Arthritis Care & Research
Vol. 72, No. S10, October 2020, pp 522–564
DOI 10.1002/acr.24216

AMERICAN COLLEGE of RHEUMATOLOGY
Empowering Rheumatology Professionals

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

Ljoudmila Busija,[1] Ilana N. Ackerman,[1] Romi Haas,[2] Jason Wallis,[2] Sandra Nolte,[3] Sharon Bentley,[4] Daisuke Miura,[1] Melanie Hawkins,[5] and Rachelle Buchbinder[2]

INTRODUCTION

The aim of this review is to provide an update of an earlier review of adult measures of general health and health-related quality of life (HRQOL) commonly used in rheumatic disease settings (1). 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 measures included in this review were further subdivided into generic health profiles (questionnaires that provide assessment of more than one dimension of health) and health utility measures (questionnaires that provide an overall measure of HRQOL).

As with our previous review, relevant measures were identified through a systematic search of publications indexed to the PubMed database. Two search queries were used: Search query 1 consisted of the following: ("quality of life" [MeSH terms] AND "rheumatic diseases" [MeSH terms]) AND "patient outcome assessment" [MeSH terms] AND ("2014/10/06" [PDat]: "2019/10/04" [PDat] AND "humans" [MeSH terms] AND "adult" [MeSH terms]). Search query 2 consisted of the following: "quality of life" [title/abstract] OR "qol" [title/abstract] OR "life quality" [title/abstract] AND "rheumatic diseases" [MeSH terms] AND "2014/10/06" [PDat]: "2019/10/04" [PDat] AND "humans" [MeSH Terms] AND "adult" [MeSH terms] AND "2014/10/06" [PDat]: "2019/10/04" [PDat] AND "humans" [MeSH terms] AND "adult" [MeSH terms] AND ("2014/10/06" [PDat]: "2019/10/04" [PDat] AND "humans" [MeSH terms] AND "adult" [MeSH terms]). Both searches were carried out on October 4, 2019, and were limited to the past 5 years.

Search query 1 identified 97 articles, and search query 2 identified 1270 articles. After the removal of 72 duplicates, the remaining 1295 articles were screened to identify measures of HRQOL for inclusion in this review (ie, those generic questionnaires that were identified by the study authors as being used for the purpose of assessing general health or HRQOL). If abstracts contained insufficient information to determine the type of measures used, full-text publications were obtained.

We identified six generic health profiles and five generic health utility measures. The generic health profiles included the Medical Outcomes Study (MOS) 36-Item Short Form Health Survey (SF-36) (402 occurrences), the MOS 12-Item Short Form Health Survey (SF-12) (86 occurrences), the Nottingham Health Profile (NHP) (15 occurrences), the World Health Organization Quality of Life short-version instrument (WHOQOL-BREF) (12 occurrences), the Patient-Reported Outcomes Measurement Information System–General Health (PROMIS-GH) (one occurrence), and the SF-8 (one occurrence). The generic health utility measures included the EQ-5D (including the EQ Visual Analog Scale [VAS]) (227 occurrences), the Assessment of Quality of Life Scale (AQoL) (14 occurrences), SF-6D (three occurrences), the Health Utilities Index Mark 3 (HUI3) (one occurrence), and the 15D (one occurrence). Measures that occurred only once in the screened studies were not included in this review because of their apparent low use in rheumatic disease settings. An exception was PROMIS-GH, which was included in this review given the growing popularity of PROMIS measures. Of the instruments included in this review, five were also included in our earlier review (the SF-36, SF-12, NHP, AQoL, and SF-6D) (1).

In assessing the psychometric properties of included questionnaires, we interpreted Cronbach's α and intraclass correlation coefficient (ICC) of 0.70 or more as indicating adequate group-level internal consistency and test-retest reliability, respectively. In evaluating responsiveness, we interpreted values of standardized effect sizes (ESs) of less than 0.20 as no meaningful effect, ESs of

[1] Ljoudmila Busija, PhD, Ilana N. Ackerman, PhD, Daisuke Miura, BBMedSc: Monash University, Melbourne, Victoria, Australia; [2] Romi Haas, PhD, Jason Wallis, PhD, Rachelle Buchbinder, FRACP, FAHMS, PhD: Cabrini Institute, Malvern, Victoria, Australia, and Monash University, Melbourne, Victoria, Australia; [3] Sandra Nolte, PhD: Charité – Universitätsmedizin Berlin and Berlin Institute of Health, Berlin, Germany, ICON GmbH, Munich, Germany, and Deakin University, Burwood, Victoria, Australia; [4] Sharon Bentley, PhD: Queensland University of Technology, Kelvin Grove, Queensland, Australia; [5] Melanie Hawkins, PhD: Deakin University, Burwood, Victoria, Australia, and Swinburne University of Technology, Melbourne, Victoria, Australia.

Drs. Haas and Wallis contributed equally to this work.

Dr. Busija received consultancy fees for Monash University from Jesuit Social Services and Charité Universitätsmedizin Berlin for projects unrelated to the current work. Dr. Nolte was a full-time employee of ICON GmbH at the time of preparing this review; as part of her role, she provided consultancy services to the pharmaceutical industry regarding patient-centered outcomes.

Address correspondence to Ljoudmila Busija, PhD, Monash University, School of Public Health and Preventive Medicine, 553 St. Kilda Road, Melbourne, VIC 3004, Australia. E-mail: lucy.busija@monash.edu.

Submitted for publication March 6, 2020; accepted in revised form April 7, 2020.

0.20 to 0.49 as a small effect, ESs of 0.50 to 0.79 as a moderate effect, and ESs of 0.80 or more as a large effect (2). Floor and ceiling effects were defined as 15% or more respondents obtaining the worst and best health state/HRQOL scores, respectively.

MEDICAL OUTCOMES STUDY 36-ITEM SHORT FORM HEALTH SURVEY

Description

Purpose. The SF-36 is a multi-item generic health profile intended to measure "general health concepts not specific to any age, disease, or treatment group" (3). The SF-36 is suitable for use in general and clinical populations and, as such, can be used to compare health between populations and between diseases.

There are two versions of the SF-36. The original version was developed by the RAND Corporation from the MOS (3). Subsequently, a group of researchers from the original study released an updated version of SF-36 (SF-36 version 2). The revised version is very similar to its original form, with major differences involving changes to item wording, revision of the response scale to incorporate a greater number of response options, and norm-based scoring (4). Unless otherwise indicated, this review focuses on the SF-36 version 2.

Content or domains. The SF-36 measures the following eight health domains: physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional, and mental health. The SF-36 can also be used to derive the following two aggregate summary measures: the physical component summary (PCS) score and the mental component summary (MCS) score.

The SF-36 originated from a the 116-item health survey developed for the MOS—a 2-year observational study of patients with chronic conditions that involved 22 462 patients in a cross-sectional phase and 2349 patients in a longitudinal phase (5). The MOS items have been adapted from pre-existing instruments used to measure function, emotional well-being, and physical health (3). The items for inclusion in the SF-36 were selected to capture the content of the full MOS health survey as much as possible (3). No specific information on the development and selection of the SF-36 items is available.

Number of items. The SF-36 consists of 36 items, 35 of which are used in the calculation of the eight scale scores. The physical functioning scale is the longest, with 10 items. The general health and mental health scales have five items each, and the vitality and role physical scales have four items each. The role emotional scale has three items, and the bodily pain and social functioning scales have two 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.

Response options/scale. The response scales for the SF-36 version 2 items vary across and within the scales. The physical functioning items are scored on a three-point response scale (1 = limited a lot, 2 = limited a little, and 3 = not limited at all). The role physical, role emotional, mental health, vitality, and social functioning items all use a five-point response scale, in which 1 = all of the time and 5 = none of the time. The general health items also use a five-point response scale, with four items scored from 1 (definitely true) to 5 (definitely false) and one item scored from 1 (excellent) to 5 (poor). One of the bodily pain items and the remaining social functioning item use a five-point response scale, in which 1 = not at all and 5 = extremely. The second pain item is scored on a six-point scale (1 = none and 6 = very severe). The health transition item is scored on a five-point scale, in which 1 = much better now than 1 year ago and 5 = much worse now than 1 year ago.

Recall period for items. The SF-36 is available in two forms: a standard form, which uses a 4-week recall period, and an acute form, which uses a 1-week recall period. 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 readministration of the instrument. The acute 1-week recall form is appropriate when more frequent administration is required and changes are likely to occur rapidly.

Cost to use. The original version of the SF-36 (the RAND 36-Item Health Survey 1.0 Questionnaire) is available cost-free. Use of the SF-36 version 2 is subject to license fee. Survey license fees vary according to whether the survey is used in a commercial or nonprofit setting. Quotations can be obtained from Optum following the completion of online information request form (https://www.optum.com/solutions/life-sciences/answer-research/patient-insights/outcomes-survey-request.html). Manuals can also be purchased from Optum.

How to obtain. The original version of the SF-36 can be obtained free of charge from the RAND Corporation (https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form.html). English and Arabic language versions are available.

The SF-36 version 2 can be obtained from Optum (https://www.optum.com/en.html). The Office of Grants and Scholarly Research at Optum provides students and researchers undertaking unfunded, noncommercial research with access to the survey royalty-free or at a modest fee. To determine eligibility, prospective users need to complete a survey license request form on the Optum website.

Practical application

Method of administration. The SF-36 can be self-administered or interviewer-administered. Multiple modes are available, including paper and pencil, online, personal digital assistant, tablet, and interactive voice response (IVR) via telephone. Several studies reported a consistent bias for the lower SF-36 scores (worse health) when self-completed compared with interviewer administration (6–10). Data quality also tends to be better in interviewer administration, with a lower proportion of missing data, lower ceiling effects, and better internal consistency estimates (7,11). Data collection costs, on the other hand, are lower (up to 77%) for self-administration (7,11).

The SF-36 can also be administered by proxy, but concordance between self and proxy ratings varies across proxy types. Generally, professional proxies (eg, occupational therapists or nurses) provide a description of an individual's health state that is closer to the individual's own ratings than lay proxies, who tend to overestimate the level of impairment (12,13).

Scoring. The SF-36 contains a mixture of positively (higher scores = better health) and negatively worded response scales, and, hence, 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 to 100 scale. The MCS and PCS scores are calculated by summing factor-weighted scores across all eight scales, with factor weights derived from a United States–based general population sample (14). Country-specific weights are also available for Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, the United Kingdom (15), Australia (16), New Zealand (17), and Switzerland (18).

In the calculation of summary component scores, each scale contributes to each component. However, in the PCS calculation, the highest weights are given to the physical functioning, role physical, bodily pain, and general health scales, whereas for the MCS, higher weights are given to the vitality, social functioning, role emotional, and mental health scales. The scoring weights are derived from an orthogonal (uncorrelated) factor analytic model, and, consequently, the PCS and MCS are uncorrelated by design. Recent work by Tucker et al found that the orthogonal approach to PCS and MCS scoring tends to produce biased scores (19) and that scoring weights derived from different countries are not equivalent (20). Current recommendations favor the use of correlated approaches to MCS and PCS scoring (19) and the use of country-specific weights when available (20).

Scoring instructions for the SF-36 original version are available free of charge from the RAND Corporation website (https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form/scoring.html). A score calculator for the SF-36 original version is also available from the Orthotoolkit website (https://www.orthotoolkit.com/sf-36/).

Scoring instructions for the SF-36 version 2 are available from Optum upon the completion of survey license request (https://www.optum.com/solutions/life-sciences/answer-research/patient-insights/outcomes-survey-request.html). SAS-based scoring instructions for the SF-36 version 2 are also publicly available from the University of California, Los Angeles (https://labs.dgsom.ucla.edu/hays/files/view/docs/programs-utilities/sf36v2-4-public.sas.txt).

Score interpretation. Scores on the SF-36 range from 0 to 100, with higher scores indicating better health (3).

Age- and sex-based norms for the SF-36 are available for several countries, including the United States (14,21), the United Kingdom (6,22), Australia (16,23,24), Sweden (25), China (26), New Zealand (27), Singapore (28), and Switzerland (18), among others. Notable cross-country differences in normative SF-36 scores have been previously reported (16), which may reflect cultural differences in health perceptions.

Norm-based scoring of the SF-36 version 2 is available and can be used to produce T scores for each scale (with a mean of 50 and an SD of 10) as well as the PCS and MCS scores (14). Further information on country-specific scoring can be requested from the Optum upon completion of a survey license request form.

Respondent time to complete. The self-reported version of the SF-36 takes only 7 to 10 minutes to complete (29), although the presence of cognitive or physical impairment and depressed mood have been shown to substantively increase the completion time (30).

Administrative burden. The SF-36 has a relatively low administrative burden. Interviewer administration of the SF-36 by telephone takes between 16 and 17 minutes (31). No specific training for the administration of the SF-36 is required, and completion instructions are self-explanatory. Computerized scoring algorithms for the original and revised versions of the SF-36 require basic knowledge of statistical software.

Translations/adaptations. The original version of the SF-36 is available in English and Arabic. The revised version is available in over 160 languages. A list of translated versions can be obtained from the Optum following registration. Cultural adaptations of the original United States version to other English-speaking countries are also available (32).

Psychometric information

Floor and ceiling effects. Both ceiling and floor effects on the SF-36 are commonly reported in the setting of rheumatic conditions. For example, among patients about to undergo joint arthroplasty, ceiling effects were present on the social functioning scale (18% hip; 23% knee), the role physical scale had floor effects (81% hip; 70% knee), and the role emotional scale had both ceiling (30% hip; 28% knee) and floor effects (47% hip; 48% knee) (33). At 6-month follow-up, floor effects were present on the role physical (35% hip; 38% knee) and role emotional (23% hip; 27% knee) scales, whereas ceiling effects were present on the role physical (35% hip; 19% knee), bodily pain (26% hip; 19% knee), social functioning (56% hip; 54% knee), role emotional (54% hip, 44% knee), and mental health (19% hip; 17% knee) scales (33).

More recently, among 45 patients who underwent hip revision arthroplasty, both floor and ceiling effects were recorded for the SF-36 bodily pain scale at 2-year follow-up, with 16% having the worst possible score and 18% recording the best possible score (34). Among patients with spinal stenosis (35) and spondyloarthritis (36), ceiling effects (highest possible scores) were present on role physical (16% or more), social functioning (25% or more), and role emotional (21% or more) scales, but there were no floor effects. On the other hand, among patients with cervical spondylotic myelopathy, none of the SF-36 scales had ceiling effects, and floor effects were reported only for the role emotional scale (16%) (37).

More broadly, a study of hospital outpatients with a range of chronic conditions found no evidence of floor effects on the SF-36 scales, whereas ceiling effects were reported on the role emotional (69%), social functioning (59%), bodily pain (27%), role physical (60%), and physical functioning (23%) scales (38). In the general population, only general health, vitality, and mental health scales appear to be free from floor effects, whereas ceiling effects for the remaining scales tend to range from 27% (physical functioning) to 70% (role emotional) (18,39).

Reliability. *Internal consistency.* The SF-36 is reported to have high internal consistency reliability in a range of rheumatic conditions. In a study of Singaporean patients with spondyloarthritis, internal consistency (Cronbach's *α*) was 0.88 for the social functioning and role emotional scales and 0.89 or more for the remaining scales (36). In 306 Chinese patients with gout, Cronbach's *α* was 0.78 or more for all scales of the SF-36 (40). Similarly, in individuals with osteoarthritis, the SF-36 scales also had high internal consistency, with Cronbach's *α* of 0.78 or more (41). A systematic review of HRQOL measures used in low back pain also concluded that all of the SF-36 scales had very good internal consistency in this setting (42).

More broadly, the SF-36 was also found to have good internal consistency in hospital outpatients with a range of chronic conditions (Cronbach's *α* of more than 0.80) (38). In the general population, internal consistency of the SF-36 also tends to be high, with Cronbach's *α* of 0.81 or more for all of the SF-36 scales except general health (*α* = 0.73).

Test-retest. Results for test-retest reliability of the SF-36 are less encouraging. In a recent study among stable patients with rheumatoid arthritis, an ICC above the recommended standard of 0.70 was recorded only for the vitality scale (ICC = 0.79), with the ICCs for the remaining scales ranging from 0.52 (role emotional and social functioning) to 0.69 (bodily pain) at 3 months test-retest interval (43). However, a 3-month interval is likely to be sufficient for a real change to occur, and, hence, the results need to be treated cautiously, especially because in an earlier study involving patients with rheumatoid arthritis (44), only social functioning scale had an ICC below 0.7 (ICC = 0.67) over a 2-week test-retest interval. In a study of individuals with stable osteoarthritis, 2-week test-retest reliability of the SF-36 was close to acceptable for all scales, with only the role emotional (ICC = 0.68) and role physical (ICC = 0.65) scales having test-retest reliability slightly below the recommended level (41). Similarly, in a sample of patients with lumbar stenosis, ICCs for the SF-36 scales ranged from 0.72 to 0.86 over a 2-week test-retest interval, supporting temporal stability of the SF-36 (35). On the other hand, general health, vitality, and mental health scales were reported to have below optimal test-retest reliability (ICC = 0.65 or less) in 69 elderly individuals drawn from the Croatian general population over a 4-week period (45).

Validity. *Face/content.* The SF-36 captures a broad range of health states applicable to individuals with rheumatic conditions and appears to have good face validity, with all items referring to health-related issues. However, the presence of 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. Additionally, item development for the SF-36 did not involve input from patients, and, hence, decisions regarding content validity of this measure in rheumatology settings need to be made on a study-by-study basis to ensure that the SF-36 is suitable for the purpose of a given study.

Construct. The SF-36 provides scores for eight separate domains of health in addition to two component scores. However, factor analytic studies generally report results that are consistent with the SF-36 being a two-dimensional scale measuring physical and mental health (14,46), with equivocal support for the existence of eight separate lower-order dimensions (46). Furthermore, a recent systematic review concluded that in low back pain, the two-dimensional higher-order model of the SF-36 was not supported (42). Results of confirmatory factor analysis studies also challenge the traditional conceptualization of the two summary scores as being uncorrelated, with the models that allow for the correlation between the summary scores providing a consistently better fit for the data (19,20).

Our earlier review (1) concluded that the evidence generally favored convergent but not discriminant validity of the SF-36, with higher than expected correlations between the SF-36 scales and dissimilar constructs in rheumatic conditions. The results of recent studies also present mixed evidence in support of the construct validity of the SF-36. For example, among Singaporean patients with spondyloarthritis, the SF-36 scales that measure physical aspects of health showed stronger correlations with disease-specific questionnaires, including the Bath Ankylosing Spondylitis Global Score and the Health Assessment Questionnaire (HAQ), than the scales measuring mental health aspects (36). Based on the observed pattern of correlations, the authors concluded that the results supported both the convergent and discriminant validity of the SF-36. However, the convergent and discriminant validity hypotheses have not been explicitly stated, and no details were provided regarding the strength of correlations that would be consistent with support for the convergent and discriminant validity of the SF-36. Therefore, the results of this study are open to interpretation. Similarly, a systematic review of generic measures of HRQOL used in low back pain concluded that the evidence indicates that the construct validity of the summary scores of the SF-36 is inadequate in patients with low back pain (42).

The evidence for the known-groups validity of the SF-36 in rheumatic conditions is more favorable. All the SF-36 scales and the component scores were able to differentiate between patients with spondyloarthritis and the general population, with patients recording lower scores (*P* < 0.001) (36). Similarly, the SF-36 scales and component scores differentiated patients with rheumatoid arthritis and psoriatic arthritis from the general population, with lower scores on all scales and components for the patients with arthritis compared with the general population (*P* < 0.001) (47). The scores of the two patient groups were very similar apart from higher (*P* < 0.001) general health and vitality scores for patients with rheumatoid arthritis. A systematic review and meta-analysis also found that the SF-36 scales and component scores were able to differentiate between patients with ankylosing spondylitis and those without the condition (48). However, in another study, contrary to the researchers' hypothesis, the SF-36 was unable to differentiate patients with spinal stenosis who had ossification of the ligamentum flavum and those who did not (35).

Responsiveness. The responsiveness of the SF-36 in rheumatic conditions has been reported for perceived changes in global health, pharmaceutical interventions, and surgery. In a cohort of 185 patients with systemic lupus erythamatosus, responsiveness of the SF-36 scales to self-reported global changes in health status (measured on a VAS) was low to moderate, with a standardized response mean (SRM) ranging from 0.32 (vitality) to 0.58 (role emotional) (49). The SRMs for the PCS and MCS were 0.44 and 0.43, respectively.

Evidence for the responsiveness of the SF-36 to pharmacological interventions is mixed. Among patients with osteoporosis and a history of vertebral fracture who were treated for 1 year with alendronate (70 mg/wk) and calcium plus vitamin D supplementation, PCS and MCS showed large improvements, with ESs of 1.67 and 1.55, respectively (50). Responsiveness of individual scales was not reported in that study. In patients with chronic knee pain who were commenced on oral analgesia, the bodily pain and physical functioning scales and the PCS showed low levels of responsiveness at 13-week follow-up, with an SRM of 0.49 (95% confidence interval [CI] 0.39-0.58), an SRM of 0.21 (95% CI 0.11-0.30), and an SRM of 0.32 (95% CI 0.22-0.42), respectively (51). The MCS and the remaining scales showed submeaningful changes, with an SRM of 0.10 or less.

Among patients who underwent hip arthroplasty, there were large changes in the bodily pain (ES = 1.4), physical functioning (ES = 0.8), and role physical (ES = 0.9) scales and a moderate change in the vitality (ES = 0.6) scale at 2-year follow-up (34). Responsiveness of the SF-36 bodily pain scale was comparable with that of the pain scale from the disease-specific Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) (ES = 1.7) (34). In another study of over 300 patients who underwent total hip replacement, 6-month ESs ranged from 0.25 (general health) to 1.54 (physical functioning), and 2-year ESs ranged from 0.09 (general health) to 1.72 (role physical) (52).

At least two studies investigated responsiveness of the SF-36 to spinal surgery. In one study involving 192 patients with spinal stenosis, significant changes were recorded in the physical functioning (*P* < 0.001), bodily pain (*P* < 0.001), role physical (*P* = 0.01), general health (*P* = 0.01), and mental health (*P* = 0.03) scales at 42 days postsurgery, whereas the vitality (*P* = 0.54), social functioning (*P* = 0.61), and role emotional (*P* = 0.39) scales did not change significantly during this period (35). No ESs for the magnitude of change were reported in this study. Among 142 patients who underwent surgery for cervical spondylotic myelopathy, the physical functioning scale showed a significant improvement at both 3 months and 1 year after surgery (*P* < 0.05), whereas the role physical, social functioning, role emotional, and mental health scales showed significant improvements at 1 year after surgery (*P* < 0.05), and these were sustained at 2-year follow-up (37). ESs were moderate for the bodily pain (ES = 0.57), vitality (ES = 0.73), social functioning (ES = 0.74), mental health (ES = 0.77), and general health (ES = 0.79) scales and large for the role emotional (ES = 0.82), physical functioning (ES = 0.91), and role physical (ES = 0.93) scales and the MCS (ES = 0.81) and PCS (ES = 0.84). However, no information is provided for the time frame over which the ESs were calculated in the study by Zhang et al (37).

Minimally important differences. Information on the group-level minimally important differences (MIDs) for the SF-36 summary scores is available for some rheumatic conditions, with results indicating that the SF-36 is a sensitive measure of change at the group level. Specifically, an MID of 3.0 points and 3.1 points on the PCS is reportedly associated with a 20% increase in 2-year mortality among patients with upper limb and lower limb arthritis, respectively (14). An MID of 6.7 and 6.9 units in the PCS is associated with a 50% increase in 2-year mortality in people with lower limb and upper limb arthritis, respectively (14).

More generally, a group-level MID of 5 points on a 100-point scale has previously been reported for the SF-36 based on the SEM derived from a normative population sample (14). The developers of the SF-36 also reported MID values of two to three points for the PCS and three points for the MCS. For individual scales, MIDs differ according to the score range. For the psychical functioning, role physical, bodily pain, general health, and vitality scales, the MID is two points for the scores in the range below 40 and three points for the scores in the range above 40. For the social functioning and mental health scales, the MID is three points across the full range of the scale, and for role emotional, the MID is four points across the range of the scale (14).

Individual-level MIDs for the SF-36 have been studied less extensively in rheumatic conditions and indicate that the SF-36 might not be suitable for monitoring individuals because of a large amount of measurement error. In patients who underwent hip or knee replacement surgery, individual-level MIDs ranged from 22% (general health) to 97% (role physical) of the total score range (33,53).

