J Rehabil Med 2011; 43: 411–419

ORIGINAL REPORT

VALIDATION OF OUTCOME MEASUREMENT INSTRUMENTS USED IN A MULTIDISCIPLINARY REHABILITATION INTERVENTION FOR PATIENTS WITH CHRONIC INFLAMMATORY ARTHRITIS: LINKING TO THE INTERNATIONAL CLASSIFICATION OF FUNCTIONING, DISABILITY AND HEALTH, CONSTRUCT VALIDITY AND RESPONSIVENESS TO CHANGE

Sofia Hagel, PT, MSc[1], Elisabet Lindqvist, MD, PhD[1], Ingemar F. Petersson, MD, PhD[1,2], Jan-Åke Nilsson, BS[1] and Ann Bremander, PT, PhD[2,3]

From the [1]Department of Clinical Sciences Lund, Section of Rheumatology, Lund University and Skåne University Hospital, [2]Department of Clinical Sciences Lund, Orthopaedics, Lund University, Lund and [3]Research and Development Center, Spenshult Hospital for Rheumatic Diseases, Halmstad, Sweden

Objective: **To determine the validity of 15 standardized instruments frequently used to measure the outcome of chronic arthritis treatment.**

Methods: **Analyses were performed on data collected at a rehabilitation programme (n=216). The outcome measures evaluated were health-related quality of life, global health, pain, physical function and aerobic capacity. The instrument items were linked to the International Classification of Functioning, Disability and Health (ICF) (content validity), construct validity was analysed based on predetermined hypothesis (Spearman's correlations, r_s), and responsiveness (after 18 days and 12 months) by the standardized response mean.**

Results: **Most instruments covered the ICF component body function and/or activity-participation, only a few covered the environmental component. The short Euroqol-5 Dimensions performed as well as the longer health-related quality of life instruments in covering the ICF and in responsiveness. The health-related quality of life instruments did not measure similar constructs as hypothesized, neither did pain measures. The Bath Ankylosing Spondylitis indices covered several components of the ICF often exhibiting a large responsiveness. Aerobic capacity had the largest responsiveness of all measures.**

Conclusion: **Many instruments are not highly correlated, although at face value they appear to measure the same construct, information also applying to content validity and responsiveness. Results from this study can assist in choosing outcome measures in the clinic and in research.**

Key words: outcome measures; rehabilitation; rheumatoid arthritis; ankylosing spondylitis; quality of life.

J Rehabil Med 2011; 43: 411–419

Correspondence address: Sofia Hagel, Department of Clinical Sciences Lund, Section of Rheumatology, Lund University and Skåne University Hospital, Lund, Sweden. E-mail; sofia.hagel@med.lu.se

Submitted May 18, 2010; accepted January 19, 2011

INTRODUCTION

Chronic inflammatory arthritis affects many aspects of life: physiological, psychological and social (1–2). Complex multidisciplinary interventions, such as team care, are thus necessary for some patients with arthritis (3). Rehabilitation has proved to be effective in different settings, but evaluating the effects of treatment is challenging, especially since the intervention is usually intended to target multifaceted problems. For instance, one-dimensional evaluations may not be capable of reflecting the complex nature of the interventions (4). Furthermore, weak associations, or a lack of associations, between the interventions performed and the measures used to evaluate the outcome do not necessarily reflect a lack of effectiveness, but could simply be a reflection of using inappropriate measures that do not address the constructs of interest adequately (5).

The outcome measures employed are often based on a general consensus among researchers and practitioners, and for patients with inflammatory arthritis these often involve self-reported questionnaires to evaluate disease activity, pain, physical function, fatigue, health-related quality of life (HRQoL) and global health (6–7). Recommendations often concern the aspects that should be evaluated; rarely do they address the use of a specific instrument. As a result, treatment outcomes have been evaluated in numerous ways in different studies. Hence, when interventions are compared the efficacy regarding certain constructs has often been evaluated with several outcome measures, yet comparisons are performed as if they were equivalent.

In 2001, the World Health Organization (WHO) published the International Classification of Functioning, Disability and Health (ICF) in order to help overcome the difficulties encountered when describing the complex relations between disease, treatment and the evaluation of outcome. Linking outcome measures to the different ICF components using a specific linking process has recently become a common method to understand the concepts that are evaluated by different outcome measures (8). It is also important to ensure that the outcome measures used evaluate all or relevant components; body structure, body function, activity and participation and environmental factors (5, 8, 9). Most of the instruments currently used in clinical practise were developed prior to the ICF, and how they cover the different ICF components has not yet been investigated adequately.

Our aim was to study the validity of a set of instruments in order to determine which instruments will provide the best information for multidisciplinary rehabilitation outcome in patients with chronic arthritis. First, we studied how well a number of instruments commonly used in outcome evaluations covered the ICF components (content validity). We also assessed construct validity based on predetermined hypotheses and responsiveness to change of the chosen instruments.

METHODS

This validation study was based on data from a rehabilitation intervention and its corresponding follow-up. The procedure and outcome of the rehabilitation programme has been published previously (10). Consecutively enrolled patients with inflammatory arthritides ($n = 216$) attended an 18-day, outpatient, team rehabilitation programme in 2002–2006. The female/male ratio was 153/63 (71%/29%), mean age and disease duration at inclusion were 50 years (standard deviation (SD) 12 years) and 15 years (SD 11 years), respectively. Peripheral arthritis (PA, mainly rheumatoid arthritis) was the primary diagnosis in 149 patients, and spondylarthritides (SpA) in 67 patients. Evaluations were performed at the start of the programme, after 18 days and after 12 months. A number of patient-reported outcome (PRO) measures, as well as measures of observed physical function, were used to evaluate the results of this complex intervention.

