until the respondent is indifferent between the certain and uncertain prospects. The data obtained from these valuations are then used in various modeling procedures to generate an algorithm to convert the SF questionnaires into SF-6D utility scores. Further details are available from the development papers (163,164).

Acceptability. The acceptability and missing values of the SF-6D are reflected in the original questionnaires (the SF-36 and SF-12), which are generally acceptable. Barton and colleagues compared the completion rates of the SF-6D with the EuroQol 5-domain (EQ-5D) measure in 1,865 general practice patients and found that individuals who were older, women, of a lower occupational skill level, from an area of lower socioeconomic status, or used prescribed medication were significantly less likely to complete the SF-6D (84%) compared with the EQ-5D (93%) (171).

Importantly, HRQOL measures are intended to provide valuations of health states that range from perfect health (1.0) to death (0.0). However, the SF-6D scale does not extend beyond 0.3, i.e., the worse health state described by the SF-6D does not extend to death. This is a serious flaw if a substantial number of subjects in a study are expected to have very poor health states.

Reliability. The reliability of the SF-6D has been tested in a variety of settings, with generally favorable results. In a small study (n = 61) of proximal humeral fractures, Slobogean et al (172) found good reliability (intraclass correlation coefficient [ICC]) for the SF-6D (0.79) and EQ-5D (0.78), but poor for the Health Utilities Index Mark 3 (0.47). Khanna et al also found the SF-6D to be reliable in a sample of patients with systemic sclerosis (ICC 0.82) (173). On the other hand, Boonen et al found that in patients with ankylosing spondylitis, the test–retest reliability of the SF-6D was only modest (ICC 0.68), and this was greatly reduced in subgroups with lower disease activity (174).

Validity. The SF-6D contains items that cover physical function, role limitation, social function, bodily pain, mental health, and vitality. Consistent with most utility scales, the SF-6D was not derived through consultations with patients and clinicians to ensure face and content validity (151). Nonetheless, the dimensions are broadly concurrent with those covered by the many disease-specific tools available in rheumatology. Support for convergent and discriminant validity of the SF-6D is evidenced by consistent findings of moderate correlations between the SD-6D and other HRQOL scales (175–179) and lower, but still substantive, correlations with disease-specific questionnaires (174,177,179,180).

The known-groups validity of the SF-6D appears to be supported. Marra et al undertook a comprehensive study in 313 people with RA to compare several disease-specific measures (Rheumatoid Arthritis Quality of Life Questionnaire and the Health Assessment Questionnaire [HAQ]) with several preference-based measures including the SF-6D (179). They found that utility scales, including the SF-6D, appeared to discriminate well across RA severity categories, although the disease-specific measures were generally more sensitive in this setting. In 167 patients with systemic lupus erythematosus, Aggarwal et al (178) found that both the EQ-5D and SF-6D tools differentiated among patient groups of varied disease severity. Importantly, very few patients in this study reported very low HRQOL, therefore the tools are more likely to appear to perform relatively well. However, in a population sample, the SF-6D has been found to be more sensitive than the EQ-5D in detecting differences between groups of individuals reporting very good, good, fair, bad, or very bad health (181).

Ability to detect change. Evidence for the ability of the SF-6D to detect change is mixed. While several studies have demonstrated that the SF-6D is capable of detecting change, findings of many other studies are less favorable. Boonen et al found that in 254 patients with ankylosing spondylitis, the smallest detectable change was smaller (i.e., more sensitive) in the SF-6D compared with the EQ-5D. However, it discriminated less well between patients with different disease severities (174). Harrison et al undertook a comparative responsiveness study of the EQ-5D and SF-6D in cohorts of patients with early inflammatory disease through to severe RA (182). As the use of the SF-6D in patients with severe progressive disease may be inappropriate due to the scale not extending lower than a utility of 0.30, the study by Harrison and colleagues (182) highlights the need for careful attention to disease severity at study onset. The SF-6D did, however, appear to be somewhat more responsive than the EQ-5D in detecting improvements in health (182). On the other hand, in a controlled trial, Barton and colleagues administered the Western Ontario and McMaster Universities Osteoarthritis Index to 389 people with knee pain and classified change score as no change, improved >20%, or declined >20% (158). The SF-6D performed poorly at detecting improvement. Similar results were obtained by Adams et al in 505 patients with RA and psoriatic arthritis and again reflect the inability of the SF-6D to detect poor health states (183).

Several studies reported on minimum clinically important difference (MCID) of the SF-6D. In rheumatology settings, Khanna et al have proposed a MCID of 0.035 units in systemic sclerosis using change in the HAQ Disability Index score as an anchor (173), and Marra et al have estimated a MCID of 0.03 for people with RA using the SF-36 health transition question as an anchor (179). More broadly, Walters and Brazier undertook a review of 11 studies across a variety of health conditions and found that the MCID for the SF-6D ranged from 0.011–0.097, with a mean of 0.041. The corresponding standardized response means ranged from 0.12–0.87, with a mean of 0.39, and were in the "small to moderate" range using Cohen's criteria (152).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The SF-6D can be a useful indicator of utility in the absence of other utility measures. A unique aspect of this tool is that, if the SF-36 and the SF-12 have been applied in a completed trial or observational study, a utility score for cost utility analyses can be derived from existing data without the need for administering further questionnaires.

Caveats and cautions. The major drawback of the SF-6D is that the scale does not cover the range from below 0.3, which would be a common health state in many rheumatic conditions. This makes the scale insensitive to changes between very poor health and moderate health. If researchers are working with a well-defined clinical condition with mild to moderately poor HRQOL, then the SF-6D may be preferred over other utility measures, such as the EQ-5D, which may be insensitive to improvements in this range.

Clinical/research usability. The SF-6D is not a tool to be used in the clinical setting since it is a utility instrument designed to inform economic evaluations. It is also useful for comparisons across conditions, and to provide estimates of relative societal burden of different conditions when national norms are used as benchmarks.

HEALTH UTILITIES INDEX MARK 3 (HUI3)

Description

Purpose. The HUI is a family of generic preference-based (utility) measures developed for measuring health-related quality of life (HRQOL) (184). The intended uses of the HUI include describing treatment processes and outcomes in clinical studies, economic evaluations of health care programs, and the measurement and monitoring of population health (185). The original version (HUI1) was developed for assessment of out-of-pocket costs and quality of life of pediatric oncology survivors. The HUI2 was developed as a revised version of HUI1 to measure the global morbidity burden of childhood cancer (184). The HUI3 was developed as a more generically applicable measure than HUI1 and HUI2. Items present in earlier versions specific to pediatrics (e.g., cognition domain items in HUI2 relate to schoolwork) were replaced with more broadly applicable items, while some domains were expanded (e.g., sensation in HUI2 was broken into 3 separate domains of vision, hearing, and speech in HUI3), and others were removed (e.g., fertility in HUI2 was removed from HUI3). Therefore, HUI3 domains largely overlap with those of HUI2. HUI1 was first published in 1982 (186), while HUI2 and HUI3 were described in the literature in the mid-1990s (187). This review focuses on HUI3, since this version is commonly used in rheumatology (117,188).

Content. The HUI3 measures 8 HRQOL domain areas including vision, hearing, speech, ambulation/mobility, pain, dexterity, emotion, and cognition. HUI2 measures the 7 domain areas of sensation, mobility, emotion, cognition, self-care, pain, and fertility.

Number of items. HUI2 and 3 require the participant to select one descriptor that most accurately reflects their condition per domain.

Response options/scale. Each domain within the HUI3 has 5–6 rank-ordered response options, while HUI2 has 3–5 response categories per domain. Descriptors of response categories may contain 1 element (e.g., the HUI3 emotion domain has a response option of "somewhat happy") or it may contain several elements (e.g., the HUI3 hearing domain has a response option of "able to hear what is said in a conversation with one other person in a quiet room with a hearing aid," and "able to hear what is said in a group conversation with at least three other people, with a hearing aid"). Therefore, if the HUI3 is being administered via telephone where the participant cannot read the entire descriptor of a response option, a series of shorter questions need to be asked to allow the individual to select a response option that is most appropriate to their situation.

To resolve this problem, the developers have produced a 15-question (15Q) survey to allow the participant to identify the appropriate response option based on a series of shorter questions. There is also a 40-question (40Q) survey comprised of even less complex, predominantly yes/no response options. The 40Q version of the HUI3 has a skip pattern so that only some questions will need to be asked of each participant.

Recall period for items. There are several versions of the HUI3 available with recall periods of 1 week, 2 weeks or 4 weeks (e.g., "Describe your ability during the past 4 weeks to . . ."). There is also a version available for "usual health," where participants are asked about their usual health (e.g., "Describe your usual ability to . . .").

Examples of use. In rheumatology, the HUI3 has been used to assess HRQOL in patients with rheumatoid arthritis (188,189), changes in HRQOL in patients with juvenile idiopathic arthritis (190), and to assess the effectiveness of hylan G-F 20 in treatment of knee osteoarthritis (191).

Practical Application

How to obtain. The HUI2 and HUI3 classification systems can be viewed online at http://www.healthutilities.com (185). The questionnaires and user manuals are only distributed under license from HUInc. The cost of HUI3 is $4,000 per study for the questionnaire and the matching user manual (185).

Method of administration. The 15Q version of the HUI3 is designed to be self-administered, while the 40Q version can be interviewer-administered (by telephone or face-to-face), on paper or using a computer (184).

Scoring. The functions to derive the scores are multiplicative and based on classical utility theory. The scoring manual contains decision tables showing all possible combinations of responses per attribute. Typically, scoring is done using a common statistical package such as SPSS or SAS. A spreadsheet such as Excel can be used, but it is not recommended by HUI developers if there are more than a few subjects and/or multiple assessment points. The decision tables of response combinations are used to determine the health-state level for each health domain and then, using the tables and the scoring algorithm, the utility scores for all attributes of health and the overall HRQOL score can be determined.

Missing responses are scored as 0. At the same time, the presence of missing responses is problematic, since at least 2 scores (1 domain and the overall score) will be missing for each subject that has 1 response missing. Nonresponse to an item on the 40Q will also cause problems with the skip pattern, making the questionnaire difficult to score.

Score interpretation. The score range for HUI3 is −0.36–1.00 and −0.03–1.00 for HUI2. A score of 1.00 signifies perfect health and 0.00 represents death. HUI allows for negative numbers for health states considered worse than death. Population normative data are available from numerous large general population surveys. Normative values by age (15+, 17+, 18+, 20−85 and 35–89 years), race ("unselected," "Hispanic non-Black," "Black non-Hispanic," "non-Black and non-Hispanic") and country (Canada, USA) are available on the HUI web site (185).

Respondent burden. HUI3 generally has low responder burden. The mean time to complete the 15Q is 5–10 minutes (151). The 40Q, which has a built-in skip pattern takes 3 minutes to complete (184).

Administrative burden. Administration burden for the HUI3 is moderately high. Interviewer administered assessments will require interviewer training, especially for the 40Q version of the HUI3. It is also recommended to review completed 15Q version questionnaires once received and to contact the respondent if there are missing answers. Scoring will require basic knowledge of statistical software.

Translations/adaptations. HUI3 was first developed in English and is now available in more than 35 languages worldwide. It has been used successfully without modification in Canada, the UK, the US, and Australia. There are 16 variations of the HUI questionnaires, which are dependent on mode of administration (self-complete or interviewer-administered), recall period (past week, 2 weeks, 4 weeks, or usual health), and assessment viewpoint (self or proxy). One or more variations of the HUI questionnaires are available in Afrikaans, Chinese (traditional and simplified characters), Croatian, Czech, Danish, Dutch, Finnish, Flemish, French (continental or European French and French-Canadian), German, Hebrew, Hungarian, Italian, Japanese, Korean, Malay, Norwegian, Polish, Portuguese (European and Brazilian), Romanian, Russian, Serbian, Slovak, Spanish (European and Mexico, Latin and South American), Swedish, Thai, and Turkish. Other versions in preparation include Serbian.

Psychometric Information

Method of development. The original items for the HUI were generated from the work of Cadman et al (192) (note, we have been unable to access this original work from 1986), who sought to determine the most important attributes of HRQOL based upon clinical experience. A random sample of adults from the general population then ranked these attributes on their desirability (193); this information was subsequently used to derive weights for the HUI2 and HUI3.