The minimal clinically important change (MCIC) for the SF-36 has been reported in at least three studies conducted in rheumatology settings. Among 606 patients undergoing surgical decompression for degenerative cervical myelopathy, the MCICs for the SF-36 PCS and MCS were 4.6 and 6.8 when 0.5 SD was used as a criterion and 2.9 and 4.3 when SEM was used as a criterion (54). MCICs in PCS and MCS were 5.52 and 3.43, respectively, in Chinese patients with cervical spondylotic myelopathy 2 years after the surgery (37).

Generalizability. The SF-36 covers a broad range of health-related domains and can potentially be used to measure health in all rheumatic conditions. The SF-36 is especially suited when comparisons between disease groups or with the general population are required. Presence of floor and ceiling effects on the SF-36 scales make this questionnaire potentially unsuitable for groups with very severe or very mild conditions.

Use in clinical trials. The SF-36 has been extensively used in randomized controlled trials in rheumatic disease settings. Some examples of recent trials that utilized the SF-36 include the assessments of the efficacy of debridement of unstable chondral lesions versus observation following arthroscopic partial meniscectomy (55), intra-articular ozone versus placebo in knee osteoarthritis (56), the effects of Tai Ji Quan training versus health promotion classes on self-reported sleep quality in elderly Chinese women with knee osteoarthritis (57), tofacitinib in combination with conventional disease-modifying antirheumatic drugs in rheumatoid arthritis (58,59), acupuncture versus sham procedure for knee osteoarthritis (60), and *andrographis paniculata* extract versus placebo in knee osteoarthritis (61). The SF-36 was also used in comparative effectiveness trials, including trials comparing muscle-stretching exercise and resistance training in fibromyalgia (62), sensory-motor training versus resistance training among patients with knee osteoarthritis (63); and quadriceps strength training alone versus quadriceps and hip abductor strength training in knee osteoarthritis (64).

Critical appraisal of overall value to the rheumatology community

Strengths. The SF-36 has strong evidence of convergent validity as a generic measure of health and can be used when the assessment of a broad range of health aspects is needed. The availability of population norms also provides context for score interpretation. The SF-36 appears to be able to differentiate between levels of disease severity in rheumatic conditions and between people with and without rheumatic conditions, and it is reasonably sensitive to change at the group level.

Caveats and cautions. The discriminant validity of the SF-36 and its eight-dimensional structure received mixed support. The role physical, role emotional, and social functioning scales are frequently reported to have low reliability, which further puts the validity of this measure into question. Floor and ceiling effects in rheumatic conditions also indicate that the SF-36 does not adequately target the full range of the health experiences of this population.

Clinical usability. In clinical settings, large intraindividual variations in the SF-36 scale scores make this measure 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.

Research usability. Ease of administration, availability of an online version, and availability of a computerized scoring algorithm support the usability of the SF-36 in research settings, in which it can be used to compare different disease groups or to compare disease groups with population norms. However, low reliability for the SF-36 scales can also dramatically increase sample size because of potentially low measurement precision. Financial costs can also limit the use of this scale in low-budget studies, although the original version of the SF-36 is available at no cost.

MEDICAL OUTCOMES STUDY 12-ITEM SHORT FORM HEALTH SURVEY

Description

Purpose. Like the SF-36, the SF-12 is a multi-item generic health profile (3). The SF-12 is a shorter version of the SF-36 that uses only 12 questions to measure functional health and well-being from the patient's perspective. The original objective was to develop a shorter measure that reproduces the two summary scores of the SF-36, the PCS and MCS (65). The questionnaire was originally published in 1996 and was revised in 2000. The revised version contains some changes to item wording, improved formatting, revision of the response scale to incorporate a consistent number of response options, and norm-based scoring (66). Moreover, the SF-12 version 2 is an improvement over the original version in that it can produce the eight domain scales as well as the two component summary scores. Because of these improvements, the revised version is recommended over the original (14). An exception is when the original version has been used in a longitudinal study.

Content or domains. The SF-12 measures the same eight health domains as the SF-36 in addition to overall physical and mental health (PCS and MCS components).

Number of items. The SF-12 consists of 12 items: 2 each for physical functioning, role physical, role emotional, and mental health and 1 each for bodily pain, general health, vitality, and social functioning.

Response options/scale. The response scales for the SF-12 items vary. Two of the physical functioning items have three response options (yes, limited a lot; yes, limited a little; and no, not limited at all). The other 10 items have 5 response options. Eight items have response options based on frequency of problems (all of the time, most of the time, some of the time, a little of the time, and none of the time). The general health item response options are excellent, very good, good, fair, and poor. The pain item response options are not at all, a little bit, moderately, quite a bit, and extremely.

Recall period for items. The standard form has a 4-week recall period, and the acute form has a 1-week recall period.

Cost to use. Quotations on annual license fees are available and vary according to study specifics and whether the survey is used in a commercial or nonprofit setting.

How to obtain. The SF-12 version 2 can be obtained from Optum (https://www.optum.com/en.html). Users are required to register. Manuals can also be purchased.

Practical application

Method of administration. As with the SF-36, the SF-12 can be self- or interviewer-administered, with multiple modes available. A consistent bias for higher scores (better health) has been reported when the SF-12 is interviewer-administered compared with self-administered (10). In a study of telephone versus mail-out administration among veterans (10), mean PCS scores and MCS scores were significantly higher for telephone administration ($P < 0.05$), and telephone administration was approximately 30% more expensive than mail-out administration, primarily because of the cost of labor. IVR reduces this cost difference. IVR and live telephone methods for administering the SF-12 have been compared in a study of patients with back pain, 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; $P < 0.01$) (67).

Scoring. As with the SF-36, the SF-12 contains a mixture of positively worded (higher scores indicate better health) and negatively worded response scales, requiring some items to be recoded prior to scoring. Scoring of the individual items of the SF-12 is identical to that for the SF-36. The domain scores are calculated by summing responses across items when there is more than one item and transforming raw scores to a 0 to 100 scale. Computerized certified scoring algorithms are specified by and available from the developer (Optum), which produce norm-based T scores for each domain (with a mean of 50 and an SD of 10) as well as the PCS and MCS (68). Norms are based on a 1998 United States general population sample. Missing data can be handled by including imputed scores calculated using purpose–developed algorithms. If using the IVR mode, data can be loaded directly into the Optum database for scoring, interpretation, and reporting in real time.

Score interpretation. Scores on the SF-12 scales range from 0 to 100, with higher scores indicating better health. Because the SF-12 uses norm-based scoring developed using normative data for the SF-36 in the United States (21,65), comparisons can be made with the SF-36. Age- and sex-based norms for the SF-12 are available for several countries (65,68,69). Data from general population surveys in nine European countries have shown little difference between standard United States–derived scoring algorithms and country-specific algorithms, and standard scoring algorithms are recommended (70). There is also a very high correlation between SF-12 summary scores derived from United States normative data and those derived from Australian normative data ($r > 0.99$) (71).

Respondent time to complete. The SF-12 takes 2 to 3 minutes to complete, which is less than one-third of the time required for the SF-36 (65).

Administrative burden. No specific training for the administration of the SF-12 is required. The computerized scoring algorithm is available for purchase from the developer (Optum) and requires basic knowledge of statistical software. Scores are not easily computed during a clinic visit.

Translations/adaptations. The SF-12 is available in English and more than 100 other languages, developed as part of the International Quality of Life Assessment Project. Versions in other languages and a translation service, if required, are available through Optum.

Psychometric information

Floor and ceiling effects. There do not appear to be either ceiling or floor effects for the SF-12 PCS and MCS scores among patients with rheumatic conditions (72,73). However, ceiling effects have been reported for the domains of role emotional (17%), vitality (18%), physical functioning (33%), and social functioning (40%), and floor effects have been reported for the physical functioning (20%), general health (20%), role emotional (22%), and bodily pain (28%) scales (74).

Reliability. *Internal consistency.* High internal consistency of the SF-12 component summary scores has been demonstrated in studies of the general population (Cronbach's α of 0.84 or more and 0.75 or more for the SF-12 PCS and MCS, respectively) (75,76) as well as in studies of back pain and rheumatoid arthritis (Cronbach's α = 0.77-0.91 and 0.80-0.91 for the SF-12 PCS and MCS, respectively) (74,77,78).

Test-retest. Although test-retest reliability of the SF-12 has not been evaluated extensively in rheumatic conditions, the results are encouraging. Specifically, among patients with rheumatoid arthritis, the 2-week test-retest reliability for all domains has been found to be high (ICC of 0.991 or more) (74). Test-retest reliability of the SF-12 administered 2 weeks apart has also been shown to be adequate in United States and United Kingdom general populations (PCS r = 0.89 and MCS r = 0.76) (65) and in a healthy community population (PCS ICC = 0.84 and MCS ICC = 0.75) (75). For both the PCS and MCS scales, average changes in scores between test and retest in people who have not reported a change have been found to be less than one point, and at the second administration, 85% score within the 95% CI of the score at the first administration (65).

Validity. *Face/content.* The SF-12 appears to have good face validity, with all items referring to health-related issues.

Criterion. Criterion-related validity of the SF-12 is difficult to establish because of the absence of a gold standard for measuring health. However, given that the primary purpose of the SF-12 was to reproduce the PCS and MCS 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 correlate 0.95 and 0.96 with the SF-36 PCS and MCS, respectively (65,70). These findings of criterion validity have been replicated in a study of the general populations of nine European countries (Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, and the United Kingdom) (70), with very high correlations between the SF-12 PCS and SF-36 PCS (r = 0.94-0.96) and SF-12 MCS and SF-36 MCS (r = 0.94-0.97), as well as in studies of Greek and Australian general populations (71,79). Clinical trials data from patients with osteoarthritis and rheumatoid arthritis also indicate good criterion validity of the SF-12 in rheumatic conditions, with strong correlations between the SF-12 PCS and SF-36 PCS and the SF-12 MCS and SF-36 MCS (r = 0.92-0.96) (72).

Construct. The two-factor structure of the SF-12 (PCS and MCS) has been confirmed using factor analysis in population-based (79,80) and clinical studies (72,74,81). However, one study has challenged the uncorrelated two-factor structure of the SF-12 among the elderly, people with Parkinsons's disease, and people with stroke (82).

Convergent and discriminant validity of the SF-12 when measured in the general population is supported by relationships found with the EQ-5D (79), in which comparable SF-12 summary scores and EQ-5D dimensions had stronger correlations than less comparable summary scores and dimensions. However, results for the convergent and discriminant validity of the SF-12 in rheumatic conditions are somewhat variable. In Danish patients with rheumatoid arthritis, the SF-12 PCS and MCS were found to have unexpectedly weak correlations with similar constructs, including the HAQ (PCS r = –0.15 and MCS r = –0.25), albeit correlations with dissimilar domains were even lower (73). In contrast, in a study of Asian patients with rheumatoid arthritis, moderately strong correlations were found with the Rheumatoid Arthritis Impact of Disease score (PCS r = –0.45 and MCS r = –0.52) (83). Similarly, in spinal clinic patients, back pain and disability have been found to be significantly moderately correlated with the SF-12 PCS (r = –0.41 and –0.63, respectively; $P < 0.0001$) and MCS (r = –0.33 and –0.55, respectively; $P < 0.0001$) (77).

Among the Greek general population, the SF-12 PCS has been found to be significantly worse among those reporting hip and knee problems compared with those not reporting such problems ($P < 0.01$) (79), supporting known-groups validity. In addition, among the Taiwanese elderly population, the SF-12 PCS has been found to be able to differentiate women with and without knee osteoarthritis ($P < 0.043$), and the SF-12 MCS has been found to be able to differentiate men with and without knee osteoarthritis ($P < 0.02$) (84). However, among Danish patients with rheumatoid arthritis, the SF-12 was unable to differentiate between groups based on disease severity and treatment options (73).

Responsiveness. There is limited information on the responsiveness of the SF-12 in rheumatic conditions. There is some evidence among those attending a spinal clinic with back pain (77), with a large ES for SF-12 PCS (ES = 0.82) and a small ES for SF-12 MCS (ES = 0.37) observed in patients whose self-reported back pain became much better after 3 to 6 months of follow-up. Similarly, small ESs for SF-12 PCS and MCS (ES = –0.46 and –0.21, respectively) were observed in patients whose self-reported back pain became much worse. The SF-12 has also been shown to be able to discriminate between patients with rheumatoid arthritis of the hands/wrists who have high pain and those with low pain (PCS ES = 0.81 and MCS ES = 0.33) (78). In addition, the responsiveness of the SF-12 to a wide range of treatments and programs for musculoskeletal conditions has been reported (85–95).

Minimally important differences. There are few investigations of MIDs for the SF-12 in rheumatic conditions, and findings are variable depending on the patient group and methodological approach. Among patients having undergone total knee arthroplasty, a minimal clinically important difference (MCID) value of 5 has been suggested for the PCS using a distribution-based method (half the SD of outcome change scores) (96) and a value of 4.5 to 4.8 has been suggested using an anchor-based method (with patient-perceived pain relief and functional outcomes as the external criterion) (97). Among patients with subacute and chronic low back pain, a value of 3.29 for PCS scores and 3.77 for MCS scores has been suggested based on four different anchor methods (with self-reported health status as the external criterion) (98).

More recently, Clement et al (99) recommended an MCID of 1.8 for the PCS after total knee arthroplasty for comparing group outcomes and a minimal important change of 2.7 for ensuring a clinical difference has been observed among a cohort of patients (based on patient-perceived improvement in quality of life as the external criterion). The MCS was found not to be an appropriate measure for assessing the group effects of total knee arthroplasty because the MCID and MID did not reach statistical significance. To assess whether or not an individual patient has experienced change, Clement et al (99) recommended a minimal detectable change (MDC) of 8.9 for the PCS and an MDC of 13.8 for the MCS (based on a distribution method using the SEM, and the MDCs should therefore be independent of the type of intervention).

Generalizability. The SF-12 has been shown to have good applicability across varying populations, including those with rheumatic conditions.

Use in clinical trials. The SF-12 has been used as an outcome measure in clinical trials to evaluate the efficacy of a broad range of interventions for rheumatic conditions, including pharmacological treatment for osteoarthritis of the hip, knee, and hand (72,89,100,101); hydrotherapy treatment for osteoarthritis (86) and fibromyalgia (85); Tai Chi for hip or knee osteoarthritis (87); hand exercises for rheumatoid arthritis (102); foot orthoses for rheumatoid arthritis (103); self-management for fibromyalgia (91,104); surgical procedures, eg, total hip arthroplasty (105–107) and total knee arthroplasty (108–111); and postsurgical rehabilitation (exercise following total knee arthroplasty) (112).

Critical appraisal of overall value to the rheumatology community

Strengths. The SF-12 is brief, appears to adequately reproduce the two summary scores of the SF-36, and is available in many languages. It is especially suitable when comparisons between disease groups or with the general population are required. Availability of population norms also provides context for score interpretation. It can differentiate between levels of disease severity in rheumatic conditions and between people with and without rheumatic conditions and can respond to treatment-related changes in the health status of people with rheumatic conditions.

Caveats and cautions. Because more items permit better representation of each domain, the domains are best represented by the SF-36. Therefore, the most useful measures derived from the SF-12 are the two aggregate summary measures: the PCS and MCS. In contrast to the extensive SF-36 literature, the psychometric properties of the SF-12 have been less well studied in rheumatic conditions. 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 also limited evidence of its responsiveness to treatment-related changes in the health status of people with rheumatic conditions.

Clinical usability. Psychometric evaluation of the SF-12 does not support interpretation of scores to make decisions for individuals and, thus, limits its clinical use. However, the SF-12 may have satisfactory ability to detect treatment-related effects at a group level.

Research usability. The SF-12 is a suitable measure for large group studies (greater than $n = 500$) in which information on the SF-36 PCS and MCS is required to measure health status and for monitoring health outcomes over time (65). Low respondent and administrative burden and the availability of multiple modes of administration, including online administration, support the usability of the SF-12 in research settings. However, licensing costs can limit its use in low-budget studies.

NOTTINGHAM HEALTH PROFILE

Description

Purpose. The NHP is a generic health profile designed for measuring perceived health status and its impact on daily life (113). It provides a profile of an individual's perceived emotional, social, and physical health and is intended for use in the general population as well as in clinical settings (114).

Content or domains. The NHP has two parts. Part 1 covers six domains of functioning that are expected to be affected by ill health, including physical mobility, energy, sleep, pain, social isolation, and emotional reactions (113). Part 2 measures the impact of ill health on the areas of daily life that are expected to be most often affected by health (113). The two parts of the NHP can be used together or separately, with Part 1 most frequently used on its own. This review will focus on Part 1.

There is no clear information within the NHP development literature on whether an explicit conceptual framework was used to guide the development and evaluation of this questionnaire. Items for the NHP were derived from interviews of 768 individuals with a range of acute and chronic health conditions (115). The interviews produced 2200 statements describing the typical effects of ill health (social, psychological, behavioral, and physical). The statements were tested on their ability to differentiate between degrees of impairment and between physical and mental health conditions (116). Following additional review of item content, 38 statements that were easy to understand, unambiguous, and easily answerable using a yes/no response format were included in the final version of the NHP Part 1 (116). No further specific information on the characteristics of the participants in the studies of item refinement and selection is available.

Number of items. The NHP Part 1 has 38 items, including 3 energy items, 8 pain items, 9 emotional reactions items, 5 sleep items, 5 social isolation items, and 8 physical mobility items. Part 2 has seven items that cover the impact of health on paid employment, ability to do jobs around the house, social life, family relationships, sex life, hobbies, and vacations/holidays (114).

Response options/scale. Responses to the NHP Parts 1 and 2 are scored on a dichotomous yes/no scale.

Recall period for items. Respondents are asked to identify whether each statement applies to them at the moment. If unsure, the respondents are instructed to select an answer that is most true at the moment.

Cost to use. The use of the NHP incurs a small administrative fee for noncommercial studies, and there is a license fee for commercial use. Any inquiries should be sent to gr@galen-research.com. To facilitate the processing of inquiries, Galen Research requests that emails requesting information about the NHP include a short summary or protocol of the study, along with any funding information.

How to obtain. Current copyright of the NHP is held by Galen Research (http://www.galen-research.com/measures-database/), and a sample copy of the NHP can be obtained by contacting Galen Research customer service (gr@galen-research.com). The first page of the NHP can also be viewed at http://www.galen-research.com/content/measures/NHP%20UK%20-%20First%20page%20sample.pdf.

Practical application

Method of administration. The NHP is designed for self-administration but has also been administered through face-to-face interviews (113,117,118). The most common mode of administration is paper and pencil format.

Scoring. A scoring algorithm is available with the purchase of the questionnaire. Scores for each of the six domains in Part 1 are computed by summing weighted values assigned to each positive response. The weights were derived using Thurstone's method of paired comparisons from interviews of 1200 patients and members of the general public (119). The sum of the weighted scores is 100 for each domain, 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 (115). Only domain scores are calculated for Part 1, with no overall score. The seven statements in Part 2 are only intended to be administered to individuals who have a health condition. Part 2 items are unweighted, with the number of "yes" responses added to produce a score that ranges from 0 to 7 (116). Part 2 items can also be reported as the proportion of individuals who selected a "yes" response on a given item (120).

Score interpretation. The scores on NHP Part 1 represent the severity of perceived dysfunction in each domain of functioning and range from 0 (best health state) to 100 (worst health state) (114).

At least two large-scale studies have been carried out to develop norms for the NHP Part 1 according to sex, age, and socioeconomic status. The participants in the first study were 2192 individuals drawn from a register of a general practice near Nottingham in the United Kingdom (121). The participants in the second study were 1753 employees of a large company (120). Both studies utilized postal surveys to collect data and had reasonable response rates: 68% (2173 usable questionnaires returned of 3200 posted) and 58% (1753 usable questionnaires returned of 3000 posted) in the first and second studies, respectively (120). Additional information on normative studies and normative data can be found in the NHP user's manual (122).

Respondent time to complete. The NHP has a low respondent burden, with an estimated completion time of 5 to 10 minutes for Part 1 (29). Although information on usability of NHP in rheumatic conditions settings has not been reported, in a field-testing study of the Hungarian NHP among 29 patients with chronic kidney disease, the average NHP completion time was 6.4 minutes (SD 2.9 minutes) (123).

Administrative burden. Scoring and administration instructions are self-explanatory and require no specific training. The administrative burden of the NHP in rheumatic conditions has not been evaluated specifically. A study of feasibility of the German NHP in nursing home residents reported an average of 12.6 (SD +6.0) minutes to complete data collection through a face-to-face interview (118).

Translations/adaptations. The NHP was initially developed in English (United Kingdom) and is now available in 28 languages and/or cultural adaptations, including an Australian English adaptation (124) and translations and adaptations into Greek (125), French (126), Danish (127), Swedish (128), Dutch (129), Hungarian (123), and Spanish (130). An extended list of translations and adaptations is available at http://www.galen-research.com/measures-database/. It should also be noted that cautions have been previously raised about the cross-cultural comparability of item weights (126,130), which could potentially impact the applicability of the NHP in multicountry settings or the comparability of scores obtained from the questionnaire's different translations and adaptations.

Short forms. Two shorter unidimensional versions of the NHP have been developed: the NHP index of Distress (NHPD) and the Spanish NHP short version. The NHPD utilizes 24 of the 38 NHP Part 1 items and was designed to enable the calculation of quality-adjusted life years (QALYs) (131). The NHPD calculation omits the eight items comprising the physical mobility section and an additional six items that have been deemed not relevant to hospitalized patients (131). The NHPD showed evidence of unidimensionality and construct validity, evaluated through the application of Rasch modeling, in patients with Parkinson's disease and patients with peripheral arterial disease (132). However, thus far, the NHPD does not appear to have been utilized or evaluated in rheumatic conditions.

The 22-item short version of the Spanish NHP has been derived through the application of Rasch modeling to a sample of 9419 individuals from the general population and clinical settings (including 1243 individuals with musculoskeletal conditions). The resultant measure contains 11 physical and 11 psychological items that can be scored as separate scales or used together as an overall score (133).

Psychometric information

Floor and ceiling effects. Although the NHP was initially developed for use in the general population, the items in Part 1 represent severe health-related problems (115). Hence, it is not surprising that subsequent applications of the NHP identified either pronounced ceiling or floor effects. Among community-dwelling individuals aged 65 years and older, ceiling effects were present for all domains, ranging from 38% for physical mobility to 68% for social isolation (134). There was a decrease in ceiling effects with age in all dimensions, and no substantive floor effects were observed in this study population. Conversely, in a sample of 409 outpatients visiting a hospital for medical examinations, floor effects were present for all domains, ranging from 44% (sleep) to 84% (social isolation) (38). Ceiling effects were negligible and ranged from 0.2% (social isolation) to 7% (energy) (38). Both ceiling and floor effects were observed in a study of 60 patients with hereditary neuromuscular disease (43% wheelchair users) (135). Only the physical mobility (9%) and pain (11%) domains were relatively free from ceiling effects in this population, with substantive ceiling effects recorded for the pain (37%), emotional reactions (40%), social isolation (59%), and sleep (71%) domains and substantive floor effects for the energy (27%) and physical mobility (33%) domains (135).

Reliability. *Internal consistency.* 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 domain was found to be acceptable in a sample of 160 people with rheumatoid arthritis (Cronbach's α of 0.83 or more) (136). In another study, conducted with a sample of 116 patients with rheumatoid arthritis, the internal consistency of the NHP energy, physical mobility, and pain domains was evaluated and was reported to be acceptable for the physical mobility (Cronbach's $\alpha = 0.82$) and pain (Cronbach's $\alpha = 0.77$) domains but not for the energy domain (Cronbach's $\alpha = 0.63$) (137). Among wheelchair-dependent individuals (including 30 with rheumatoid arthritis), the internal consistency of the physical mobility and social isolation domains was suboptimal (Cronbach's $\alpha = 0.69$ and 0.64, respectively) but was good for the remaining domains (Cronbach's α of 0.76 or more) (138).

In the general population, the sleep (Cronbach's $\alpha = 0.67$) (38) and social isolation (Cronbach's $\alpha = 0.65$-0.66) (38,139) domains have been reported to have internal consistency slightly below the lower limit of 0.70 recommended for group assessment. Internal consistency of the remaining domains was reported to be acceptable in at least two population-based studies (38,139) (Cronbach's α of 0.71 or more).