Measures used for the evaluation of treatment outcome

Different instruments for measuring the outcome of treatment, some of which evaluated similar aspects of disease and rehabilitation outcome, were chosen in consensus at the rheumatology clinic with the intention of obtaining as complete a picture as possible of both the subjects and the treatment outcome. All the outcome measures used were standardized instruments measuring aspects included in consensus recommendations (11–13) with acceptable validity and reliability. Experienced physiotherapists and occupational therapists performed the observed tests. Patients in need of assistance with the PRO measures were aided by experienced health professionals.

Physical functioning. The self-administered Health Assessment Questionnaire (HAQ) was used to evaluate physical disability. It covers the ability to perform 20 activities, and the total score range from 0 to 3 (best to worst) (14, 15). In the SpA group, the self-administered disease-specific Bath Ankylosing Spondylitis (BAS) Indices for Disease Activity (BASDAI) and Function (BASFI) were used to obtain additional information on disease activity and functional ability. The BAS instruments consist of visual analogue scales (VAS); the BASDAI has 6 items, and the BASFI 10 items. The total score can range from 0 to 10 (best to worst) (16–19).

Health and pain. VAS were used to assess global health and pain (0–100, best to worst) (20). In addition, the Bath Indices for global health, 1 VAS for each item, measuring global health last week (BASG-1), and global health during the past 6 months (BASG-2) were used in the SpA group, based on the recommendations of the Assessment of Spondylarthritis International (ASAS) (0–10, best to worst) (21).

Health-related quality of life. Three different measures of HRQoL were administered: the Nottingham Health Profile (NHP) (22–24), the Short Form-36 Health Survey (SF-36) (25), and the Euroqol-5 Dimensions (EQ-5D) (26).

The NHP, part I, is a generic questionnaire including 38 items which cover 6 subscales: emotional reactions (9 items), energy level (3 items), pain (8 items), physical mobility (8 items), sleep (5 items) and social isolation (5 items). Each subscale and the total score range from 0 to 100 (best to worst) (22–24).

The SF-36, is a generic questionnaire including 8 dimensions of health covered by 36 items: physical functioning (PF, 10 items), physical role limitations (RP, 4 items), bodily pain (BP, 2 items), general health perceptions (GH, 6 items), vitality (VT, 4 items), social functioning (SF, 2 items), emotional role limitations (RE, 3 items), and mental health (MH, 5 items). The scores range from 0 to 100 (worst to best) (25).

In the self-reported, generic EQ-5D questionnaire 5 questions are posed, 1 each on mobility, self-care, pain, usual activities, and psychological status. The instrument presents an index value for health status (0 = death, 1 = full health) (26).

Aerobic capacity. Aerobic capacity (maximal oxygen consumption, VO_{2max}) was determined using an 8-min, sub-maximal treadmill walking test. Age, sex, self-selected walking speed (km/h), and working heart rate were used to calculate the individual's oxygen uptake, expressed as $ml \cdot kg^{-1} \cdot min^{-1}$(27, 28). Participants taking β-blockers were excluded from this test.

Grip strength. The GRIPPIT dynamometer was used to measure grip strength. In this standardized test the patient was seated in a standardized position and instructed to press the handle of the instrument for 10 s with each hand. Three values were measured (in Newtons), maximal strength, mean strength and final strength. In this study we used mean strength, calculated as left plus right hand strength divided by 2 (29, 30).

Performance of shoulder, arm and hand. The shoulder, arm and hand test was used to evaluate the performance of the upper extremities. Five different tasks were used to evaluate the range of movement of the shoulder, arm and hand, resulting in a total score ranging from 0 to 60 (worst to best) (31).

Composite score of observed function. The Signals of Functional Impairment (SOFI) index was used to evaluate the observed function of the upper (8 items) and lower (4 items) limbs in the PA group only. The total score ranged from 0 to 48 (best to worst) (32).

Composite score of observed axial status. The range of spinal movement was evaluated, in the SpA group only, with the Bath Ankylosing Spondylitis Metrology Index (BASMI). Five clinical measures in the cervical and lumbar area provide a total score from 0 to 10 (best to worst) (33).

Analysis

The analyses were performed in 3 steps. First, the 15 outcome measures were linked to the ICF. The linking was performed to validate the instruments' ability to cover the different ICF components (content validity). Secondly, construct validity was assessed based on hypotheses of convergent and divergent validity. Finally, the instruments responsiveness to change was calculated.

Linking the outcome measures to the ICF. The linking process was done by identifying each item in the Swedish version of all instruments included. The meaningful concepts of the question, including the response options and examples given, were identified according to previously published linking rules (8). Each meaningful concept was linked to the most precise third-level ICF category. The representation of the categories was then linked to the ICF component(s): body function, body structure, activity and participation and environmental factors (5, 9). The outcome measures coverage of the ICF components was analysed on a total score or on a subscale level, depending on how the instrument was constructed.

The ICF model was developed to describe and capture aspects of health, and thus questions aimed only at assessing "health" must be linked to the encompassing term instead of to the concepts of body structure/function, activity and participation and environmental factors (5, 8, 34).

The meaningful concepts were identified by one of the authors, SH, who also performed the linking of the scales to the ICF. In the second step, AB critically reviewed the proposed linking, and after discussions

and consensus between SH and AB the linking was presented to EL and IFP, who reviewed it thoroughly, disagreements were discussed and thereafter SH and AB finally reached consensus.