Acceptability. Given the complexity of some of the response option descriptors within domains of the HUI3, the 15Q and 40Q (with simplified response options) have been developed to make it easier for participants (or the interviewer administering the HUI3) to select an appropriate response option descriptor. Missing data on HUI3 in rheumatology studies are not frequent. A study among 114 rheumatology outpatients found that there were no missing responses at a baseline face-to-face assessment on HUI3 administered by a trained nurse interviewer. In a telephone-based followup interview 2 weeks later, <5% of the respondents had missing data (194). Similarly, floor and ceiling effects are not commonly encountered in rheumatology populations. Only 4 subjects (3.5%) in the above study obtained the highest possible health rating.

Reliability. Results for reliability of the HUI vary considerably. Cronbach's alpha ($\alpha = 0.71–0.79$) was reported for the Spanish version of the HUI3 in the general population (195). For a cohort of heart-failure patients, Cronbach's alpha for the total score of the HUI3 was reported as $\alpha = 0.51$ (196). We have not been able to identify any studies that assessed internal consistency of the HUI3 in rheumatology-specific populations. Test–retest reliability of interviewer-administered HUI3 in a study of 114 rheumatology outpatients was intraclass correlation coefficient (ICC) 0.75 (95% confidence [95% CI] 0.65, 0.83) over a 2-week period (194). However, these results are difficult to interpret, since the first interview was done face-to-face while the second interview took place over the telephone.

More favorable results were obtained in a study of 50 rheumatoid arthritis patients (randomly selected from a larger study), where 3 months test–retest reliability of HUI3 was found to be acceptable (ICC 0.81, 95% CI 0.66, 0.90) (197). Similarly, in a stratified random sample of people completing the Canadian General Social Survey (n = 506), the test–retest reliability for the HUI3 of ICC 0.77 was recorded for telephone assessments conducted 1 month apart (198).

Validity. The results of studies investigating construct validity of the HUI3 are mixed. An observational study of 144 rheumatology outpatients (194) found that HUI3 did not discriminate between people with and without chronic health conditions. Despite the hypothesized high to moderate correlations, the correlations between HUI3 and Short Form 36 (SF-36) scores were in a low to moderate range ($\rho = 0.29-0.49$, $P < 0.01$ for all), with the SF-36 physical functioning and bodily pain scales showing the lowest and highest correlations with the HUI3 score, respectively. When compared to the EuroQoL 5-domain (EQ-5D) instrument, median EQ-5D and HUI3 scores were very similar. The correlation between EQ-5D and HUI3 scores for all patients was $\rho = 0.45$ for baseline interviews and $\rho = 0.57$ for followup interviews (Spearman's rho, $P < 0.001$ for both).

On the other hand, a study of 114 osteoarthritis patients on the waiting list to see an orthopedic surgeon (199) found support for construct (convergent and discriminant) validity of both HUI2 and HUI3. Of the 87 a priori hypotheses examined, 75% were confirmed by zero-order correlations, suggesting that the constructs within the HUI2 and HUI3 were, in general, related to similar constructs in other conceptually related measures (SF-36, Harris Hip Scale [HHS], Western Ontario and McMaster Universities Osteoarthritis Index [WOMAC], McMaster Toronto Arthritis Patient Preference Questionnaire, the State-Trait Anxiety Inventory, and the 6-Minute Walk Test).

The HUI3 was used in the 1990 Ontario Health Survey and found to be able to differentiate people with stroke or arthritis from those who had neither of these conditions (200). The highest mean score, indicative of best health, was for people without a history of arthritis or stroke (0.93), followed by those who had arthritis (0.77) and stroke (0.54).

Ability to detect change. Results for the ability of the HUI3 to detect change are also mixed. In a study of 99 patients on a waiting list for total hip arthroplasty who had completed the HUI3 before and after the surgery (201), the HUI3 showed improvement in the overall summary score and various domains following surgery. There was a large standardized effect size (ES) for the overall summary score (1.19) and pain (1.30), and a moderate ES for ambulation (0.56). There was no change in vision, hearing, speech, dexterity, and cognition, which would be expected in this population. Although the HUI3 was not as responsive to change after total hip arthroplasty as the disease specific measures considered in the same study, (HHS, WOMAC), it was the most responsive of the generic measures considered (SF-36, EQ-5D, and HUI2).

Less favorable results for the responsiveness of the HUI3 were obtained in a study of 320 rheumatoid arthritis patients recruited from private rheumatology practices (197). The study compared responsiveness to change over time (disease progression) of a number of generic HRQOL measures (HUI2, HUI3, SF-6D, EQ-5D) as well as some disease-specific measures (the Health Assessment Questionnaire Disability Index and the Rheumatoid Arthritis Quality of Life Questionnaire). The HUI3 appeared to be poorly responsive to deterioration but was able to identify those classified as "better" on global assessment of disease severity at 3- and 6-months followup. Of all the measures used, the HUI3 and SF-6D were found to be the most responsive between baseline and 6 months for measuring improvement ("worse" HUI3 = ES −0.10, 95% CI −0.31, 0.13; "same" HUI3 = ES 0.12, 95% CI −0.03, 0.26; "better" HUI3 = ES 0.23, 95% CI 0.08, 0.41) (197).

Information on minimum clinically important difference (MCID) for HUI3 in rheumatology is limited. In a study with individuals who had stroke or arthritis, drawn from the 1990 Ontario Health Survey, MCID on the HUI3 was defined as a difference of 1 level within HUI3 attributes, which equates to a change of ≥0.03 units in the HUI3 score. More generally, Drummond reported that a difference of ≥0.03 in mean HUI overall HRQOL scores were clinically important, and differences as little as 0.01 may be meaningful and important in some contexts (202). However, it is not clear how these values were derived.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The instrument appears to measure some aspects of quality of life that are affected by rheumatic diseases, although there are several items (e.g., hearing) within the scale that are not relevant to this field. As a utility measure, HUI3 can be used in health economic analyses. The instrument appears to be sensitive to positive changes brought about by some treatments for rheumatic conditions (e.g., hip replacement). However, it appears to be poorly responsive to deterioration, and therefore may not be suitable for individual followup. This instrument appears to be widely applicable to most patient populations; however, research to date in rheumatology has been primarily in rheumatoid arthritis and total hip replacement populations.

Caveats and cautions. This instrument does not appear to be as sensitive to change brought about by treatment of disease as other disease- or joint-specific instruments. There may be difficulties using this instrument among older adult populations or persons with cognitive impairment due to the complexity of some of the items. The psychometric robustness of the HUI3, especially its temporal stability and construct validity, have also received mixed support in rheumatology.

Clinical/research usability. The interpretation of HUI3 scores in clinical settings is hampered by the lack of information on cutoff scores for what is considered to be meaningful change in HRQOL for patients with rheumatic conditions. For example, there is no cutoff threshold indicative of when joint replacement may be required or whether such surgery has been successful at improving an individual's HRQOL. The HUI appears to have a moderate administrative burden, although use of the computerized scoring algorithm may compensate for the extra interviewer training necessary for the administration of the 40Q version. Respondent burden does not appear to be a problem that would limit clinical or research use. The cost of the questionnaire and scoring algorithms may limit the use of the HUI3 for clinician-initiated unfunded research projects.

QUALITY OF WELL-BEING SCALE (QWB)

Description

Purpose. The QWB scale was developed more than 30 years ago as a measure of health-related quality of life (HRQOL) in the general population (203). The QWB is a preference-based measure that combines functioning and symptoms to produce a well-being index ranging from 0 (death) to 1 (full, symptom-free functioning) (204). The QWB can also be used to calculate quality-adjusted life years, which combine life expectancy with HRQOL to produce a summary measure of quality and quantity of life lived.

The QWB was initially developed for interviewer administration, but the use of this measurement tool has been low due to its length and difficulty of administration (205). The self-administered version, Quality of Well-Being Scale-Self Administered (QWB-SA) was developed to address these limitations of the QWB. The QWB-SA was released in 1997 (204). This review focuses on the QWB-SA version of the instrument.

Content. The QWB-SA includes a wide range of physical and mental symptoms that people might experience in daily life (205). The symptoms assessed by QWB-SA reflect different aspects of health and cover different degrees of severity. Most items focus on a specific problem related to one body system, such as visual problems (e.g., blindness) or central nervous system functioning (e.g., paralysis).

The QWB-SA has 5 parts, including a symptoms checklist and 4 function sections. The symptoms section incorporates assessment of chronic (e.g., speech problems, physical deformities) and acute symptoms. Acute symptoms include physical (e.g., headache, pain) and mental health symptoms (e.g., sadness, anxiety). The function sections of the QWB-SA include self-care, mobility (including use of transportation), physical activity (e.g., climbing stairs), and usual activity (e.g., work, home, or recreation) (205).

Number of items. The QWB-SA consists of 74 items. The symptom checklist has 58 symptoms, including 19 chronic symptoms, 25 acute physical symptoms, and 14 mental health symptoms. Self-care is assessed by 2 items, the mobility and usual activity sections have 3 items each, and the physical activity section has 8 items.

Response options/scale. The presence/absence of 19 chronic symptoms is measured on a dichotomous scale (yes/no), with participants asked to indicate whether they are currently experiencing any of the symptoms or problems listed. For the remaining items, participants are asked to indicate which days over the past 3 days they experienced each of the health problems listed, using a 4-point scale with response options including "no days," "yesterday," "2 days ago," and "3 days ago." Respondents are able to select more than one response option if they experienced the symptom on more than one of the days (for example, yesterday and 3 days ago). Responses are scored according to the number of days that a health problem was experienced (0, 1, 2, or 3).

Recall period for items. With the exception of the chronic symptoms section, QWB-SA asks patients about symptoms and function over 3 days prior to the day of administration. The format of the chronic symptoms questions does not use the 3-days recall period since it is expected that chronic conditions do not vary much over the 3-day assessment period (205).

Endorsements. Approved by the Scientific Advisory Committee of the Medical Outcomes Trust (http://www.outcomes-trust.org/instruments.htm).

Examples of use. In rheumatology, the QWB/QWB-SA had been previously used to measure HRQOL in osteoarthritis (206), to measure the impact of total hip or knee replacement on HRQOL (81,207), and to assess the impact of an active drug treatment relative to placebo on HRQOL in a randomized controlled trial in rheumatoid arthritis (208).

Practical Application

How to obtain. An inspection copy of the QWB-SA can be obtained from https://hoap.ucsd.edu/qwb-info/. For nonprofit organizations, the scale and scoring instructions are available free of charge, although the researchers are required to sign a copyright agreement with the Health Services Research Centre (HSCR), the University of California, San Diego. Profit organizations are required to purchase a yearly license at $1,000 per year, with an additional charge of $0.25 for each questionnaire administered.

Method of administration. The QWB-SA was designed for self-administration and is available in paper and pencil or web-based formats. The QWB-SA can also be administered by telephone or in a face-to-face interview, although the psychometric properties of these methods of administration have not been specifically studied (205).

Scoring. The QWB-SA requires computerized scoring. A scoring algorithm (SPSS syntax) is available for purchase from the scale developers for $240. The QWB-SA scoring algorithm assumes that missing responses are equivalent to the absence of a problem.

Score interpretation. Symptoms and the 4 function scores are combined into a total preference-weighted score of well-being that ranges from 0 (death) to 1.0 (symptom-free, optimal functioning). Normative data are available for clinical and nonclinical samples by age, sex, and ethnicity. However, these normative data, especially for nonclinical samples, are based on relatively small numbers of participants, with a total normative sample of 843 people. Participant numbers across subgroups range from 1 (e.g., Native Americans age ≤30 years) to 235 (whites age ≥71 years) (205). Another recent study also presents means and SEs for QWB-SA scores derived from a probability sample of 3,844 US adults ages 35–89 years by sex and 5-year age groups (169).

Respondent burden. The QWB-SA takes ~10 minutes to complete in paper and pencil format. Completion instructions are self-explanatory. In a telephone-administered interview of 3,844 US residents, completion time for the QWB-SA varied from 7.7 to 17.5 minutes with an average of 11.1 minutes (169).