Test-retest reliability. Test-retest reliability of the NHP in rheumatic conditions has been previously evaluated, although a number of studies utilized Pearson's or Spearman's correlation coefficients to assess test-retest reliability (114,120,128). Both of these coefficients provide a suboptimal indication of temporal stability because of their inability to capture systematic changes in scores over time, thereby potentially overestimating the true test-retest reliability of a measure.

Studies that used ICCs to assess test-retest reliability of NHP reported mixed results. In a sample of 49 employed individuals with stable musculoskeletal disorders, 3-week test-retest reliability of NHP domains was satisfactory, ranging from 0.76 (physical mobility) to 0.87 (pain) (140). On the other hand, in a study of 116 individuals with stable rheumatoid arthritis, test-retest reliability was acceptable for physical mobility (ICC = 0.73), borderline for pain (ICC = 0.68), and suboptimal for energy (ICC = 0.53) (137). However, the study by Bouchet et al (137) used a test-retest interval of 1 year, thus the results might reflect real change in the NHP domains over time.

Validity. *Face/content.* The NHP appears to have good face validity as a measure of perceived health, with all items referring to an aspect of health. The NHP covers a broad range of health-related functions (physical abilities, pain, and sleep) that could be expected to be affected in rheumatic conditions. In the development of the NHP, items were derived from patient interviews, thus increasing the relevance of questionnaire items to patients. However, the content validity of the

NHP in rheumatic settings has only been evaluated for low back pain and was found to be low for capturing core outcomes (42).

Construct. Factorial validity of the NHP Part 1 has not been evaluated in rheumatic conditions and received limited support in community settings (141). At least three studies examined construct validity of the NHP in rheumatoid arthritis. Although interpretation of results of these studies is hampered by the absence of clear convergent and discriminant validity hypotheses, broadly, as might be expected, the physical mobility, pain, and energy domains show stronger correlations with disease-specific measures (including grip strength, pain, stiffness, the Ritchie Articular Index, the HAQ, and the Modified Disease Activity Score) than emotional reactions, social isolation, and sleep do (142). The emotional reactions domain was also moderately correlated with the Beck Depression Inventory ($r = 0.54$; $P < 0.001$) (142), supporting convergent validity of this domain.

Construct validity of the NHP was also examined in at least two studies involving patients with knee osteoarthritis, with the results lending mixed support for the convergence between its physical domains and disease activity measures. In a sample of 50 outpatients, all domains of the NHP were reported to have strong and significant correlations (r of 0.9 or more; $P < 0.01$) with the range of motion and the Kellgren-Lawrence score (143). However, in an earlier study involving 140 individuals with knee osteoarthritis, only the energy and physical mobility domains were weakly correlated with either the Kellgren-Lawrence or range of motion scores (r of 0.31 or less; $P < 0.01$), whereas the remaining domains were uncorrelated with disease activity measures (144).

Support for the known-groups validity of NHP is more consistent, with the evidence to support the ability of the NHP to differentiate people with rheumatoid arthritis from community-dwelling adults in good health (142,145) and from patients with other chronic diseases (145,146). The NHP was also able to differentiate individuals with osteoarthritis from healthy controls (143,144) and between individuals with and without fibromyalgia (147).

Responsiveness. Information about the ability of the NHP to detect change in rheumatic conditions is limited and does not consistently support the responsiveness of individual domains. Most of the information about the responsiveness of the NHP comes from observational studies, and the ability of this questionnaire to detect intervention effects in rheumatic conditions is currently not known. In a longitudinal study of recovery among patients with hip fracture, physical mobility (ES = 0.61) and pain (ES = 0.55) showed moderate improvement, and social isolation showed a small improvement (ES = 0.21) at 1 month postfracture compared with 1 week postfracture. At 4 months, improvements were large in physical mobility (ES = 1.48) and pain (ES = 0.95) and were small in sleep (ES = 0.35) and energy (ES = 0.37) (148). No clinically meaningful improvements were recorded for the remaining domains of the NHP at either 1 or 4 months postfracture.

Studies that have compared responsiveness of the NHP with that of other questionnaires indicate 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 the health status of individuals with rheumatoid arthritis was compared with that of one generic (Functional Limitations Profile) and two arthritis-specific (the Arthritis Impact Measurement Scales and the HAQ) questionnaires (149). The NHP had the lowest ability to detect self-reported change in physical mobility (ES = 0.27), pain (ES = 0.38), and emotional reactions (ES = 0.59), with only small to moderate changes recorded. Standardized ESs for other questionnaires measuring similar concepts were moderate to high, ranging from 0.69 to 0.83. In the social domain, the NHP (ES = 0.24) was worse at detecting change than the Functional Limitations Profile (ES = 0.60) but better than the Arthritis Impact Measurement Scales (ES = 0.06). Similarly, in a study of individuals with a range of musculoskeletal disorders who were attending injury rehabilitation clinics, the NHP showed small to moderate changes (ESs ranged from 0.32 [sleep] to 0.57 [pain]) on all domains except emotional reactions (ES = 0.18) (140). The magnitude of changes was generally comparable with those recorded for the Duke Health Profile, the Sickness Impact Profile, and the SF-36, although SF-36 recorded a larger standardized change in pain (ES = 0.99) than the NHP (ES = 0.57).

Minimally important differences. No information on MIDs for the NHP is currently available.

Generalizability. The presence of floor effects in populations with high levels of disability and ceiling effects in relatively well populations indicate that the utility of the NHP might be optimal in population groups with moderate levels of health impairment. Additionally, most psychometric studies of the NHP in rheumatic conditions settings were undertaken with individuals with rheumatoid arthritis or osteoarthritis, and information on the functioning of this questionnaire in other rheumatic populations is currently limited.

Use in clinical trials. Examples of the uses of the NHP in randomized controlled trials in rheumatic conditions include evaluation of Kinesio taping versus sham taping (150) and self-administered superficial local heat versus cold (151) in knee osteoarthritis and exercise interventions in fibromyalgia (152,153). In all of these trials, the NHP was utilized as a measure of HRQOL, despite its intended use as a measure of perceived health.

Critical appraisal of overall value to the rheumatology community

Strengths. Being a generic measure, the main advantage of the NHP is that it can be used to compare the impact of rheumatic conditions with that of other illnesses or with the general population. The NHP is available in a number of languages and cultural adaptations, which facilitates its applications in multicountry and multicultural settings. Other strengths of the NHP include its relative brevity and ease of administration and scoring; its items are easy to understand, there is an unambiguous response scale, and population norms are available. The NHP domains are also intended to be independent of each other, and therefore researchers can select and administer only those domains that are relevant to their study aims.

Caveats and cautions. Although studies assessing construct validity of the NHP generally produce favorable results, the strongest support exists for the convergent validity of the physical domains of the NHP and known-groups validity in rheumatology settings, with limited information on factorial validity. Consistently suboptimal internal consistency and test-retest reliability of some domains (emotional reactions and social isolation) raise doubts about the reproducibility and measurement precision of the NHP in rheumatic conditions. Although the NHP is available in several translations and cultural adaptations, the NHP item weights tend to differ between countries, and this has the potential to affect the meaning of domain scores, and, consequently, the cross-cultural comparability of the NHP. Furthermore, the NHP appears to be less sensitive to change than other generic health status measures, and there is currently no information on what constitutes the MID for the NHP domains.

Clinical usability. The NHP is easy to use and score. Part 1 can be used to derive a moderately detailed picture of the patient's current health, and responses to Part 2 could serve to flag areas for further assessment of disease impact. However, financial costs associated with the use of the NHP could limit its usefulness in clinical settings.

Research usability. Sensitivity to change of the NHP is lower than that of other instruments measuring similar aspects of health; thus its use in longitudinal studies and clinical trials is less assured. As a measure of general health, the NHP may be of greater interest in epidemiologic research, although ceiling effects in well populations may limit its usefulness in population-based studies.

WORLD HEALTH ORGANIZATION QUALITY OF LIFE SHORT VERSION INSTRUMENT

Description

Purpose. The WHOQOL-BREF is a non–disease-specific quality of life measure that forms part of a suite of instruments developed by the WHOQOL Group (154). The WHOQOL instruments provide a broad profile of an individual's life rather than health status or HRQOL and were developed to enable a cross-cultural assessment of quality of life by considering an individual's beliefs, life situation, and environmental context (155,156). The WHOQOL-BREF also uniquely considers the degree of satisfaction with aspects of an individual's life and health that are relevant to rheumatology settings.

Content or domains. The instrument covers the following four domains: physical health, psychological, social relationships, and environment.

The WHOQOL-BREF was derived from the longer WHOQOL-100 quality of life instrument, using data from 15 international collaborating centers (156).

Number of items. The WHOQOL-BREF contains 26 items relating to overall quality of life and general health.

Response options/scale. Each item uses a five-point Likert scale, with endpoint descriptors ranging from either very poor to very good, from very dissatisfied to very satisfied, from not at all to an extreme amount, from not at all to extremely, from not at all to completely, from very poor to very good, and from never to always, depending on the individual item.

Recall period for items. Items 1 and 2 assess overall quality of life and overall satisfaction with one's health and do not specify a recall period. Items 3 to 26 use a 2-week recall period.

Cost to use. There is no cost associated with the use of the WHOQOL-BREF.

How to obtain. The United States English version of the instrument and scoring syntax can be freely obtained after completing a user agreement (http://depts.washington.edu/seaqol/WHOQOL-BREF). Translated versions can be obtained by contacting the WHOQOL instrument coordinator (whoqol@who.int).

Practical application

Method of administration. The WHOQOL-BREF instrument can either be self-administered or administered by an interviewer.

Scoring. A score for each of the four WHOQOL-BREF domains is obtained by summing the item responses within each domain and then calculating the mean. Raw domain scores are transformed to a 4 to 20 scale but may also be converted to a 0 to 100 scale. Domain scores can also be generated using an SPSS syntax file, which is available on request (154).

A score cannot be calculated when more than 20% of the items have missing responses (154). When responses to two items or fewer within a domain are missing, the missing response can be substituted with the mean of the remaining items within the domain. When more than two items are missing, the domain score is generally not calculated (except for the three-item social relationships domain, for which calculation is only recommended if there is one missing item response or no missing responses). The social relationships domain is known to have issues with missing data, which likely reflects the inclusion of an item concerning sexual relationships. Only items 3 to 26 are used in the calculation of domain scores.

Score interpretation. The WHOQOL-BREF domain scores are positively scaled, with higher scores indicating a higher quality of life for that domain.

Age- and sex-adjusted normative domain scores are available for 23 countries from a large psychometric evaluation involving over 11 000 people (156). Country-specific population norms are also available for some countries (eg, Thailand, Israel, India, Australia, India, Panama, the United States, the Netherlands, Croatia, Japan, Zimbabwe, Spain, the United Kingdom, Russia, and France) (154,157).

Respondent time to complete. Administration of the WHOQOL-BREF is generally time efficient, with one study of 751 randomly selected people in Brazil reporting a mean completion time of only 8 minutes (158). A WHOQOL-BREF validation study involving 1200 people from an elderly Taiwanese community found response times to range from 4 to 20 minutes (mean of 11 minutes) (159). Electronic administration does not appear to greatly reduce respondent time. A cross-sectional study of 98 elderly people in Brazil found mean respondent times of the Web-based WHOQOL-BREF to only be 33 seconds shorter than the paper version (mean of 12 minutes and 9 seconds) (160). WHOQOL-BREF studies undertaken in rheumatic conditions have not reported completion times (161).

Administrative burden. There is minimal administrative burden associated with using the WHOQOL-BREF. Interviewer-administered times in a study of elderly members of a Taiwanese community ranged from 8 to 32 minutes (mean of 15 minutes) (159). Administration instructions are outlined in the WHOQOL-BREF manual and are easy to follow (154). No specific software is required to administer the WHOQOL-BREF, and a clear scoring algorithm is provided in the user manual.

Translations/adaptations. The WHOQOL-BREF has been translated into 19 different languages, and a list of translated language versions can be obtained by emailing the WHOQOL Group (whoqol@who.ch). Cross-cultural equivalence was evaluated in a pilot study of the Somali version of the WHOQOL-BREF involving 303 Somali refugees in the United States. This study involved pretesting a direct translation of the original WHOQOL-BREF on a smaller sample of three non–English-speaking Somali individuals. Following their feedback, some items were reworded and compared with the original WHOQOL-BREF to maximize equivalence (162). Concerns have been raised regarding the sensitive nature of one of the items in the social relationships domain regarding sexual activity (163,164), with greater missing data for this item reported in one Bangladeshi study (164). This has implications for the cultural appropriateness of this item in settings where sexual relationships are less openly discussed.

Short forms. A shortened five-item version of the WHOQOL-BREF (the WHOQOL-5) has also been described in the literature (165) but does not appear to have undergone extensive psychometric evaluation.

Psychometric information

Floor and ceiling effects. Overall, floor effects were minimal in studies evaluating the WHOQOL-BREF questionnaire. In one study that assessed patients undergoing hip or knee replacement surgery, negligible floor effects for the WHOQOL-BREF were reported preoperatively (0%-2.2%; n = 279) and at 3 months postoperatively (0%-1.4%; n = 74) (166). Although other studies on the WHOQOL-BREF in rheumatic conditions are limited, in a study involving elderly people with various health conditions in Taiwan, Hwang et al reported floor effects ranging from 0.0% to 0.3% across the four domains (159).

Likewise, ceiling effects were also found to be minimal for the WHOQOL-BREF. Ackerman et al (166) reported negligible to low ceiling effects for patients undergoing joint replacement surgery across all domains, ranging from 0% to 9% at baseline and from 0% to 8.1% at 3 months. Low ceiling effects were detected across all domains (0.0%-0.8%) in the Taiwanese study of elderly people (159).

Reliability. *Internal consistency.* In a study involving people undergoing joint replacement surgery for arthritis, good internal consistency was reported for the WHOQOL-BREF, with all domains demonstrating a Cronbach's α of 0.70 or more (166). Only one study, which involved 214 patients with rheumatoid arthritis, reported poor internal consistency for the social relationships domain, with an α of 0.64 (161).

More broadly, the WHOQOL-BREF has been shown to have good internal consistency in the community-based samples. A study involving 4628 people in the United Kingdom with and without a range of health conditions (including arthritis) also reported high internal consistency (Cronbach's α = 0.92) (167). Similarly, a cross-sectional Brazilian study of quality of life involving 278 older people (mean age of 64 years) reported an overall Cronbach's α of 0.83 (168). Another study assessing the psychometric properties of WHOQOL-BREF in 1316 Singaporean adults reported a Cronbach's α of 0.70 or more for all domains (169).

Test-retest reliability. Evidence of test-retest reliability for the WHOQOL-BREF among people with rheumatic conditions is limited, with only one study involving patients with rheumatoid arthritis. This study found that the WHOQOL-BREF had good temporal stability (mean test-retest interval of 3.5 days), as indicated by high ICCs for the physical health (ICC = 0.79), psychological (ICC = 0.86), social relationships (ICC = 0.91), and environment (ICC = 0.72) domains (161). Similarly, good test-retest reliability of the WHOQOL-BREF has been demonstrated for other chronic conditions. In a study of 199 individuals with traumatic brain injury, test-retest reliability was acceptable across the physical health (ICC = 0.86), psychological (ICC = 0.95), social relationships (ICC = 0.74), and environment (ICC = 0.90) domains over a 2-week interval (170). In another study involving 39 individuals with depression, excellent test-retest reliability (1-week test-retest interval) was reported, with an ICC score of 0.92 for the overall score based on 26 items (171). However, ICC scores for each domain were not reported.

Validity. *Face/content.* The WHOQOL-BREF has good face validity for the assessment of quality of life in rheumatic conditions given its ability to broadly address the physical health, psychological, social relationships, and environmental aspects of health. Content validity is also supported by the relatively low prevalence of floor and ceiling effects, highlighting the questionnaire's ability to capture a wide continuum of health states.

Construct. Factorial validity of the WHOQOL-BREF is yet to be evaluated in those with rheumatic conditions. Convergent validity was generally demonstrated for the WHOQOL-BREF, although there is limited evidence from rheumatology studies. Coulbourn et al found a significant correlation between higher disability weight (derived from the Global Burden of Disease 2004 study) and lower scores on all the domains of the WHOQOL-BREF (172). The WHOQOL-BREF has also demonstrated good correlation with other health measures, including the General Health Questionnaire-28 and Life Satisfaction Index (173). However, a Brazilian study with older adults found there to be only a fair correlation between WHOQOL-BREF and SF-36 scores (Pearson's *r* of 0.6 or less) (168).

In studies comparing quality of life between those with musculoskeletal disease and healthy populations, WHOQOL-BREF scores differentiated between the two groups. A prospective study of 76 patients with adhesive capsulitis and 72 healthy controls also found between-group differences in the WHOQOL-BREF physical health, psychological, and environmental domain scores (174). Other studies also support known-groups validity of the WHOQOL-BREF, with significantly higher quality of life scores reported for patients in relatively good health versus those in poorer health (158,159,172,173,175,176).

Responsiveness. The evidence regarding the responsiveness of the WHOQOL-BREF in rheumatic conditions is mixed. A study involving patients with rheumatoid arthritis reported moderate to high responsiveness for the WHOQOL-BREF domains, with acceptable SRMs reported for the physical health (1.05) and psychological (0.59) domains and suboptimal SRM values for the social relationships (0.46) and environment (0.50) domains (161). The SRM for the WHOQOL-BREF physical health domain was found to be higher than that of the disease-specific HAQ, indicating that the WHOQOL-BREF can detect changes in quality of life among people with rheumatoid arthritis (161).

In contrast, studies using the relative efficiency (RE) approach to assessing responsiveness have found the WHOQOL-BREF to be less responsive to change than disease-specific instruments. A study involving patients undergoing hip or knee replacement surgery reported the WHOQOL-BREF domains to have low to moderate responsiveness compared with the disease-specific WOMAC scales. RE values ranged from 0.01 for the social relationships domain to 0.50 for the physical health domain; in comparison, RE values ranged from 0.66 for the WOMAC stiffness scale to 1.00 (assigned as the reference) for the WOMAC physical function scales (166). A similar pattern was seen for the ESs, with smaller values reported for the WHOQOL-BREF domains (0.13-0.98) than for the WOMAC scales (1.24-1.69) (166). However, in that study, the WHOQOL-BREF physical health domain demonstrated a larger ES than other generic measures (including the AQoL and the Kessler Psychological Distress Scale) and an equivalent ES to that reported for the Modified HAQ.

Minimally important differences. A minimally important difference for the WHOQOL-BREF instrument has not been reported for people with musculoskeletal conditions or for general populations.

Generalizability. Although the psychometric properties of the WHOQOL-BREF have not been well studied in rheumatic conditions, the questionnaire has promise for use as a measure of quality of life in rheumatic conditions. The concepts covered by the WHOQOL-BREF are broadly applicable to most rheumatic conditions, and there appear to be no floor or ceiling effects, indicating that this measure can be used with both severely and mildly impacted patient groups.

Use in clinical trials. The WHOQOL-BREF has been used to evaluate quality of life outcomes in at least one relevant randomized controlled trial. The study investigated the use of assistive devices for hand osteoarthritis compared with the provision of written information (177).

Critical appraisal of overall value to the rheumatology community

Strengths. The WHOQOL-BREF demonstrates good convergent validity and is able to measure non–disease-specific quality of life. Translated versions, as well as population norms for a range of countries, are available. From the limited rheumatology studies, the WHOQOL-BREF appears to be a reliable (with the exception of the social relationships domain) and reasonably responsive (particularly the physical health domain) measurement tool. It is free to use and is easy to administer and score.

Caveats and cautions. Although the WHOQOL-BREF has been extensively validated in general populations internationally, the literature is still limited regarding the psychometric properties of the WHOQOL-BREF in rheumatic conditions. There are also concerns about the low internal consistency reported for the social relationships domain, and this likely relates to the few items contained within this domain and likelihood of missing data. No MIDs have been reported for rheumatic conditions or for general populations, and hence the utility of the WHOQOL-BREF for assessing change in quality of life in longitudinal studies is currently not known.

Clinical usability. In clinical settings, the WHOQOL-BREF offers a method of measuring overall quality of life that can be compared with available population norms. Although it contains 26 items, reports indicate that these can be completed relatively quickly. It can be completed independently or with assistance in numerous translated versions.

Research usability. In research settings, the WHOQOL-BREF can be used to measure overall quality of life, providing complementary information to disease-specific measures. The low administrative burden regarding accessing the instrument and generating domain scores further supports its research usability. The WHOQOL-BREF is easy to administer and score and is freely available, which may enhance its appeal for low-budget studies.

PROMIS GLOBAL HEALTH

Description

Purpose. The PROMIS-GH is a generic health status measure that was developed by the PROMIS initiative. The initiative was established in the United States in 2004 with the aim to standardize the assessment of patient-reported outcomes (PROs) by developing and validating item banks on mental, physical, and social domains of health (178).

The original PROMIS-GH, version 1.0, has been updated over the years, and it is recommended that the most recent version be used (currently version 1.2). According to the PROMIS-GH scoring manual (http://www.healthmeasures.net/index.php?option=com_instruments&view=measure&id=778&Itemid=992), the PROMIS Adult Global version 1.0 and version 1.1 are identical except for minor revisions of wording for one item. The PROMIS Adult Global version 1.2 was revised to enable automatic item response theory (IRT) scoring, which included revision of reverse-scored items to ensure uniform direction of scoring. Of note, most studies summarized in this review did not provide information on the version used. Where it was available, the version number is stated.

Content or domains. PROMIS-GH contains the following two scales: global physical health (GPH) (four items) and global mental health (GMH) (four items). It also contains two additional single items that assess general health and satisfaction with social roles. The authors of the original paper recommend producing scores for the GPH and GMH scales and scoring the other two items as single items (179). Therefore, although some of the studies included in this review undertook analyses on a total PROMIS-GH score, we only focus on the two domain scores.

During the first PROMIS wave, 11 item banks were developed along with the 10-item PROMIS-GH. Respective PROMIS measures generally consist of items from existing PRO instruments, but PROMIS investigators also wrote new items to complement the initial item pools. Patient input was not sought during the item development process, but patient interviews were undertaken on the PROMIS response options. Furthermore, both experts and patients reviewed the items before the quantitative evaluation and calibration of the final item banks (178). Although most PROMIS measures capture specific, unidimensional constructs, including emotional distress, fatigue, or physical function, the PROMIS-GH is intended to combine global health items from the different PROMIS domains to provide a global assessment of overall health status (178,179).

Number of items. The PROMIS-GH consists of 10 items.

Response options/scale. Responses are scored on a Likert-type scale. Five-category response scales are used, except in the case of average rating of pain, which is rated from 0 (no pain) to 10 (worst pain imaginable). Of the remaining nine items, six (general health, general quality of life, general physical health, general mental health, satisfaction with social activities and relationships, and degree of carrying out usual activities and roles) are scored from excellent to poor, one (ability to carry out physical activities) is scored from completely to not at all, one (bother with emotional problems) is scored from never to always, and one (average fatigue) is scored from none to very severe.

Recall period for items. No recall period is specified for the first seven items (ie, general health, general quality of life, general physical health, general mental health, satisfaction with social activities and relationships, degree of carrying out usual activities and roles, and ability to carry out physical activities). The recall period for the remaining three items (ie, items related to bother with emotional problems, average fatigue, and average pain) is the past 7 days.

Cost to use. All English and Spanish PROMIS measures are publicly available for individual, not-for-profit use without licensing or royalty fees. Commercial users must seek permission to use, reproduce, or distribute measures. Integration into proprietary technology requires written permission. For more information, please visit http://www.healthmeasures.net/explore-measurement-systems/promis/obtain-administer-measures and see http://www.healthmeasures.net/images/LearnMore/Pricing_Info/HealthMeasures__Overview_of_Free_and_Fee-Based_Services__v1.6__Final__101618.pdf.

How to obtain. A pdf version of the PROMIS-GH version 1.2 may be downloaded from HealthMeasures at http://www.healthmeasures.net/index.php?option=com_instruments&view=measure&id=778&Itemid=992.

For any queries regarding the PROMIS-GH version 1.2, prospective users can contact HealthMeasures (help@healthmeasures.net).

Practical application

Method of administration. The PROMIS-GH is available as e-version or paper and pencil version. All PROMIS measures are intended to be completed by the respondent. If respondents are unable to answer (eg, people with cognitive or communication deficits), a proxy can report on their behalf.