Construct validity of the outcome measures. To determine the relationship among the various instruments of physical function, HRQoL and pain of the disease, construct validity was analysed on baseline values and we focused on convergent vs divergent validity according to predefined hypotheses (35):

- We hypothesized that outcome measures constructed to measure patient-reported pain (VAS pain, NHP pain, SF-36 pain, BASDAI) would be highly related, $r_s \geq 0.8$ (convergent validity).
- We further hypothesized a relationship of $r_s \geq 0.8$ (convergent validity) in measures of global health (BASG1, BASG-2, VAS global and SF-36 GH).
- Subscales of HRQoL measures describing similar constructs, such as energy levels (NHP energy, NHP sleep and SF-36 VT), mental conditions (NHP emotion, SF-36 MH), social aspects (NHP social, SF-36 SF), and physical aspects (SF-36 PF, NHP physical) were expected to be related, meeting requirements of convergent validity ($r_s \geq 0.8$).
- Summary scales of the HRQoL measures NHP and EQ-5D were expected to be highly related $r_s \geq 0.8$ (convergent validity).
- Measures of patient-reported physical functioning as captured by the HAQ and the BASFI, were expected to be related, $r_s \geq 0.8$ (convergent validity).
- Observed outcome instruments measuring hand and arm functioning: grip strength, the SOFI and the Shoulder-arm-hand test were expected to show a convergent validity of $r_s \geq 0.8$.
- Aerobic capacity and BASMI were expected to have low relationships, $r_s \leq 0.2$ (divergent validity) with all other outcome instruments measuring observed physical function.

Responsiveness. We wanted to compare the magnitude of change after the intervention. After completion of the intervention and the subsequent 12 month follow-up, a non-parametric standardized response mean (SRM_{np}) was calculated for each instrument or its subscales (36).

Statistics

Data analyses were performed on all patients as well as for the PA and the SpA groups separately. As results were similar in both groups the results from the total group of 216 patients are represented herein. Where differences occurred between the two groups, complementary subgroup data have been provided. Statistical analyses were performed with SPSS. Non-parametric statistics were used for analyses and to

Table I. *Description of the linking to International Classification of Functioning, Disability and Health (ICF) of outcome instruments included to evaluate function, pain and global health*

ICF category	Aerobic capacity	BAS-DAI	BAS-FI	BAS-MI	BASG-1 BASG-2	HAQ	Grip strength	SOFI	Shoulder-arm-hand function	VAS pain	VAS global
b126 Temp and Personal											
b130 Energy		×									
b134 Sleep											
b152 Emotional reactions											
b270 Sensory functions		×									
b280 Pain		×						×		×	
b455 Exercise tolerance	×										
b710 Mobility of joint functions		×	×	×			×	×	×		
b730 Muscle power functions							×	×			
d410 Changing basic body position			×			×		×			
d415 Maintaining a body position			×								
d430 Lifting and carrying			×			×					
d440 Fine hand use			×			×					
d445 Hand and arm use			×			×					
d450 Walking			×			×					
d455 Moving around			×			×					
d460 Moving around in different locations											
d510 Washing oneself						×					
d520 Caring for body parts						×					
d540 Dressing			×			×					
d550 Eating						×					
d620 Acquisition of goods and service						×					
d630 Preparing meals						×					
d640 Doing housework			×			×					
d650 Caring for household objects			×								
d720 Complex interpersonal interactions											
d750 Informal social relationships											
d760 Family relationships											
d820 School education											
d850 Remunerative employment			×								
d920 Recreation and leisure			×								
e115 Products and technology for personal use in daily living			×			×					
HEALTH					×						×

BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; BASFI: Bath Ankylosing Spondylitis Functional Index; BASMI: Bath Ankylosing Spondylitis Metrology Index; BASG-1: Bath Ankylosing Spondylitis Global health last week; BASG-2: Bath Ankylosing Spondylitis Global health the past 6 months; HAQ: Health Assessment Questionnaire; SOFI: Signals of Functional Impairment; VAS: visual analogue scale.

determine changes over time, since the distribution was skewed. Construct validity was analysed by Spearman's correlations (r_s). Baseline values were regarded as fulfilling criteria for divergent validity when the correlation coefficient was ≤0.2 and for convergent validity when the correlation coefficient was ≥0.8. Responsiveness was analysed using change values (Δ), (between the start of the intervention, after 18 days, and after 12 months) expressed as the median (md) and inter quartile range (IQR). The non-parametric SRM_{np} was calculated as the median change in score divided by the interquartile range of change in scores, to account for the fact that the data were skewed. The magnitude of change due to intervention (responsiveness) was classified as small (0–0.2), moderate (0.3–0.5) or large (>0.5) (37).

Ethics

Approval was obtained from the Regional Ethical Review Board, (No. 405/2008).

RESULTS

ICF components

The 15 outcome measures investigated comprised 14 subscales, rendering a total of 27 measurement scales. Using the ICF linking rules, we found that all outcome measures included at least one ICF component. Body function was the most well-represented ICF component; 19 out of 27 outcome measures or subscales included items that covered this component, followed by the component of activity and participation (11 outcome measures/subscales). Environmental factors were covered by 4 outcome measures/subscales investigated in this study (Tables I–II).

The overall construct of health was covered by the VAS global, BASG-1 and BASG-2, and also by a single item in EQ-5D and in 4/8 subscales of the SF-36 (GH, PF, RP and SF) (Tables I and II).

