Measuring functional disability in early rheumatoid arthritis: the validity, reliability and responsiveness of the Recent-Onset Arthritis Disability (ROAD) index

F. Salaffi[1], A. Stancati[1], R. Neri[2], W. Grassi[1], S. Bombardieri[2]

[1]*Department of Rheumatology, Università Politecnica delle Marche, Jesi (AN), Italy,* [2]*Department of Internal Medicine, Rheumatology Unit, Università di Pisa, Italy*

Please address correspondence to: F. Salaffi, MD, PhD, Department of Rheumatology, Università Politecnica delle Marche, Ospedale A. Murri, ASL5, Via dei Colli, 52, 60035 Jesi (ANCONA), Italy. E-mail: fsalaff@tin.it

Clin Exp Rheumatol 2005; 23 (Suppl. 39): S31-S42.

Key words: Recent-Onset Disability Questionnaire, early rheumatoid arthritis, functional status, disability questionnaire, outcome.

ABSTRACT

Objective. *Disability has been identified as a core outcome measure in rheumatoid arthritis (RA). The aim of this study was to test the Recent-Onset Arthritis Disability (ROAD) questionnaire for validity, reliability and responsiveness in Italian patients with early RA.*

Methods. *The psychometric properties of ROAD were tested in 159 patients with early RA, mean age 54.7 (±8.8), 74.3% women, mean disease duration 14.5 months (±1.9 months). All completed the ROAD, the Medical Outcomes Study SF-36 Health Survey (SF-36), the Health Assessment Questionnaire (HAQ) and the patient global assessment (PGA) of functional disability twice, in order to test for validity and responsiveness. Of the 159 patients who completed the health status instruments on two occasions, 121 were included in the responsiveness analyses. The test–retest reliability of the ROAD questionnaire was calculated using intraclass correlation coefficients (ICCs) and the Bland and Altman method on 77 patients who completed the questionnaire twice over an interval of one week. Construct validity was assessed using Spearman's correlations, while responsiveness was evaluated by 3 different methods: (1) effect size (the mean difference between the baseline scores and the follow-up scores divided by the standard deviation of the baseline scores); (2) standardized response mean (the mean change in scores divided by the standard deviation of the change in scores); (3) receiver operating characteristics (ROC) curve analysis.*

Results. *ROAD fulfilled the established criteria for validity, reliability and responsiveness. In comparison with the SF-36, the expected correlations were found when comparing items measuring similar constructs, thus supporting the convergent construct validity. Significant correlations were seen between ROAD scores and HAQ scores (rho = 0.372), SF-36 physical component summary (PCS) (rho = -0.413), PGA functional disability (rho = 0.417), pain (rho = 0.639), Ritchie index (rho = 0.357), number of swollen joints (rho = 0.387), patient and physician assessment of disease activity (rho = 0.467 and 0.323, respectively), and Disease Activity Score (rho = 0.476). Test-retest reliability was satisfactory, with ICCs of 0.927 (upper extremity function), 0.892 (lower extremity function), and 0.851 (activity of daily living/work). Bland-Altman plots confirmed this finding. The results of responsiveness analysis indicate that the ROAD subscales were slightly more sensitive to perceived change in functional disability than those of HAQ, SF-36 PCS, and PGA of functional disability.*

Conclusion. *Our data suggest that the ROAD index is a reliable, valid and responsive tool for measuring physical functioning in patients with early RA, and is suitable for use in clinical trials and daily clinical practice. Its generalizability and utility for assessing aggressive treatment and functional outcomes must now be evaluated in broader settings.*

Introduction

Functional disability testing is proving to be a valuable resource in the assessment of the long-term outcomes in rheumatoid arthritis (RA) patients. Functional disability and quality-of-life indices, which measure the effects of RAon patients'lives, are used to evaluate therapies and to assess the course of disease (1-5). Moreover, demonstration of preserved physical function is a prerequisite for proving that a drug has disease-controlling capacity (4). This is especially pertinent, given the new rea-

lities of more aggressive and early management of RAin the 1990s with costly interventions that could potentially have severe adverse effects.