Scoring. PROMIS measures use standardized scoring based on IRT, by which individual items are linked to an underlying latent trait using information on the amount of a trait represented by an item. PROMIS scores are reported in the T score metric (population mean 50, SD ±10). The mean is based on a United States general population sample, drawn to be representative of age, sex, and race distribution from the 2000 United States Census (178,180). It is recommended that the PROMIS-GH be scored by using the HealthMeasures scoring service or one of the data collection tools (eg, assessment center or Research Electronic Data Capture [REDCap] autoscore). Alternatively, simple summed raw scores of the GPH and GMH scales (respective score range: 4-20) can be converted to T scores using lookup tables that are provided in the PROMIS global scoring manual (http://www.healthmeasures.net/index.php?option=com_instruments&view=measure&id=778&Itemid=992).

Score interpretation. PROMIS-GH T scores are calibrated relative to the mean of the US general population, with higher scores indicating more of the concept being measured. For example, a GPH T score of more than 50 corresponds with a better-than-average GPH, and a GPH T score of less than 50 corresponds with a worse-than-average GPH compared with the mean of the US general population.

Respondent time to complete. The original authors describe a 2-minute completion time for the PROMIS-GH (179). In rheumatology settings, comparable average completion times have been reported (181,182).

Administrative burden. The administrative burden varies. As with other PRO measures, the burden depends on the method of administration (electronic or paper/pencil) and mode of administration (self versus interviewer). HealthMeasures offers various data collection tools (http://www.healthmeasures.net/resource-center/data-collection-tools).

Translations/adaptations. Various translations of the PROMIS-GH exist. Terwee et al describe results of the translation and cultural adaptation of 17 PROMIS item banks into Dutch-Flemish, including the PROMIS-GH (183), whereas Zumpano et al give a detailed account of the translation and cultural adaptation of PROMIS-GH into Portuguese, including psychometric testing (184). Because the number of translations is constantly growing, please see http://www.healthmeasures.net/explore-measurement-systems/promis/intro-to-promis/available-translations/117-available-translations for the most current list of available translations. For permission to use the translations or permission to translate, please contact HealthMeasures (translations@healthmeasures.net).

Psychometric information

Floor and ceiling effects. There is little evidence that the PROMIS-GH exhibits floor or ceiling effects in rheumatology settings. For example, in a study of 161 individuals with shoulder arthritis, no floor or ceiling effects were recorded for either GPH or GMH (185). In a study of 119 patients with ankylosing spondylitis, Hwang et al reported that the PROMIS-GH version 1.1 GPH scale had no floor effects. Of note, this study focuses on GPH only (186). Finally, Kahan et al did not find floor or ceiling effects among 62 patients with lateral epicondylitis for the GPH scale but identified a minor ceiling effect for the GMH scale in one subgroup (181).

Reliability. *Internal consistency.* The internal consistency of the PROMIS-GH in rheumatic conditions has only been assessed in one study. A Cronbach's α of 0.91 has been reported for the PROMIS-GH version 1.1 GPH scale in a study with 119 patients with ankylosing spondylitis (186). More broadly, the original authors of the PROMIS-GH report internal consistencies of 0.81 and 0.86 for the GPH and GMH scales, respectively (179).

Test-retest. Test-retest reliability of PROMIS-GH in rheumatic diseases was assessed in at least two studies. In 90 stable patients with systemic lupus erythematosus, the PROMIS-GH version 1.1 had ICCs of 0.89 for the GPH scale and of 0.85 for the GMH scale over a 7-day test-retest interval (182). The PROMIS-GH version 1.1 GPH had an ICC of 0.90 (95% CI 0.76-0.96); however, it needs to be noted that the median retest period among the 24 individuals with ankylosing spondylitis was 1 day (interquartile range, 1-2) only (186). Therefore, it is plausible that respondents were able to recall their answers from the previous day, thus inflating the estimated test-retest reliability of the PROMIS-GH.

Validity. *Face/content.* Although content validity of the PROMIS-GH has not been specifically evaluated in rheumatology settings, the PROMIS-GH appears to have good face and content validity as a generic measure of health. The 10 items cover the 5 primary PROMIS domains (ie, physical function, fatigue, pain, emotional distress, and social health) that are of relevance to a number of chronic health conditions, including rheumatic conditions.

Construct. We were unable to identify studies that evaluated the factorial structure of the PROMIS-GH in rheumatic conditions. However, the results of the original publication did not fully support a two-factor structure of PROMIS-GH in either the exploratory or confirmatory factor analyses but also clearly rejected a one-factor solution (179). The factorial structure of the PROMIS-GH in rheumatic conditions remains to be evaluated in future studies.

Several studies of rheumatic diseases assessed the convergent and discriminant validity of the PROMIS-GH with favorable results. For example, a study of 204 individuals with systemic lupus erythematosus reported large correlations between the PROMIS-GH version 1.1 GPH scale and the physical function and pain domains of legacy instruments such as the SF-36 physical function, the SF-36 PCS, the SF-36 bodily pain, the Lupus Quality of Life (LupusQoL) pain, and various PROMIS computer-adaptive test (CAT) versions of PROMIS physical domains (absolute values $r = 0.71$-0.80). Correlations between the GMH scale and emotional health domains of legacy instruments were also generally good, particularly with SF-36 mental health, SF-36 MCS, LupusQoL emotional health, and PROMIS depression CAT (absolute values $r = 0.70$-0.73). Discriminant validity was supported by weaker correlations (absolute values r of less than 0.60) between the GPH and GMH scales and divergent legacy instrument domains (eg, substantially lower correlations between the GPH scale and legacy instruments measuring mental health and between the GMH scale and legacy instruments measuring physical health, respectively) (182).

Although the above study included measures of both physical and mental health (182), other validation studies of the PROMIS-GH only included measures of disease-specific physical health (181,185,187). These studies again largely support the construct validity of the PROMIS-GH; however, the results are less consistent. For example, two studies (n = 112 patients with upper-extremity disorders; n = 62 patients with lateral epicondylitis) (187,181). showed moderate to strong correlations between the GPH scale and established rheumatology scales and either no correlation or substantially lower correlations between these scales and the GMH scale. In contrast, a study of 161 individuals with shoulder arthritis by Saad et al only showed a moderate correlation of the GPH scale with the American Shoulder and Elbow Surgeons (ASES) assessment form ($r = 0.57$; $P < 0.0001$). However, this study also reported low or nonsignificant correlations of GPH scale with two other osteoarthritis-specific measures, namely the Single Assessment Numeric Evaluation (SANE) instrument ($r = 0.23$; $P = 0.0045$) and the Western Ontario Osteoarthritis of the Shoulder (WOOS) index ($r = 0.11$; $P = 0.3743$). Correlation coefficients between the GMH scale and the ASES assessment, SANE, and WOOS index were low ($r = 0.09$-0.26) (185).

Known-groups validity of the PROMIS-GH in rheumatic diseases has been assessed in several studies with favorable results. For example, in a study of 204 patients with systemic lupus erythematosus, patients reported a mean T score of more than 0.5 SD below the general population mean of 50 (182). Furthermore, Gouttebarge et al (188) compared mean PROMIS-GH version 1.2 scores of 361 current and 396 retired professional football players with or without lower extremity osteoarthritis. Mean GPH scores of retired players without osteoarthritis and current players were almost identical (mean = 52.7 versus mean = 52.6) and five points (0.5 SD) higher than the scores of retired players with osteoarthritis (mean = 47.6). Differences were less pronounced in the mean GMH scores, which were slightly above the general population mean of 50 for retired players without osteoarthritis and current players (mean = 52.2 versus mean = 51.7), whereas retired players with osteoarthritis were almost exactly at the general population mean of 50 (mean = 49.9) (188). Of note, although there is no empirically derived MID threshold for the PROMIS-GH, Jensen et al consider a three-point difference in T scores in a study on patients with cancer as clinically meaningful across various PROMIS measures (189).

In a study of 119 patients with ankylosing spondylitis, Hwang et al also showed strong support for the known-groups validity of the GPH scale, with clear differences between patients with different disease activity levels (GPH T scores: mean = 53.5 for inactive disease, mean = 47.3 for moderate disease activity, and mean = 38.4 for high disease activity) (186).

Finally, the PROMIS-GH version 1.1, as measured in 156 patients with rheumatoid arthritis, was able to detect significant differences between groups based on disease activity (GPH T scores: mean = 45.3 for low disease activity, mean = 42.6 for moderate disease activity, and mean = 38.8 for high disease activity; GMH T scores: mean = 51.2 for low disease activity, mean = 49.3 for moderate disease activity, and mean = 46.1 for high disease activity) (190).

Responsiveness. Four studies explored the responsiveness of the PROMIS-GH in rheumatology settings. Patients undergoing knee arthroscopy ($N = 50$; $n = 45$ for responsiveness analysis) showed moderate improvement in GPH scores at 3 to 6 months of follow-up (ES = 0.51, SRM = 0.72), whereas change in GMH scores was trivial during the same period (ES = 0.06, SRM = 0.08). Patients undergoing total knee replacement surgery (*n* = 721) showed large changes in GPH scores (ES = 1.30, SRM = 1.20) at 6-month follow-up and trivial changes in GMH scores (ES = 0.04, SRM = 0.04) (191). Responsiveness was further investigated in 186 patients with systemic lupus erythematosus, using anchors based on self- and clinician-reported global assessments of change. Small to moderate decreases in scores were observed in patients reporting worsening (GPH: ES = –0.27, SRM = –0.37; GMH: ES = –0.54, SRM = –0.68), whereas small increases in scores (GPH: ES = 0.29, SRM = 0.41; GMH: ES = 0.29, SRM = 0.36) were observed in patients reporting improvement (192). Finally, the PROMIS-GH version 1.1, as measured in a population of 156 patients with rheumatoid arthritis, (*n* = 106 at 12-week follow-up visit) showed significant associations between GPH change scores and changes in several disease activity measures, including the Clinical Disease Activity Index (CDAI), tender joints, patient global assessment, and assessor global assessment (standardized *β* coefficients range from –0.21 to –0.30). In contrast, changes in swollen joint count were not associated with changes in GPH scores. For the GMH scale, significant associations were found with the CDAI, tender joints, and clinician global assessment (standardized *β* coefficients range from –0.22 to –0.34) but not with swollen joints or patient global assessment (190).

Minimally important differences. We were unable to locate empirically derived MID thresholds for the PROMIS-GH. The only study referring to an MID for the PROMIS-GH in rheumatic diseases was a conference abstract in which Husni et al mention a threshold of five points, but it remains unclear how this MID was derived (193). As noted earlier, a three-point difference in T scores has been considered clinically meaningful across various PROMIS measures in oncology (189). See also http://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/meaningful-change.

Generalizability. All PROMIS measures were developed with the aim to measure common generic symptoms and experiences that are assumed to be relevant to people in a variety of contexts and disease areas. Hence, the PROMIS-GH should be applicable in a wide range of rheumatic conditions if the aim is to assess global health. However, there are limited comprehensive psychometric evaluations of the PROMIS-GH in rheumatic conditions, and its applicability in this setting requires further studies. Although the PROMIS-GH has not been used frequently in the past, we observed a trend toward increased popularity of this measure in rheumatic conditions. That is, the articles included in this review were all published recently (generally within the last 2 years), and there are several ongoing trials with the PROMIS-GH included either as a primary or secondary endpoint (see Use in clinical trials section below).

Use in clinical trials. We were not able to locate any published clinical trials in rheumatic diseases using the PROMIS-GH. However, several relevant ongoing trials have been registered on www.clinicaltrials.gov that are currently using the PROMIS-GH as either a primary or secondary outcome measure. Studies that use PROMIS-GH as a primary outcome measure include a trial of immediate accelerated shoulder rehabilitation versus a standard rehabilitation protocol following reverse total shoulder arthroplasty (NCT03804853) and a trial comparing outcomes for patients undergoing total hip replacement surgery with robotic-arm assistance with those undergoing traditional total hip replacement surgery (NCT03891199). The PROMIS-GH is further listed as a secondary outcome measure in trials assessing pharmaceutical (NCT02958267), surgical (NCT02947321), physiotherapy (NCT02763488), and self-management (NCT03245463) interventions for knee osteoarthritis.

Critical appraisal of overall value to the rheumatology community

Strengths. The PROMIS-GH is a short measure of overall health status (179) that patients should be able to complete in about 2 minutes, enabling a fast and easy way to assess a person's global health. T scoring of the PROMIS-GH enables score interpretation relative to the United States general population. Finally, even though evidence from psychometric studies on the PROMIS-GH in rheumatic disease settings is still limited, our review suggests that there is good support for its face and construct validity in rheumatic conditions.

Caveats and cautions. The application of IRT-based scoring offers increased measurement precision compared with simple sum scores. However, this approach to scoring might be less intuitive to users who are not familiar with modern test theory methods. For example, many PRO instruments are traditionally transformed into a score ranging from 0 to 100. Therefore, having a T score centered around a mean of 50 (SD +10) could pose initial difficulties with score interpretation for users not familiar with this scoring approach. The use of IRT-based scoring also increases the administrative burden of the PROMIS-GH because users either need to engage the HealthMeasures scoring service, use other electronic applications, or use the lookup tables provided in the PROMIS-GH scoring manual. Finally, despite supportive evidence presented herein, the fact remains that there are limited comprehensive psychometric evaluations of PROMIS-GH in rheumatic conditions, and its applicability requires further studies, particularly on internal consistency and factorial structure, in this setting.

Clinical usability. The PROMIS-GH offers an efficient way to assess a person's global health status, and evidence is accumulating to support its use in rheumatic conditions. Despite potential challenges regarding scoring, HealthMeasures is providing a wide range of resources to facilitate the scoring of the measure.

The interpretation of patients' scores relative to general population norm data further adds intrinsic meaning to the obtained scores, making this measure, like all PROMIS measures, easy to interpret in terms of number of SDs above or below the general population mean of 50, that is, once users are familiar with this scoring approach (see Caveats and cautions section).

Research usability. In research settings, a reasonably short measure of global health is often needed to reduce respondent burden; hence, the PROMIS-GH might be preferable to longer instruments such as the SF-36. Further, all English and Spanish PROMIS measures are publicly available, which is an important aspect in research settings in which funding is often scarce. HealthMeasures offers a wide range of resources to facilitate the scoring of the measure, and the number of available translations is constantly increasing. Although the applicability of the PROMIS-GH in rheumatic conditions requires further studies, it appears to be a promising short measure of general health for use in rheumatology research settings.

EQ-5D AND EQ VAS

Description

Purpose. The EQ-5D, a standardized measure of HRQOL developed by the European Quality of Life (EuroQol) Research Foundation, is used for a wide range of conditions and populations, including rheumatic conditions (194). The EQ-5D family of instruments consists of two parts: the descriptive system, which comes in three versions, including EQ-5D-3L (3L), EQ-5D-5L (5L), and EQ-5D-Y (Youth); and a VAS (EQ VAS). The EQ 5D and EQ-VAS can be used in clinical trials, population health surveys, routine outcome measurement, and many other types of studies in which a generic measure of health status is useful (195). The descriptive system can be presented as a health profile or converted to a single summary index value that reflects health state preferences by using data from the general population. This index value can then be used to calculate a cost utility ratio in an economic analysis. Many different terms are used in the literature for the index value, including utility score, preference weight, and preference-base value. Here, we use the term utility score.

Content or domains. The 3L, introduced in 1990 (196), consists of five domains of health, including mobility, self-care, usual activities, pain/discomfort, and anxiety/depression, with a three-level descriptive system (see Response options/scale). Mobility refers to walking about, self-care refers to washing or dressing themselves, and usual activities refers to work, study, housework, family activities, or leisure activities. The fourth and fifth domains comprise pain or discomfort and anxious or depressed, respectively. The 5L was developed to improve the sensitivity of the EQ-5D by increasing the number of descriptive levels from three to five (see Response options/scale section below) (197).

Number of items. The EQ-5D consists of five items representing five health domains. The EQ VAS is a single item that records the respondent's self-rated health on a vertical visual analog scale.

Response options/scale. The 3L descriptive system consists of three response levels: no problems, some problems, and extreme problems. The 5L descriptive system consists of five response levels: no problems, slight problems, moderate problems, severe problems, and unable to/extreme problems. The response options on the EQ VAS range from 0 (worst health imaginable) to 100 (best health imaginable).

Recall period for items. All EQ-5D and EQ VAS items request the user to describe their health state on the day of assessment (today). This lessens the cognitive demand by eliminating recall bias.

How to obtain. Sample United Kingdom English versions of the 3L and 5L are available for download from the EuroQol office support section of the website (https://euroqol.org/support/how-to-obtain-eq-5d/). Sample versions in other languages are available for inspection on request. Written permission from the EuroQol Research Foundation is required prior to using the EQ-5D. A request to use the EQ-5D and the EQ VAS can be made by completion of a registration form available from the EuroQol website (https://euroqol.org/eq-5d-registration-form/).

Cost to use. A user license policy (https://euroqol.org/wp-content/uploads/2020/03/EQ-5D-User-License-Policy-18MAR2020.pdf) outlines the process and costs involved when a license fee and/or agreement is required to use the EQ-5D and EQ VAS. In general, no license fee is charged for noncommercial use except when data are collected with the intention to charge a fee for access. However, a license agreement is required for substantial research projects involving 100 000 or more patients, for projects of 5 years or longer, or for requests to use the digital versions on unsupported digital platforms, for which no digital modules are available. A license agreement will be set up and a license fee will be proposed for requests made by or on behalf of a pharmaceutical company, medical device manufacturer, or any other for-profit stakeholder.

Practical application

Method of administration. The EQ-5D and EQ VAS can be administered using self-complete modes via paper or digital software (eg, REDCap) for laptops, desktops, tablets, personal digital assistants, and smartphones. The questionnaires can also be administered via interview (eg, by telephone or face-to-face), via proxy (eg, a caregiver), and via IVR technology.

Scoring. EQ VAS scoring is based on a visual analog scale ranging from 0 (worst health imaginable) to 100 (best health imaginable).

EQ-5D response levels for the five domains provide a five-digit code that can be summarized descriptively or converted to a utility score. For example, a descriptive summary for the five-digit code '32354' using the 5L indicates that the person perceives they have moderate problems in walking about, slight problems washing or dressing him/herself, moderate problems doing their usual activities, and extreme pain or discomfort and are severely anxious or depressed. The EQ-5D utility scores are derived from country-/region-specific value sets containing all possible health states. The 5L version has 3125 (equivalent to 5^5) possible health states, and the 3L has 243 (equivalent to 3^5) possible health states. For example, a utility score for the 5L for an England-based population can be derived using an England-based value set (198) with the five-digit code 32354, providing a utility score of 0.191. Using the 32354 example (198), the utility score is based on the scoring algorithm 1 – (0.076 + 0.050 + 0.063 + 0.335 + 0.285), with a weight derived for each response level and each domain, and deducted from the value of 1. For example, a weight of 0.285 is used for the response level 4 (severely anxious or depressed) on the anxiety/depression domain.

A current list of available value sets for the 3L (33 countries) and 5L (20 countries) can be found on the EuroQol Research Foundation website (https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/valuation/). Value sets are available from published validation studies or via the validation study authors. If value sets are not available for a specific country, then a value set that most closely resembles that specific country is recommended. Alternatively, if a value set is required for the 5L and is only available for the 3L, then a crosswalk value set can be used to derive a utility score for the 5L. Crosswalk value sets are based on a study that administered both the 3L and 5L versions of the EQ-5D in order to adapt the 3L value set to fit the five response options for the 5L (199). These value sets are available for the following countries from the EuroQol website (https://euroqol.org/eq-5d-instruments/eq-5d-5l-about/valuation-standard-value-sets/crosswalk-index-value-calculator/): Denmark, France, Germany, Japan, the Netherlands, Spain, Thailand, the United Kingdom, the United States, and Zimbabwe. Crosswalk value sets are an interim solution until more empirical value sets are available for the 5L.

Score interpretation. EQ-5D utility scores are defined as "the estimates of the preference for a given state of health, expressed as a score of 1 or less" (200). A higher utility score indicates a better HRQOL or health state, with a utility score of 1 (coded as 11111) indicating the value of a perfect health state. A negative utility score is possible, indicating a health state worse than death (200).

For EQ VAS scoring, a higher score indicates better HRQOL. Scoring interpretation for the EQ VAS is different from the EQ-5D utility score. The EQ VAS score (range: 0-100) represents an individual's perspective of their own health state and is appropriate for use in clinical practice and research. In contrast, EQ-5D utility scores are linked to a general population's perspective about the value of a health state and are preferred in health economic analyses (195). However, compared with the 3L, the 5L mean utility scores move up toward best imaginable health and condenses mean scores into a smaller range (201). This has implications for economic analyses as improvements in quality of life are valued less using the 5L than using the 3L with substantially different estimates of cost-effectiveness (201).

Respondent time to complete. Respondent time is a key advantage of the EQ-5D and the EQ VAS, taking only a few minutes to complete (194,202). The EQ-5D and EQ VAS questionnaires were perceived to be completed successfully in less than 5 minutes in elderly participants (more than 79 years old) in the United Kingdom (203). However, respondents found the EQ VAS to be a difficult item to complete (203).

Administrative burden. The EQ-5D questionnaire has low administrator burden, taking a few minutes to complete via interview. Determining the descriptive summary for the EQ-5D and the score for the EQ VAS is simple and quick. Although locating and selecting the appropriate value set or crosswalk to determine a utility score can be challenging and time consuming, once the appropriate value set has been located, determining the utility score is also simple and quick.

Translations/adaptations. The EQ-5D was originally developed simultaneously in five languages: Dutch, English, Finnish, Norwegian, and Swedish (https://euroqol.org/). To date, the EQ-5D has undergone over 200 adaptations and is available in 97 languages, representing 109 countries (https://euroqol.org/eq-5d-instruments/all-eq-5d-versions/).

Psychometric information

Floor and ceiling effects. Although the EQ-5D utility score and the EQ VAS have showed minimal floor and ceiling effects in adults with rheumatic conditions (204–207), ceiling effects have been observed at the domain level of the EQ-5D. For example, a ceiling effect (more than 15% of sample with lowest rating) was observed for all domains of the 5L except for pain/discomfort in a population of patients with spondyloarthritis (208). In addition, a ceiling effect was observed in the domains of self-care, usual activities, and anxiety/depression in a cohort of 758 Spanish people with hip or knee osteoarthritis (206). However, the 5L utility score showed negligible floor (0.27%, 55555) and ceiling effects (2.5%, 11111) in this population (206) and no ceiling effect (0%, 11111) in a cohort of 176 Canadian patients with osteoarthritis referred for hip or knee replacement (207). No floor effect was observed at the domain level of the 5L in a cohort of patients with spondyloarthritis (208) pre and post total hip replacement (204). The EQ VAS demonstrated negligible floor and ceiling effects (only 1.47% and 3.21% of participants, respectively) in a study of Spanish patients with hip or knee osteoarthritis (206).

The 5L has demonstrated decreased ceiling effects compared with the 3L in a cohort of preoperative and postoperative total hip replacement patients (204), rheumatologic rehabilitation patients from an inpatient setting (209), and patients with arthritis (205). For example, in a cohort of 371 people with arthritis from three countries (Denmark, England, and Scotland), the ceiling effect (percentage of people with the code 11111) was 6.5% for the 3L versus 1.9% for the 5L (205).

Reliability. *Internal consistency.* The EQ-5D and EQ VAS have shown favorable internal consistency (Cronbach's α of more than 0.8) (210) in patients with rheumatic conditions (78,206,208). For example, Cronbach's α was 0.84 for both the 3L and EQ VAS in a cohort of 488 people with rheumatoid arthritis who had pain and dysfunction of the hands and/or wrists (78). Similarly, Cronbach's α for the 5L was 0.86 at baseline, 0.89 at 6-month follow-up in a cohort of 758 Spanish people with hip and knee osteoarthritis (206), and 0.84 in a sample of 220 Chinese people with spondyloarthritis (208).

Test-retest. The EQ-5D and EQ VAS have shown moderate test-retest reliability (211) in rheumatic conditions (207,208,212–216). The ICCs for test-retest reliability of the 5L 2 weeks apart ranged from 0.61 (mobility) to 0.77 (anxiety/depression) for individual domains, and the ICC was 0.87 for the utility score and 0.73 for the EQ VAS in patients with osteoarthritis referred for hip and knee replacement surgery (207). For the 3L, measured in a cohort of 82 people with knee osteoarthritis 1 week apart, the ICC was 0.70 (95% CI 0.58-0.80) for the utility score and 0.73 (95% CI 0.61-0.82) for the EQ VAS (213). In a systematic review including three studies that directly compared the test-retest reliability of the 3L with the 5L, there was no clear pattern of better reliability for either the 3L or the 5L (217). However, none of these studies were conducted in rheumatic conditions.