Measures of pain

All measures of patient-reported pain (VAS pain, NHP pain, SF-36 pain, BASDAI) were linked to the ICF component body function. The NHP pain and the SF-36 pain also represented activity and participation covering two components of the ICF (Table I–II). In the BASDAI, 3 out of 6 questions include pain estimated on a VAS and correlation to a single measure of VAS pain was r_s 0.8, indicating a large relationship between these two outcome measures (convergent validity) (Table III). BASDAI showed a larger SRM_{np} after 18 days than did the VAS (0.8 vs 0.5), while both measures had values of SRM_{np} close to zero 12 months later (SRM_{np} 0.1 vs 0.2) demonstrating the BASDAI to be superior to a single VAS pain measure in short-term evaluation of outcome in patients with SpA (Table IV). No other measures of pain showed a convergent validity according to our predefined hypotheses (Table III). The SF-36 BP (SRM_{np} 0.5) and the NHP pain (SRM_{np} 0.4) showed more consistent responsiveness than did the VAS pain (Table IV).

Measures of global health

Outcome measures of global health (VAS global, SF-36 GH, and, in the SpA group also BASG-1 and BASG-2) were linked to the ICF overall construct health. These instruments did not show convergent validity, contrary to our hypothesis (r_s 0.5–0.7) (Table III). Concerning the magnitude of change due to intervention, VAS global, SF-36 GH and BASG-1 had similar responsiveness after 18 days (SRM_{np} 0.5–0.7), and after 12 months (SRM_{np} of 0.1–0.4), with SF-36 GH being the most consistent measure (Table IV). The results of the two groups of patients diverged regarding responsiveness of the SF-36 GH, where it was larger and more consistent in the SpA group compared with the PA group (SRM_{np} of 0.6 and 0.6 vs 0.4 and 0.2, respectively).

Measures of energy levels, mental, social and physical aspects of health-related quality of life

The HRQoL subscales describing energy levels (NHP energy, SF-36 VT and NHP sleep), mental aspects (NHP emotion, SF-36 MH), and social aspects (NHP social, SF-36 SF) were all linked to the ICF component body function, both NHP social and SF-36 SF also covered activity and participation, with the SF-36 SF subscale also covering the overall construct health. Subscales describing physical aspects (SF-36 PF, NHP physical) were linked to the activity and participation component, whereas the SF-36 PF also covered the health construct (Table II). However, neither of these subscales met the criteria for convergent validity hypothesized *a priori* (Table III).

The SF-36 VT was found to be the most responsive subscale measuring energy levels, with an SRM_{np} of 0.7 at 18 days and SRM_{np} 0.2 after 12 months. The SF-36 MH was the most responsive subscale measuring change due to intervention (SRM_{np} 0.7 and 0.2) while subscales measuring social aspects (NHP social, SF-36 SF) had an SRM_{np} of 0 at all points of evaluation (Table IV). SF-36 PF had a larger responsiveness after 18 days (SRM_{np} 0.5) compared with NHP physical (SRM_{np} 0.3), but after 12 months both subscales had an SRM_{np} of 0.

Total scores of Euroqol-5 Dimensions and Nottingham Health Profile

Both the EQ-5D and the NHP questionnaires provide total scores. Linking the EQ-5D to the ICF it captured body function, activity and participation, environmental aspects and health. The NHP total score does not cover health, but otherwise covers the same aspects as EQ-5D. The total scores of the EQ-5D and the NHP showed moderate correlation (r_s 0.6) not high enough to fulfil our *a priori* hypotheses (Table III). These two measures of HRQoL outcome were comparable in responsiveness over time (NHP SRM_{np} 0.6 and 0.3, and EQ-5D SRM_{np} 0.4 and 0.2) (Table IV).

Measures of patient-reported physical function

Patient-reported physical function, as measured by the HAQ and the BASFI, showed similar linking to the ICF components activity and participation and environmental factors, but the BASFI could also be linked to the component body function. The two questionnaires had a correlation coefficient of r_s 0.8, implying measures of related constructs (convergent validity) (Table III). The BASFI was superior to the HAQ in reflecting responsiveness, with a SRM_{np} of 0.7 and 0.6 (18 days and 12 months later) vs 0 at both time points for the HAQ (Table IV).

Table II. *Description of the linking to the International Classification of Functioning, Disability and Health (ICF) of outcome instruments included to evaluate health-related quality of life*

	Outcome instrument														
ICF category	EQ5-D	NHP emo	NHP energ	NHP pain	NHP phys	NHP sleep	NHP soc	SF-36 BP	SF-36 MH	SF-36 PF	SF-36 RE	SF-36 RP	SF-36 SF	SF-36 VT	SF-36 GH
b126 Temp and Personal		×					×								
b130 Energy		×	×											×	
b134 Sleep		×				×									
b152 Emotional reactions	×	×					×		×		×		×		
b270 Sensory functions															
b280 Pain	×			×				×							
b455 Exercise tolerance															
b710 Mobility of joint functions															
b730 Muscle power functions															
d410 Changing basic body position				×	×					×					
d415 Maintaining a body position					×					×					
d430 Lifting and carrying										×					
d440 Fine hand use															
d445 Hand and arm use					×										
d450 Walking	×			×	×					×					
d455 Moving around				×	×					×					
d460 Moving around in different locations					×										
d510 Washing oneself	×									×					
d520 Caring for body parts															
d540 Dressing	×				×					×					
d550 Eating															
d620 Acquisition of goods and service															
d630 Preparing meals															
d640 Doing housework	×							×							
d720 Complex interpersonal interact													×		
d750 Informal social relationships													×		
d760 Family relationships	×														
d820 School education	×														
d850 Remunerative employment	×							×			×	×			
d920 Recreation and leisure	×									×	×	×	×		
e115 Products and technology for personal use in daily living	×				×										
HEALTH	×									×		×	×	×	×

EQ-5D: Euroqol-5 Dimensions; NHP: Nottingham Health Profile; SF-36: Short-Form-36 Health Survey; BP: bodily pain; MH: mental helath; PF: physical functioning; RE: emotional role limitations; RP: physical role limitations; SF: social functioning; VT: vitality; GH: general health perceptions.