Several self-report questionnaires have been constructed over the past decades for use in clinical practice and research, health policy evaluation, and general population surveys (6-9). Despite their potential to provide useful and unique information on the patient's functional status and disease outcome, patient self-reported data are still used relatively infrequently in clinical practice (4). Therefore, benefit/risk assessment in standard clinical care is based predominantly on empirical, rather than quantitative assessments, with very limited documentation of the clinical status and the possible improvement or deterioration of the patient's condition over time (4). Some concerns include insufficient exposure to questionnaires, extra time needed to administer and complete the questionnaires, and perceived potential disruption of the clinic's workflow (4). Furthermore, self-report questionnaires are regarded as ''subjective,'' while ''objective'', highly technical data are thought to provide the best information for clinical decision making, monitoring, and prognosis.

Multiple studies have demonstrated that self-report questionnaires used to test functional disability are useful only if they provide valid, reliable information and are sensitive to changes over the course of the period of observation (2, 7). In rheumatology, the most widely used self-report questionnaires are the Health Assessment Questionnaire (HAQ) (10), and its modified versions the MHAQ (11) or multi-dimensional (MDHAQ) (12), RA-HAQ (13), and HAQ II (14) which were developed to measure patient-perceived disability in RA. Each has different metric properties (5). These questionnaires are a valuable tool in routine clinical care, as more accurate information about a patient's current functional status is derived from a self-report than from laboratory, imaging, or other tests. Furthermore, the HAQ has been shown to predict severe outcomes of RA, such as mortality and work disability (15, 16), as well as hip replacement surgery (17). However, this instrument can be influenced by the variables of large joints such as the hips, knees and shoulders, and presents some limitations in detecting changes in hand and foot function (18). This is especially true in patients with early RA. The results from early RAregistries and epidemiological studies suggested that the most frequently involved joints in terms of tenderness and for swelling were the wrists, the metacarpophalangeal (MCP) joints, and the proximal interphalangeal (PIP) joints (19, 20). Moreover, a controlled clinical investigation of the interrelationship between synovitis and bone in early RA concluded that MCP joint bone oedema is present in the majority of patients with RA at presentation and that joint structural changes are secondary to synovitis (21).

In order to develop a measurement tool for the assessment of physical disability in patients with early RA, we have recently proposed a self-administered questionnaire – the Recent-Onset Arthritis Disability (ROAD) questionnaire (22). The ROAD has 12 items assessing three reported patient-relevant dimensions: upper extremity function, lower extremity function, and activities of daily living/work (Appendix). These items represent a combination of symptoms that are common, frequently recurring and of general importance to early RA patients (22). Our goal was to assess the reliability, construct validity, and responsiveness of the ROAD index.

Patients and methods

Patient population

From January 2004 to March 2005, a total of 159 patients (119 women, 40 men) with recent-onset (symptom duration <2 years) RA, seen at the care facilities of the Department of Rheumatology of Università Politecnica delle Marche, and who gave their informed consent, participated in the longitudinal study over a 6-month period. All patients took part in an ongoing prospective study in which we compared the efficacy and tolerability of continuous treatment with a combination of disease-modifying anti-rheumatic drugs (DMARDs) versus a single-drug therapy. At study entry, 154 of the 159 (96.8%) patients were on therapy with DMARDs, including methotrexate (10-15 mg weekly; n = 57; 37%), sulphasalazine (n = 31; 20.1%), antimalarials (n = 38; 24.7%), leflunomide (n= 17; 11.0%), etanercept (n=6; 3.9%) and adalimumab (n = 5; 3.2%); 59 patients (38.3%) were on combination DMARD therapy. Sixty-seven patients (43.5%) had additional therapy with corticosteroids (4-16 mg prednisolone equivalent daily), and all patients were treated with NSAIDs as needed. The patient selection criteria were as follows: fulfilment of the American College of Rheumatology (ACR; formerly the American Rheumatism Association) 1987 revised criteria for RA (23), age 18-75 years, duration of symptoms less than 2 years, and active disease, with ≥ 3 swollen joints, and at least 3 of the following 4 features: either an erythrocyte sedimentation rate (ESR) ≥ 28 mm/hour or a C-reactive protein (CRP) level > 19 mg/litre, morning stiffness ≥ 29 minutes, > 5 swollen joints, and > 10 tender joints (24). The study was performed according to the principles of the Declaration of Helsinki. The protocols were approved by the ethics committees.