Administrative burden. Administration instructions for the paper and pencil version of the QWB-SA are self-explanatory. Scoring requires access to a computer. Apart from some knowledge of SPSS statistical software, no specific training is required for administration and scoring of the QWB-SA. The QWB-SA form is designed for optical scanning, and the HSRC also provides data cleaning, as well as entry and scoring services for the QWB-SA for $57 per hour.

Translations/adaptations. The instrument is available in English, German, French, Dutch, Italian, and Spanish. Translations to other language are available upon request, with fees determined by the languages requested and project timelines/needs.

Psychometric Information

Method of development. The original, interviewer-administered version of the QWB was developed for the purpose of defining "the universe of all possible health states between optimum function and death" (203). Items for the inclusion into the QWB were generated from specialty-by-specialty review of medical reference works. The initial tool included the assessment of 3 dimensions of functioning, reflecting different levels of mobility, physical activity, and social activity, as well as 36 different health symptoms. The 3 dimensions generated 100 theoretical combinations of health states, of which 43 were observed in a pragmatic study of more than 10,000 people. Open-ended questions administered to the observational sample identified no additional health states or symptoms (203).

In the development of the QWB-SA, the symptom checklist was expanded to 58 symptoms, including at least 12 mental health symptoms (205). These additional symptoms were identified through focus groups conducted with physicians. Preference weights for the QWB-SA were derived from 435 English-speaking adults drawn from primary care clinics and college campuses in San Diego, California. Participants were presented with descriptions of hypothetical health states defined by the scale items and asked to provide numerical ratings (on a scale of 0–100) for how undesirable each health state was. These ratings were analyzed with regression analysis using levels of functioning and symptoms as predictors. Regression coefficients were subsequently used to generate weights for the scale scores (205). No specific patient groups were involved in item and weight generation for either the QWB or the QWB-SA.

Acceptability. Readability of the QWB-SA could potentially be problematic, as the scale contains words and phrases that might not be commonly understood by people with lower education levels (e.g., pelvic cramping, usual activities). The sentence structure of the QWB-SA is also rather complicated, with each item containing several concepts (e.g., "Because of any physical or emotional health reasons, on which days did you avoid or feel limited in doing some of your usual activities, such as visiting family or friends, hobbies, shopping, recreational, or religious activities."). The complicated wording and sentence structure of the QWB-SA could potentially lead to difficulties with understanding the meaning of the question, as well as difficulties with the selection of the appropriate response option, especially when used with the elderly or unwell individuals.

In a study conducted in Germany with 264 rehabilitation inpatients with musculoskeletal (n = 106), cardiovascular (n = 88), or psychosomatic disorders (n = 70), no missing data were observed for the QWB-SA (209). For comparison, the proportion of missing data on other generic HRQOL measures used in the same study was 1.3% for the EuroQol 5-domain (EQ-5D) measure and 15D, 1.9% for the Health Utilities Index Mark 3 (HUI3), and 6.1% for the Short Form 6D (SF-6D). In a random sample of the US general population (n = 3,844) (169), the proportion of missing data for the QWB-SA was 2.2%, 0.7% for the EQ-5D, 7% for the HUI3, and 2.7% for the SF-6D.

In a population-based sample of 293 adults age ≥65 years, the proportion of missing data on QWB-SA items ranged from 0.3% (hearing and skin problems items in symptoms section) to 14.6% for loss of sexual interest or performance (also a symptoms section item) (136). Nearly 50% of all respondents skipped at least 1 symptom on the QWB-SA 3-day recall section, with the mean number of missing items being unrelated to age, but higher in men than in women.

The QWB-SA appears to have good range of score functioning, with no floor or ceiling effects observed in a random sample of the US general population (n = 3,844) (169), as well as in a sample of patients from Germany with musculoskeletal conditions (n = 106) (209). While none of the HRQOL measures used in the second study showed evidence of floor effects, 5.7% of patients obtained the maximum EQ-5D score (ceiling effect), while 2 patients (1.9%) achieved the maximum possible score on the 15D and 1 patient (0.9%) on the SF-6D (209).

In a study of performance of the QWB-SA in 293 people age ≥65 years, the scale appears to have been well received by the respondents, with 60% reporting that they were very or somewhat satisfied (95% confidence interval 54.2–65.4%) with the scale. The satisfaction ratings for the QWB-SA were similar to those for the Sickness Impact Profile (SIP) (69% were very or somewhat satisfied) and the SF-36 (67% were very or somewhat satisfied) (136).

Reliability. Reliability of the QWB-SA has not been well studied in general (210), and there appears to be no published reliability data for rheumatology populations. In other clinical populations, the QWB-SA was reported to have low temporal stability, with an intraclass correlation coefficient (ICC) of only 0.59 between 1- and 6-months postoperative scores of 265 cataract surgery patients (211). However, it is possible that during the 5-months followup, real changes in the individual's HRQOL might have occurred. In an earlier study of 218 adults with stable health conditions recruited from primary care clinics, QWB-SA scores were only moderately stable over a 1-month test–retest period (Pearson's r = 0.77) (204).

Validity. The QWB-SA appears to have good face validity, with items appearing to capture health-related symptoms. Although the content validity of a measure is influenced by the nature of the construct that is being measured, the original version of the QWB was reported to have good content validity for capturing health-related symptoms. In a sample of more than 10,000 people drawn from a variety of clinical settings (203), open-ended questions (designed to elicit additional information about health-related problems that people might experience in daily life) yielded no health states or symptoms in addition to those already listed in the scale. The involvement of physicians into focus groups during QWB-SA construction (to identify aspects of health that are understood by physicians to be signs/predictors of various diseases) increased the likelihood that the scale has good content validity for use in clinical settings. The QWB-SA also appears to have good ability to capture the full range of HRQOL impairment as indicated by no or low floor/ceiling effects in the general population and musculoskeletal patients (169,209). Although the scale also contains items that are not specifically related to rheumatic conditions, retention of these items is justifiable since they represent part of HRQOL and are potentially relevant indicators of the overall well-being of people with rheumatic conditions.

Criterion-related validity of HRQOL questionnaires is difficult to establish due to the absence of a "gold standard" measure of HRQOL. Evidence for the construct validity of QWB-SA in rheumatic conditions is generally positive. While agreement between QWB-SA and other generic measures of HRQOL in musculoskeletal patients was reported to be poor to moderate (with an ICC ranging from 0.26 for agreement between QWB-SA and EQ-5D and 0.48 for agreement between QWB-SA and 15D [209]), in a community sample of older adults, QWB-SA was found to have moderate correlations with physical health components of the SIP and SF-36 ($r = \geq 0.42$) and weaker correlations with the SIP psychosocial dimension and the SF-36 summary mental health score (136). All correlations were of expected magnitude and direction.

Support for the construct validity of QWB-SA in musculoskeletal conditions was also provided by a report of significant correlations with the scores on arthritis-specific measures (Rapid Assessment of Disease Activity in Rheumatology [RADAR], Arthritis Impact Measurement Scales, and the Health Assessment Questionnaire [HAQ]) (212). The correlations were in the low ($r = -0.28$ for QWB-SA with RADAR) to moderate ($r = -0.62$ for QWB-SA with HAQ) range. However, while correlations were in the hypothesized direction, the authors did not provide specific predictions about the strength of the expected correlations, which makes it difficult to draw robust conclusions about the convergent and discriminant validity of the QWB-SA in musculoskeletal diseases.

The QWB-SA also appears to have good ability to differentiate patients with and without musculoskeletal conditions and between severity levels of musculoskeletal conditions, further supporting its construct validity. Patients with arthritis ($n = 334$) were reported to have significantly lower QWB-SA scores and significantly higher HAQ scores than those without arthritis ($n = 562$) (212). In another study, QWB scores were sensitive to different levels of osteoarthritis severity (206), although no effect sizes (ES) for the magnitude of the differences in QWB scores for different levels of osteoarthritis severity have been provided.

Ability to detect change. Information on the ability of QWB-SA to detect change in rheumatic conditions is limited. QWB, on the other hand, was reported to be sensitive to changes in HRQOL of people with osteoarthritis following education and self-management intervention (standardized response mean 0.24) (206). In another study, the QWB also had modest ability to detect change in the health status of 330 patients with rheumatoid arthritis, with only a small standardized ES recorded (0.23) following pharmaceutical treatment, although this was similar to the ES found for other measures used in the same study, including the HAQ (ES 0.25) and tender joint count (ES 0.24) (208). There appears to be no published data on minimal clinical important differences for either the QWB or QWB-SA in either rheumatic populations or broader literature.

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. The QWB-SA offers comprehensive coverage of health state levels, with no compelling evidence of floor or ceiling effects, which makes this scale potentially useful across a broad range of HRQOL impairment levels. There is also good evidence for the construct validity of the QWB-SA in musculoskeletal conditions, although its appropriateness in specific disease groups and in various treatment interventions awaits further evaluations.

Caveats and cautions. Limited information is currently available on psychometric properties of the QWB-SA or QWB in musculoskeletal conditions. Since no patient samples were involved in the development or weighting of scale items, the relevance of different health states to different clinical populations is not known. In addition, the score range on QWB-SA does not allow for health states worse than death, which might make this instrument insensitive for measuring very poor health. The scoring algorithm assumes that missing responses are equivalent to absence of a problem; however, validity of this assumption is not certain. Normative scale values are only available for the US, and further studies are required to develop cross-cultural norms as well as norms for clinical populations. While there appears to be a substantial amount of evidence that support the construct validity of the QWB-SA, most psychometric evaluation studies were carried out by the scale developers, therefore further inquiries into psychometric properties of the QWB-SA by independent groups are warranted.

Clinical/research usability. Overall, the QWB-SA appears to have good support for its construct validity in rheumatic conditions, which supports its use in clinical and research settings. However, given the limited evidence for the reliability of this scale, information on its validity needs to be interpreted with caution. While the scale could potentially be useful for comparing HRQOL in rheumatic conditions with other populations (clinical or general), the complicated wording and sentence structure may limit the utility of this scale in clinical settings, where individuals may be expected to be unwell or with those who are elderly or have low levels of education. A complicated scoring system may further limit the use of the scale in clinical settings. Absence of appropriate norms and lack of information on reliability, ability to detect change, and minimal detectible change could also potentially limit the use of scale in clinical and research settings due to the difficulty with interpreting change in scale scores.

ASSEMSSMENT OF QUALITY OF LIFE (AQoL)

Description

Purpose. The AQoL instruments are multi-attribute utility measures of health-related quality of life (HRQOL) (213). In a similar way to the other utility measures (EuroQol 5-domain [EQ-5D], Short Form 6D [SF-6D], and Health Utilities Index Mark 3 [HUI3]) the AQoL was designed for use across health conditions to enable health economic evaluation studies. The AQoL allows assessment of the impact of interventions on HRQOL, comparing HRQOL in different populations and disease settings, and monitoring longitudinal changes in a broad range of health conditions. The AQoL was originally published in 1999 (214), with 4 versions developed to date: AQoL-4D (the original version), AQoL-6D (with additional elements of pain and coping), AQoL-7D (with emphasis on vision), and AQoL-8D (with emphasis on mental health) (213). This review focuses on the original version, AQoL-4D since it is the one that has been previously used in rheumatic diseases. Where possible, we also review information on the AQoL-6D due its potential relevance in rheumatology settings. A version with 8 items has also been published (215) although this version has not undergone specific validation studies in musculoskeletal conditions.

Content. The AQoL-4D covers 4 domains of independent living, mental health, relationships, and senses. The AQoL-6D has 2 additional domains of coping and pain.

Number of items. The AQoL-4D has 12 items, with 3 items per dimension. The AQoL-6D has 20 items; the additional dimensions of coping and pain have 4 items each.

Response options/scale. The AQoL items have variable numbers of response levels, ranging from 4–7. Response options are on a Guttman scale, with higher scores indicative of progressively higher levels of disability. A visual analog scale version of the AQoL is also available.

Recall period for items. The AQoL asks respondents to evaluate their health state over the previous week.

Examples of use. In rheumatology, the AQoL has been previously used in a probability sample of the general population to compare the HRQOL of people with arthritis to those who have no arthritis (216), to assess the HRQOL of people on a waiting list for joint replacement surgery (217), to evaluate the impact of self-management (218,219) and exercise-based interventions on HRQOL in arthritis (220–222), as well as in a randomized controlled trial of vertebroplasty for osteoporotic vertebral fractures (223).