When analysing the responsiveness of the HAQ we found a subgroup difference where the SRM_{np} was 0.2 after 12 months in the SpA group vs 0 in the PA group.

Measures of hand and arm functioning

Measures of hand and arm functioning (grip strength, the SOFI and the Shoulder-arm-hand test) were linked to the ICF component body function. Contrary to our hypothesis construct validity among these measures was not shown (Table V). The SOFI had the largest responsiveness, SRM_{np} 0.7 and 0.3, while grip strength and shoulder-arm-hand function had lower responsiveness, SRM_{np} 0.2–0.4 (Table IV).

Aerobic capacity and Bath Ankylosing Spondylitis Metrology Index

Both aerobic capacity and the BASMI were linked to the body function component. In accordance with our hypothesis, both instruments showed a divergent validity to all other observed physical outcome measures, with correlations of $r_s \leq 0.2$ (Table V). Both instruments had large responsiveness at all time points (aerobic capacity SRM_{np} 1.1 and 1.2 and BASMI SRM_{np} 0.8 and 0.5) (Table IV).

DISCUSSION

In this methodological study we found that outcome instruments commonly used in rehabilitation practice and research covered the ICF components body function, activity and participation, whereas the environmental component was covered to a lesser extent. In the clinic as well as in research, knowledge of what ICF components the different outcome measures cover can be helpful in order to choose the right outcome measure for a specific intervention. Our findings also showed that a short questionnaire with 5 items, such as the EQ-5D, may cover more ICF components than a more extensive measure. Aerobic capacity and the BAS indices were highly responsive measures over time and can be recommended when applicable to the intervention performed.

The choice of HRQoL outcome measure to use depends on the context, and several aspects will have to be considered. If a multidimensional instrument is appropriate the subscales of the SF-36 may be preferable to

Table III. *Baseline correlations of the patient-reported outcome measures according to predetermined hypothesis of construct validity, convergent validity* $r_s \geq 0.8$ *and divergent validity* $r_s \leq 0.2$

Outcome instrument	VAS pain	BAS-DAI	NHP pain	SF-36 BP	BASG-1	BASG-2	VAS global	SF-36 GH	NHP energy	NHP sleep	SF-36 VT	NHP emo	SF-36 MH	NHP social	SF-36 SF	NHP phys	SF-36 PF	NHP tot	EQ-5D	HAQ	BAS-FI
VAS pain		0.8	0.6	0.6																	
BASDAI	0.8		0.6	0.6																	
NHP pain	0.6	0.6		0.6																	
SF-36 BP	0.6	0.6	0.6																		
BAS-G1						0.7	0.7	0.5													
BAS-G2					0.7		0.6	0.6													
VAS global					0.7	0.6		0.5													
SF-36 GH					0.5	0.6	0.5														
NHP energy										0.2	0.5										
NHP sleep									0.2		0.1										
SF-36 VT									0.5	0.1											
NHP emo													0.7								
SF-36 MH												0.7									
NHP soc															0.3						
SF-36 SF														0.3							
NHP phys																	0.3				
SF-36 PF																0.3					
NHP tot																			0.6		
EQ-5D																		0.6			
HAQ																					0.8
BASFI																				0.8	

All correlations were statistically significant $p \leq 0.002$, except for SF-36 GH and BASG-1 $p = 0.007$.

VAS: visual analogue scale; BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; NHP: Nottingham Health Profile; SF-36: Short-Form-36 Health Survey; BP: bodily pain; BASG-1: Bath Ankylosing Spondylitis Global health last week; BASG-2: Bath Ankylosing Spondylitis Global health the past 6 months; GH: general health perceptions; VT: vitality; MH: mental health; SF: social functioning; PF: physical functioning; EQ-5D: Euroqol-5 Dimensions; HAQ: Health Assessment Questionnaire; BASFI: Bath Ankylosing Spondylitis Functional Index.

Table IV. *Values of instruments used described as median and interquartile range (IQR) at baseline, median change (Δ) compared with baseline, IQR of change and non-parametric standardised response mean (SRMnp) calculated as median change in score divided by the interquartile range of change in scores*