Clinical and outcome measures

The following parameters or variables were considered: the number of swollen joints (out of a total of 44 diarthrodial joints); the physician's assessment of disease activity assessed on an 11-point numerical rating scale (NRS) with anchors of 0 "asymptomatic" and 10 "very severe"; patient's assessment of disease activity (same as for the physician's assessment); intensity of pain, on an NRS with anchors of 0 "no pain" and 10 "pain as bad as it could be", the patient's global assessment (PGA) of functional disability on an NRS with anchors of 0 "best possible" and 10 "worst possible". The Ritchie articular index (RAI) was used to assess the level of tenderness (25). The RAI tenderness score was obtained according to the original description (53 joints in 26 units, graded for tenderness on pressure, where 0 = no pain, 1= patient complains of pain, 2 = patient com-

plains of pain and winces, 3 = patient complains, winces, and withdraws; maximum score 78) (25). The European League Against Rheumatism (EULAR) criteria for improvement/response were computed using an index of activity in RA, the Disease Activity Score (DAS) (26). The DAS combines information from the RAI, the swollen joint count, the ESR, and the patient's global assessment of his/her disease activity (27-29). To calculate the DAS, we used the following formula (27):

$$DAS = 0.53938 \times \sqrt{(RAI)} + 0.06465 \times (44SJC) + 0.330 \times \ln(ESR) + 0.00722 \times GH.$$

where RAI is the Ritchie articular index and ranges from 0 to 78, 44SJS is the ungraded count of joints with swelling due to synovitis (maximum score 44), lnESR is the natural logarithm of the ESR (mm/hour; Westergren), and GH is general health, as assessed by the patient using a 100-mm visual analogue scale. The range of DAS is 0-10 (27-29). A 1.2 point change in the DAS values from baseline was considered a statistically significant change. High disease activity was defined as a DAS > 3.7. Low disease activity was defined as a DAS ≤ 2.4. Good response was defined as a significant decrease in DAS (> 1.2) and a low disease activity (≤ 2.4). Non-response was defined as a decrease ≤ 0.6, or a decrease >0.6 and ≤ 1.2 with a DAS > 3.7. Any other scores were regarded as moderate responses (29, 30).

Questionnaires

Two disability questionnaires, the ROAD (22) and HAQ (10), and a generic instrument, the SF-36 (31, 32), were administered twice, at baseline and at 6 months.

Recent-Onset Arthritis Disability (ROAD) questionnaire. The ROAD is a self-administered instrument that consists of 12 items assessing a patient's level of functional ability and includes questions related to fine movements of the upper extremities (5 items), locomotor activities of the lower extremities (4 items), and activities that involve both the upper and lower extremities (3 items) (22). For each item, patients are asked to rate the level of difficulty they experienced over the past week on a 5-point scale ranging from 0 (without any difficulty) to 4 (unable to do). Missing data were treated as follows: one or two missing values were substituted with the average value for the dimension. If more than two items were omitted, the response was considered invalid. In order to express these scores in a more clinically meaningful format, a simple mathematical normalization procedure was then performed so that all the scores could be expressed in the range 0-10, with 0 representing better status and 10 representing poorer status. In this way, 3 physical function sub-scores ranging from 0-10 could be presented graphically as a ROAD disability profile. The total score was the mean of the 3 sub-scores (22). The ROAD can be self-administered in 3-4 minutes and scored in less than one minute. This index and the normalization formula are presented in the appendix.