Practical Application

How to obtain. The AQoL questionnaires and scoring algorithms are available at no cost from http://www.aqol.com.au/. However, the use of the AQoL is subject to copyright restrictions and the users are asked to complete a registration form (using web-based or paper format).

Method of administration. The AQoL can be self- (paper and pencil or online) or interviewer-administered. The agreement between self- and interviewer-administered (by telephone) versions of the AQoL was high with an intraclass correlation coefficient of 0.83 (95% confidence interval [95% CI] 0.76–0.88), with the 2 versions producing comparable mean scores (224). However, in another study, the correlation between mail and telephone administration of the AQoL was only 0.66, indicating that different methods of AQoL administration should not be used interchangeably (225).

Scoring. The AQoL instruments can be used to obtain an overall utility score as well as to separate scores for each dimension. The health states described between the items are initially weighted using values obtained from the general population from Time Trade Off interviews, a common procedure in the health economics field. The scores across the scales are combined using a multiplicative scoring procedure. Scoring algorithms are available from the AQoL web site (www.aqol.com) in SPSS and STATA readable formats. The AQoL developers also provide an online scoring service for their questionnaires. The scoring algorithm allows for only 1 missing value per dimension for dimensions with 3 or 4 items and 2 missing values per dimension for longer scales. Missing values are imputed from the mean of the nonmissing items in the dimension (213).

Score interpretation. The AQoL utility score ranges from −0.04 (health state worse than death) to 0.00 (death) and 1.00 (full health) (226). Normative values, broken down by age (in 10-year age groups) and sex, are available for AQoL-4D from the AQoL web site (http://www.aqol.com.au/documents/AQoL-4D-Population-Norm.pdf). The norms have been derived from a probability sample of 3,010 Australian residents (213).

Respondent burden. The AQoL has a low respondent burden. The scale developers estimate completion time for the AQoL-4D to be 1 to 2 minutes, although a more realistic estimate for a 12-item questionnaire that uses the Guttmann response scale might be 5–10 minutes, which is still quite low (213). Completion instructions are self-explanatory and easy to follow. The questionnaire uses simple language and is easy to understand and complete. The developers reported that in interview settings, ~2% of respondents tend to seek clarification about an item or a response option. Detailed information about items for which clarification is commonly sought can be found in the user manual (225), which can be downloaded from http://www.psychiatry.unimelb.edu.au/centres-units/cpro/aqol/instruments/AQoL_User_Manual.pdf. Some items describing poor HRQOL were also found to be distressing for some participants (214).

Administrative burden. The AQoL appears to have low administrator burden. Administering AQoL by interview requires basic training in interviewing technique. The use of the computerized scoring algorithm requires basic knowledge of statistical software.

Translations/adaptations. No translations or adaptations of the AQoL were identified at the time of preparing this review.

Psychometric Information

Method of development. The conceptual model for the initial version of the AQoL was based on the World Health Organization's definition of health. The 2 major sources of items for the AQoL were focus groups of clinicians and the review of the content of existing HRQOL questionnaires. No patients took part in item generation. The 61 draft items of the AQoL were administered to a sample of 255 individuals recruited from community and hospital settings. The final selection of items to be included in the AQoL-4D was made based on exploratory and confirmatory factor analysis and reliability analyses (214). The additional items for the later version of the AQoL were developed from focus groups with clinicians and review of existing questionnaires (213).

Acceptability. The AQoL appears to have high acceptability overall. In community-based studies, the proportion of missing data varies from <1% (for either self- or interviewer-administered) (224) to 2.5% (self-administered version) (226). The ability of the AQoL to adequately cover the full range of HRQOL states appears to be good in rheumatology, with no floor or ceiling effects recorded in a sample of 222 osteoarthritis patients recruited from clinical and community settings (227).

Reliability. Internal consistency (Cronbach's alpha) of the AQoL utility score is good and is generally reported to be ~0.80 in samples consisting of hospital patients and community-dwelling adults (225,226). Although the 3-item domains of the original version of the AQoL were reported to have much lower internal consistency estimates, with coefficients ranging from 0.52 (psychological well-being) to 0.77 (independent living) (214), the AQoL was intended to be used primarily as an overall utility score, rather than as single domain scores.

Information on test–retest reliability of the AQoL is currently limited. The user manual reports test–retest reliability, measured by Pearson's correlation coefficient, as 0.80 (225). However, Pearson's correlation coefficient tends to be a poor indicator of temporal stability, due to the insensitivity to systematic (rather than random) changes over time. Systematic differences in questionnaire scores over time could occur for a variety of reasons, including real change in a health state, change in internal frame of reference for the severity of one's health condition (response shift), reactivity, or learning effect. Evidence indicates that the AQoL may be subject to such systematic biases. Over repeated administrations, the AQoL-6D scores were somewhat higher for the second administration, which suggests that the individuals tend to re-apprise the severity of their condition after some reflection (228).

Validity. The AQoL instruments cover a broad range of health domains, not all of which (e.g., vision) are relevant to rheumatology. Nonetheless, these domains represent important elements of overall generic HRQOL and permit comparisons across diseases and populations. The AQoL appears to have good face and content validity for measuring HRQOL, although content validity is largely dependent upon the nature of the construct being measured. The absence of floor or ceiling effects in osteoarthritis further supports content validity of the AQoL in rheumatology since it indicates that the AQoL is able to adequately capture the full range of HRQOL experiences in this population (227). Criterion-related validity of HRQOL measures is difficult to establish due to the absence of a "gold standard" for measuring HRQOL.

Evidence for construct validity of the AQoL is good, with results thus far supporting its factorial, convergent discriminant, and known-groups validity. Factorial structure of the AQoL-4D, including the 4 first-order factors and 1 higher-order factor, was examined in the initial construction study (214) using confirmatory factor analysis, with no evidence of misfit between the hypothesized model and the data. At least one other study each subsequently supported the 4-dimensional structure of the AQoL-4D using exploratory factor analysis (229) and the 6-dimensional structure of the AQoL-6D (228). While these results provide strong support for the factorial validity of the AQoL, it should be noted that none of these studies were specifically concerned with rheumatology populations.

In rheumatology settings, convergent validity of AQoL-4D was tested in a study of 222 individuals with osteoarthritis (227), where AQoL utility had high to moderate correlations with the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) scales (r = −0.51, −0.63) and the Lequesne Index (r = −0.76). All correlations were of hypothesized magnitude and direction. More broadly, in a sample of 606 individuals drawn from community and hospital settings, correlations between the AQoL-6D and other generic measures of HRQOL, including the HUI3, EQ-5D, 15D, and the SF-36 were 0.73 or higher (228), indicating good convergent validity. The AQoL-4D utility scores also correlated well with health care costs in an 18-month followup of more than 1,500 individuals with a range of chronic conditions. While these results support convergent validity of the AQoL-4D, less is known about its discriminant validity, which needs further study.

The AQoL has good ability to differentiate between people with and without rheumatic conditions, as well as between severity levels in rheumatic conditions. In a large probability sample of the general population (n = 2,840), the AQoL-4D was able to differentiate people with chronic joint conditions (self-reported doctor-diagnosed arthritis and chronic joint symptoms) from those who had no joint problems, with the lowest mean AQoL scores for arthritis group (mean 0.72; 99% CI 0.70–0.74), followed by chronic joint symptoms group (mean 0.75; 99% CI 0.72–0.78), and those who had no joint problems (mean 0.85; 99% CI 0.84–0.87) (215). The AQoL-4D was also able to differentiate between severity levels of osteoarthritis, with the utility score exhibiting moderate effect size (ES) of 0.66 for the difference in HRQOL between people with osteoarthritis recruited from the general community and those who were on a waiting list for joint replacement surgery for their osteoarthritis (227). Similar results were reported in at least one other study (217). More broadly, in a sample of

996 individuals selected to cover a very broad range of health conditions from those who were healthy to those who were terminally ill, the AQoL was reported to have better ability to differentiate between the levels of HRQOL impairments than other utility instruments, including HUI3, EQ-5D, 15D, and SF6D (230).

Ability to detect change. The ability of the AQoL to detect change in rheumatic populations has not been well studied. More generally, a minimum clinically important difference (MCID) of 0.06 for the AQoL-4D utility score had been recorded for self-reported change in health state (226). This finding was based on the results of 4 longitudinal studies (with approximate followup time of 12 months), 2 of which were community trials of coordinated care for people at risk of hospitalization, 1 involved a followup of community-dwelling elderly people, and 1 was an evaluation of health services for acute conditions in a hospital emergency department. As this study did not specifically target individuals with rheumatic conditions, transferability of this finding to rheumatology settings is currently not known.

The results for the ability of the AQoL to detect treatment effects in rheumatology settings are mixed. In a randomized controlled trial of the efficacy of physiotherapy and exercise program for chronic rotator cuff disease, the mean change in AQoL-4D utility score following 22 weeks of treatment was 0.00 (SD 0.20) (220). At the same time, condition-specific measures of pain and movement (assessed using the Shoulder Pain and Disability Index) showed large improvements during the course of intervention (standardized response mean 0.90 for movement and 1.05 for pain). Nonetheless, in the same study the AQoL-4D was able to distinguish between intervention and placebo groups at 22 weeks of followup, with significantly higher scores recorded for the intervention group (mean difference 0.07; 95% CI 0.04–0.10). In a randomized controlled trial of self-management intervention for people on a waiting list for joint replacement surgery, the intervention group had a slightly higher AQoL utility score at the end of the study (ES 0.21) (218). Although the improvement was small and not statistically significant ($P = 0.23$), similar results were obtained on the WOMAC (ES 0.09, 0.36, and 0.26 for pain, stiffness, and physical functioning scales, respectively).

Critical Appraisal of Overall Value to the Rheumatology Community

Strengths. As with all the HRQOL scales, the AQoL covers a range of issues important to rheumatology. The AQoL appears to have good ability to differentiate between people with and without arthritis and between the levels of arthritis severity. Overall, the evidence supports the use of the AQoL when comparisons with the general population are required. The ability of the AQoL to detect treatment effects is promising but requires further research in a broader range of interventions with treatment effects of known magnitude.

Caveats and cautions. Only a handful of studies examined the psychometric properties of the AQoL in rheumatic conditions, with generally positive results. However, the more definitive conclusions about the psychometric robustness of this questionnaire in rheumatology await further investigations.

Clinical/research usability. The AQoL is a relatively new instrument for rheumatology, and information about its psychometric properties is still accumulating. The questionnaires have low respondent and administrator burden and are available at no cost, which greatly enhances their usability in clinical and research settings. Availability of population norms also provides context for score interpretation, which further facilitates the usefulness of the AQoL. However, only Australian norms are currently available and cross-cultural applicability of these norms is currently not known. Usability of the AQoL in different countries is also affected by the lack of AQoL in languages other than English. Like all generic HRQOL tools designed to generate utilities, it is unlikely to detect small clinical changes but should be useful for comparison with other diseases and for health economic appraisals such as cost utility assessments.

DISCUSSION

The results of this review indicate that there is currently no single "best" measure of general health and health-related quality of life in rheumatology, with psychometric weaknesses identified in all measures considered. Although this review also identified several gaps in the information available on measurement properties of the reviewed questionnaires, the available evidence identifies the Sickness Impact Profile (136) as the worst performing measure, with relatively high administrative burden and questionable reliability of subscale scores. At the other end of the spectrum is the Assessment of Quality of Life Scale, with very low administrative burden and good evidence of reliability and validity thus far, indicating that it is a promising measure. The results of this review suggest that there is an urgent need for systematic investigations of the psychometric properties of many instruments currently used to assess health and health-related quality of life in rheumatology.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published.

REFERENCES

1. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36) I. Conceptual framework and item selection. Med Care 1992; 30:473–83.
2. 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;34:220–33.
3. Ware JE. SF-36 Health Survey update. In: Maruish ME, editor. The use of psychological testing for treatment planning and outcomes assessment. Mahwah (NJ): Lawrence Earlbaum; 2004. p. 693–718.
4. Ware JE, Kosinski MA, Gandek B. SF-36 Health Survey: manual and interpretation guide. Lincoln (RI): Quality Metric; 2005.
5. Ware JE Jr, Gandek B, Kosinski M, Aaronson NK, Apolone G, Brazier J, et al. The equivalence of SF-36 summary health scores estimated using standard and country-specific algorithms in 10 countries: results

from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1167–70.