		Baseline		18 days			12 months		
Outcome instrument	Worst–Best	Median	IQR	Δ	IQR	SRMnp	Δ	IQR	SRMnp
Aerobic capacity, *n*=176	0–100	25.2	7.9	5.0	4.8	1.1	5.4	4.6	1.2
Shoulder-arm-hand function, *n*=211	0–60	56	9	1	3	0.2	1	3	0.3
Grip strength, *n*=187	0–300	118.5	142	13.5	38	0.4	10	41	0.2
SOFI, *n*=141	48–0	12	12	–2	3	–0.7	–1	3.5	–0.3
BASMI, *n*=65	10–0	3.4	2.4	–0.6	0.8	–0.8	–0.4	0.8	–0.5
VAS pain, *n*=210	100–0	48	38	–13	26	–0.5	–6	30	–0.2
NHP pain, *n*=215	100–0	55	55	–11	29	–0.4	–11	26	–0.4
SF-36 BP, *n*=97	0–100	41	29	10	21	0.5	9	19	0.5
BASDAI, *n*=65	10–0	4.4	3.3	–1.4	1.7	–0.8	–0.3	3	–0.1
NHP emotion, *n*=215	100–0	19	38	–9	27	–0.3	–9	–23	–0.4
SF-36 MH, *n*=98	0–100	72	32	12	18	0.7	4	20	0.2
NHP social, *n*=214	100–0	0	24	0	0	0	0	0	0
SF-36 SF, *n*=98	0–100	62	50	0	25	0	0	31	0
NHP sleep, *n*=215	100–0	33	42	0	–22	0	0	31	0
NHP energy, *n*=215	100–0	61	100	–24	61	–0.4	0	–39	0
SF-36 VT, *n*=98	0–100	40	31	20	30	0.7	5	30	0.2
HAQ, *n*=214	3–0	0.9	0.8	0	0.2	0	0	0.4	0
NHP physical, *n*=215	100–0	28	32	–4	15	–0.3	–0.2	14	0
SF-36 PF, *n*=97	0–100	50	32	7	15	0.5	0	19	0
SF-36 RP, *n*=96	0–100	25	50	0	50	0	0	50	0
BASFI, *n*=65	10–0	3.7	2.9	–1.0	1.4	–0.7	–0.8	1.3	–0.6
SF-36 RE, *n*=96	0–100	67	100	0	33	0	0	33	0
NHP total, *n*=215	100–0	36	26	–11	19	–0.6	–7	19	–0.3
EQ-5D, *n*=211	0–1	0.6	0.2	0.07	0.2	0.4	0.04	0.17	0.2
VAS global, *n*=210	100–0	52	36	–15	32	–0.5	–6	30	–0.2
SF-36 GH, *n*=98	0–100	35	28	10	20	0.5	10	28	0.4
BASG-1, *n*=65	10–0	3.8	4.9	–1.5	2.2	–0.7	–0.5	3.5	–0.1
BASG-2, *n*=65	10–0	5.2	3.9	–0.1	2.4	–0.04	–1.2	2.5	–0.5

SOFI: Signals of Functional Impairment; BASMI: Bath Ankylosing Spondylitis Metrology Index; VAS: visual analogue scale; NHP: Nottingham Health Profile; SF-36: Short-Form-36 Health Survey; BP: bodily pain; BASDAI: Bath Ankylosing Spondylitis Disease Activity Index; MH: mental health; SF: social functioning; VT: vitality; HAQ: Health Assessment Questionnaire; PF: physical functioning; RP: physical role limitations; BASFI: Bath Ankylosing Spondylitis Functional Index; RE: emotional role limitations; EQ-5D: Euroqol-5 Dimensions; GH: general helath perceptions; BASG-1: Bath Ankylosing Spondylitis Global health last week; BASG-2: Bath Ankylosing Spondylitis Global health the past 6 months.

the NHP subscales; however, the EQ-5D had several advantages to both the NHP and the SF-36 in our study. For instance, the 5 items covered body function, activity and participation, environmental factors and health components of the ICF, whereas the NHP and the SF-36 covered only 3 out of these 4 components. Furthermore, the EQ-5D showed a stable responsiveness over time at levels comparable to the subscales of the NHP and the SF-36. The largest responsiveness over time was seen in the NHP total score, but when analysing the subscales, several of them showed no responsiveness at all. Furthermore, the EQ-5D is short and takes only a few minutes to complete.

Most of the outcome instruments studied showed moderate correlations. This implies that outcome measures, often used clinically and assumed to evaluate similar aspects of the disease and rehabilitation outcome, only partially reflect and measure the same construct. Although similar findings have been reported previously (38, 39), pain and HRQoL are often evaluated using different instruments. A standard core set of outcome instruments would make comparisons among studies more feasible and could prevent future comparisons of "apples and oranges".

The BAS indices covered many aspects of health according to the ICF classification and showed a comparably large responsiveness even after 12 months, except for the BASDAI and BAS-G1. According to our findings it is redundant to administer both BASDAI and VAS pain when measuring pain in SpA patients. The same finding applies for the BASFI and the HAQ, which were highly related and should not be administered together. The BASFI is preferred due to greater and more consistent responsiveness, it also covers an additional ICF component compared with the HAQ. The preference for the BAS indices compared with the HAQ in our study support previous knowledge of the accuracy of the BAS (40) and the low responsiveness of the HAQ, which has been described earlier (39, 41). BASMI proved to be an important measure for patients with SpA, since it had low correlations with upper extremity range of motion and strength and also had a large responsiveness up to 12 months after the intervention.

Our findings also showed that aerobic capacity is an important measure of physical function, when applicable. It provided an aspect of body function that was not detected by any other of the outcome measures in this component. In addition, it had considerable responsiveness. Given current knowledge of cardiovascular co-morbidity and subsequent recommendations about physical activity, it is important to

Table V. *Baseline correlations of the observed patient reported outcome measures according to predetermined hypothesis of construct validity, convergent validity* $r_s \geq 0.8$ *and divergent validity* $r_s \leq 0.2$

Outcome instrument	Grip strength	SOFI	Shoulder-arm-hand	Aerobic capacity	BASMI
Grip strength		0.3	0.4	0.5	0.1
SOFI	0.3		0.7	0.1	
Shoulder-arm-hand	0.4	0.7		0.2	0.3
Aerobic capacity	0.5	0.1	0.2		0.001
BASMI	0.1		0.3	0.001	

All correlations were statistically significant $p \leq 0.001$, except for BASMI and Shoulder-arm-hand ($p = 0.005$), BASMI and Aerobic capacity ($p = 1.0$), and BASMI and Grip strength ($p = 0.3$).
SOFI: Signals of Functional Impairment; BASMI: Bath Ankylosing Spondylitis Metrology Index.

ensure that the measures used capture what they are intended to capture.