Health Assessment Questionnaire. HAQ assesses the degree of difficulty a person has in accomplishing tasks in 8 functional areas: dressing and grooming, arising, eating, walking, hygiene, reach, grip, and activities (10). For each item, patients are asked to rate the level of difficulty experienced over the past week on a 4-point scale ranging from 0 (no difficulty) to 3 (unable to perform). To calculate the disability dimension score, a score is calculated for each of the subscales, these subscale scores are summed, and the sum is then divided by 8. The disability score ranges from 0 to 3, with a higher score indicating more disability. A version adapted for use among Italian patients was utilized in the present study (33).

Medical Outcomes Study SF-36 Health Survey (SF-36). The SF-36 general health questionnaire is a generic instrument with scores that are based on responses to individual questions, which are summarized in 8 scales, each of which measures a health concept (31). These 8 health concepts are Physical Function (PF), Role Function-Physical aspect (RP), Bodily Pain (BP), General Health perception (GH), Vitality (VT), Social Function (SF), Role function Emotional aspect (RE), and Mental Health (MH) (31). For each of the SF-36 scales, higher values indicate better health. The summed scores are transformed to a 0-100 scale following the designated scoring algorithm, with higher scores reflecting a better quality of life. SF-36 results can also be described as 2 aggregate scores, the physical component summary (PCS) and the mental component summary (MCS) (34). The SF-36 has been validated for use in Italy (32), and most patients can complete it within 15 minutes.

Data processing

The data are stored in a FileMaker 7.0 relational database and have been processed with SPSS 11.0 (SPSS Inc, Chicago, Illinois, USA), and MedCalc 7.1.02 (MedCalc Software, Belgium) statistical software packages for Windows XP.

Statistical analysis

Parametric techniques may be applicable for data of a certain ordinal level; however, our data were generally not normally distributed (Kolmogorov-Smirnov test for normal distribution), and therefore, the use of non-parametric techniques provided a more conservative estimate of their statistical significance. Where appropriate, median and interquartile ranges are presented, as well as means and standard deviations. The differences among the groups were computed by the Wilcoxon test for independent samples, as appropriate. Following standard guidelines for the evaluation of measurement properties of quality of life instruments (35-37), we tested the validity, reliability, and responsiveness of the ROAD questionnaire. The operational qualities or feasibility of the ROAD questionnaire were analyzed according to the percentage of patients who were able to complete the questionnaire by themselves and the time required to fill it out.

Validity. The construct validity of the ROAD questionnaire, as a measure of disability in patients with early RA, was investigated in two ways. First, we examined its convergent validity by

correlating the scores of the index subscales with the other measures applied in the study. A particular subscale may be expected to converge with the scores of those instruments targeting the same construct, and to deviate from the scores yielded by instruments or scales assessing a different construct, i.e. divergent validity. To quantify these relationships, Spearman's rho correlation coefficients were obtained. Due to multiple comparisons with an increasing risk of type I errors, the level of statistical significance was set at 0.01.

The SF-36 PCS subscale measures limitations in the patient's ability to perform general physical activities, a corresponding construct to the one the ROAD subscales are intended to measure. HAQ assesses the degree of difficulty a person has in accomplishing tasks in 8 functional areas, a construct similar to the one the ROAD subscale is designed to measure. We expected the highest correlations to emerge when comparing scales that are supposed to measure the same or similar constructs. Furthermore, the 8 subscales of SF-36 have been shown to produce valid indices of the Physical Component Summary (PCS) and the Mental Component Summary (MCS) (34). Since the ROAD questionnaire is designed to measure physical rather than mental health, we expected to observe higher correlations between the HAQ and the subscales of Physical Function, Role Function–Physical aspect, Bodily Pain, and General Health perception (convergent construct validity) than between the SF-36 subscales of Vitality, Social Functioning, Role function–Emotional aspect, and Mental Health (divergent construct validity). We also calculated the percentage of the sample who achieved the lowest (floor effect) and highest (ceiling effect) possible scores. Changes in scores for all instruments were normally distributed. Baseline and follow-up scores were compared using non-parametric Wilcoxon signed rank tests. Differences were considered significant when $p < 0.05$.