6. Hawthorne G, Osborne RH, Taylor A, Sansoni J. The SF-36 version 2: critical analyses of population weights, scoring algorithms and population norms. Qual Life Res 2007;16:661–73.
7. Boardman DL, Dorey F, Thomas BJ, Lieberman JR. The accuracy of assessing total hip arthroplasty outcomes: a prospective correlation study of walking ability and 2 validated measurement devices. J Arthroplasty 2000;15:200–4.
8. Dass S, Bowman SJ, Vital EM, Ikeda K, Pease CT, Hamburger J, et al. Reduction of fatigue in Sjögren's syndrome with rituximab: results of a randomised, double-blind, placebo-controlled pilot study. Ann Rheum Dis 2008;67:1541–4.
9. Gladman DD, Urowitz MB, Gough J, MacKinnon A. Fibromyalgia is a major contributor to quality of life in lupus. J Rheumatol 1997;24: 2145–8.
10. Harrison MJ, Tricker KJ, Davies L, Hassell A, Dawes P, Scott DL, et al. The relationship between social deprivation, disease outcome measures, and response to treatment in patients with stable, long-standing rheumatoid arthritis. J Rheumatol 2005;32:2330–6.
11. Soderlin MK, Lindroth Y, Turesson C, Jacobsson LT. A more active treatment has profound effects on the health status of rheumatoid arthritis (RA) patients: results from a population-based RA register in Malmo, Sweden, 1997-2005. Scand J Rheumatol 2010;39:206–11.
12. Heiberg MS, Nordvag BY, Mikkelsen K, Rodevand E, Kaufmann C, Mowinckel P, et al. The comparative effectiveness of tumor necrosis factor-blocking agents in patients with rheumatoid arthritis and patients with ankylosing spondylitis: a six-month, longitudinal, observational, multicenter study. Arthritis Rheum 2005;52:2506–12.
13. Becker MA, Schumacher HR, Benjamin KL, Gorevic P, Greenwald M, Fessel J, et al. Quality of life and disability in patients with treatment-failure gout. J Rheumatol 2009;36:1041–8.
14. Busija L, Osborne RH, Nilsdotter A, Buchbinder R, Roos EM. Magnitude and meaningfulness of change in SF-36 scores in four types of orthopedic surgery. Health Qual Life Outcomes 2008;6:55.
15. Dervin GF, Stiell IG, Rody K, Grabowski J. Effect of arthroscopic debridement for osteoarthritis of the knee on health-related quality of life. J Bone Joint Surg Am 2003;85A:10–9.
16. Dierick F, Aveniere T, Cossement M, Poilvache P, Lobet S, Detrembleur C. Outcome assessment in osteoarthritic patients undergoing total knee arthroplasty. Acta Orthop Belg 2004;70:38–45.
17. Maini RN, Breedveld FC, Kalden JR, Smolen JS, Furst D, Weisman MH, et al, Anti-Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. Sustained improvement over two years in physical function, structural damage, and signs and symptoms among patients with rheumatoid arthritis treated with infliximab and methotrexate. Arthritis Rheum 2004;50:1051–65.
18. Linde K, Weidenhammer W, Streng A, Hoppe A, Melchart D. Acupuncture for osteoarthritic pain: an observational study in routine care. Rheumatology (Oxford) 2006;45:222–7.
19. Ferrara PE, Rabini A, Maggi L, Piazzini DB, Logroscino G, Magliocchetti G, et al. Effect of pre-operative physiotherapy in patients with end-stage osteoarthritis undergoing hip arthroplasty. Clin Rehabil 2008;22:977–86.
20. Sutbeyaz ST, Sezer N, Koseoglu F, Kibar S. Low-frequency pulsed electromagnetic field therapy in fibromyalgia: a randomized, double-blind, sham-controlled clinical study. Clin J Pain 2009;25:722–8.
21. Wang C, Schmid CH, Hibberd PL, Kalish R, Roubenoff R, Rones R, et al. Tai Chi is effective in treating knee osteoarthritis: a randomized controlled trial. Arthritis Rheum 2009;61:1545–53.
22. Coleman S, Briffa NK, Carroll G, Inderjeeth C, Cook N, McQuade J. Effects of self-management, education and specific exercises, delivered by health professionals, in patients with osteoarthritis of the knee. BMC Musculoskelet Disord 2008;9:133.
23. Carmona L, Ballina J, Gabriel R, Laffon A. The burden of musculoskeletal diseases in the general population of Spain: results from a national survey. Ann Rheum Dis 2001;60:1040–5.
24. Hill CL, Gill T, Taylor AW, Daly A, Grande ED, Adams RJ. Psychological factors and quality of life in arthritis: a population-based study. Clin Rheumatol 2007;26:1049–54.
25. Gandhi SK, Salmon JW, Zhao SZ, Lambert BL, Gore PR, Conrad K. Psychometric evaluation of the 12-item Short-Form Health Survey (SF-12) in osteoarthritis and rheumatoid arthritis clinical trials. Clin Ther 2001;23:1080–98.
26. Theiler R, Bischoff HA, Good M, Uebelhart D. Rofecoxib improves quality of life in patients with hip or knee osteoarthritis. Swiss Med Wkly 2002;132:566–73.
27. Foley A, Halbert J, Hewitt T, Crotty M. Does hydrotherapy improve strength and physical function in patients with osteoarthritis: a randomised controlled trial comparing a gym based and a hydrotherapy based strengthening programme. Ann Rheum Dis 2003;62:1162–7.
28. Calandre EP, Rodriguez-Claro ML, Rico-Villademoros F, Vilchez JS, Hidalgo J, Delgado-Rodriguez A. Effects of pool-based exercise in fibromyalgia symptomatology and sleep quality: a prospective randomised comparison between stretching and Ai Chi. Clin Exp Rheumatol 2009;27:S21–8.
29. Fransen M, Nairn L, Winstanley J, Lam P, Edmonds J. Physical activity for osteoarthritis management: a randomized controlled clinical trial evaluating hydrotherapy or Tai Chi classes. Arthritis Rheum 2007;57: 407–14.
30. McCalden RW, MacDonald SJ, Rorabeck CH, Bourne RB, Chess DG, Charron KD. Wear rate of highly cross-linked polyethylene in total hip arthroplasty: a randomized controlled trial. J Bone Joint Surg Am 2009;91:773–82.
31. Thomas S, Kinninmonth AW, Kumar CS. Long-term results of the modified Hoffman procedure in the rheumatoid forefoot: surgical technique. J Bone Joint Surg Am 2006;88 Suppl 1:149–57.
32. Schick M, Stucki G, Rodriguez M, Meili EO, Huber E, Michel BA, et al. Haemophilic arthropathy: assessment of quality of life after total knee arthroplasty. Clin Rheumatol 1999;18:468–72.
33. Stockl KM, Shin JS, Lew HC, Zakharyan A, Harada AS, Solow BK, et al. Outcomes of a rheumatoid arthritis disease therapy management program focusing on medication adherence. J Manag Care Pharm 2010;16:593–604.
34. Bowling A, Bond M, Jenkinson C, Lamping DL. Short Form 36 (SF-36) Health Survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. J Public Health Med 1999;21:255–70.
35. McHorney CA, Kosinski M, Ware JE Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. Med Care 1994; 32:551–67.
36. Lyons RA, Wareham K, Lucas M, Price D, Williams J, Hutchings HA. SF-36 scores vary by method of administration: implications for study design. J Public Health Med 1999;21:41–5.
37. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? J Clin Epidemiol 1996;49:135–40.
38. Jones D, Kazis L, Lee A, Rogers W, Skinner K, Cassar L, et al. Health status assessments using the Veterans SF-12 and SF-36: methods for evaluating otucomes in the Veterans Health Administration. J Ambul Care Manage 2001;24:68–86.
39. Perkins JJ, Sanson-Fisher RW. An examination of self- and telephone-administered modes of administration for the Australian SF-36. J Clin Epidemiol 1998;51:969–73.
40. Millard RW, Carver JR. Cross-sectional comparison of live and interactive voice recognition administration of the SF-12 health status survey. Am J Manag Care 1999;5:153–9.
41. Ball AE, Russell EM, Seymour DG, Primrose WR, Garratt AM. Problems in using health survey questionnaires in older patients with physical disabilities: can proxies be used to complete the SF-36? Gerontology 2001;47:334–40.
42. Yip JY, Wilber KH, Myrtle RC, Grazman DN. Comparison of older adult subject and proxy responses on the SF-36 health-related quality of life instrument. Aging Ment Health 2001;5:136–42.
43. Kosinski M, Bayliss M, Bjorner JB, Ware JE. Improving estimates of SF-36-1 Health Survey scores for respondents with missing data. Med Outcome Trust Mon 2000;5:8–10.
44. Ware JE Jr, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: summary of results from the Medical Outcomes Study. Med Care 1995;33:AS264–79.
45. Jenkinson C, Coulter A, Wright L. Short form 36 (SF-36) health survey questionnaire: normative data for adults of working age. BMJ 1993; 306:1437–40.
46. Watson EK, Firman DW, Baade PD, Ring I. Telephone administration of the SF-36 health survey: validation studies and population norms for adults in Queensland. Aust N Z J Public Health 1996;20:359–63.
47. Australian Bureau of Statistics. National health survey: SF36 population norms, Australia, 1995. Cat. no. 4399.0. Canberra (Australia): ABS; 1997.
48. Sullivan M, Karlsson J, Ware JE. SF-36 Swedish manual and interpretation guide. Gothenburg: Gothenburg University; 1994.
49. Thumboo J, Chan SP, Machin D, Soh CH, Feng PH, Boey ML, et al. Measuring health-related quality of life in Singapore: normal values for the English and Chinese SF-36 health survey. Ann Acad Med Singapore 2002;31:366–74.
50. Scott KM, Tobias MI, Sarfati D, Haslett SJ. SF-36 health survey reliability, validity and norms for New Zealand. Aust N Z J Public Health 1999;23:401–6.
51. Hanmer J, Lawrence WF, Anderson JP, Kaplan RM, Fryback DG. Report of nationally representative values for the noninstitutionalized US adult population for 7 health-related quality-of-life scores. Med Decis Making 2006;26:391–400.