Validity. *Face/content.* Rasch modeling of the EQ-5D (both the 3L and 5L) items showed acceptable goodness of fit, indicating unidimensionality for measuring HRQOL in patients with back and neck pain (218). In a qualitative assessment of the content validity of the 5L, clinical and research professionals in the United Kingdom and Australia viewed the 5L as offering good coverage of health determinants of quality of life (219). Content validity has also been demonstrated during the development of the 5L versions of UK English and Spanish in a population of healthy participants and those with chronic illness (197).

Construct. Adequate construct validity of the EQ-5D utility score and the EQ VAS have been demonstrated in a number of rheumatology settings, including patients with hip or knee osteoarthritis, rheumatoid arthritis, systemic lupus erythematosus, and chronic pain (205–207,212,215,220–222). For example, in people with rheumatoid arthritis, there was adequate construct validity between the 3L utility score and the EQ VAS when compared with pain level on a VAS (Spearman's ρ = 0.73 and 0.63 for 3L and EQ VAS, respectively) (215).

Although the EQ-5D has shown favorable convergent and discriminant validity in people with rheumatic conditions (205–207,223), convergent validity coefficients for the 5L have been shown to be slightly higher compared with the 3L in people with rheumatic conditions (205,207). For example, in people with osteoarthritis, convergent validity coefficients between the EQ-5D mobility domain and SF-12 physical component domain were 0.65 for the 5L compared with 0.38 for the 3L (207). This was likely due to changing the 3L mobility descriptive response of "confined to bed" to "unable to walk about" in the 5L.

The EQ-5D and EQ VAS have shown favorable known-groups validity (205,206). For example, in patients with rheumatic and other conditions, both the 5L and 3L have shown lower scores on the EQ-5D domains associated with older age and lower education levels, with the exception of anxiety/depression (205).

Responsiveness. The EQ-5D instruments have demonstrated different levels of responsiveness in rheumatic conditions depending on the treatment effects (eg, higher responsiveness for treatments expected to have large treatment effects), the measure used (the 3L, 5L, or EQ VAS), and type or severity of the rheumatic condition (224). For example, the 3L utility score showed moderate responsiveness (SRM = 0.50) for patients with rheumatoid arthritis treated with tumor necrosis factor blockers and low responsiveness (SRM = 0.20) for patients with knee osteoarthritis treated nonsurgically in a hospital rheumatology clinic (225). In comparison, responsiveness was better in surgical populations, with very high responsiveness of the 5L utility score demonstrated for patients with knee osteoarthritis 1 year following knee replacement (SRM = 1.04) (212) and for patients 6 months following a hip or knee replacement (SRM = 1.48) (206). These results indicate that the EQ-5D is better able to detect change in surgical compared with nonsurgical rheumatic populations.

The 5L utility score has shown better responsiveness to change (ES = 0.39, SRM = 0.42) for nonsurgical patients with hip or knee osteoarthritis at 6-month follow-up with worsened health compared with the EQ VAS (ES = 0.27, SRM = 0.24) (206). For patients with knee osteoarthritis 1 year following knee replacement, the 5L also showed better responsiveness to change (ES = 1.19, SEM = 1.0) compared with the EQ VAS (ES = 0.63, SRM = 0.55) (212). Therefore, the EQ-5D utility score may be more useful for detecting change compared with the EQ VAS.

In one study comparing the responsiveness of the 5L and 3L in a rheumatic inpatient rehabilitation population, the 5L had greater ability to detect both improved and worsened health changes over time in all EQ-5D domains (209). Therefore, the 5L may be more useful for detecting change compared with the 3L.

As expected for generic quality of life instruments, the EQ-5D is less responsive than disease-specific instruments used in rheumatic populations. For example, in improved patients who underwent hip or knee replacement surgery in Spain, the responsiveness parameters for the WOMAC pain (standardized effect size [SES] = 2.27, SRM = 1.75) and function (SES = 2.40, SRM = 1.79) scores were higher than those found for the 5L (SES = 1.48, SRM = 1.48) and the EQ VAS (SES = 0.82, SRM = 0.90) (206).

Minimally important differences. Variability in the estimates of the MCID for EQ-5D utility scores and the EQ VAS have been observed in rheumatology populations. This variability is likely to be due to differences in study populations and the different methods utilized for calculation (212,226). For example, in a review of the estimation of the MCID for the 3L utility score based on the UK scoring algorithm, estimates ranged from 0.03 for patients with chronic low back pain in the Netherlands to 0.52 for patients undergoing fusion for same-level recurrent lumbar stenosis (226). Twelve of the 18 studies included in this review focused on musculoskeletal conditions. Similarly, estimates of the MCID for the 3L utility score ranged between 0.05 and 0.33 in patients with chronic widespread pain using data from a multicenter randomized controlled trial in the UK (227) and between 0.06 and 0.20 (improvement) for patients with rheumatoid arthritis (228).

The MCIDs for changes in the 5L utility score have been calculated as 0.07 points (improvement) and –0.05 points (worsening) in 514 Spanish patients with hip and knee osteoarthritis receiving nonsurgical management (206). In comparison, MCID for the 5L utility score among surgical patients appears to be substantially higher, with values of 0.32 points calculated for 120 Spanish improved patients 6 months following hip and knee replacement surgery (206) and 0.41 and 0.28 points for patients 12 months following total hip and knee replacement surgery, respectively, with improved health (212).

The MDC of the 5L utility score has been calculated and compared with the MCID in the same population of Spanish surgical and nonsurgical patients with hip and knee osteoarthritis referred to above (206). In this study, the 95% confidence level of the minimal detectable change (MDC95%) was 0.30 and 0.01 points at the individual and group level (*n* = 644), respectively. Among patients who underwent joint replacement surgery, the ratio of the MCID to the $\text{MDC}_{95\%}$ was greater than 1 at both the individual and group level, indicating that the MCID can be discriminated from measurement error with 95% confidence. However, among nonsurgical patients, the ratio of the MCID to the $\text{MDC}_{95\%}$ was less than 1 (as a result of smaller MCID values) for both improved and worsened patients at the individual level, indicating that the MCID cannot be discriminated from measurement error at the individual (but not group) level with 95% confidence.

The MCID for the EQ VAS has been calculated as 9.34 points and 7.75 points for patients 12 months after total hip replacement and total knee replacement surgery, respectively, using self-rated improved health as the anchor (212). In comparison, the MCID for the Dutch translation of the EQ VAS has been calculated as 10.5 points, using the optimal cutoff point under the receiver operating characteristics curve in a sample of 151 patients with nonspecific chronic low back pain in the Netherlands (229).

Compared with the 3L, the 5L has shown improved discriminatory power to detect small to moderate differences in HRQOL in patients with hip and knee osteoarthritis, following total hip replacement surgery, and in a mixed-patient sample that included participants with rheumatoid arthritis, osteoarthritis, orthopedic accidents, and back pain (204,205,209). In one study involving patients with rheumatic conditions undergoing rehabilitation in Germany, the proportion of patients reporting no changes was smaller in the 5L than in the 3L instrument for all five domains, indicating that the 5L was better at detecting both positive and negative changes in HRQOL (209).

Generalizability. The EQ-5D instruments are useful in a broad range of health conditions (including, but not limited to, rheumatology), settings, and countries (195,217,224). EQ-5D utility scores can be used to compare treatment effectiveness and cost-effectiveness across a variety of different conditions and settings. The EQ-5D has also been used to specifically assess HRQOL among people with noncommunicable chronic diseases compared with reference values from the general population (195). In this systematic review that included 48 studies that examined this tool in people with rheumatic conditions, the mean EQ-5D utility scores ranged from 0.26 units for juvenile idiopathic arthritis to 0.94 units for arthritis (in general) (195).

Use in clinical trials. The EQ-5D instruments have been used as an outcome measure in many randomized controlled trials to evaluate the efficacy of a broad range of interventions for rheumatic conditions, including magnetic resonance imaging–guided treatment for rheumatoid arthritis (230); self-management for gout (231); spa-exercise therapy for ankylosing spondylitis (232); surgery for osteoarthritis of the ankle (233); walking for osteoarthritis of the knee (234); neuromuscular exercise for osteoarthritis of the hip (235); knitting for osteoarthritis of the hands (236); and pharmacological treatment for psoriatic arthritis (237), fibromyalgia (238), primary Sjögren’s syndrome (239), and systemic lupus erythematosus (240).

Critical appraisal of overall value to the rheumatology community

Strengths. The EQ-5D instruments are quick and easy to use and can be used in a broad range of rheumatic conditions for the measurement of HRQOL. The 5L version of the descriptive system may be the preferred choice for use in clinical, research, and economic evaluations because of its improved measurement properties compared with the 3L. These include lower ceiling effects and better construct validity, responsiveness, and discriminatory power to detect small to moderate differences in HRQOL.

Caveats and cautions. Prior to using the EQ-5D and EQ VAS, written permission from the EuroQol Research Foundation is required. Locating and selecting the appropriate value set or crosswalk to determine a utility score can be challenging and time consuming. More value sets for the 5L are also needed to produce reliable calculations of a utility score that are applicable to a wide range of settings, populations, and countries. The EQ-5D and EQ VAS are both less responsive than disease-specific measures. It is therefore recommended that they be used in conjunction with disease-specific measures when it is important to determine the effectiveness of interventions. Estimates of the MCID have varied depending upon the rheumatic condition, indicating that caution may be warranted in interpreting treatment effects. Depending on the value set and version used (3L or 5L), substantially different estimates of cost-effectiveness can be produced (201). It may therefore be prudent to ensure consistency in value sets and versions when making comparisons between different versions of the EQ-5D.

Clinical usability. The descriptive summary generated by EQ-5D and EQ VAS scores may be appropriate in a clinical setting and aid clinical decision making together with appropriate disease-specific measures. The 5L may be more relevant to apply in most rheumatic conditions because the additional response levels allow for better discrimination to detect difference in levels of health.

Research usability. The utility score is highly relevant for use in economic evaluations. Both 5L and 3L versions can be used, although, provided that a value set or crosswalk is available, the 5L may be preferred based on its advantages in psychometric properties.

SHORT FORM 6-DIMENSION HEALTH INDEX

Description

Purpose. The SF-6D is a health utility measure derived from items within the widely used SF-36 and SF-12 instruments and is intended for use in health economic evaluations and to derive QALYs. Of note, it is not recommended that the SF-6D be administered as a standalone instrument (https://www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d/faqs). Rather, SF-6D scores should be derived following the administration of either the SF-36 or the SF-12.

Content or domains. The SF-6D covers the following six domains: physical function, role participation (role physical and role emotional), social function, bodily pain, mental health, and vitality.

Number of items. The SF-6D utility score is derived from 11 items of the SF-36 or 7 items of the SF-12 (241).

Response options/scale. Items are scored on a Likert-type scale, with the number of response options varying between items. The physical functioning items are scored on a three-point scale (yes, limited a lot; yes, limited a little; and no, not limited at all), and the role participation, social functioning, vitality, and mental health items are scored on a five-point response scale (all of the time, most of the time, some of the time, a little of the time, and none of the time). Pain is scored on a five-point response scale (not at all, a little bit, moderately, quite a bit, and extremely). If the SF-6D score is based on the SF-36, the pain domain includes an additional item scored on a six-point scale (none, very mild, mild, moderate, severe, and very severe).

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

Cost to use. The SF-6D scores are calculated from the SF-36 or SF-12 questionnaires; therefore, it is necessary to obtain these questionnaires first. For details, see the Cost to use sections for the SF-36 and SF-12 in this article.

Programs that convert the SF-36 or SF-12 data into SF-6D scores are available free of charge to noncommercial researchers, whereas commercial researchers attract a license fee per study. Further information about the costs of the SF-6D scoring algorithms can be obtained from https://licensing.sheffield.ac.uk/i/health-outcomes/SF-6D.html.

How to obtain. The SF-6D scores are calculated from SF-36 or SF-12 questionnaires. For details, see the How to obtain sections for the SF-36 and SF-12 in this article.

Practical application

Method of administration. Please see the Method of administration sections for the SF-36 and SF-12 in this article.

Scoring. The SF-6D utility score is calculated as a function of weighted scores across the selected items of the SF-36 or SF-12. The algorithm to calculate the SF-6D scores from the SF-36 and SF-12 questionnaires can be obtained through three types of licenses, as described on the University of Sheffield website (www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d) (https://licensing.sheffield.ac.uk/i/health-outcomes/SF-6D.html):

- A license is available free of charge for all noncommercial applications including work funded by research councils, government agencies and charities.
- Commercial applications (eg, clinical trial) will require a license for each study, though an open license for a fixed period is available.
- The SF-6D can be calculated using purpose-developed software available from Optum Insight (https://www.optum.com).

Score interpretation. The SF-6D represents 18 000 possible health state combinations. The best health state (code 111111) characterizes an individual whose health does not limit them in vigorous work, social, and other daily activities; who has no pain; who does not feel tense or downhearted; and who has a lot of energy all of the time. The health states represented by the SF-6D are subsequently converted into an overall utility score ranging from 0.30 (poorest HRQOL) to 1.0 (perfect HRQOL).

Population norms for the SF-6D utility scores are available for a number of countries, including the United States (242), the United Kingdom (243), Brazil (244), Portugal (245), Japan (246), and Australia (247).

Respondent time to complete. See the Respondent time to complete sections for the SF-36 and SF-12 in this article.

Administrative burden. For administration instruction, please refer to the Administrative burden sections for the SF-12 and SF-36 instruments in this article. Some additional effort is required to derive the SF-6D score, as outlined in the scoring section.

Translations/adaptations. The parent questionnaire (the SF-36) is available in over 170 languages (https://www.optum.com/solutions/life-sciences/answer-research/patient-insights/sf-health-surveys/sf-36v2-health-survey.html). However, the preference weights used to calculate the SF-6D utility score may vary between countries and are currently only available for the United Kingdom, Australia, Brazil, Hong Kong, Japan, Portugal, and Singapore. Information on how to obtain these preference weights is available online (https://www.sheffield.ac.uk/scharr/sections/heds/mvh/sf-6d/faqs). A recent systematic review concluded that value sets for the SF-6D differ between countries and identified a need to develop value sets for more countries to take cultural differences into account (248).

Psychometric information

Although administration of the SF-6D as a standalone instrument is discouraged, a number of studies evaluated the psychometric properties of the items that comprise the SF-6D, following the administration of one of the parent questionnaires (the SF-12 or SF-36).

Floor and ceiling effects. Floor and ceiling effects for the SF-6D utility score are minimal (less than 5%) or nonexistent in studies involving people with systemic lupus erythamatosus (249), low back pain (250), rheumatoid arthritis (251–253) and psoriatic arthritis (254). A further study involving people with rheumatoid arthritis also found no ceiling effect for the SF-6D, although floor effects were not specifically reported (255).

Reliability. *Internal consistency.* The items that make up the SF-6D have good evidence of internal consistency in both the general population and rheumatic conditions. In a sample of over 5000 individuals drawn from the general population in Brazil (256), the SF-6D was reported to have good internal consistency, with a Cronbach's α of 0.86. In a sample of 488 individuals with rheumatoid arthritis involving the hands, Cronbach's α was 0.83 and was comparable with that of the EQ-5D and the SF-12 PCS (both Cronbach's α = 0.84) but slightly lower than that of the SF-12 MCS (Cronbach's α = 0.87) (78). However, in a study of 84 patients undergoing arthroscopic partial meniscectomy, the internal consistency of the SF-6D (Cronbach's α = 0.74) was slightly higher than that of the EQ-5D (Cronbach's α = 0.70) (257).

Test-retest. The test-retest reliability of the SF-6D has been assessed in a range of rheumatic conditions, with generally favorable results. In a small study (n = 61) of people with proximal humeral fractures (258), the SF-6D had good test-retest reliability over a 4-week period (ICC = 0.79) that was comparable with that of the EQ-5D (ICC = 0.78) and better than that of the HUI3 (ICC = 0.47). Similarly, Khanna et al also found that the SF-6D had good 4-week test-retest reliability in a sample of 168 patients with systemic sclerosis (ICC = 0.82) (259). However, Boonen et al found that in patients with ankylosing spondylitis, the 4-week test-retest reliability of the SF-6D was only modest (ICC = 0.68) but was higher than that of the EQ-5D (ICC = 0.55) (260).

Validity. *Face/content.* Although like most utility scales, the SF-6D was not derived through consultations with patients and clinicians (261), the dimensions of health captured by this measure are broadly applicable to rheumatic conditions. In a study of 172 patients with low back conditions, the SF-6D had good targeting ability and good coverage of health states at both the highest and lowest ends of the scale (262).

Construct. The unidimensionality of the SF-6D has been assessed through the application of IRT modeling, with the results providing support for the hypothesized one-dimensional structure of this measure (262). Convergent and discriminant validity of the SF-6D in rheumatology settings has been evaluated in several studies. These have generally produced consistent findings of moderate correlations between the SF-6D and other scales measuring HRQOL (249,263–266) and lower, but still substantive, correlations with disease-specific questionnaires (241,260,265–267), supporting the convergent validity of this measure. Furthermore, absolute agreement between the SF-6D and the EQ-5D tends to be only moderate (268,269), indicating that these two instruments measure different concepts and thus support the discriminant validity of the SF-6D.

The known-groups validity of the SF-6D appears to be supported. Marra et al undertook a study in 313 people with rheumatoid arthritis to compare several disease-specific measures (the Rheumatoid Arthritis Quality of Life Questionnaire and the HAQ) with several preference-based measures, including the SF-6D (266). They found that utility scales, including the SF-6D, appeared to discriminate well across rheumatoid arthritis severity categories, although the disease-specific measures were generally more sensitive in this setting. In 167 patients with systemic lupus erythematosus, Aggarwal et al (249) found that both the EQ-5D and SF-6D differentiated among patient groups of varied disease severity. However, two recent studies, one involving 272 Chinese patients with low back pain (267) and another involving 356 Thai hospital outpatients with a variety of chronic conditions (including musculoskeletal conditions) (269), found the SF-6D to be less efficient than the EQ-5D at discriminating between groups based on disease severity.

Responsiveness. Earlier evidence for the ability of the SF-6D to detect change in HRQOL (either improvement or deterioration) is mixed and varies according to the condition studied and the severity of disease (249,270–273).

In later studies, the SF-6D was only moderately responsive to improvement following treatment with tumor necrosis factor blockers in people with rheumatoid arthritis (SRM = 0.67), although it was more sensitive to change than the EQ-5D in this setting (SRM = 0.50) (251). Similarly, in a study of 813 people with early rheumatoid arthritis, the SF-6D was found to be more responsive to improvement in disease activity (SRM = 0.83) than the EQ-5D (SRM = 0.57), but the SF-6D was less responsive to deterioration in disease activity (SRM = −0.11) than the EQ-5D (SRM = −0.20) (274). Other studies have found the SF-6D to be less responsive to improvement (mean change from baseline to endpoint of 0.05 utility units) than the EQ-5D (mean change of 0.15 utility units) when used in the assessment of people with knee osteoarthritis or low back pain (221). In particular, it was less sensitive to change among people who had more severe knee symptoms at baseline (SRM = 0.87-1.33 for the SF-6D versus SRM = 1.38-2.02 for the EQ-5D) (221). A study involving people with rheumatoid arthritis with hand symptoms found that the SF-6D was markedly less sensitive to both improvement and deterioration (SRM = 0.15 for improvement; SRM = −0.05 for deterioration) than a hand-specific measure (Michigan Hand Outcome Questionnaire; SRM = 0.56 for improvement; SRM = −0.08 for deterioration) and the generic EQ-5D (SRM for improvement = 0.31; SRM for deterioration = −0.16) (78).

A study involving 1104 people with low back conditions undergoing operative treatment found that the SF-6D was only slightly less responsive to postoperative change at 2 years (SRM = 0.70) than the back-specific Oswestry Disability Index (SRM = 0.73) but was more responsive than the physical composite summary of the SF-36 instrument (SRM = 0.57) (275). Given the measure's variable ability to detect change over time (and in comparison with other measures), this could impact the outcomes of comparative effectiveness research or health economic analyses in which data collected using different outcome measures are used. Furthermore, because the SF-6D may not be able to detect deterioration in patients with severe progressive disease (because the scale has a relatively high lower bound of 0.30 utility units), findings of past studies highlight the need for careful attention to disease severity in studies that aim to measure change in health status over time.

Minimally important differences. Evidence for a relatively small MCID for the SF-6D comes from several studies with relevance to rheumatology settings. Applying anchor-based methods, Khanna et al proposed an MCID of 0.035 utility units in systemic sclerosis (259), whereas Marra et al estimated an MCID of 0.03 utility units for people with rheumatoid arthritis (270). 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 to 0.097 utility units, with a mean of 0.041 units (276). Using an 80% specificity cut point, the MCID for improvement was estimated at 0.07 to 0.09 utility units in a study examining short-term outcomes from disease-modifying antirheumatic drug treatment for people with rheumatoid arthritis, psoriatic arthritis, or ankylosing spondylitis (277). Most recently, the MID for improvement was reported to be 0.03 utility units in a study of people with rheumatoid arthritis who had pain and dysfunction of the hands or wrists (78). Taken together, these findings indicate that only a small absolute change in the SF-6D utility score (for example, a change from 0.65 to 0.68 utility units) can represent a meaningful improvement in quality of life for patients (78).

Generalizability. The psychometric properties of the SF-6D have been assessed for a wide range of rheumatic conditions, with the results generally supporting the applicability of this measure in rheumatology settings. At present, there are no indications that the SF-6D is inappropriate for use with particular patient populations.

Use in clinical trials. The SF-6D has been used to assess the cost-effectiveness of various treatment options in a number of randomized controlled trials in the field of rheumatic diseases, including acupuncture for persistent low back pain (278), early surgery versus conservative care for sciatica (279), and exercise versus usual care to improve function in rheumatoid arthritis (280). The SF-6D was also used to capture health utilities in a randomized controlled trial that compared different methods of alleviating knee pain (273).

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 its ability to be derived from longer instruments (the SF-12 or SF-36). This means that if either of these measures has been administered in a clinical trial or observational study, a utility score can be obtained from existing data without the need for licensing or administering further questionnaires. A further strength is that a relatively small change in SF-6D utility score corresponds with change that is meaningful to patients.

Caveats and cautions. The major drawback of the SF-6D is that the scale does not cover the range below 0.30, which would be a common health state in many rheumatic conditions. This can potentially make the scale insensitive to changes among individuals with poor health. The evidence for the responsiveness of the SF-6D is mixed, and studies that aim to measure change over time (and particularly deterioration in HRQOL) should exercise caution when considering the use of the SF-6D.

Clinical usability. The SF-6D is not a tool to be used in the clinical setting as it is a utility instrument designed to inform economic evaluations.

Research usability. In research settings, its generic (non–disease-specific) nature enables the SF-6D to be used for comparisons across health conditions and to provide estimates of relative societal burden of different conditions when population norms are used as benchmarks. Similarly, it can also be used to inform health care resource allocation decisions across health care conditions through cost utility analysis and the calculation of incremental cost-effectiveness ratios.

ASSESSMENT OF QUALITY OF LIFE SCALE

Description

Purpose. The AQoL instruments are multiattribute utility measures of HRQOL (https://www.aqol.com.au). In a similar way to other health utility measures (EQ-5D and SF-6D), the AQoL instruments were designed for use across health conditions to enable health economic evaluation studies to be undertaken. The AQoL was originally published in 1999 (281), and six versions have been developed to date: the AQoL-4D (the original version), the AQoL-6D (which includes pain and coping domains), the AQoL-7D (the AQoL-6D plus a vision domain), the AQoL-8D (with domains for mental health) (see https://www.aqol.com.au), and an abridged version of the AQoL-4D called the AQoL-8 (which has only eight items) (https://www.aqol.com.au/index.php/aqolinstruments/2-uncategorised/74-aqol-8). The focus of this review is on the AQoL-4D and the AQoL-6D, both of which have been used in rheumatic conditions settings.

Content or domains. The AQoL-4D consists of the following four domains: independent living, mental health, relationships, and senses (281). The AQoL-6D includes the same four domains (but with revised items and response options) and the following two additional domains: coping and pain (282).

The conceptual model for the original version of the AQoL was based on the WHO definition of health (281). In the development of both the AQoL-4D and the AQoL-6D, the items were sourced from focus groups with clinicians and review of the content of existing HRQOL questionnaires. Final item selection was based on exploratory and confirmatory factor analyses and reliability analyses (281).