Reliability. Reliability embraces the concept that the repeated administration of a measurement tool in stable subjects will yield the same results. After a one-week interval, patients were asked by the same data collector to repeat the measurement on a ROAD questionnaire, without access to any previous ROAD ratings. Because it was possible for a patient's condition to change over a one-week interval, a global rating of change questionnaire was concurrently administered to the subjects. The concept of the "transition" method was originally developed by Jaeschke *et al*. (38) and has since been applied in different rheumatologic settings (39-43). The so-called "transition questionnaire" investigated the current health status compared to that when the first questionnaire was completed (Question: Compared to when you completed the questionnaire regarding your functional disability a week ago, how is your health now ?) The possible response options were "much better," "slightly better", "no change," "slightly worse," or "much worse" (38). Subjects who reported no change were considered stable and those who reported a change were eliminated from this analysis.

In this study, the test–retest reliability of the ROAD index was analysed in a group of 77 patients who reported no change in their arthritis using intra-class correlation coefficients (ICCs). The ICC reflects both systematic and random differences in test scores. Values of ICC thus may vary from 1 (perfectly reliable) to 0 (totally unreliable). The ICC was chosen in preference to the Pearson correlation, which may overestimate reliability. One suggestion for acceptable test–retest reliability for the assessment of an individual is an intra-class correlation coefficient of 0.85 (37). According to Bland and Altman (44), repeatability can be demonstrated by plotting the difference against the mean of the two assessments for each subject; 95% of the differences are expected to be less than two standard deviations.

Responsiveness. Responsiveness was defined as "the ability of an instrument to accurately detect change when it has occurred" (45). Internal responsiveness characterises the capacity of a questionnaire to change over time, and external responsiveness compares the change in scores with the patients'global assessment of change. Internal responsiveness was assessed for each health measure using the effect size (ES) and standardized response mean (SRM) (46-49). The change in score with standard error of the mean was calculated for each instrument. Change due to intervention was assessed by Wilcoxon's signed rank test. The level of significance was chosen as $\alpha = 0.05$. Conceptually, ES and SRM are closely connected, relating the magnitude of the change in score to the variability in score (35) and can be thought of as indicators of the ability to distinguish "signal" from "noise" (51). The ES is calculated as the mean change in score from baseline divided by the standard deviation of baseline scores (46-49). The variation in the baseline score is a reference against which to judge change. The SRM is the mean change in score between assessments divided by the standard deviation of these changes (46-49). Although there are no absolute standards for ES and SRM, it has been suggested that in comparative studies examples of small, medium, and large effect sizes might have values of 0.2, 0.5, 0.8, respectively (52).

A third method of demonstrating external responsiveness uses receiver operating characteristics (ROC) curves (53). ROC curves have been used with an external criterion as a dichotomous gold standard to evaluate responsiveness (53). This method has the advantage of synthesizing information on sensitivity and specificity for detecting improvement by an external criterion (40, 41). This is important because some score variability is unrelated to true changes in clinical status. The area under the ROC curve (AUC) in this setting can be interpreted as the probability of correctly identifying patients showing improvement from those not showing improvement (53). The area ranges from 0.5 (no accuracy in distinguishing responders from non-responders) to 1.0 (perfect accuracy). In addition, the EULAR definition of response or improvement (26, 29, 30) was used as an external criterion. Since ROC analysis requires external criteria to be dichotomous, the three-point scale was

Table I. Descriptive statistics and features of score distributions for health status measures in RApatients* (n = 159).