52. Ware JE, Kosinski MA, Turner-Bowker DM, Gandek B. SF-12: how to score version 2 of the SF-12 Health Survey (with a supplement documenting version 1). Lincoln (RI): QualityMetric; 2002.
53. Gandek B, Ware JE, Aaronson NK, Apolone G, Bjorner JB, Brazier JE, et al. Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51: 1171–8.
54. Coons SJ, Rao S, Keininger DL, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000;17:13–35.
55. Parker SG, Bechinger-English D, Jagger C, Spiers N, Lindesay J. Factors affecting completion of the SF-36 in older people. Age Ageing 2006; 35:376–81.
56. Adamson J, Gooberman-Hill R, Woolhead G, Donovan J. 'Questerviews': using questionnaires in qualitative interviews as a method of integrating qualitative and quantitative health services research. J Health Sci Res Pol 2004;9:139–45.
57. Fowler RW, Congdon P, Hamilton S. Assessing health status and outcomes in a geriatric day hospital. Public Health 2000;114:440–5.
58. DeBrota DJ, Bradt EW, Andrejasich CM, Kosinski M, Ware JE. Comparison of interactive voice response SF-36 to self-administered SF-36 and personal interview via telephone SF-36: Boston: The Health Institute, New England Medical Center; 1996.
59. Sanson-Fisher RW, Perkins JJ. Adaptation and validation of the SF-36 Health Survey for use in Australia. J Clin Epidemiol 1998;51:961–7.
60. Tarlov AR, Ware JE Jr, Greenfield S, Nelson EC, Perrin E, Zubkoff M. The Medical Outcomes Study: an application of methods for monitoring the results of medical care. JAMA 1989;262:925–30.
61. Koh ET, Leong KP, Tsou IY, Lim VH, Pong LY, Chong SY, et al. The reliability, validity and sensitivity to change of the Chinese version of SF-36 in oriental patients with rheumatoid arthritis. Rheumatology (Oxford) 2006;45:1023–8.
62. Loge JH, Kaasa S, Hjermstad MJ, Kvien TK. Translation and performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. I. Data quality, scaling assumptions, reliability, and construct validity. J Clin Epidemiol 1998;51:1069–76.
63. Parker SG, Peet SM, Jagger C, Farhan M, Castleden CM. Measuring health status in older patients: the SF-36 in practice. Age Ageing 1998;27:13–8.
64. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Rasmussen C, Jensen DV, et al. What factors influence the health status of patients with rheumatoid arthritis measured by the SF-12v2 Health Survey and the Health Assessment Questionnaire? J Rheumatol 2009;36:2183–9.
65. Soderman P, Malchau H. Validity and reliability of Swedish WOMAC osteoarthritis index: a self-administered disease-specific questionnaire (WOMAC) versus generic instruments (SF-36 and NHP). Acta Orthop Scand 2000;71:39–46.
66. Saleh KJ, Radosevich DM, Kassim RA, Moussa M, Dykes D, Bottolfson H, et al. Comparison of commonly used orthopaedic outcome measures using palm-top computers and paper surveys. J Orthop Res 2002;20:1146–51.
67. Kvien TK, Kaasa S, Smedstad LM. Performance of the Norwegian SF-36 Health Survey in patients with rheumatoid arthritis. II. A comparison of the SF-36 with disease-specific measures. J Clin Epidemiol 1998;51:1077–86.
68. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Hetland ML. 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;35:1528–37.
69. Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE Jr. The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability. Med Care 1999;37: MS10–22.
70. Ruta DA, Hurst NP, Kind P, Hunter M, Stubbings A. Measuring health status in British patients with rheumatoid arthritis: reliability, validity and responsiveness of the short form 36-item health survey (SF-36). Br J Rheumatol 1998;37:425–36.
71. Davey RC, Edwards SM, Cochrane T. Test-retest reliability of lower extremity functional and self-reported measures in elderly with osteoarthritis. Adv Physiother 2003;5:155–61.
72. Brazier JE, Harper R, Munro J, Walters SJ, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology (Oxford) 1999;38:870–7.
73. Jakobsson U, Westergren A, Lindskov S, Hagell P. Construct validity of the SF-12 in three different samples. J Eval Clin Pract. E-pub ahead of print.
74. Lenert LA. The reliability and internal consistency of an Internet-capable computer program for measuring utilities. Qual Life Res 2000; 9:811–7.
75. Lim LL, Fisher JD. Use of the 12-item Short-Form (SF-12) Health Survey in an Australian heart and stroke population. Qual Life Res 1999;8:1–8.
76. Luo X, Lynn George M, Kakouras I, Edwards CL, Pietrobon R, Richardson W, et al. Reliability, validity, and responsiveness of the short form 12-item survey (SF-12) in patients with back pain. Spine (Phila Pa 1976) 2003;28:1739–45.
77. Resnick B, Nahm ES. Reliability and validity testing of the revised 12-item Short-Form Health Survey in older adults. J Nurs Meas 2001; 9:151–61.
78. Salyers MP, Bosworth HB, Swanson JW, Lamb-Pagone J, Osher FC. Reliability and validity of the SF-12 health survey among people with severe mental illness. Med Care 2000;38:1141–50.
79. Keller SD, Ware JE Jr, Bentler PM, Aaronson NK, Alonso J, Apolone G, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA Project. International Quality of Life Assessment. J Clin Epidemiol 1998;51:1179–88.
80. Wolinsky FD, Stump TE. A measurement model of the Medical Outcomes Study 36-Item Short-Form Health Survey in a clinical sample of disadvantaged, older, black, and white men and women. Med Care 1996;34:537–48.
81. Shields RK, Enloe LJ, Leo KC. Health related quality of life in patients with total hip or knee replacement. Arch Phys Med Rehabil 1999;80: 572–9.
82. Kontodimopoulos N, Pappa E, Niakas D, Tountas Y. Validity of SF-12 summary scores in a Greek general population. Health Qual Life Outcomes 2007;5:55.
83. Montazeri A, Vahdaninia M, Mousavi SJ, Omidvari S. The Iranian version of 12-item Short Form Health Survey (SF-12): factor structure, internal consistency and construct validity. BMC Public Health 2009; 9:341.
84. Maurischat C, Ehlebracht-Konig I, Kuhn A, Bullinger M. Factorial validity and norm data comparison of the Short Form 12 in patients with inflammatory-rheumatic disease. Rheumatol Int 2006;26:614–21.
85. Fleishman JA, Selim AJ, Kazis LE. Deriving SF-12v2 physical and mental health summary scores: a comparison of different scoring algorithms. Qual Life Res 2010;19:231–41.
86. Veehof MM, ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Comparison of internal and external responsiveness of the generic Medical Outcome Study Short Form-36 (SF-36) with disease-specific measures in rheumatoid arthritis. J Rheumatol 2008;35:610–7.
87. Fan ZJ, Smith CK, Silverstein BA. Assessing validity of the QuickDASH and SF-12 as surveillance tools among workers with neck or upper extremity musculoskeletal disorders. J Hand Ther 2008;21:354–65.
88. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Comm Health 1980;34:281–6.
89. Jenkinson C, Fitzpatrick R, Argyle M. The Nottingham Health Profile: an analysis of its sensitivity in differentiating illness groups. Soc Sci Med 1988;27:1411–4.
90. Hunt SM, McKenna SP, McEwen J, Williams J, Papp E. The Nottingham Health Profile: subjective health status and medical consultations. Soc Sci Med A 1981;15:221–9.
91. Hunt SM, McEwen J, McKenna SP. Measuring health status: a new tool for clinicians and epidemiologists. J R Coll Gen Pract 1985;35: 185–8.
92. Baillet A, Payraud E, Niderprim VA, Nissen MJ, Allenet B, François P, et al. A dynamic exercise programme to improve patients' disability in rheumatoid arthritis: a prospective randomized controlled trial. Rheumatology (Oxford) 2009;48:410–5.
93. Ekici G, Bakar Y, Akbayrak T, Yuksel I. Comparison of manual lymph drainage therapy and connective tissue massage in women with fibromyalgia: a randomized controlled trial. J Manipulative Physiol Ther 2009;32:127–33.
94. Yurtkuran M, Alp A, Nasircilar A, Bingol U, Altan L, Sarpdere G. Balneotherapy and tap water therapy in the treatment of knee osteoarthritis. Rheumatol Int 2006;27:19–27.
95. Evcik D, Kavuncu V, Yeter A, Yigit I. The efficacy of balneotherapy and mud-pack therapy in patients with knee osteoarthritis. Joint Bone Spine 2007;74:60–5.
96. Knahr K, Korn V, Kryspin-Exner I, Jagsch R. Quality of life five years after total or partial knee arthroplasty. Z Orthop Ihre Grenzgeb 2003; 141:27–32. In German.
97. Atroshi I, Ornstein E, Franzen H, Johnsson R, Stefansdottir A, Sundberg M. Quality of life after hip revision with impaction bone grafting on a par with that 4 years after primary cemented arthroplasty. Acta Orthop Scand 2004;75:677–83.
98. Uutela T, Hannonen P, Kautiainen H, Hakala M, Paananen ML, Hakkinen A. Positive treatment response improves the health-related quality of life of patients with early rheumatoid arthritis. Clin Exp Rheumatol 2009;27:108–11.