Number of items. The AQoL-4D has 12 items, with 3 items in each of the 4 domains. The AQoL-6D has 20 items, with the 2 additional domains of independent living and mental health each having 4 items.

Response options/scale. Guttman scaling is used for the AQoL response options, with higher scores indicative of progressively higher levels of disability/difficulty. The AQoL-4D items each have four response options, and the AQoL-6D items each have between four and seven options. Each item uses its own response scale.

Recall period for items. The instructions in the AQoL instruments ask respondents to rate their health state over the previous week.

Cost to use. The AQoL instruments and scoring algorithms are available at no cost.

How to obtain. All AQoL instruments are available from http://www.aqol.com.au/ and from Monash University Centre for Health Economics (https://www.monash.edu/business/che/aqol). The AQoL is subject to copyright restrictions, and online registration of study details is requested before downloading the AQoL instruments (https://www.monash.edu/business/che/aqol/aqol-registration-form). Modification of the AQoL instruments is not permitted.

Practical application

Method of administration. The AQoL instruments can be self-administered (paper and pencil or electronically) or administered by an interviewer (face-to-face or by telephone). The agreement between self- and interviewer-administered (by telephone) versions of the original AQoL-4D was high, with an ICC of 0.83 (95% CI 0.76-0.88) and with the two versions producing comparable mean scores (283). However, in another study, the correlation between mail and telephone administration of the AQoL-4D was only 0.66, indicating that different methods of AQoL administration should not be used interchangeably (284).

Scoring. Domain scores and an overall utility score are calculated for each instrument. The health states described by the items are initially weighted using values obtained from the general population from Time Trade-Off interviews (https://www.aqol.com.au/index.php/online-tto). Domain scores within an AQoL instrument are combined using a multiplicative scoring procedure. Scoring algorithms are available for download from the AQoL website in SPSS and STATA readable formats (https://www.aqol.com.au/index.php/scoring-algorithms). The scoring algorithm allows for only one missing value per domain for domains with three to four items and two missing values per domain for domains with more than four items. Missing values are imputed from the mean of the nonmissing items in the domain. Alternatively, assistance with scoring can be sought from the AQoL developers (https://www.aqol.com.au/index.php/contact-aqol-group).

Score interpretation. The AQoL utility score ranges from –0.04 (health state worse than death) to 0.00 (death) and 1.00 (full health) (285,286). Population norms are available only for the Australian general population. Studies reporting the normative values for the AQoL-4D (287) and the AQoL-6D (288) are available from the AQoL website (https://www.aqol.com.au/index.php/norms).

Respondent time to complete. The AQoL developers estimate completion time for the AQoL-4D to be 1 or 2 minutes and for the AQoL-6D to be 2 to 3 minutes (https://www.aqol.com.au/index.php/aqolinstruments).

Administrative burden. The AQoL instruments appear to have low administrator burden for pen and paper administration and electronic administration. In the original 1999 AQoL paper, the developers reported that most respondents found the items easy to answer (281). In less than 1% of cases, participants rated items as difficult because they already experienced a low quality of life and it was "upsetting and difficult" to reflect on this. The AQoL instruments use simple language and are easy to administer and score. The use of the computerized scoring algorithm requires basic knowledge of statistical software. No special equipment is needed to administer the AQoL instruments.

Translations/adaptations. Translations of the AQoL in Spanish, German, Danish, Chinese, and Italian are available at no cost at https://www.aqol.com.au/index.php/aqol-translations, but there is no information provided on the translation methods and processes. The study by Si et al appears to have used the simplified Chinese AQoL-4D, although this is not explicitly stated in the paper (289). The AQoL-4D has also been translated into the South Indian Kannada language (290).

Psychometric information

Floor and ceiling effects. Information on floor and ceiling effects for the AQoL instruments in rheumatic diseases settings is limited. No floor or ceiling effects were found for the AQoL-4D in a 2006 study of 222 patients with osteoarthritis recruited from clinical and community settings (291). More broadly, in a publication reporting AQoL-4D population norms (using data from the 2007 Australian National Survey of Mental Health and Wellbeing), there were negligible floor effects (1.4%) and substantial ceiling effects (47.2%). Furthermore, men (49.6%) were more likely than women (45.1%) to obtain scores in the ceiling decile (0.91-1.00) (287).

Reliability. *Internal consistency.* Information on the internal consistency of the AQoL instruments in rheumatic conditions is also limited. In a study of 139 individuals with rheumatoid arthritis (292), the AQoL-4D utility score had a Cronbach's α of 0.71 at baseline and of 0.61 at 2 weeks follow-up testing. In the general population, there has been mixed support for the internal consistency of the AQoL instruments. Two population-based studies reported a Cronbach's α of about 0.80 for the AQoL-4D (281,285). However, in a web-based sample of 385 individuals, Cronbach's α was 0.47 for the AQoL-4D and was 0.94 for the AQoL-6D (293).

The three-item domains of the original AQoL-4D had low to acceptable internal consistency in the general population, with Cronbach's α ranging from 0.52 (mental health) to 0.77 (independent living) (281). Similarly, in two separate longitudinal community cohort studies, Cronbach's α for the AQoL-6D domains ranged from 0.50 (senses) to 0.86 (independent living) (294). The relationships domain was found to have low internal consistency using Cronbach's α (0.63) but an acceptable internal consistency using coefficient H (0.76). However, it is important to remember that these instruments were intended to be primarily used as an overall utility score rather than as single domain scores.

Test-retest reliability. In a web-based sample of 385 individuals, very high test-retest reliability was indicated for both the AQoL-4D (0.83 and 0.85) and AQoL-6D (0.88 and 0.85) with three completions (2-week and 1-month intervals) (293). In a study of 39 patients rheumatoid arthritis from community-based private rheumatology practices in Australia, Spearman's correlation between AQoL-4D utility scores administered 2 weeks apart was 0.87, further supporting test-retest reliability of this measure (292).

Validity. *Face/content.* There appear to be no specific studies into face validity and the AQoL-4D and AQoL-6D, except as described for the development of the AQoL instruments (https://www.aqol.com.au/index.php/research-papers). Richardson et al describe the use of instrument construction theory to assess the content validity of the AQoL-6D (282). The authors concluded that compared with the AQoL-4D, the AQoL-6D captures a broader range of quality-of-life states. The absence of floor or ceiling effects in osteoarthritis also lends supports for the content validity of the AQoL-4D in rheumatic conditions because it indicates that the AQoL-4D can adequately capture the full range of HRQOL experiences in this population (291).

Construct. The evidence for the construct validity of the AQoL is generally favorable, with several publications providing evidence to support the factorial validity of these instruments (281,282,294–296). However, the factorial structure of the AQoL instruments has not yet been evaluated among patients with rheumatic conditions.

In rheumatology settings, the convergent validity of AQoL-4D was tested in a study of 222 individuals with osteoarthritis (291), and AQoL utility had high to moderate correlations with the WOMAC scales (r = –0.51 to –0.63) and the Lequesne index (r = –0.76). All correlations were of hypothesized magnitude and direction. The study of 139 patients with rheumatoid arthritis from community-based private rheumatology practices in Australia showed correlations of expected direction and magnitude between the AQoL-4D and the HAQ (r = –0.76) and the SF-36 PCS (r = –0.72) (292).

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, the EQ-5D, the 15D, and the SF-36 were 0.73 or higher (297), indicating good convergent validity. The AQoL-4D utility scores also correlated well with health care costs in an 18-month follow-up of more than 1500 individuals with a range of chronic conditions (296). Although these results support convergent validity of the AQoL instruments, their discriminant validity needs further study.

The AQoL has a 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 = 2840), 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 the arthritis group (mean = 0.72; 99% CI 0.70-0.74) followed by the 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) (298). The AQoL-4D was also able to differentiate between severity levels of osteoarthritis, with the utility score exhibiting a moderate ES (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 (291). Similar results were reported in at least one other study (299). A 2015 study demonstrated a much lower HRQOL in younger people (aged 20-55) with hip or knee OA compared with the population norm (mean difference of –0.35 AQoL units; 95% CI –0.40 to –0.31) (286).

Responsiveness. There is little recent research into the ability of the AQoL instruments to detect change and treatment effects in rheumatology settings. However, one longitudinal study found a significant improvement in AQoL-4D scores 2 years after total knee arthroplasty in patients with osteoarthritis (before surgery: 0.70 ± 0.11, 1 year after surgery: 0.71 ± 0.17, and 2 years: 0.75 ± 0.13; $P < 0.01$) (289).

Minimally important differences. The estimated MCID for self-reported health transition over time for the AQoL-4D utility score is 0.06 (285), but these results were derived from population-based respondents in Australia (n = 3010), not from studies in rheumatology settings.

Generalizability. Although the psychometric methods used to construct the AQoL instruments provide strong evidence for content and construct validity, the AQoL instruments have not been extensively evaluated in rheumatology settings, and hence the generalizability of these instruments to specific rheumatic conditions is not known at present.

Use in clinical trials. Since our 2011 review (1), the AQoL instruments have been used as a secondary outcome measure in knee osteoarthritis clinical trials to evaluate a physical therapist–delivered combined pain-coping skills and exercise intervention (AQoL-6D) (300), an internet-delivered home exercise and pain-coping skills training intervention (AQoL-6D) (301), the efficacy of unloading shoes for self-management of knee osteoarthritis symptoms (AQoL-6D) (302), and the benefit of high- versus low-dose fish oil for symptomatic and structural outcomes in knee osteoarthritis (AQoL-4D) (303).

Critical appraisal of overall value to the rheumatology community

Strengths. The primary strength of the AQoL instruments is the use of stringent psychometric methods and a theoretical model of HRQOL for their development (281). The domains covered by the AQoL instruments are relevant and useful for clinical and research purposes in rheumatology. Studies conducted with the general population and some rheumatology populations indicate that the AQoL instruments have good evidence of construct validity and good temporal stability, and there are population norms available. The AQoL instruments also have a low administration burden and are available at no cost.

Caveats and cautions. Research on psychometric properties of the AQoL-4D in rheumatology is currently limited. Furthermore, evidence indicates that the AQoL domains have low internal consistency, and it might be preferable to only report the overall utility scores. Translations and cultural adaptations are available in only a few languages. As with all generic health utility tools, the data from the AQoL instruments are unlikely to detect small clinical changes.

Clinical usability. Studies of psychometric properties of the AQoL instruments in rheumatic conditions are limited, and so their clinical usability in this setting is not known. Furthermore, given the relatively low internal consistency of the domain scores, these instruments do not appear to be suitable for conducting assessments at the individual level. These instruments can potentially be used to track changes for groups of patients over time, although more research into their responsiveness is needed.

Research usability. The AQoL instruments can be used for the evaluation of HRQOL pre/post interventions, measurement (and potentially comparison) of HRQOL in different populations and disease settings, and monitoring longitudinal changes in HRQOL in a range of health conditions, although most studies supporting these are not conducted in rheumatology settings. The availability of the population norms for both the 4D and 6D versions of the AQoL provides context for score interpretation and facilitates the usefulness of these tools for research purposes. However, only Australian norm data are available. There is the potential for data from the AQoL to be useful for comparisons of quality of life across rheumatology diseases and populations and for health economic appraisals such as cost utility assessments.

CONCLUSIONS

The results of this review concur with our previous findings and indicate that there is currently no single best measure of general health or HRQOL in rheumatology, with strengths and weaknesses identified in all measures considered. When a relatively brief measure of overall physical and mental health status is required, the PROMIS-GH or SF-12 might be recommended, whereas the SF-36, WHOQOL-BREF, or NHP would be suitable for a more detailed profiling of health. The NHP might be the least preferred measure of health status in either mildly or severely impaired subpopulations because of floor and ceiling effects, whereas the WHOQOL-BREF appears to have good ability to capture the full range of health states in rheumatic conditions with no substantive floor or ceiling effects. Of the SF-6D, the EQ-5D, and the AQoL, the EQ-5D has been evaluated most extensively in rheumatic conditions, with the five response level (EQ-5D-5L) version being preferable to the three response level (EQ-5D-3L) version. The AQoL, with very low administrative burden and good evidence of construct validity, is a promising measure, but its psychometric properties in rheumatic conditions require further study. The advantage of the SF-6D utility score is that it can be derived from either the SF-12 or the SF-36, eliminating the need for administering a separate utility measure for use in health economic analysis.

For the questionnaires reviewed in this paper, the strongest psychometric evidence pertains to their face, convergent, and known-groups validity, whereas evidence of factorial validity is generally weak for all measures apart from the AQoL. The content validity of a measure is largely dependent upon the nature of the construct being measured, and users are encouraged to carefully assess the intended measures to determine suitability for the study aims. Information on the MIDs/MCIDs is also largely lacking for the rheumatic conditions. The results of this review call for further systematic investigations of the psychometric properties of instruments currently used to assess health and HRQOL in rheumatic conditions. Table 1 and 2.

AUTHOR CONTRIBUTIONS

All authors drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

REFERENCES

1. Busija L, Pausenberger E, Haines TP, Haymes S, Buchbinder R, Osborne RH. 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). Arthritis Care Res (Hoboken) 2011;63 Suppl 11:S383–412.
2. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale (NJ): Lawrence Erlbaum Associates; 1988.
3. 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.
4. Ware JE. SF-36 health survey update. Spine (Phila Pa 1976) 2000;25:3130–9.
5. Tarlov AR, Ware JE, 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.
6. 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.
7. McHorney CA, Kosinski M, Ware JE. 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.
8. 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.
9. Weinberger M, Oddone EZ, Samsa GP, Landsman PB. Are health-related quality-of-life measures affected by the mode of administration? [clinical trial]. J Clin Epidemiol 1996;49:135–40.
10. 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.
11. 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.
12. 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? [comparative study]. Gerontology 2001;47:334–40.

13. 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.
14. Ware JE, Kosinski MA, Gandek B. SF-36 Health Survey: manual and interpretation guide. Lincoln (RI): QualityMetric Inc.; 2005.
15. Ware JE, 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. J Clin Epidemiol 1998;51:1167–70.
16. Hawthorne G, Osborne RH, Taylor A, Sansoni J. The SF36 Version 2: critical analyses of population weights, scoring algorithms and population norms. Qual Life Res 2007;16:661–73.
17. Frieling MA, Davis WR, Chiang G. The SF-36v2 and SF-12v2 health surveys in New Zealand: norms, scoring coefficients and cross-country comparisons. Aust N Z J Public Health 2013;37:24–31.
18. Roser K, Mader L, Baenziger J, Sommer G, Kuehni CE, Michel G. Health-related quality of life in Switzerland: normative data for the SF-36v2 questionnaire. Qual Life Res 2019;28:1963–77.
19. Tucker G, Adams R, Wilson D. Observed agreement problems between sub-scales and summary components of the SF-36 version 2: an alternative scoring method can correct the problem. PLoS One 2013;8:e61191.
20. Tucker G, Adams R, Wilson D. The case for using country-specific scoring coefficients for scoring the SF-12, with scoring implications for the SF-36. Qual Life Res 2016;25:267–74.
21. Ware JE, 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 Suppl:AS264–79.
22. 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.
23. 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.
24. Australian Bureau of Statistics. National health survey: SF36 population norms, Australia, 1995. Canberra: Australian Bureau of Statistics; 1997.
25. Sullivan M, Karlsson J, Ware JE. SF-36 Swedish manual and interpretation guide. Gothenburg: Gothenburg University; 1994.
26. 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.
27. 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.
28. Sow WT, Wee HL, Wu Y, Tai ES, Gandek B, Lee J, et al. Normative data for the Singapore English and Chinese SF-36 version 2 health survey. Ann Acad Med Singapore 2014;43:15–23.
29. Coons S, Rao S, Keininger D, Hays RD. A comparative review of generic quality-of-life instruments. Pharmacoeconomics 2000;17:13–35.
30. 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.
31. 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. Eli Lilly and Company and The Health Institute, New England Medical Center; 1996.
32. 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.
33. 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.
34. Zampelis V, Ornstein E, Franzen H, Atroshi I. A simple visual analog scale for pain is as responsive as the WOMAC, the SF-36, and the EQ-5D in measuring outcomes of revision hip arthroplasty. Acta Orthop 2014;85:128–32.
35. Klosinski M, Tomaszewski KA, Tomaszewska IM, Klosinski P, Skrzat J, Walocha JA. Validation of the Polish language version of the SF-36 Health Survey in patients suffering from lumbar spinal stenosis. Ann Agric Environ Med 2014;21:866–70.
36. Kwan YH, Fong WW, Lui NL, Yong ST, Cheung YB, Malhotra R, et al. Validity and reliability of the Short Form 36 Health Surveys (SF-36) among patients with spondyloarthritis in Singapore. Rheumatol Int 2016;36:1759–65.
37. Zhang Y, Zhou F, Sun Y. Assessment of health-related quality of life using the SF-36 in Chinese cervical spondylotic myelopathy patients after surgery and its consistency with neurological function assessment: a cohort study. Health Qual Life Outcomes 2015;13:39.
38. Krantz E, Wide U, Trimpou P, Bryman I, Landin-Wilhelmsen K. Comparison between different instruments for measuring health-related quality of life in a population sample, the WHO MONICA Project, Gothenburg, Sweden: an observational, cross-sectional study. BMJ Open 2019;9:e024454.
39. Thumboo J, Wu Y, Tai ES, Gandek B, Lee J, Ma S, et al. Reliability and validity of the English (Singapore) and Chinese (Singapore) versions of the Short-Form 36 version 2 in a multi-ethnic urban Asian population in Singapore. Qual Life Res 2013;22:2501–8.
40. Guo Q, Li Q, Zheng B, Yang T, Li Y, Liu B, et al. The reliability and validity of Short Form-36 questionnaire in patients with gout [abstract]. Chinese Journal of Rheumatology 2018;22:446–51.
41. Busija L, Osborne RH, Tatangelo G, Niutta S, Buchbinder R. Psychometric evaluation supported construct validity, temporal stability, and responsiveness of the Osteoarthritis Questionnaire. J Clin Epidemiol 2019;114:11–21.
42. Chiarotto A, Terwee CB, Kamper SJ, Boers M, Ostelo RW. Evidence on the measurement properties of health-related quality of life instruments is largely missing in patients with low back pain: a systematic review. J Clin Epidemiol 2018;102:23–37.
43. Kaya BB, İçağasıoğlu A. Reliability and validity of the Turkish version of Short Form 36 (SF-36) in patients with rheumatoid arthritis. J Surg Med 2018;2:11–6.
44. 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.
45. Kuzmanić B, Ivančić N, Paušić J. Test-retest reliability of the 36-item health survey (SF-36) as quality of life measures in elderly Croatian population. Acta Kinesiologica 2017;11:104–7.
46. LoMartire R, Ang BO, Gerdle B, Vixner L. Psychometric properties of Short Form-36 Health Survey, EuroQol 5-dimensions, and Hospital Anxiety and Depression Scale in patients with chronic pain. Pain 2020;161:83–95.
47. Michelsen B, Uhlig T, Sexton J, van der Heijde D, Hammer HB, Kristianslund EK, et al. Health-related quality of life in patients with psoriatic and rheumatoid arthritis: data from the prospective

multicentre NOR-DMARD study compared with Norwegian general population controls. Ann Rheum Dis 2018;77:1290–4.

48. Yang X, Fan D, Xia Q, Wang M, Zhang X, Li X, et al. The health-related quality of life of ankylosing spondylitis patients assessed by SF-36: a systematic review and meta-analysis. Qual Life Res 2016;25:2711–23.

49. Devilliers H, Amoura Z, Besancenot JF, Bonnotte B, Pasquali JL, Wahl D, et al. Responsiveness of the 36-item Short Form Health Survey and the Lupus Quality of Life questionnaire in SLE. Rheumatology (Oxford) 2015;54:940–9.

50. Yilmaz F, Dogu B, Sahin F, Sirzai H, Kuran B. Investigation of responsiveness indices of generic and specific measures of health related quality of life in patients with osteoporosis. J Back Musculoskelet Rehabil 2014;27:391–7.

51. Papou A, Hussain S, McWilliams D, Zhang W, Doherty M. Responsiveness of SF-36 Health Survey and Patient Generated Index in people with chronic knee pain commenced on oral analgesia: analysis of data from a randomised controlled clinical trial. Qual Life Res 2017;26:761–6.

52. Quintana JM, Escobar A, Bilbao A, Arostegui I, Lafuente I, Vidaurreta I. Responsiveness and clinically important differences for the WOMAC and SF-36 after hip joint replacement. Osteoarthritis Cartilage 2005;13:1076–83.

53. 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.

54. Badhiwala JH, Witiw CD, Nassiri F, Akbar MA, Jaja B, Wilson JR, et al. Minimum clinically important difference in SF-36 scores for use in degenerative cervical myelopathy. Spine (Phila Pa 1976) 2018;43:E1260–6.

55. Bisson LJ, Kluczynski MA, Wind WM, Fineberg MS, Bernas GA, Rauh MA, et al. Patient outcomes after observation versus debridement of unstable chondral lesions during partial meniscectomy: the Chondral Lesions And Meniscus Procedures (ChAMP) randomized controlled trial. J Bone Joint Surg Am 2017;99:1078–85.

56. Lopes de Jesus CC, Dos Santos FC, de Jesus LM, Monteiro I, Sant'Ana M, Trevisani VF. Comparison between intra-articular ozone and placebo in the treatment of knee osteoarthritis: a randomized, double-blinded, placebo-controlled study. PLoS One 2017;12:e0179185.

57. Lu J, Huang L, Wu X, Fu W, Liu Y. Effect of Tai Ji Quan training on self-reported sleep quality in elderly Chinese women with knee osteoarthritis: a randomized controlled trail. Sleep Med 2017;33:70–5.

58. Strand V, Kremer JM, Gruben D, Krishnaswami S, Zwillich SH, Wallenstein GV. Tofacitinib in combination with conventional disease-modifying antirheumatic drugs in patients with active rheumatoid arthritis: patient-reported outcomes from a Phase III randomized controlled trial. Arthritis Care Res (Hoboken) 2017;69:592–8.

59. Li Z, An Y, Su H, Li X, Xu J, Zheng Y, et al. Tofacitinib with conventional synthetic disease-modifying antirheumatic drugs in Chinese patients with rheumatoid arthritis: patient-reported outcomes from a Phase 3 randomized controlled trial. Int J Rheum Dis 2018;21:402–14.

60. Au KY, Chen H, Lam WC, Chong CO, Lau A, Vardhanabhuti V, et al. Sinew acupuncture for knee osteoarthritis: study protocol for a randomized sham-controlled trial. BMC Complement Altern Med 2018;18:133.

61. Hancke JL, Srivastav S, Caceres DD, Burgos RA. A double-blind, randomized, placebo-controlled study to assess the efficacy of Andrographis paniculata standardized extract (ParActin®) on pain reduction in subjects with knee osteoarthritis. Phytother Res 2019;33:1469–79.

62. Assumpcao A, Matsutani LA, Yuan SL, Santo AS, Sauer J, Mango P, et al. Muscle stretching exercises and resistance training in fibromyalgia: which is better? A three-arm randomized controlled trial. Eur J Phys Rehabil Med 2018;54:663–70.

63. Gomiero AB, Kayo A, Abraao M, Peccin MS, Grande AJ, Trevisani VF. Sensory-motor training versus resistance training among patients with knee osteoarthritis: randomized single-blind controlled trial. Sao Paulo Med J 2018;136:44–50.

64. Xie Y, Zhang C, Jiang W, Huang J, Xu L, Pang G, et al. Quadriceps combined with hip abductor strengthening versus quadriceps strengthening in treating knee osteoarthritis: a study protocol for a randomized controlled trial. BMC Musculoskelet Disord 2018;19:147.

65. Ware J, 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.

66. Ware JE. SF-36 Health Survey update. In: Maruish ME, editor. The use of psychological testing for treatment planning and outcomes assessment. 3rd ed. Mahwah (NJ): Lawrence Earlbaum Associates; 2004. p. 693–718.

67. 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.

68. 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 Inc.; 2002.

69. 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.

70. 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. J Clin Epidemiol 1998;51:1171–8.

71. Sanderson K, Andrews G. The SF-12 in the Australian population: cross-validation of item selection. Aust N Z J Public Health 2002;26:343–5.

72. 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.

73. 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? [research article]. J Rheumatol 2009;36:2183–9.