	Mean score	S.D.	Median	Percentiles			% at floor or ceiling	
				25	50	75	% at floor	% at ceiling
ROAD scales (0-10)								
ROAD - Upper Extremity Function	4.40	1.96	5.00	3.00	5.00	6.00	1.7	0.2
ROAD - Lower Extremity Function	3.65	2.30	3.12	2.50	3.12	5.00	2.5	0.1
ROAD - Activities of Daily Living/Work	4.52	1.98	4.17	3.33	4.17	5.83	1.1	0.8
ROAD - Overall Score	4.19	1.51	4.29	3.26	4.19	5.30	0.8	0.2
SF- 36 scales (0-100)								
Physical Function (PF)	43.25	21.08	45.00	25.00	45.00	60.00	18.3	1.0
Role function Physical aspect (RP)	25.31	2.79	25.00	25.00	25.00	25.00	24.9	10.4
Bodily Pain (BP)	22.22	17.53	22.00	11.25	22.00	32.00	6.9	2.8
General Health (GH)	42.99	19.89	40.00	25.00	40.00	57.00	1.3	1.0
Vitality (VT)	45.50	19.92	45.00	25.00	45.00	50.00	2.5	0
Social Function (SF)	49.37	23.52	50.00	37.50	50.00	62.50	20.5	15.0
Role function Emotional aspect (RE)	35.53	39.29	33.33	0	33.33	66.66	24.0	27.0
Mental Health (MH)	49.33	24.27	52.00	28.00	52.00	68.00	11.3	2.3
SF-36 PCS	33.53	15.34	30.50	23.00	30.50	39.25		
SF-36 MCS	43.68	21.02	37.33	26.75	37.33	58.83		
Health Assessment Questionnaire (0-3)	1.32	0.62	1.37	0.87	1.37	1.87	11.9	0.8
PGAfunctional disability (0-100)	9.51	23.11	50.00	40.00	50.00	60.00	4.8	1.9

*Increasing scores (from 0 to 100) in the SF-36 scales reflect better health status. For the ROAD and HAQ scores, increasing scores (from 0 to 10 and from 0 to 3, respectively) reflect poor health status.

collapsed to a dichotomous variable (good/moderate responder; non-responder). The non-parametric Wilcoxon signed ranks test was used to calculate and compare the areas under the ROC curves derived from the sample of patients, as suggested by Hanley and McNeil (54).

Results

Patient characteristics

Of the 159 patients who completed the health status instruments on two occasions, 121 were included in the responsiveness analyses. The remaining 38 patients had information missing from one or more of the study measures. No significant differences were seen in baseline characteristics between the 121 patients included in the longitudinal analyses and the 38 patients who were excluded because of missing information. At the first assessment, the mean age was 54.7 ± 8.8 years (range 20-79). Ninety of the respondents were female (74.3%) and 31 were male (25.7%). The mean disease duration (i.e. symptomatic) was 14.5±1.9 months. 74% were seropositive for rheumatoid factor (RF by nephelometry), with titres of ≥ 20 IU/ml, and 28% were positive for antinuclear antibody (ANA) (1:40 on Hep-2 cells). Their formal education level was generally low: 55.3% had received only a primary school education, and only 18.2% had completed high school. The majority of patients (62.2%) were married and living with their family; 31% were housewives. Of the 121 subjects who completed the follow-up, 45 (45.5%) reported one or more medical co-morbidities, mostly cardiovascular (30.2%), respiratory