99. Wigers SH, Finset A. Rehabilitation of chronic myofascial pain disorders. Tidsskr Nor Laegeforen 2007;127:604–8. In Norwegian.
100. Kersten P, Mullee MA, Smith JA, McLellan L, George S. Generic health status measures are unsuitable for measuring health status in severely disabled people. Clin Rehabil 1999;13:219–28.
101. Vidalis A, Syngelakis M, Papathanasiou M, Whalley D, McKenna SP. The Greek version of the Nottingham Health Profile: features of its adaptation. Hippokratia 2002;6 Suppl 1:75–8.
102. Bucquet D, Condon S, Ritchie K. The French version of the Nottingham Health Profile: a comparison of items weights with those of the source version. Soc Sci Med 1990;30:829–35.
103. Wiklund I, Romanus B, Hunt SM. Self-assessed disability in patients with arthrosis of the hip joint: reliability of the Swedish version of the Nottingham Health Profile. Disabil Rehabil 1988;10:159–63.
104. Essink-Bot ML, Krabbe PF, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures: the Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Med Care 1997;35:522–37.
105. Alonso J, Prieto L, Anto JM. The Spanish version of the Nottingham Health Profile: a review of adaptation and instrument characteristics. Qual Life Res 1994;3:385–93.
106. Hunt SM, McEwen J, McKenna SP. Measuring health status. London: Croom Helm; 1986.
107. Post MW, Gerritsen J, Diederiks JP, De Witte LP. Measuring health status of people who are wheelchair-dependent: validity of the sickness impact profile 68 and the Nottingham health profile. Disabil Rehabil 2001;23:245–53.
108. Nagyova I, van den Heuvel W, Steward R, Macejova Z, van Dijk J. Predictors of change in self-rated health: a longitudinal analysis in patients with rheumatoid arthritis. Netherlands: University of Groningen; 2005.
109. VanderZee KI, Sanderman R, Heyink J. A comparison of two multidimensional measures of health status: the Nottingham Health Profile and the RAND 36-Item Health Survey 1.0. Qual Life Res 1996;5:165–74.
110. Hunt SM, McKenna SP, Williams J. Reliability of a population survey tool for measuring perceived health problems: a study of patients with osteoarthrosis. J Epidemiol Comm Health 1981;35:297–300.
111. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol 1997;50:79–93.
112. Bouchet C, Guillemin F, Briancon S. Comparison of 3 quality of life instruments in the longitudinal study of rheumatoid arthritis. Rev Epidemiol Sante Publique 1995;43:250–8. In French.
113. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A, Mowat A. A generic health status instrument in the assessment of rheumatoid arthritis. Br J Rheumatol 1992;31:87–90.
114. Houssien DA, McKenna SP, Scott DL. The Nottingham Health Profile as a measure of disease activity and outcome in rheumatoid arthritis. Br J Rheumatol 1997;36:69–73.
115. Fitzpatrick R, Ziebland S, Jenkinson C, Mowat A. Importance of sensitivity to change as a criterion for selecting health status measures. Qual Health Care 1992;1:89–93.
116. Bachrach-Lindstrom M, Karlsson S, Pettersson LG, Johansson T. Patients on the waiting list for total hip replacement: a 1-year follow-up study. Scand J Caring Sci 2008;22:536–42.
117. Lillegraven S, Kvien TK. Measuring disability and quality of life in established rheumatoid arthritis. Best Pract Res Clin Rheumatol 2007;21:827–40.
118. De Bruin AF, Buys M, De Witte LP, Diederiks JP. The sickness impact profile: SIP68, a short generic version. First evaluation of the reliability and reproducibility. J Clin Epidemiol 1994;47:863–71.
119. Gilson BS, Gilson JS, Bergner M, Bobbit RA, Kressel S, Pollard WE, et al. The sickness impact profile: development of an outcome measure of health care. Am J Public Health 1975;65:1304–10.
120. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The sickness impact profile: development and final revision of a health status measure. Med Care 1981;19:787–805.
121. De Bruin AF, Diederiks JP, De Witte LP, Stevens FC, Philipsen H. The development of a short generic version of the sickness impact profile. J Clin Epidemiol 1994;47:407–18.
122. Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine 1983;8:141–4.
123. Sullivan M, Ahlmen M, Bjelle A, Karlsson J. Health status assessment in rheumatoid arthritis. II. Evaluation of a modified Shorter Sickness Impact Profile. J Rheumatol 1993;20:1500–7.
124. Knutsson S, Engberg IB. An evaluation of patients' quality of life before, 6 weeks and 6 months after total hip replacement surgery. J Adv Nurs 1999;30:1349–59.
125. Tak E, Staats P, Van Hespen A, Hopman-Rock M. The effects of an exercise program for older adults with osteoarthritis of the hip. J Rheumatol 2005;32:1106–13.
126. Ahlmen M, Sullivan M, Bjelle A. Team versus non-team outpatient care in rheumatoid arthritis: a comprehensive outcome evaluation including an overall health measure. Arthritis Rheum 1988;31:471–9.
127. SIP (Sickness Impact Profile). Mapi Research Trust Education Information Dissemination, 2010. URL: http://www.mapi-trust.org/services/questionnairelicensing/cataloguequestionnaires/118-sip.
128. Agel J, Swiontkowski MF. Guide to outcomes instruments for musculoskeletal trauma research. J Orthop Trauma 2006;20:S1–146.
129. De Bruin AF, De Witte LP, Stevens F, Diederiks JP. Sickness impact profile: the state of the art of a generic functional status measure. Soc Sci Med 1992;35:1003–14.
130. Gilson BS, Bergner M, Bobbitt RA, Carter WB. The Sickness Impact Profile: final development and testing. Hyattsville (MD): National Center for Health Services Research; 1979.
131. Rothman ML, Hedrick S, Inui T. The Sickness Impact Profile as a measure of the health status of noncognitively impaired nursing home residents. Med Care 1989;27:S157–67.
132. Anonymous. The Sickness Impact Profile 68 (SIP 68). Spinal Cord Injury Rehabilitation Evidence 2010. URL: http://www.scireproject.com/outcome-measures/sickness-impact-profile-68-sip-68.
133. Sickness Impact Profile (SIP). 2010. URL: http://www.proqolid.org/instruments/sickness_impact_profile_sip.
134. Patrick D, Peach H, editors. Disablement in the community. Oxford: Oxford University Press; 1989.
135. Bergner M, Bobbitt RA, Kressel S. The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. Int J Health Serv 1976;6:393–415.
136. Andresen EM, Rothenberg BM, Kaplan RM. Performance of a self-administered mailed version of the quality of well-being (QWB-SA) questionnaire among older adults. Med Care 1998;36:1349–60.
137. Van Straten A, De Haan RJ, Limburg M, Schuling J, Bossuyt PM, van den Bos GA. A stroke-adapted 30-item version of the sickness impact profile to assess quality of life (SA-SIP30). Stroke 1997;28:2155–61.
138. Nanda U, McLendon PM, Andresen EM, Armbrecht E. The SIP68: an abbreviated sickness impact profile for disability outcomes research. Qual Life Res 2003;12:583–95.
139. Kessler S, Jaeckel W, Cziske R. Assessing health in musculoskeletal disorders: the appropriateness of a German version of the Sickness Impact Profile. Rheumatol Int 1997;17:119–25.
140. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures: statistics and strategies for evaluation. Control Clin Trials 1991;12:142S–58S.
141. Pollard WE, Bobbitt RA, Bergner M, Martin DP, Gilson BS. The Sickness Impact Profile: reliability of a health status measure. Med Care 1976;14:146–55.
142. Andresen EM, Nanda U, McLendon P, Meyer A, Armbrech E. SIP68: an abbreviated Sickness Impact Profile for disability outcomes research? [abstract]. Qual Life Res 2000;9:343.
143. Weinberger M, Samsa GP, Tierney WM, Belyea MJ, Hiner SL. Generic versus disease specific health status measures: comparing the Sickness Impact Profile and the Arthritis Impact Measurement Scales. J Rheumatol 1992;19:543–6.
144. Deyo RA, Patrick DL. The significance of treatment effects: the clinical perspective. Med Care 1995;33:AS286–91.
145. Post MW, Gerritsen J, van Leusen ND, Paping MA, Prevo AJ. Adapting the Nottingham Health Profile for use in people with severe physical disabilities. Clin Rehabil 2001;15:103–10.
146. Deyo RA, Inui TS. Toward clinical applications of health status measures: sensitivity of scales to clinically important changes. Health Serv Res 1984;19:275–89.
147. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917–25.
148. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopedic evaluation. Med Care 1990;28:632–42.
149. Bryan S, Longworth L. Measuring health-related utility: why the disparity between EQ-5D and SF-6D? Eur J Health Econ 2005;6:253–60.
150. Grieve R, Grishchenko M, Cairns J. SF-6D versus EQ-5D: reasons for differences in utility scores and impact on reported cost-utility. Eur J Health Econ 2009;10:15–23.
151. Hawthorne G, Richardson J. Measuring the value of program outcomes: a review of multiattribute utility measures. Expert Rev Pharmacoeconomics Outcome Res 2001;1:215–28.
152. Walters SJ, Brazier JE. Comparison of the minimally important difference for two health state utility measures: EQ-5D and SF-6D. Qual Life Res 2005;14:1523–32.
153. Wolfe F, Michaud K, Wallenstein G. Scale characteristics and mapping accuracy of the US EQ-5D, UK EQ-5D, and SF-6D in patients with rheumatoid arthritis. J Rheumatol 2010;37:1615–25.
154. Joore M, Brunenberg D, Nelemans P, Wouters E, Kuijpers P, Honig A, et al. The impact of differences in EQ-5D and SF-6D utility scores on

the acceptability of cost-utility ratios: results across five trial-based cost-utility studies. Value Health 2010;13:222–9.

155. Konerding U, Moock J, Kohlmann T. The classification systems of the EQ-5D, the HUI II and the SF-6D: what do they have in common? Qual Life Res 2009;18:1249–61.

156. Marra CA, Marion SA, Guh DP, Najafzadeh M, Wolfe F, Esdaile JM, et al. Not all "quality-adjusted life years" are equal. J Clin Epidemiol 2007;60:616–24.

157. Osnes-Ringen H, Kvamme MK, Kristiansen IS, Thingstad M, Henriksen JE, Kvien TK, et al. Cost-effectiveness analyses of elective orthopaedic surgical procedures in patients with inflammatory arthropathies. Scand J Rheumatol 2011;2:108–15.

158. Barton GR, Sach TH, Avery AJ, Doherty M, Jenkinson C, Muir KR. Comparing the performance of the EQ-5D and SF-6D when measuring the benefits of alleviating knee pain [abstract]. Cost Eff Resour Alloc 2009;7:12.

159. Barton GR, Sach TH, Doherty M, Avery AJ, Jenkinson C, Muir KR. An assessment of the discriminative ability of the EQ-5Dindex, SF-6D, and EQ VAS, using sociodemographic factors and clinical conditions. Eur J Health Econ 2008;9:237–49.

160. Sach TH, Barton GR, Jenkinson C, Doherty M, Avery AJ, Muir KR. Comparing cost-utility estimates: does the choice of EQ-5D or SF-6D matter? Med Care 2009;47:889–94.

161. Ruchlin HS, Insinga RP. A review of health-utility data for osteoarthritis: implications for clinical trial-based evaluation. Pharmacoeconomics 2008;26:925–35.

162. Seymour J, McNamee P, Scott A, Tinelli M. Shedding new light onto the ceiling and floor? A quantile regression approach to compare EQ-5D and SF-6D responses. Health Econ 2010;19:683–96.

163. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ 2002;21:271–92.

164. Brazier JE, Roberts J. The estimation of a preference-based measure of health from the SF-12. Med Care 2004;42:851–9.

165. Boonen A, Patel V, Traina S, Chiou CF, Maetzel A, Tsuji W. Rapid and sustained improvement in health-related quality of life and utility for 72 weeks in patients with ankylosing spondylitis receiving etanercept. J Rheumatol 2008;35:662–7.

166. Zijlstra TR, Braakman-Jansen LM, Taal E, Rasker JJ, van de Laar MA. Cost-effectiveness of spa treatment for fibromyalgia: general health improvement is not for free. Rheumatology (Oxford) 2007;46:1454–9.

167. Van den Hout WB, de Jong Z, Munneke M, Hazes JM, Breedveld FC, Vliet Vlieland TP. Cost-utility and cost-effectiveness analyses of a long-term, high-intensity exercise program compared with conventional physical therapy in patients with rheumatoid arthritis. Arthritis Rheum 2005;53:39–47.

168. Uhlig T, Loge JH, Kristiansen IS, Kvien TK. Quantification of reduced health-related quality of life in patients with rheumatoid arthritis compared to the general population. J Rheumatol 2007;34:1241–7.

169. Fryback DG, Dunham NC, Palta M, Hanmer J, Buechner J, Cherepanov D, et al. US norms for six generic health-related quality-of-life indexes from the National Health Measurement study. Med Care 2007;45: 1162–70.

170. Furlong W, Feeny D, Torrance GW, Barr R, Horsman J. Guide to design and development of health state utility instrumentation. Hamilton (Ontario): Centre for Health Economics and Policy Analysis, McMaster University; 1990.

171. Barton GR, Sach TH, Avery AJ, Jenkinson C, Doherty M, Whynes DK, et al. A comparison of the performance of the EQ-5D and SF-6D for individuals aged >or= 45 years. Health Econ 2008;17:815–32.

172. Slobogean GP, Noonan VK, O'Brien PJ. The reliability and validity of the Disabilities of Arm, Shoulder, and Hand, EuroQol-5D, Health Utilities Index, and Short Form-6D outcome instruments in patients with proximal humeral fractures. J Shoulder Elbow Surg 2010;19: 342–8.

173. Khanna D, Furst DE, Wong WK, Tsevat J, Clements PJ, Park GS, et al. Reliability, validity, and minimally important differences of the SF-6D in systemic sclerosis. Qual Life Res 2007;16:1083–92.

174. Boonen A, van der Heijde D, Landewe R, van Tubergen A, Mielants H, Dougados M, et al. How do the EQ-5D, SF-6D and the well-being rating scale compare in patients with ankylosing spondylitis? Ann Rheum Dis 2007;66:771–7.

175. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Farragher TM, Verstappen SM, et al. Why do patients with inflammatory arthritis often score states "worse than death" on the EQ-5D? An investigation of the EQ-5D classification system. Value Health 2009;12:1026–34.

176. Kontodimopoulos N, Pappa E, Papadopoulos AA, Tountas Y, Niakas D. Comparing SF-6D and EQ-5D utilities across groups differing in health status. Qual Life Res 2009;18:87–97.

177. Goncalves Campolina A, Bruscato Bortoluzzo A, Bosi Ferraz M, Mesquita Ciconelli R. Validity of the SF-6D index in Brazilian patients with rheumatoid arthritis. Clin Exp Rheumatol 2009;27:237–45.

178. Aggarwal R, Wilke CT, Pickard AS, Vats V, Mikolaitis R, Fogg L, et al. Psychometric properties of the EuroQol-5D and Short Form-6D in patients with systemic lupus erythematosus. J Rheumatol 2009;36: 1209–16.

179. Marra CA, Woolcott JC, Kopec JA, Shojania K, Offer R, Brazier JE, et al. A comparison of generic, indirect utility measures (the HUI2, HUI3, SF-6D, and the EQ-5D) and disease-specific instruments (the RAQoL and the HAQ) in rheumatoid arthritis. Soc Sci Med 2005;60:1571–82.

180. Brazier J, Roberts J, Tsuchiya A, Busschbach J. A comparison of the EQ-5D and SF-6D across seven patient groups. Health Econ 2004;13: 873–84.

181. Petrou S, Hockley C. An investigation into the empirical validity of the EQ-5D and SF-6D based on hypothetical preferences in a general population. Health Econ 2005;14:1169–89.

182. Harrison MJ, Davies LM, Bansback NJ, McCoy MJ, Verstappen SM, Watson K, et al. The comparative responsiveness of the EQ-5D and SF-6D to change in patients with inflammatory arthritis. Qual Life Res 2009;18:1195–205.

183. Adams R, Walsh C, Veale D, Bresnihan B, FitzGerald O, Barry M. Understanding the relationship between the EQ-5D, SF-6D, HAQ and disease activity in inflammatory arthritis. Pharmacoeconomics 2010; 28:477–87.

184. Horsman J, Furlong W, Feeny D, Torrance G. The Health Utilities Index(HUI®): concepts, measurement properties and applications. Health Qual Life Outcome 2003;1:54.

185. The Health Utilities Index. 2010. URL: http://www.healthutilities.com/.

186. Torrance GW, Boyle MH, Horwood SP. Application of multi-attribute utility theory to measure social preferences for health states. Oper Res 1982;30:1043–69.

187. Boyle M, Furlong W, Torrance G, Feeny D. Reliability of the Health Utilities Index-mark III used in the 1991 cycle 6 general social survey health questionnaire. Ontario: Center for Health Economics and Policy Analysis; 1994.