Our results on observed functioning correspond largely to the findings of Adams et al. (42). In their study, instruments with interval scales were found to be more responsive to change than those with ordinal scales. Aerobic capacity and the BAS indices should be regarded as interval scales. The low to moderate responsiveness over time of the HAQ and the HRQoL measures included in this study (the NHP, the SF-36 and the EQ5D) is also in accordance with the findings of Adams et al. (42).

Over the past decades, the development of methods of evaluation has been extended to encompass the consequences of the disease that are relevant to the patient, as well as to the healthcare system and society (43). In ongoing work on the ICF structure, environmental factors have been stressed as being of great importance, as are personal factors (5, 44). In accordance with others, we found that environmental factors were only superficially targeted in the outcome measures included in this study (38, 45). Instruments targeting environmental aspects are under development (44). Some of the instruments studied have previously been linked to the ICF (5, 46, 47) and repeating the linking in every culturally adapted version could be considered time-consuming and unnecessary. Nevertheless, discrepancies between the English and Swedish versions of the outcome measures emerged during the linking process, both in the HAQ and the SF-36. This indicates that national differences are to be expected on the national versions of the outcome instruments; hence, one particular ICF linking process is not valid in all countries (5). Linking to the ICF is, in the end, based on a subjective decision, which might explain the differences between published results.

Because of the skewed distribution of the data, the new SRM_{np} was used to analyse responsiveness. Contrary to the definition of the parametric SRM (mean change/SD of change), the SRM_{np} was described as the median change/the interquartile range of change. The SRM_{np} is thus, by definition, a more robust measure of responsiveness than the original SRM for non-normally distributed data. The SRM_{np} can be expected to produce smaller estimates, since the IQR is usually wider than the SD in most distributions.

The selection of the included outcome measures was primarily based on clinical reality, thus other important or useful outcome measures might have been omitted. Furthermore, the total number of included patients generating data for this study differed among the analysed instruments, which may affect the magnitude of change due to intervention and is a possible limitation to the study. The smallest number of participants occurred in the BAS indices ($n = 69$); however, these instruments performed better than instruments with a larger number of included patients.

In conclusion, in order to compare results across different intervention studies, the same instruments need to be used, as many of the instruments are not highly correlated, even though at face value they appear to measure the same construct.

REFERENCES

1. Boonen A. A review of work-participation, cost-of-illness and cost-effectiveness studies in ankylosing spondylitis. Nat Clin Pract Rheumatol 2006; 2: 546–553.
2. Lillegraven S, Kvien TK. Measuring disability and quality of life in established rheumatoid arthritis. Best Pract Res Clin Rheumatol 2007; 21: 827–840.
3. Vliet Vlieland TP, Pattison D. Non-drug therapies in early rheumatoid arthritis. Best Pract Res Clin Rheumatol 2009; 23: 103–116.
4. Cieza A, Stucki G. Understanding functioning, disability, and health in rheumatoid arthritis: the basis for rehabilitation care. Curr Opin Rheumatol 2005; 17: 183–189.
5. Cieza A, Geyh S, Chatterji S, Kostanjsek N, Ustun B, Stucki G. ICF linking rules: an update based on lessons learned. J Rehabil Med 2005; 37: 212–218.
6. Strand V, Gladman D, Isenberg D, Petri M, Smolen J, Tugwell P. Endpoints: consensus recommendations from OMERACT IV. Outcome Measures in Rheumatology. Lupus 2000; 9: 322–327.
7. van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. Assessments in ankylosing spondylitis. J Rheumatol 1999; 26: 951–954.
8. Cieza A, Brockow T, Ewert T, Amman E, Kollerits B, Chatterji S, et al. Linking health-status measurements to the international classification of functioning, disability and health. J Rehabil Med 2002; 34: 205–210.
9. Weigl M, Cieza A, Andersen C, Kollerits B, Amann E, Stucki G. Identification of relevant ICF categories in patients with chronic health conditions: a Delphi exercise. J Rehabil Med 2004; Suppl 44: 12–21.
10. Hagel S, Lindqvist E, Bremander A, Petersson IF. Team-based rehabilitation improves long-term aerobic capacity and health-related quality of life in patients with chronic inflammatory arthritis. Disabil Rehabil 2010; 32: 1686–1696.
11. Boonen A, Braun J, van der Horst Bruinsma IE, Huang F, Maksymowych W, Kostanjsek N, et al. ASAS/WHO ICF Core Sets for ankylosing spondylitis (AS): how to classify the impact of AS on functioning and health. Ann Rheum Dis 2010; 69: 102–107.
12. Tugwell P, Boers M, Strand V, Simon LS, Brooks P, Tugwell P, et al. OMERACT 9 – 9th International Consensus Conference on outcome measures in rheumatology clinical trials. J Rheumatol 2009; 36: 1765–1768.
13. Zochling J. Assessment and treatment of ankylosing spondylitis: current status and future directions. Curr Opin Rheumatol 2008; 20: 398–403.
14. Ekdahl C, Eberhardt K, Andersson SI, Svensson B. Assessing disability in patients with rheumatoid arthritis. Use of a Swedish version of the Stanford Health Assessment Questionnaire. Scand