74. Islam N, Khan IH, Ferdous N, Rasker JJ. Translation, cultural adaptation and validation of the English "Short Form SF 12v2" into Bengali in rheumatoid arthritis patients. Health Qual Life Outcomes 2017;15:109.

75. Lenert LA. The reliability and internal consistency of an Internet-capable computer program for measuring utilities. Qual Life Res 2000;9:811–7.

76. Montazeri A, Vahdaninia M, Mousavi SJ, Asadi-Lari M, Omidvari S, Tavousi M. The 12-item Medical Outcomes Study Short Form Health Survey version 2.0 (SF- 12v2): a population-based validation study from Tehran, Iran. Health Qual Life Outcomes 2011;9:12.

77. Luo X, George ML, 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.

78. Dritsaki M, Petrou S, Williams M, Lamb SE. An empirical evaluation of the SF-12, SF-6D, EQ-5D and Michigan Hand Outcome Questionnaire in patients with rheumatoid arthritis of the hand. Health Qual Life Outcomes 2017;15:20.

79. 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.

80. 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.

81. 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.

82. Jakobsson U, Westergren A, Lindskov S, Hagell P. Construct validity of the SF-12 in three different samples. J Eval Clin Pract 2012;18:560–6.

83. Cheung PP, Lahiri M, March L, Gossec L. Patient-reported outcomes in Asia: evaluation of the properties of the Rheumatoid Arthritis Impact of Disease (RAID) score in multiethnic Asian patients with rheumatoid arthritis. Clin Rheumatol 2017;36:1149–54.

84. Fang WH, Huang GS, Chang HF, Chen CY, Kang CY, Wang CC, et al. Gender differences between WOMAC index scores, health-related quality of life and physical performance in an elderly Taiwanese population with knee osteoarthritis. BMJ Open 2015;5:e008542.

85. 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 Suppl 56:S21–8.

86. 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.

87. 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.

88. 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.

89. 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.

90. Webster KE, Feller JA. Comparison of the Short Form-12 (SF-12) health status questionnaire with the SF-36 in patients with knee osteoarthritis who have replacement surgery. Knee Surg Sports Traumatol Arthrosc 2016;24:2620–6.

91. Bourgault P, Lacasse A, Marchand S, Courtemanche-Harel R, Charest J, Gaumond I, et al. Multicomponent interdisciplinary group intervention for self-management of fibromyalgia: a mixed-methods randomized controlled trial. PLoS One 2015;10:e0126324.

92. Manoy P, Yuktanandana P, Tanavalee A, Anomasiri W, Ngarmukos S, Tanpowpong T, et al. Vitamin D supplementation improves quality of life and physical performance in osteoarthritis patients. Nutrients 2017;9:799.

93. Musumeci A, Pranovi G, Masiero S. Patient education and rehabilitation after hip arthroplasty in an Italian spa center: a pilot study on its feasibility. Int J Biometeorol 2018;62:1489–96.

94. Zhang L, Lix LM, Ayilara O, Sawatzky R, Bohm ER. The effect of multimorbidity on changes in health-related quality of life following hip and knee arthroplasty. Bone Joint J 2018;100-B:1168–74.

95. Zhai H, Geng H, Bai B, Wang Y. Differences in 1-year outcome after primary total hip and knee arthroplasty: a cohort study in older patients with osteoarthritis. Orthopade 2019;48:136–43.

96. Berliner JL, Brodke DJ, Chan V, SooHoo NF, Bozic KJ. Can preoperative patient-reported outcome measures be used to predict meaningful improvement in function after TKA? [conference proceedings]. Clin Orthop Relat Res 2017;475:149–57.

97. Clement ND, MacDonald D, Simpson AH. The minimal clinically important difference in the Oxford knee score and Short Form 12 score after total knee arthroplasty [published erratum appears in Knee Surg Sports Traumatol Arthrosc 2016;24:3696]. Knee Surg Sports Traumatol Arthrosc 2014;22:1933–9.

98. Diaz-Arribas MJ, Fernandez-Serrano M, Royuela A, Kovacs FM, Gallego-Izquierdo T, Ramos-Sanchez M, et al. Minimal clinically important difference in quality of life for patients with low back pain. Spine (Phila Pa 1976) 2017;42:1908–16.

99. Clement ND, Weir D, Holland J, Gerrand C, Deehan DJ. Meaningful changes in the Short Form 12 physical and mental summary scores after total knee arthroplasty. Knee 2019;26:861–8.

100. Kingsbury SR, Tharmanathan P, Keding A, Ronaldson SJ, Grainger A, Wakefield RJ, et al. Hydroxychloroquine effectiveness in reducing symptoms of hand osteoarthritis: a randomized trial. Ann Intern Med 2018;168:385–95.

101. Simental-Mendia M, Vilchez-Cavazos JF, Pena-Martinez VM, Said-Fernandez S, Lara-Arias J, Martinez-Rodriguez HG. Leukocyte-poor platelet-rich plasma is more effective than the conventional therapy with acetaminophen for the treatment of early knee osteoarthritis. Arch Orthop Trauma Surg 2016;136:1723–32.

102. Williamson E, McConkey C, Heine P, Dosanjh S, Williams M, Lamb SE. Hand exercises for patients with rheumatoid arthritis: an extended follow-up of the SARAH randomised controlled trial. BMJ Open 2017;7:e013121.

103. Reina-Bueno M, Vazquez-Bautista MD, Perez-Garcia S, Rosende-Bautista C, Saez-Diaz A, Munuera-Martinez PV. Effectiveness of custom-made foot orthoses in patients with rheumatoid arthritis: a randomized controlled trial. Clin Rehabil 2019;33:661–9.

104. McBain H, Shipley M, Olaleye A, Moore S, Newman S. A patient-initiated DMARD self-monitoring service for people with rheumatoid or psoriatic arthritis on methotrexate: a randomised controlled trial. Ann Rheum Dis 2016;75:1343–9.

105. 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.

106. Beaupre LA, Al-Houkail A, Johnston DW. A randomized trial comparing ceramic-on-ceramic bearing vs ceramic-on-crossfire-polyethylene bearing surfaces in total hip arthroplasty. J Arthroplasty 2016;31:1240–5.

107. Parratte S, Ollivier M, Lunebourg A, Flecher X, Argenson JN. No benefit after THA performed with computer-assisted cup placement: 10-year results of a randomized controlled study. Clin Orthop Relat Res 2016;474:2085–93.

108. 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.

109. Babazadeh S, Dowsey MM, Vasimalla MG, Stoney JD, Choong PF. Gap balancing sacrifices joint-line maintenance to improve gap symmetry: 5-year follow-up of a randomized controlled trial. J Arthroplasty 2018;33:75–8.

110. Bernal-Fortich LD, Aguilar CA, Rivera-Villa AH, Galindo-Avalos J, Aguilera-Martinez P, Torres-Gonzalez R, et al. A prospective randomized trial of total synovectomy versus limited synovectomy in primary total knee arthroplasty: evaluation of bleeding, postoperative pain, and quality of life with SF-12 v2. Eur J Orthop Surg Traumatol 2018;28:701–6.

111. Parratte S, Ollivier M, Lunebourg A, Verdier N, Argenson JN. Do stemmed tibial components in total knee arthroplasty improve

outcomes in patients with obesity? [clinical trial]. Clin Orthop Relat Res 2017;475:137–45.

112. Fransen M, Nairn L, Bridgett L, Crosbie J, March L, Parker D, et al. Post-acute rehabilitation after total knee replacement: a multicenter randomized clinical trial comparing long-term outcomes. Arthritis Care Res (Hoboken) 2017;69:192–200.

113. Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. A quantitative approach to perceived health status: a validation study. J Epidemiol Community Health 1980;34:281–6.

114. 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.

115. Hunt SM, McEwen J. The development of a subjective health indicator. Sociol Health Illn 1980;2:231–46.

116. Hunt SM, McEwen J, McKenna SP. Measuring health states. Surry Hills: Croom Helm Ltd; 1986.

117. Baro E, Ferrer M, Vazquez O, Miralles R, Pont A, Esperanza A, et al. Using the Nottingham Health Profile (NHP) among older adult inpatients with varying cognitive function. Qual Life Res 2006;15:575–85.

118. Tabali M, Jeschke E, Dassen T, Ostermann T, Heinze C. The Nottingham Health Profile: A feasible questionnaire for nursing home residents? [research article]. Int Psychogeriatr 2012;24:416–24.

119. McKenna SP, Hunt SM, McEwen J. Weighting the seriousness of perceived health problems using Thurstone's method of paired comparisons. Int J Epidemiol 1981;10:93–7.

120. McEwen J. The Nottingham Health Profile. In: Walker SR, Rosser RM, editors. Quality of life assessment: key issues in the 1990s. Dordrecht: Springer; 1993.

121. Hunt SM, McEwen J, McKenna SP. Perceived health: age and sex comparisons in a community. J Epidemiol Community Health 1984;38:156–60.

122. Hunt SM, McKenna SP. The Nottingham Health Profile user's manual, revised. Manchester: Galen Research and Consultancy; 1991.

123. Lovas K, Kaló Z, McKenna SP, Whalley D, Péntek M, Genti G. Establishing a standard for patient-completed instrument adaptations in Eastern Europe: experience with the Nottingham Health Profile in Hungary. Health Policy 2003;63:49–61.

124. Baum FE, Cooke RD. Community-health needs assessment: use of the Nottingham health profile in an Australian study. Med J Aust 1989;150:581–90.

125. Vidalis A. The Greek version of the Nottingham Health Profile: features of its adaptation. Hippokratia 2009;6:79.

126. 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.

127. Thorsen H, McKenna SP, Gottschalck L. The Danish version of the Nottingham Health Profile: its adaptation and reliability. Scand J Prim Health Care 1993;11:124–9.

128. 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. Int Disabil Stud 1988;10:159–63.

129. Erdman RA, Passchier J, Kooijman M, Stronks DL. The Dutch version of the Nottingham Health Profile: investigations of psychometric aspects. Psychol Rep 1993;72:1027–35.

130. Alonso J, Prieto L, Antó JM. The Spanish version of the Nottingham Health Profile: a review of adaptation and instrument characteristics. Qual Life Res 1994;3:385–93.

131. McKenna SP, Hunt SM, Tennant A. The development of a patient-completed index of distress from the Nottingham Health Profile: a new measure for use in cost-utility studies. Br J Med Econ 1993;6:13–24.

132. Wann-Hansson C, Klevsgard R, Hagell P. Cross-diagnostic validity of the Nottingham Health Profile Index of Distress (NHPD). Health Qual Life Outcomes 2008;6:47.

133. Prieto L, Alonso J, Lamarca R, Wright BD. Rasch measurement for reducing the items of the Nottingham Health Profile. J Outcome Meas 1998;2:285–301.

134. Sharples LD, Todd CJ, Caine N, Tait S. Measurement properties of the Nottingham Health Profile and Short Form 36 health status measures in a population sample of elderly people living at home: results from ELPHS. Br J Health Psychol 2000;5:217–33.

135. Boyer F, Novella JL, Bertaud S, Delmer F, Vesselle B, Etienne JC. Hereditary neuromuscular disease and multicomposite subjective health status: feasibility, internal consistency and test-retest reliability in the French version of the Nottingham Health Profile, the ISPN. Clin Rehabil 2005;19:644–53.

136. 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. Groningen: University of Groningen; 2005.

137. 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.

138. Post MW, Gerritsen J, Diederiks JP, DeWitte 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.

139. 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.

140. 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.

141. Juhel J, Gaillot AC. Structural validity and age-based differential item functioning of the French Nottingham Health Profile in a sample of surgery patients. Adv Psychol Stud 2012;1:14–21.

142. 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.

143. Nawito ZO, El-Azkalany GS, El-Sayad M. Nottingham Health Profile assessment of health-related quality of life in primary knee osteoarthritis patients: relation to clinical features and radiologic score. The Egypt Rheumatologist 2018;40:265–8.

144. Yildiz N, Topuz O, Gungen GO, Deniz S, Alkan H, Ardic F. Health-related quality of life (Nottingham Health Profile) in knee osteoarthritis: correlation with clinical variables and self-reported disability. Rheumatol Int 2010;30:1595–600.

145. 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.

146. 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.

147. Garip Y, Oztas D, Guler T. Prevalence of fibromyalgia in Turkish geriatric population and its impact on quality of life. Agri 2016;28:165–70.

148. Van Balen R, Essink-Bot ML, Steyerberg E, Cools H, Habbema DF. Quality of life after hip fracture: a comparison of four health status measures in 208 patients. Disabil Rehabil 2003;25:507–19.

149. 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.

150. Kocyigit F, Turkmen MB, Acar M, Guldane N, Kose T, Kuyucu E, et al. Kinesio taping or sham taping in knee osteoarthritis? A

randomized, double-blind, sham-controlled trial. Complement Ther Clin Pract 2015;21:262–7.

151. Aciksoz S, Akyuz A, Tunay S. The effect of self-administered superficial local hot and cold application methods on pain, functional status and quality of life in primary knee osteoarthritis patients. J Clin Nurs 2017;26:5179–90.

152. Ekici G, Unal E, Akbayrak T, Vardar-Yagli N, Yakut Y, Karabulut E. Effects of active/passive interventions on pain, anxiety, and quality of life in women with fibromyalgia: Randomized controlled pilot trial. Women Health 2017;57:88–107.

153. Kibar S, Yildiz HE, Ay S, Evcik D, Ergin ES. New approach in fibromyalgia exercise program: a preliminary study regarding the effectiveness of balance training. Arch Phys Med Rehabil 2015;96:1576–82.

154. World Health Organization. WHOQOL-BREF introduction, administration, scoring and generic version of the assessment. 1996. URL: https://www.who.int/mental_health/media/en/76.pdf.

155. The WHOQOL Group. Development of the World Health Organization WHOQOL-BREF quality of life assessment. Psychol Med 1998;28:551–8.

156. Skevington SM, Lotfy M, O'Connell KA. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial: a report from the WHOQOL group. Qual Life Res 2004;13:299–310.

157. Hawthorne G, Herrman H, Murphy B. Interpreting the WHOQOL-BREF: preliminary population norms and effect sizes. Soc Indic Res 2006;77:37–59.

158. Cruz LN, Polanczyk CA, Camey SA, Hoffmann JF, Fleck MP. Quality of life in Brazil: normative values for the WHOQOL-BREF in a southern general population sample. Qual Life Res 2011;20:1123–9.

159. Hwang HF, Liang WM, Chiu YN, Lin MR. Suitability of the WHOQOL-BREF for community-dwelling older people in Taiwan. Age Ageing 2003;32:593–600.

160. Casamali FF, Schuch FB, Scortegagna SA, Legnani E, de Marchi AC. Accordance and reproducibility of the electronic version of the WHOQOL-BREF and WHOQOL-OLD questionnaires. Exp Gerontol 2019;125:110683.

161. Taylor WJ, Myers J, Simpson RT, McPherson KM, Weatherall M. Quality of life of people with rheumatoid arthritis as measured by the World Health Organization Quality of Life Instrument, short form (WHOQOL-BREF): score distributions and psychometric properties. Arthritis Rheum 2004;51:350–7.

162. Redko C, Rogers N, Bule L, Siad H, Choh A. Development and validation of the Somali WHOQOL-BREF among refugees living in the USA. Qual Life Res 2015;24:1503–13.

163. Saqib Lodhi F, Raza O, Montazeri A, Nedjat S, Yaseri M, Holakouie-Naieni K. Psychometric properties of the Urdu version of the World Health Organization's quality of life questionnaire (WHOQOL-BREF). Med J Islam Repub Iran 2017;31:129.

164. Uddin MN, Islam FM. Psychometric evaluation of an interview-administered version of the WHOQOL-BREF questionnaire for use in a cross-sectional study of a rural district in Bangladesh: an application of Rasch analysis. BMC Health Serv Res 2019;19:216.

165. Balalla SK, Medvedev ON, Siegert RJ, Krageloh CU. Validation of the WHOQOL-BREF and shorter versions using Rasch analysis in traumatic brain injury and orthopedic populations. Arch Phys Med Rehabil 2019;100:1853–62.

166. Ackerman IN, Graves SE, Bennell KL, Osborne RH. Evaluating quality of life in hip and knee replacement: psychometric properties of the World Health Organization Quality of Life short version instrument. Arthritis Rheum 2006;55:583–90.

167. Skevington SM, McCrate FM. Expecting a good quality of life in health: assessing people with diverse diseases and conditions using the WHOQOL-BREF. Health Expect 2012;15:49–62.

168. Castro PC, Driusso P, Oishi J. Convergent validity between SF-36 and WHOQOL-BREF in older adults. Rev Saude Publica 2014;48:63–7.

169. Cheung YB, Yeo KK, Chong KJ, Khoo EY, Wee HL. Reliability and validity of the English-, Chinese- and Malay-language versions of the World Health Organization Quality of Life (WHOQOL-BREF) Questionnaire in Singapore. Ann Acad Med Singapore 2017;46:461–9.

170. Chiu WT, Huang SJ, Hwang HF, Tsauo JY, Chen CF, Tsai SH, et al. Use of the WHOQOL-BREF for evaluating persons with traumatic brain injury. J Neurotrauma 2006;23:1609–20.

171. Naumann VJ, Byrne GJ. WHOQOL-BREF as a measure of quality of life in older patients with depression. Int Psychogeriatr 2004;16:159–73.

172. Colbourn T, Masache G, Skordis-Worrall J. Development, reliability and validity of the Chichewa WHOQOL-BREF in adults in Lilongwe, Malawi. BMC Res Notes 2012;5:346.

173. Ginieri-Coccossis M, Triantafillou E, Tomaras V, Soldatos C, Mavreas V, Christodoulou G. Psychometric properties of WHOQOL-BREF in clinical and health Greek populations: incorporating new culture-relevant items. Psychiatriki 2012;23:130–42.

174. Toprak M, Erden M. Sleep quality, pain, anxiety, depression and quality of life in patients with frozen shoulder. J Back Musculoskelet Rehabil 2019;32:287–91.

175. Chachamovich E, Trentini C, Fleck MP. Assessment of the psychometric performance of the WHOQOL-BREF instrument in a sample of Brazilian older adults. Int Psychogeriatr 2007;19:635–46.

176. Kim WH, Hahn SJ, Im HJ, Yang KS. Reliability and validity of the Korean World Health Organization Quality of Life (WHOQOL)-BREF in people with physical impairments. Ann Rehabil Med 2013;37:488–97.

177. Amaral DS, Duarte AL, Barros SS, Cavalcanti SV, Ranzolin A, Leite VM, et al. Assistive devices: an effective strategy in non-pharmacological treatment for hand osteoarthritis-randomized clinical trial. Rheumatol Int 2018;38:343–51.

178. Cella D, Riley W, Stone A, Rothrock N, Reeve B, Yount S, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63:1179–94.

179. Hays RD, Bjorner JB, Revicki DA, Spritzer KL, Cella D. Development of physical and mental health summary scores from the Patient-Reported Outcomes Measurement Information System (PROMIS) global items. Qual Life Res 2009;18:873–80.

180. Liu H, Cella D, Gershon R, Shen J, Morales LS, Riley W, et al. Representativeness of the Patient-Reported Outcomes Measurement Information System Internet panel. J Clin Epidemiol 2010;63:1169–78.

181. Kahan JB, Kassam HF, Nicholson AD, Saad MA, Kovacevic D. Performance of PROMIS Global-10 to legacy instruments in patients with lateral epicondylitis. Arthroscopy 2019;35:770–4.

182. Kasturi S, Szymonifka J, Burket JC, Berman JR, Kirou KA, Levine AB, et al. Feasibility, validity, and reliability of the 10-item Patient Reported Outcomes Measurement Information System Global Health Short Form in outpatients with systemic lupus erythematosus. J Rheumatol 2018;45:397–404.

183. Terwee CB, Roorda LD, de Vet HC, Dekker J, Westhovens R, van Leeuwen J, et al. Dutch-Flemish translation of 17 item banks from the Patient-Reported Outcomes Measurement Mnformation System (PROMIS). Qual Life Res 2014;23:1733–41.

184. Zumpano CE, Mendonca TM, Silva CH, Correia H, Arnold B, Pinto RM. Cross-cultural adaptation and validation of the PROMIS Global Health scale in the Portuguese language. Cad Saude Publica 2017;33:e00107616.

185. Saad MA, Kassam HF, Suriani RJ, Pan SD, Blaine TA, Kovacevic D. Performance of PROMIS Global-10 compared with legacy instruments in patients with shoulder arthritis. J Shoulder Elbow Surg 2018;27:2249–56.

186. Hwang MC, Ogdie A, Puravath A, Reveille JD. Reliability and validity of Patient-Reported Outcomes Measurement Information System short forms in ankylosing spondylitis. J Rheumatol 2020;47: 190201.

187. Stoop N, Menendez ME, Mellema JJ, Ring D. The PROMIS Global Health Questionnaire correlates with the QuickDASH in patients with upper extremity illness. Hand (N Y) 2018;13:118–21.

188. Gouttebarge V, Aoki H, Kerkhoffs GM. Lower extremity osteoarthritis is associated with lower health-related quality of life among retired professional footballers. Phys Sportsmed 2018;46:471–6.

189. Jensen RE, Potosky AL, Moinpour CM, Lobo T, Cella D, Hahn EA, et al. United States population-based estimates of Patient-Reported Outcomes Measurement Information System symptom and functional status reference values for individuals with cancer. J Clin Oncol 2017;35:1913–20.

190. Wohlfahrt A, Bingham CO, Marder W, Phillips K, Bolster MB, Moreland LW, et al. Responsiveness of Patient-Reported Outcomes Measurement Information System measures in rheumatoid arthritis patients starting or switching a disease-modifying antirheumatic drug. Arthritis Care Res (Hoboken) 2019;71:521–9.

191. Shim J, Hamilton DF. Comparative responsiveness of the PROMIS-10 Global Health and EQ-5D questionnaires in patients undergoing total knee arthroplasty. Bone Joint J 2019;101-B:832–7.

192. Kasturi S, Szymonifka J, Berman JR, Kirou KA, Levine AB, Sammaritano LR, et al. Responsiveness of PROMIS® Global Health Short Form in outpatients with systemic lupus erythematosus. Arthritis Care Res 2019. E-pub ahead of print.

193. Husni ME, Deal C, Calabrese LH, Strnad G, Bena J, Abelson A. Using patient reported outcomes at point of care in immune mediated diseases: minimal clinically important differences [abstract]. Arthritis Rheumatol 2018;70 Suppl 10. URL: https://acrabstracts.org/abstract/using-patient-reported-outcomes-at-point-of-care-in-immune-mediated-diseases-minimal-clinically-important-differences/.

194. EuroQol Research Foundation. EQ-5D-5L user guide. URL: https://euroqol.org/publications/user-guides/.2019.

195. Van Wilder L, Rammant E, Clays E, Devleesschauwer B, Pauwels N, de Smedt D. A comprehensive catalogue of EQ-5D scores in chronic disease: results of a systematic review. Qual Life Res 2019;28:3153–61.

196. EuroQol Group. EuroQol: a new facility for the measurement of health-related quality of life. Health Policy 1990;16:199–208.

197. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.

198. Devlin NJ, Shah KK, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. Health Econ 2018;27:7–22.

199. Van Hout B, Janssen M, Feng YS, Kohlmann T, Busschbach J, Golicki D, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health 2012;15:708–15.

200. Rabin R, de Charro F. EQ-SD: a measure of health status from the EuroQol Group. Ann Med 2001;33:337–43.

201. Hernandez Alava M, Wailoo A, Grimm S, Pudney S, Gomes M, Sadique Z, et al. EQ-5D-5L versus EQ-5D-3L: the impact on cost effectiveness in the United Kingdom. Value Health 2018;21:49–56.

202. EuroQol Research Foundation. EQ-5D-3L User Guide 2018. URL: https://euroqol.org/publications/user-guides/.

203. Holland R, Smith RD, Harvey I, Swift L, Lenaghan E. Assessing quality of life in the elderly: a direct comparison of the EQ-5D and AQoL. Health Econ 2004;13:793–805.

204. Greene ME, Rader KA, Garellick G, Malchau H, Freiberg AA, Rolfson O. The EQ-5D-5L improves on the EQ-5D-3L for health-related quality-of-life assessment in patients undergoing total hip arthroplasty. Clin Orthop Relat Res 2015;473:3383–90.

205. Janssen M, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, et al. Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. Qual Life Res 2013;22:1717–27.