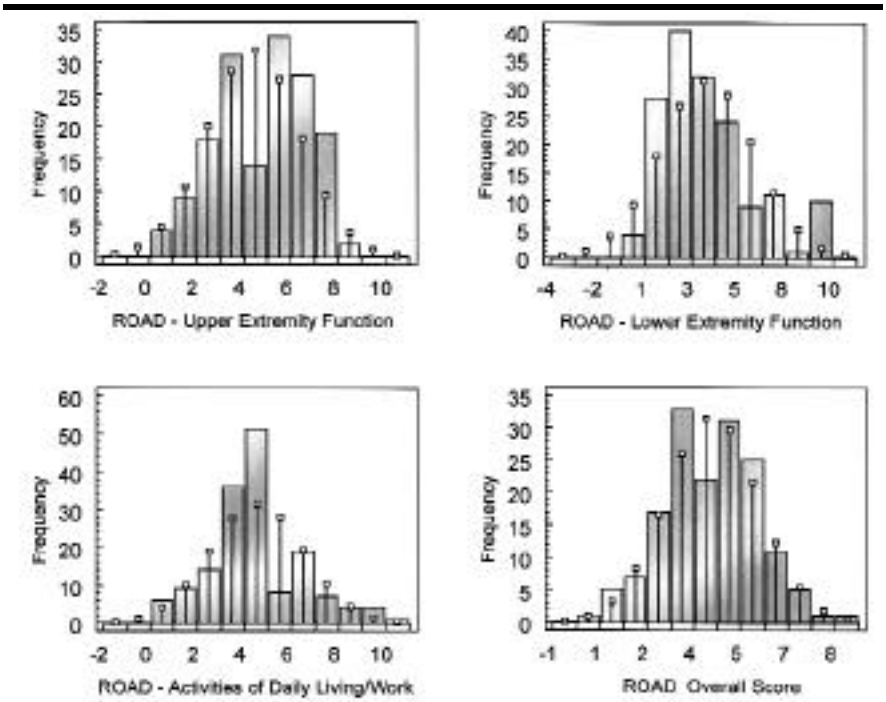

Fig. 1. Distribution of the scores in the ROAD (upper extremity function, lower extremity function, activities of daily living/work, and overall score).

(16.6%), and metabolic (11.1%) disorders.

Score distributions

Table I shows the mean±standard deviation, median, percentiles and the percentages of the sample showing the floor and ceiling effects for health status and disability measures at the initial assessment. As reported for several other health status instruments, the ROAD scale scores were not normally distributed (Kolmogorov-Smirnov test for normal distribution) (34, 55,56). The distributions of the scores in the ROAD (upper extremity function, lower extremity function, activities of daily living/work, and overall score) are presented in Figure 1. The bar on the left of each graph represents the number of subjects with a score of 0 (floor effect); the bar on the right represents the number of subjects with a maximum possible score (ceiling effect). The ROAD had negligible floor and ceiling effects in patients with early RA. Data from the SF-36 scales, HAQ and PGA of functional disability are included in Table I for comparison. For the SF-36, noteworthy floor effects were observed for the two role-disability scales (24.9% for role limitation due to physical health problems and 24% for role limitations due to emotional problems) and for physical function (18.3%). Substantial ceiling effects were also observed for social function (20.5%) and mental health (11.3%). For the HAQ, the percentage of respondents with minimum (floor) scores was 11.9% whereas modest ceiling effects were observed for the PGAof functional disability (4.8%).

Validity study

Establishing the validity of an instrument to measure functional status is difficult, as no established "gold standards" are available for comparison. In this study, the construct validity was examined in terms of convergence with variables that should have a converging relationship. As expected, higher significant correlations were seen when comparing the ROAD scales to the SF-36 scales for physical health (convergent construct validity), than those for mental health (divergent construct validity) (Table II). Negative correlations between the measures result from higher scores on the SF-36 associated with greater well being, whereas higher scores on the ROAD indicate lower well being. Of special interest are the correlations between ROAD and HAQ (rho = 0.372; p < 0.0001) and between ROAD and SF-36 PCS dimension (rho

Table II. Construct validity analysis: correlation matrix of ROAD scores versus SF-36 dimensions, HAQ, and PGAfunctional disability.