188. Strand V, Singh JA. Newer biological agents in rheumatoid arthritis: impact on health-related quality of life and productivity. Drugs 2010; 70:121–45.

189. Mittendorf T, Dietz B, Sterz R, Kupper H, Cifaldi MA, von der Schulenburg JM. Improvement and longterm maintenance of quality of life during treatment with adalimumab in severe rheumatoid arthritis. J Rheumatol 2007;34:2343–50.

190. Prince FH, Geerdink LM, Borsboom GJ, Twilt M, van Rossum MA, Hoppenreijs EP, et al. Major improvements in health-related quality of life during the use of etanercept in patients with previously refractory juvenile idiopathic arthritis. Ann Rheum Dis 2010;69:138–42.

191. Raynauld JP, Torrance GW, Band PA, Goldsmith CH, Tugwell P, Walker V, et al. A prospective, randomized, pragmatic, health outcomes trial evaluating the incorporation of hylan G-F 20 into the treatment paradigm for patients with knee osteoarthritis (Part 1 of 2): clinical results. Osteoarthritis Cartilage 2002;10:506–17.

192. Cadman D, Goldsmith C, Torrance G, Boyle M, Furlong W. Development of a health status index for Ontario children. Hamilton (Ontario): McMaster University, Centre for Health Economics and Policy Analysis; 1986. Final Report to Ontario Ministry of Health Research, grant DM648: (00633).

193. Feeny DH, Torrance GW, Furlong WJ. Health Utilities Index. In: Bert Spilker, editor. Quality of life and pharmacoeconomics in clinical trials: part 2. Vol. 26. Philadelphia: Lippincott-Raven Press; 1996. pp. 239–52.

194. Luo N, Chew LH, Fong KY, Koh DR, Ng SC, Yoon KH, et al. A comparison of the EuroQol-5D and the Health Utilities Index mark 3 in patients with rheumatic disease. J Rheumatol 2003;30:2268–74.

195. Ruiz MA, Rejas J, Soto J, Pardo A, Rebollo I. Adaptation and validation of the Health Utilities Index Mark 3 into Spanish and correction norms for Spanish population. Qual Life Res 2002;11:12–8.

196. Pressler SJ, Eckert GJ, Morrison GC, Murray MD, Oldridge NB. Evaluation of the Health Utilities Index Mark-3 in heart failure. J Card Fail 2011;2:143–50.

197. Marra CA, Rashidi AA, Guh D, Kopec JA, Abrahamowicz M, Esdaile JM, et al. Are indirect utility measures reliable and responsive in rheumatoid arthritis patients? Qual Life Res 2005;14:1333–44.

198. Boyle MH, Furlong W, Feeny D, Torrance GW, Hatcher J. Reliability of the Health Utilities Index-mark III used in the 1991 cycle 6 Canadian General Social Survey Health Questionnaire. Qual Life Res 1995;4: 249–57.

199. Blanchard C, Feeny D, Mahon JL, Bourne R, Rorabeck C, Stitt L, et al. Is the Health Utilities Index valid in total hip arthroplasty patients? Qual Life Res 2004;13:339–48.

200. Grootendorst P, Feeny D, Furlong W. Health Utilities Index Mark 3: evidence of construct validity for stroke and arthritis in a population health survey. Med Care 2000;38:290–9.

201. Blanchard C, Feeny D, Mahon JL, Bourne R, Rorabeck C, Stitt L, et al. Is the Health Utilities Index responsive in total hip arthroplasty patients? J Clin Epidemiol 2003;56:1046–54.

202. Drummond M. Introducing economic and quality of life measurements into clinical studies. Ann Med 2001;33:344–9.

203. Kaplan RM, Bush JW, Berry CC. Health status: types of validity and the index of well-being. Health Serv Res 1976;11:478–507.

204. Kaplan RM, Sieber WJ, Ganiats TG. The quality of well-being scale: comparison of the interviewer-administered version with a self-administered questionnaire. Psych Health 1997;12:783–91.

205. Seiber WJ, Groessl EJ, David KM, Ganiats TG, Kaplan RM. Quality of Well Being Self-Administered (QWB-SA) Scale: user's manual. San Diego: Health Services Research Center, University of California; 2008.

206. Groessl EJ, Kaplan RM, Cronan TA. Quality of well-being in older people with osteoarthritis. Arthritis Rheum 2003;49:23–8.

207. Ganiats TG, Muhlen DG, Kaplan RM, Barrett-Connor E. Gender differences in quality of life in geriatric orthopaedic patients [abstract]. Qual Life Res 1997;6:648.

208. Bombardier C, Raboud J, and the Auranofin Cooperating Group. A comparison of health-related quality-of-life measures for rheumatoid arthritis research. Control Clin Trials 1991;12:243S–56S.

209. Moock J, Kohlmann T. Comparing preference-based quality-of-life measures: results from rehabilitation patients with musculoskeletal, cardiovascular, or psychosomatic disorders. Qual Life Res 2008;17: 485–95.

210. Haywood KL, Garratt AM, Fitzpatrick R. Quality of life in older people: a structured review of generic self-assessed health instruments. Qual Life Res 2005;14:1651–68.

211. Palta M, Chen HY, Kaplan RM, Feeny D, Cherepanov D, Fryback DG. Standard error of measurement of 5 health utility indexes across the range of health for use in estimating reliability and responsiveness. Med Decis Making 2011;2:260–9.

212. Frosch DL, Kaplan RM, Ganiats TG, Groessl EJ, Sieber WJ, Weisman MH. Validity of self-administered quality of well-being scale in musculoskeletal disease. Arthritis Rheum 2004;51:28–33.

213. AQoL instruments. 2009. URL: http://www.aqol.com.au/.

214. Hawthorne G, Richardson J, Osborne R. The Assessment of Quality of Life (AQoL) instrument: a psychometric measure of health-related quality of life. Qual Life Res 1999;8:209–24.

215. Hawthorne G. Assessing utility where short measures are required: development of the short Assessment of Quality of Life-8 (AQoL-8) instrument. Value Health 2009;12:948–57.

216. Busija L, Buchbinder R, Osborne RH. Quantifying the impact of transient joint symptoms, chronic joint symptoms, and arthritis: a population-based approach. Arthritis Rheum 2009;61:1312–21.

217. Ackerman IN, Graves SE, Wicks IP, Bennell KL, Osborne RH. Severely compromised quality of life in women and those of lower socioeconomic status waiting for joint replacement surgery. Arthritis Rheum 2005;53:653–8.

218. Crotty M, Prendergast J, Battersby MW, Rowett D, Graves SE, Leach G, et al. Self-management and peer support among people with arthritis on a hospital joint replacement waiting list: a randomised controlled trial. Osteoarthritis Cartilage 2009;17:1428–33.

219. Osborne RH, Buchbinder R, Ackerman IN. Can a disease-specific education program augment self-management skills and improve health-related quality of life in people with hip or knee osteoarthritis? BMC Musculoskelet Disord 2006;7:90.

220. Bennell K, Wee E, Coburn S, Green S, Harris A, Staples M, et al. Efficacy of standardised manual therapy and home exercise programme for chronic rotator cuff disease: randomised placebo controlled trial. BMJ 2010;340:c2756.

221. Bennell KL, Hinman RS, Metcalf BR, Buchbinder R, McConnell J, McColl G, et al. Efficacy of physiotherapy management of knee joint osteoarthritis: a randomised, double blind, placebo controlled trial. Ann Rheum Dis 2005;64:906–12.

222. Galea MP, Levinger P, Lythgo N, Cimoli C, Weller R, Tully E, et al. A targeted home- and center-based exercise program for people after total hip replacement: a randomized clinical trial. Arch Phys Med Rehabil 2008;89:1442–7.

223. Buchbinder R, Osborne RH, Ebeling PR, Wark JD, Mitchell P, Wriedt C, et al. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. N Engl J Med 2009;361:557–68.

224. Hawthorne G. The effect of different methods of collecting data: mail, telephone and filter data collection issues in utility measurement. Qual Life Res 2003;12:1081–8.

225. Hawthorne G, Richardson J, Day NA. Using the Assessment of Quality of Life (AQoL) Instrument, version 1.0. Report no. 12. Melbourne: Centre for Health Program Evaluation, University of Melbourne; 2000.

226. Hawthorne G, Osborne R. Population norms and meaningful differences for the Assessment of Quality of Life (AQoL) measure. Aust N Z J Public Health 2005;29:136–42.

227. Whitfield K, Buchbinder R, Segal L, Osborne RH. Parsimonious and efficient assessment of health-related quality of life in osteoarthritis research: validation of the Assessment of Quality of Life (AQoL) instrument. Health Qual Life Outcomes 2006;4:19.

228. Richardson J, Day NA, Peacock S, Iezzi A. Measurement of the quality of life for economic evaluation and the Assessment of Quality of Life (AQoL) Mark 2 instrument. Aust Health Econ Rev 2004:1;62–88.

229. Osborne RH, Hawthorne G, Lew EA, Gray LC. Quality of life assessment in the community-dwelling elderly: validation of the Assessment of Quality of Life (AQoL) instrument and comparison with the SF-36. J Clin Epidemiol 2003;56:138–47.

230. Hawthorne G, Richardson J, Day NA. A comparison of the Assessment of Quality of Life (AQoL) with four other generic utility instruments. Ann Med 2001;33:358–70.

Summary Table for Adult Quality of Life Measures*

Scale	Purpose/content	Method of administration	Respondent burden	Administrative burden	Score interpretation	Reliability evidence	Validity evidence	Ability to detect change	Strengths	Cautions
SF-36	Generic health profile	Self-report Self-administration	Moderate to low	Low	Range 0–100 High scores = better health	Poor–good	Generally good	Good	Availability of norms provides context for score interpretation	Large amount of measurement error for some subscales
SF-12	Generic health profile	Self-report Self-administration	Low	Low	Range 0–100 High scores = better health	Good	Generally good	Fair but need more data	Low respondent burden and availability of norms	Not for monitoring individuals
NHP	Generic health profile	Self-administration	Low	Low	Range 0–100 High scores = worse health	Poor–good	Generally good	Generally good	Relatively simple to complete, low respondent/ administrative burden	Questionable psychometric properties, particularly among people with minor disabilities
SIP	Generic health profile	Self-report Self-administration	High	Moderate	Range 0–100 High scores = worse health	Overall score: good Dimensions: good Subscales: poor to good	Generally good	Generally good, but need more data	Able to detect change among a range of rheumatology interventions	High respondent burden, particularly for SIP136 Poor reliability in some subscales
SF-6D	Health utility measure	Self-report Self-administration	Moderate to low	Moderate to low	Range 0.30–1.00 High scores = better HRQOL	Very good	Very good	Fair	Low respondent burden and availability of norms; scores can be computed whenever SF-36 or SF-12 have been administered	Not suitable for use with populations with severely impaired HRQOL
HUI3	Health utility measure	Self-report Self-administration	High to moderate	High to moderate	Range 0–1 High scores = better HRQOL	Fair–good	Generally good	Fair to good, but need more data	Utility instrument, can be used in cost utility analysis; ability to detect improvement due to rheumatology treatments	Cost Limited sensitivity to deterioration
QWB-SA	Health utility measure	Self-administration	Moderate to low	Low	Range 0–1 High scores = better HRQOL	Insufficient information	Good	Generally good, but need more data	Comprehensive coverage of health state levels with no evidence of floor or ceiling effects	Need further psychometric evaluations in rheumatic conditions
AQoL	Health utility measure	Self-report Self-administration	Very low	Very low	Range −0.04 to 1.00 High scores = better HRQOL	Very good	Very good	Generally good, but need more data	Can be used to make comparisons with the general population	Need further psychometric evaluations in rheumatic conditions

* SF-36 = Medical Outcomes Study Short Form 36; SF-12 = Medical Outcomes Study Short Form 12; NHP = Nottingham Health Profile; SIP = Sickness Impact Profile; SF-6D = Medical Outcomes Study Short Form 6D; HRQOL = health-related quality of life; HUI3 = Health Utilities Index Mark 3; QWB-SA = Quality of Well-Being Scale Self-Administered; AQoL = Assessment of Quality of Life.