J Rheumatol 1988; 17: 263–271.
15. Fries JF, Spitz P, Kraines RG, Holman HR. Measurement of patient outcome in arthritis. Arthritis Rheum 1980; 23: 137–145.
16. Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994; 21: 2286–2291.
17. Waldner A, Cronstedt H, Stenstrom CH. The Swedish version of the Bath ankylosing spondylitis disease activity index. Reliability and validity. Scand J Rheumatol Suppl 1999; 111: 10–16.
18. Calin A, Garrett S, Whitelock H, Kennedy LG, O'Hea J, Mallorie P, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994; 21: 2281–2285.
19. Cronstedt H, Waldner A, Stenstrom CH. The Swedish version of the Bath ankylosing spondylitis functional index. Reliability and validity. Scand J Rheumatol Suppl 1999; 111: 1–9.
20. Joos E, Peretz A, Beguin S, Famaey JP. Reliability and reproducibility of visual analogue scale and numeric rating scale for therapeutic evaluation of pain in rheumatic patients. J Rheumatol 1991; 18: 1269–1270.
21. Jones SD, Calin A, Steiner A. An update on the Bath Ankylosing Spondylitis Disease Activity and Functional Indices (BASDAI, BASFI): excellent Cronbach's alpha scores. J Rheumatol 1996; 23: 407.
22. 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.
23. Wiklund I, Dimenas E. [The Swedish version of the Nottingham Health Profile. A questionnaire for the measurement of health-related quality of life]. Lakartidningen 1990; 87: 1575–1576 (in Swedish).
24. 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: 159–163.
25. Ware JE Jr., Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992; 30: 473–483.
26. 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: 551–559.
27. Ebbeling CB, Ward A, Puleo EM, Widrick J, Rippe JM. Development of a single-stage submaximal treadmill walking test. Med Sci Sports Exerc 1991; 23: 966–973.
28. Minor MA, Johnson JC. Reliability and validity of a submaximal treadmill test to estimate aerobic capacity in women with rheumatic disease. J Rheumatol 1996; 23: 1517–1523.
29. Nordenskiold UM, Grimby G. Grip force in patients with rheumatoid arthritis and fibromyalgia and in healthy subjects. A study with the Grippit instrument. Scand J Rheumatol 1993; 22: 14–19.
30. Lagerstrom C, Nordgren B. On the reliability and usefulness of methods for grip strength measurement. Scand J Rehabil Med 1998; 30: 113–119.
31. Bostrom C, Harms-Ringdahl K, Nordemar R. Clinical reliability of shoulder function assessment in patients with rheumatoid arthritis. Scand J Rheumatol 1991; 20: 36–48.
32. Eberhardt KB, Svensson B, Mortiz U. Functional assessment of early rheumatoid arthritis. Br J Rheumatol 1988; 27: 364–371.
33. Jones SD, Porter J, Garrett SL, Kennedy LG, Whitelock H, Calin A. A new scoring system for the Bath Ankylosing Spondylitis Metrology Index (BASMI). J Rheumatol 1995; 22: 1609.
34. Prodinger B, Cieza A, Williams DA, Mease P, Boonen A, Kerschan-Schindl K, et al. Measuring health in patients with fibromyalgia: content comparison of questionnaires based on the International Classification of Functioning, Disability and Health. Arthritis Rheum 2008; 59: 650–658.
35. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007; 60: 34–42.
36. Mokkink LB, Terwee CB, Knol DL, Stratford PW, Alonso J, Patrick DL, et al. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BMC Med Res Methodol 2010; 10: 22.
37. Cohen J, editor. Statistical power analysis for the behavioral sciences. 2:nd edn. Lawrence Erlbaum Associates: New Jersey; 1988.
38. Cieza A, Stucki G. Content comparison of health-related quality of life (HRQOL) instruments based on the international classification of functioning, disability and health (ICF). Qual Life Res 2005; 14: 1225–1237.
39. Linde L, Sorensen J, Ostergaard M, Horslev-Petersen K, Hetland ML. Health-related quality of life: validity, reliability, and responsiveness of SF-36, 15D, EQ-5D [corrected] RAQoL, and HAQ in patients with rheumatoid arthritis. J Rheumatol 2008; 35: 1528–1537.
40. Landewe R, Dougados M, Mielants H, van der Tempel H, van der Heijde D. Physical function in ankylosing spondylitis is independently determined by both disease activity and radiographic damage of the spine. Ann Rheum Dis 2009; 68: 863–867.
41. Veehof MM, ten Klooster PM, Taal E, van Riel PL, van de Laar MA. Comparison of internal and external responsiveness of the generic Medical Outcome Study Short Form-36 (SF-36) with disease-specific measures in rheumatoid arthritis. J Rheumatol 2008; 35: 610–617.
42. Adams J, Mullee M, Burridge J, Hammond A, Cooper C. Responsiveness of self-report and therapist-rated upper extremity structural impairment and functional outcome measures in early rheumatoid arthritis. Arthritis Care Res (Hoboken) 2010; 62: 274–278.
43. Kirwan JR, Newman S, Tugwell PS, Wells GA, Hewlett S, Idzera L, et al. Progress on incorporating the patient perspective in outcome assessment in rheumatology and the emergence of life impact measures at OMERACT 9. J Rheumatol 2009; 36: 2071–2076.
44. Boonen A, Stucki G, Maksymowych W, Rat AC, Escorpizo R, Boers M. The OMERACT-ICF Reference Group: integrating the ICF into the OMERACT process: opportunities and challenges. J Rheumatol 2009; 36: 2057–2060.
45. World Health Organization (WHO). International Classification of Functioning, Disability and Health. [cited 2001 March 9]. Geneva: WHO. Available from: www.who.int/icidh.
46. Cieza A, Stucki G, Weigl M, Disler P, Jackel W, van der Linden S, et al. ICF Core Sets for low back pain. J Rehabil Med 2004 Suppl 44: 69–74.
47. Hakkinen A, Arkela-Kautiainen M, Sokka T, Hannonen P, Kautiainen H. Self-report functioning according to the ICF model in elderly patients with rheumatoid arthritis and in population controls using the multidimensional health assessment questionnaire. J Rheumatol 2009; 36: 246–253.