206. Bilbao A, Garcia-Perez L, Arenaza JC, Garcia I, Ariza-Cardiel G, Trujillo-Martin E, et al. Psychometric properties of the EQ-5D-5L in patients with hip or knee osteoarthritis: reliability, validity and responsiveness. Qual Life Res 2018;27:2897–908.

207. Conner-Spady BL, Marshall DA, Bohm E, Dunbar MJ, Loucks L, al Khudairy A, et al. Reliability and validity of the EQ-5D-5L compared to the EQ-5D-3L in patients with osteoarthritis referred for hip and knee replacement. Qual Life Res 2015;24:1775–84.

208. Tsang HH, Cheung JP, Wong CK, Cheung PW, Lau CS, Chung HY. Psychometric validation of the EuroQoL 5-dimension (EQ-5D) questionnaire in patients with spondyloarthritis. Arthritis Res Ther 2019;21:41.

209. Buchholz I, Thielker K, Feng YS, Kupatz P, Kohlmann T. Measuring changes in health over time using the EQ-5D-3L and 5L: a head-to-head comparison of measurement properties and sensitivity to change in a German inpatient rehabilitation sample. Qual Life Res 2015;24:829–35.

210. Cronbach LJ. Coefficient α and the internal structure of test. Psychometrika 1951;16:297–334.

211. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research [published correction appears in J Chiropr Med 2017;16:346]. J Chiropr Med 2016;15:155–63.

212. Conner-Spady BL, Marshall DA, Bohm E, Dunbar MJ, Noseworthy TW. Comparing the validity and responsiveness of the EQ-5D-5L to the Oxford hip and knee scores and SF-12 in osteoarthritis patients 1 year following total joint replacement. Qual Life Res 2018;27:1311–22.

213. Fransen M, Edmonds J. Reliability and validity of the EuroQol in patients with osteoarthritis of the knee. Rheumatology (Oxford) 1999;38:807–13.

214. Luo N, Chew LH, Fong KY, Koh DR, Ng SC, Yoon KH, et al. Validity and reliability of the EQ-5D self-report questionnaire in English-speaking Asian patients with rheumatic diseases in Singapore. Qual Life Res 2003;12:87–92.

215. Hurst NP, Kind P, Ruta D, Hunter M, Stubbings A. Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ-5D). Br J Rheumatol 1997;36:551–9.

216. Kim MH, Cho YS, Uhm WS, Kim S, Bae SC. Cross-cultural adaptation and validation of the Korean version of the EQ-5D in patients with rheumatic diseases. Qual Life Res 2005;14:1401–6.

217. Buchholz I, Janssen MF, Kohlmann T, Feng YS. A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. Pharmacoeconomics 2018;36:645–61.

218. Wahlberg M, Zingmark M, Stenberg G, Munkholm M. Rasch analysis of the EQ-5D-3L and the EQ-5D-5L in persons with back and neck pain receiving physiotherapy in a primary care context. Eur J Physiother 2019. E-pub ahead of print.

219. Keeley T, Al-Janabi H, Lorgelly P, Coast J. A qualitative assessment of the content validity of the ICECAP-A and EQ-5D-5L and their appropriateness for use in health research. PLoS One 2013;8:e85287.

220. Linde L, Sørensen J, Ostergaard M, Hørslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, EQ-15D, EQ-5D, RAQoL, and HAQ in patients with rheumatoid arthritis [published erratum appears in J Rheumatol 2008;35:1688]. J Rheumatol 2008;35:1528–37.

221. Obradovic M, Lal A, Liedgens H. Validity and responsiveness of EuroQol-5 dimension (EQ-5D) versus Short Form-6 dimension (SF-6D) questionnaire in chronic pain. Health Qual Life Outcomes 2013;11:110.

222. Wang SL, Wu B, Zhu LA, Leng L, Bucala R, Lu LJ. Construct and criterion validity of the Euro Qol-5D in patients with systemic lupus erythematosus. PLoS One 2014;9:e98883.

223. Fang H, Farooq U, Wang D, Yu F, Younus MI, Guo X. Reliability and validity of the EQ-5D-3L for Kashin–Beck disease in China. Springerplus 2016;5:1924.

224. Payakachat N, Ali MM, Tilford JM. Can the EQ-5D detect meaningful change? A systematic review. Pharmacoeconomics 2015;33:1137–54.

225. Brazier JE, Harper R, Munro J, Walters S, Snaith ML. Generic and condition-specific outcome measures for people with osteoarthritis of the knee. Rheumatology (Oxford) 1999;38:870–7.

226. Coretti S, Ruggeri M, McNamee P. The minimum clinically important difference for EQ-5D index: a critical review. Expert Rev Pharmacoecon Outcomes Res 2014;14:221–33.

227. Coretti S, Ruggeri M, McNamee P. From the minimum clinically important difference to the minimum cost effective difference for EQ-5D in patients with chronic widespread pain. Value Health 2014;17:A50–1.

228. Kitchen H, Hansen BB, Abetz L, Højbjerre L, Strandberg-Larsen M, et al. Patient-reported outcome measures for rheumatoid arthritis: minimal important differences review [abstract]. Arthritis Rheumatol 2013.URL: https://acrabstracts.org/abstract/patient-reported-outcome-measures-for-rheumatoid-arthritis-minimal-important-differences-review/.

229. Soer R, Reneman MF, Speijer BL, Coppes MH, Vroomen PC. Clinimetric properties of the EuroQol-5D in patients with chronic low back pain. Spine J 2012;12:1035–9.

230. Moller-Bisgaard S, Horslev-Petersen K, Ejbjerg B, Hetland ML, Ornbjerg LM, Glinatsi D, et al. Effect of magnetic resonance imaging vs conventional treat-to-target strategies on disease activity remission and radiographic progression in rheumatoid arthritis: the IMAGINE-RA randomized clinical trial. JAMA 2019;321:461–72.

231. Day RO, Frensham LJ, Nguyen AD, Baysari MT, Aung E, Lau AY, et al. Effectiveness of an electronic patient-centred self-management tool for gout sufferers: a cluster randomised controlled trial protocol. BMJ Open 2017;7:e017281.

232. Van Tubergen A, Landewé R, van der Heijde D, Hidding A, Wolter N, Asscher M, et al. Combined spa-exercise therapy is effective in patients with ankylosing spondylitis: a randomized controlled trial. Arthritis Rheum 2001;45:430–8.

233. Goldberg AJ, Zaidi R, Thomson C, Dore CJ, Skene SS, Cro S, et al. Total ankle replacement versus arthrodesis (TARVA): protocol for a multicentre randomised controlled trial. BMJ Open 2016;6:e012716.

234. Wallis JA, Webster KE, Levinger P, Singh PJ, Fong C, Taylor NF. A walking program for people with severe knee osteoarthritis did not reduce pain but may have benefits for cardiovascular health: a phase II randomised controlled trial. Osteoarthritis Cartilage 2017;25:1969–79.

235. Fernandes L, Roos EM, Overgaard S, Villadsen A, Sogaard R. Supervised neuromuscular exercise prior to hip and knee replacement: 12-month clinical effect and cost-utility analysis alongside a randomised controlled trial. BMC Musculoskelet Disord 2017;18:5.

236. Guitard P, Brosseau L, Wells GA, Paquet N, Paterson G, Toupin-April K, et al. The knitting community-based trial for older women with osteoarthritis of the hands: design and rationale of a randomized controlled trial. BMC Musculoskelet Disord 2018;19:56.

237. Mease PJ, Woolley JM, Singh A, Tsuji W, Dunn M, Chiou CF. Patient-reported outcomes in a randomized trial of etanercept in psoriatic arthritis. Journal Rheumatol 2010;37:1221–7.

238. Fayed N, Olivan-Blazquez B, Herrera-Mercadal P, Puebla-Guedea M, Perez-Yus MC, Andres E, et al. Changes in metabolites after treatment with memantine in fibromyalgia: a double-blind randomized controlled trial with magnetic resonance spectroscopy with a 6-month follow-up. CNS Neurosci Ther 2014;20:999–1007.

239. Bowman SJ, Everett CC, O'Dwyer JL, Emery P, Pitzalis C, Ng WF, et al. Randomized controlled trial of rituximab and cost-effectiveness analysis in treating fatigue and oral dryness in primary Sjogren's syndrome. Arthritis Rheumatol 2017;69:1440–50.

240. Strand V, Levy RA, Cervera R, Petri MA, Birch H, Freimuth WW, et al. Improvements in health-related quality of life with belimumab, a B-lymphocyte stimulator-specific inhibitor, in patients with autoantibody-positive systemic lupus erythematosus from the randomised controlled BLISS trials. Ann Rheum Dis 2014;73:838–44.

241. 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.

242. 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.

243. Van den Berg B. SF-6D population norms. Health Econ 2012;21:1508–12.

244. Ciconelli RM, Ferraz MB, Kowalski S, Pinheiro Gda R, Sato EI. Brazilian urban population norms derived from the health-related quality of life SF-6D. Qual Life Res 2015;24:2559–64.

245. Ferreira PL, Ferreira LN, Pereira LN. SF-6D Portuguese population norms. Eur J Health Econ 2015;16:235–41.

246. Shiroiwa T, Fukuda T, Ikeda S, Igarashi A, Noto S, Saito S, et al. Japanese population norms for preference-based measures: EQ-5D-3L, EQ-5D-5L, and SF-6D. Qual Life Res 2016;25:707–19.

247. Norman R, Church J, van den Berg B, Goodall S. Australian health-related quality of life population norms derived from the SF-6D. Aust N Z J Public Health 2013;37:17–23.

248. Poder TG, Gandji EW. SF6D value sets: a systematic review. Value Health 2016;19:A282.

249. 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.

250. Cheung PW, Wong CK, Cheung JP. Differential psychometric properties of EQ-5D-5L and SF-6D utility measures in patients with low back pain. Spine (Phila Pa 1976) 2019;44:E679–86.

251. Buitinga L, Braakman-Jansen LM, Taal E, Kievit W, Visser H, van Riel PL, et al. Comparative responsiveness of the EuroQol-5D and Short Form 6D to improvement in patients with rheumatoid arthritis treated with tumor necrosis factor blockers: results of the Dutch Rheumatoid Arthritis Monitoring registry. Arthritis Care Res 2012;64:826–32.

252. Gaujoux-Viala C, Rat AC, Guillemin F, Flipo RM, Fardellone P, Bourgeois P, et al. Comparison of the EQ-5D and the SF-6D utility measures in 813 patients with early arthritis: results from the ESPOIR cohort. J Rheumatol 2011;38:1576–84.

253. Sorensen J, Linde L, Ostergaard M, Hetland ML. Quality-adjusted life expectancies in patients with rheumatoid arthritis: comparison of index scores from EQ-5D, 15D, and SF-6D. Value Health 2012;15:334–9.

254. Leung YY, Png ME, Wee HL, Thumboo J. Comparison of EuroQol-5D and Short Form-6D utility scores in multiethnic Asian patients with psoriatic arthritis: a cross-sectional study. J Rheumatol 2013;40:859–65.

255. Ferreira LN, Ferreira PL, Pereira LN. Comparing the performance of the SF-6D and the EQ-5D in different patient groups. Acta Med Port 2014;27:236–45.

256. Campolina AG, López RV, Nardi EP, Ferraz MB. Internal consistency of the SF-6D as a health status index in the Brazilian urban population. Value Health Reg Issues 2018;17:74–80.

257. Goodwin PC, Ratcliffe J, Morris J, Morrissey MC. Using the knee-specific Hughston Clinic Questionnaire, EQ-5D and SF-6D following arthroscopic partial meniscectomy surgery: a comparison of psychometric properties. Qual Life Res 2011;20:1437–46.

258. 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.

259. 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.

260. 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? [research article]. Ann Rheum Dis 2007;66:771–7.

261. Hawthorne G, Richardson J. Measuring the value of program outcomes: a review of multiattribute utility measures. Expert Rev Pharmacoecon Outcomes Res 2001;1:215–28.

262. Johnsen LG, Hellum C, Nygaard OP, Storheim K, Brox JI, Rossvoll I, et al. Comparison of the SF6D, the EQ5D, and the Oswestry Disability Index in patients with chronic low back pain and degenerative disc disease. BMC Musculoskelet Disord 2013;14:148.

263. 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.

264. 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.

265. 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.

266. 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.

267. Ye Z, Sun L, Wang Q. A head-to-head comparison of EQ-5D-5 L and SF-6D in Chinese patients with low back pain. Health Qual Life Outcomes 2019;17:57.

268. Kwakkenbos L, Fransen J, Vonk MC, Becker ES, Jeurissen M, van den Hoogen FH, et al. A comparison of the measurement properties and estimation of minimal important differences of the EQ-5D and SF-6D utility measures in patients with systemic sclerosis. Clin Exp Rheumatol 2013;31 Suppl 76:50–6.

269. Sakthong P, Munpan W. A head-to-head comparison of UK SF-6D and Thai and UK EQ-5D-5L value sets in Thai patients with chronic diseases. Appl Health Econ Health Policy 2017;15:669–79.

270. 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? [research article]. Qual Life Res 2005;14:1333–44.

271. 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.

272. 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.

273. 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. Cost Eff Resour Alloc 2009;7:12.

274. Gaujoux-Viala C, Rat AC, Guillemin F, Flipo RM, Fardellone P, Bourgeois P, et al. Responsiveness of EQ-5D and SF-6D in patients with early arthritis: results from the ESPOIR cohort. Ann Rheum Dis 2012;71:1478–83.

275. Carreon LY, Berven SH, Djurasovic M, Bratcher KR, Glassman SD. The discriminative properties of the SF-6D compared with the SF-36 and ODI. Spine (Phila Pa 1976) 2013;38:60–4.

276. 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.

277. Kvamme MK, Kristiansen IS, Lie E, Kvien TK. Identification of cutpoints for acceptable health status and important improvement in patient-reported outcomes, in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis. J Rheumatol 2010;37:26–31.

278. Ratcliffe J, Thomas KJ, MacPherson H, Brazier J. A randomised controlled trial of acupuncture care for persistent low back pain: cost effectiveness analysis. BMJ 2006;333:626.

279. Van den Hout WB, Peul WC, Koes BW, Brand R, Kievit J, Thomeer RT, et al. Prolonged conservative care versus early surgery in patients with sciatica from lumbar disc herniation: cost utility analysis alongside a randomised controlled trial. BMJ 2008;336:1351–4.

280. Lamb SE, Williamson EM, Heine PJ, Adams J, Dosanjh S, Dritsaki M, et al. Exercises to improve function of the rheumatoid hand (SARAH): a randomised controlled trial. Lancet 2015;385:421–9.

281. 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.

282. Richardson JR, Peacock SJ, Hawthorne G, Iezzi A, Elsworth G, Day NA. Construction of the descriptive system for the assessment of quality of life AQoL-6D utility instrument. Health Qual Life Outcomes 2012;10:38.

283. 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.

284. Hawthorne G, Richardson J, Day N. Technical report 12: using the Assessment of Quality of Life (AQoL) version 1. Centre for Health Program Evaluation, Monash University. 2000.URL: https://www.aqol.com.au/paper s/techr eport 12.pdf.

285. 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.

286. Ackerman IN, Bucknill A, Page RS, Broughton NS, Roberts C, Cavka B, et al. The substantial personal burden experienced by younger people with hip or knee osteoarthritis. Osteoarthritis Cartilage 2015;23:1276–84.

287. Hawthorne G, Korn S, Richardson J. Population norms for the AQoL derived from the 2007 Australian National Survey of Mental Health and Wellbeing. Aust N Z J Public Health 2013;37:7–16.

288. Maxwell A, Özmen M, Iezzi A, Richardson J. Deriving population norms for the AQoL-6D and AQoL-8D multi-attribute utility instruments from web-based data. Qual Life Res 2016;25:3209–19.
289. Si HB, Zeng Y, Zhong J, Zhou ZK, Lu YR, Cheng JQ, et al. The effect of primary total knee arthroplasty on the incidence of falls and balance-related functions in patients with osteoarthritis. Sci Rep 2017;7:1–9.
290. Thammaiah S, Manchaiah V, Easwar V, Krishna R. Translation and adaptation of five English language self-report health measures to South Indian Kannada language. Audiol Res 2016;6:153.
291. 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.
292. Hawthorne G, Buchbinder R, Defina J. Functional status and health-related quality of life assessment in patients with rheumatoid arthritis. Centre for Health Program Evaluation. 2000.URL: https://www.monash.edu/__data/assets/pdf_file/0010/1882558/wp116.pdf.
293. Richardson J, Chen G, Iezzi A, Khan MA. Transformations between the assessment of quality of life AQoL instruments and test-retest reliability. Centre for Health Economics, Monash University. 2011. URL: https://www.aqol.com.au/papers/researchpaper66.pdf.
294. Allen J, Inder KJ, Lewin TJ, Attia JR, Kelly BJ. Construct validity of the Assessment of Quality of Life-6D (AQoL-6D) in community samples. Health Qual Life Outcomes 2013;11:61.
295. Richardson J, Peacock SJ, Iezzi A, Day N, Hawthorne G. Research Paper 2007 (24): construction and validation of the Assessment of Quality of Life (AQoL) mark II instrument. Centre for Health Economics, Monash University. 2007.URL: https://www.aqol.com.au/papers/researchpaper24.pdf.
296. 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.
297. 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 Econ Rev 2004;37:62–88.
298. 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.
299. 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.
300. Bennell KL, Ahamed Y, Jull G, Bryant C, Hunt MA, Forbes AB, et al. Physical therapist–delivered pain coping skills training and exercise for knee osteoarthritis: randomized controlled trial. Arthritis Care Res (Hoboken) 2016;68:590–602.
301. Bennell KL, Nelligan R, Dobson F, Rini C, Keefe F, Kasza J, et al. Effectiveness of an internet-delivered exercise and pain-coping skills training intervention for persons with chronic knee pain: a randomized trial. Ann Intern Med 2017;166:453–62.
302. Hinman RS, Wrigley TV, Metcalf BR, Campbell PK, Paterson KL, Hunter DJ, et al. Unloading shoes for self-management of knee osteoarthritis: a randomized trial. Ann Intern Med 2016;165:381–9.
303. Hill CL, March LM, Aitken D, Lester SE, Battersby R, Hynes K, et al. Fish oil in knee osteoarthritis: a randomised clinical trial of low dose versus high dose. Ann Rheum Dis 2016;75:23–9.

Table 1. Practical applications*

Measure	Number of Items	Content/Domains	Method of Administration	Recall Period	Response Format	Range of Scores	Score Interpretation	Availability of Normative Data	Cross-Cultural Validation
SF-36	36	Physical functioning, role physical, bodily pain, general health, social functioning, role emotional, mental health, and vitality	Self- or interviewer-administered; paper, electronic, or telephone format	4 weeks (standard form) or 1 week (acute form)	Likert-type with varying number of response options	0-100	Higher score = better health	Yes	Yes
SF-12	12	Physical component summary score and mental component summary score	Self- or interviewer-administered; paper, electronic, or telephone format	4 weeks (standard form) or 1 week (acute form)	Likert-type; 2 items with 3 response options and 10 items with 5 response options	0-100	Higher score = better health	Yes	Yes
Nottingham Health Profile	38 (Part 1)	Energy, pain, emotional reactions, sleep, social isolation, and physical mobility	Self- or interviewer-administered; paper format	At the moment	Yes/no	0-100	Higher score = worse health	Yes	Yes
WHOQOL-BREF	26	Physical health, psychological, social relationships, and environment	Self- or interviewer-administered; paper or electronic format	2 weeks	Five-point Likert-type	4-20 or 0-100	Higher score = better quality of life	Yes	Yes
PROMIS-GH	10	Global physical health, global mental health, and two single items on general health and social roles	Self- or interviewer-administered; paper or electronic format	No recall for seven items; 7-day recall for three items	Nine items are rated on a five-point Likert-type scale; one item (pain) is rated 0-10.	T score (population mean: 50; SD: 10)	Higher score = better health	The population mean of 50 is based on the US general population	Yes
EQ-5D	6	Mobility, self-care, usual activities, pain/discomfort, anxiety/depression, and VAS (overall health)	Self- or interviewer-administered; paper, telephone, or electronic format	Today	Version 3L (no, some, and extreme problems); version 5L (no, , moderate, severe, and extreme problems); vertical VAS	−1 to 1 (utility score); 0-100 (VAS)	Higher score = better HRQOL	Yes	Yes

(Continued)

Table 1. (*Cont'd*)

Measure	Number of Items	Content/Domains	Method of Administration	Recall Period	Response Format	Range of Scores	Score Interpretation	Availability of Normative Data	Cross-Cultural Validation
SF-6D	11	Physical function, role performance, social function, bodily pain, mental health, and vitality	Self- or interviewer-administered; paper, telephone, or electronic format	4 weeks or 1 day	Likert-type; number of response options varies across items	0.30 to 1.00	Higher score = better HRQOL	Yes	Yes
AQoL	4D = 12; 6D = 20	4D = 4; 6D = 6	Self- or interviewer-administered; paper or electronic format	The past week	Four-point Guttman scale	−0.40 to 1.00	Higher score = better HRQOL	Yes	None identified

* 3L = EQ-5D-3L; 5L = EQ-5D-5L; AQoL = Assessment of Quality of Life Scale; HRQOL = health-related quality of life; PROMIS-GH = Patient-Reported Outcomes Measurement Information System–General Health; SF-12 = 12-Item Short Form Health Survey; SF-36 = 36-Item Short Form Health Survey; VAS = visual analog scale; WHOQOL-BREF = World Health Organization Quality of Life short version instrument.

Table 2. Psychometrics*

Measure	Floor/Ceiling Effects	Reliability	Validity	Responsiveness	Minimally Important Differences	Generalizability	Used in RCTs
SF-36	Floor and ceiling effects	Good evidence of internal consistency and moderate evidence of temporal stability	Good face validity. Content validity in rheumatic conditions is not known. Good evidence of convergent and known-groups validity and weaker evidence of discriminant and factorial validity	Generally good	2-3 points	Good	Yes
SF-12	No	Good	Good	Adequate but limited evidence	Limited evidence	Good	Yes
Nottingham Health Profile	Floor effects in severely impaired populations and ceiling effects in well populations	Suboptimal for emotional reactions and social isolation domains; good for the remaining domains	Good face validity; support for the convergent and known-groups validity; limited information on factorial validity	Low to moderate	Not available	Works best in population subgroups with moderate levels of disability	Yes
WHOQOL-BREF	Minimal floor effects or ceiling effects among healthy populations	Good internal consistency for most domains except for social relationships; good test-retest reliability.	Good face validity for assessing general health; no information available on construct validity for rheumatology patients	Limited evidence indicating good responsiveness overall for musculoskeletal conditions	Not available	Good overall, but culture-specific issues have been reported for the item relating to satisfaction with sex life	Yes
PROMIS-GH	No	Good but limited evidence in rheumatology, especially internal consistency reliability	Good face validity and good evidence of construct validity except for replication of factor structure in rheumatology settings	Good	Limited evidence	Good	Yes
EQ-5D	Minimal floor and ceiling effects in utility score (less ceiling effect in the 5L version compared with the 3L) and EQ VAS; ceiling effects at the domain level	Good internal consistency and moderate test-retest reliability for the 3L and 5L versions	Adequate content and construct validity; the 5L version has better construct validity compared with the 3L version	Acceptable responsiveness (version 5L better than 3L; utility better than VAS); EQ 5D and EQ-VAS are less responsive than disease-specific measures	Version 3L utility score: 0.03-0.52 points; version 5L utility score: 0.05-0.41 points; EQ VAS: 7.75-10.5 points	Good	Yes

(Continued)

Table 2. (*Cont'd*)

Measure	Floor/Ceiling Effects	Reliability	Validity	Responsiveness	Minimally Important Differences	Generalizability	Used in RCTs
SF-6D	No	Good	Good face validity for measuring general HRQOL; no patient involvement in content development; good support for construct validity	Variable across studies; may be more sensitive at detecting improvement than deterioration	Relatively small, 0.01-0.10 utility units	Good	Yes
AQoL	No	Good	Very good but needs more data in rheumatology	Good but needs more data in rheumatology	0.06	Good	Yes

* 3L = EQ-5D-3L; 5L = EQ-5D-5L; AQoL = Assessment of Quality of Life Scale; HRQOL = health-related quality of life; PROMIS-GH = Patient-Reported Outcomes Measurement Information System–General Health; RCT = randomized controlled trial; SF-12 = 12-Item Short Form Health Survey; SF-36 = 36-Item Short Form Health Survey; VAS = visual analog scale; WHOQOL-BREF = World Health Organization Quality of Life short version instrument.