		ROAD Upper Extremity Function	ROAD Lower Extremity Function	ROAD Activities of Daily Living/Work	ROAD Overall Score
SF-36 questionnaire					
General Health (GH)	Spearman's rho	-.222	-.200	-.373	-.234
	Sig. (2-tailed)	.005	.011	.0001	.0003
Social Function (SF)	Spearman's rho	-.146	-.169	-.173	-.338
	Sig. (2-tailed)	.065	.033	.027	.044
Vitality (VT)	Spearman's rho	-.105	-.235	-.265	-.298
	Sig. (2-tailed)	.188	.003	.001	.001
Mental Health (MH)	Spearman's rho	-.168	-.091	-.202	-.196
	Sig. (2-tailed)	.034	.421	.011	.005
Bodily Pain (BP)	Spearman's rho	-.195	-.239	-.355	-.223
	Sig. (2-tailed)	.014	.003	.0001	.005
Role function Emotional aspect (RE)	Spearman's rho	-.165	-.051	-.162	-.185
	Sig. (2-tailed)	.038	.523	.040	.011
Role function Physical aspect (RP)	Spearman's rho	-.161	-.202	-.322	-.341
	Sig. (2-tailed)	.042	.006	.0001	.0002
Physical Function (PF)	Spearman's rho	-.178	-.260	-.403	-.323
	Sig. (2-tailed)	.025	.001	.0001	.0005
SF36 - MCS	Spearman's rho	-.192	-.220	-.291	-.305
	Sig. (2-tailed)	.015	.005	.001	.001
SF36 - PCS	Spearman's rho	-.258	-.260	-.388	-.413
	Sig. (2-tailed)	.001	.001	.0001	.0001
HAQ	Spearman's rho	.287	.167	.323	.372
	Sig. (2-tailed)	.0001	.035	.0001	.0001
PGAfunctional disability	Spearman's rho	.249	.157	.158	.244
	Sig. (2-tailed)	.004	.048	.047	.002

HAQ, and PGAfunctional disability.

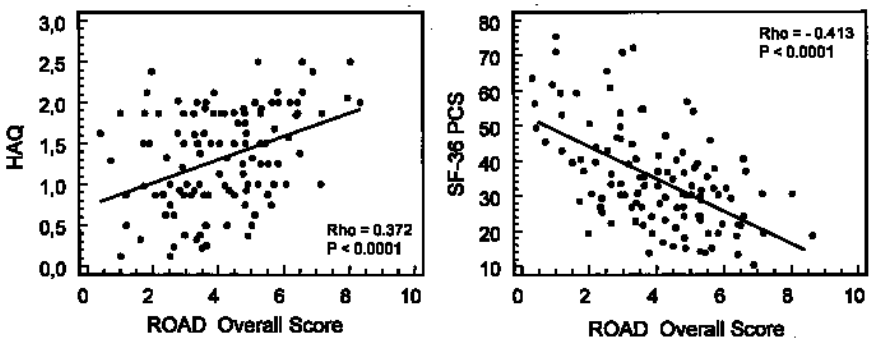

Fig. 2. Scattter plot of a patient's ROAD function against the HAQ and physical functioning score of the SF-36. For "best fit" purposes, the descriptive linear regression line has been superimposed.

Table III. Construct validity analysis: correlation matrix of ROAD scores versus clinical variables.

		ROAD Upper Extremity Function	ROAD Lower Extremity Function	ROAD Activities of Daily Living/Work	ROAD Overall Score
Pain (on a 11 points NRS)	Spearman's rho	.563	.473	.370	.639
	Sig. (2-tailed)	.0001	.0001	.0001	.0001
Morning stiffness	Spearman's rho	.272	.215	.203	.328
	Sig. (2-tailed)	.001	.006	.012	.001
Joint swelling (44 joints)	Spearman's rho	.315	.254	.287	.387
	Sig. (2-tailed)	.0001	.001	.0001	.0001
Ritchie articular index (RAI)	Spearman's rho	.205	.267	.284	.357
	Sig. (2-tailed)	.010	.001	.0001	.0001
Patient's assessment of disease activity	Spearman's rho	.405	.272	.350	.467
	Sig. (2-tailed)	.0001	.001	.0001	.0001
Physician's assessment of disease activity	Spearman's rho	.258q	.259	.226	.323
	Sig. (2-tailed)	.001	.001	.004	.0001
C-reactive protein	Spearman's rho	.051	.103	.043	.070
	Sig. (2-tailed)	.521	.196	.593	.384
Erythrocyte sedimentation rate	Spearman's rho	.133	.063	.022	.089
	Sig. (2-tailed)	.094	.427	.785	.265
Disease activity score	Spearman's rho	.363	.319	.344	.476
	Sig. (2-tailed)	.0001	.0001	.0001	.0001

= -0.413; $p < 0.0001$) (Fig. 2). Correlations of the ROAD scales with the patient global assessment of functional disability were lower than with other SF-36 subscales assessing physical health and with the HAQ.

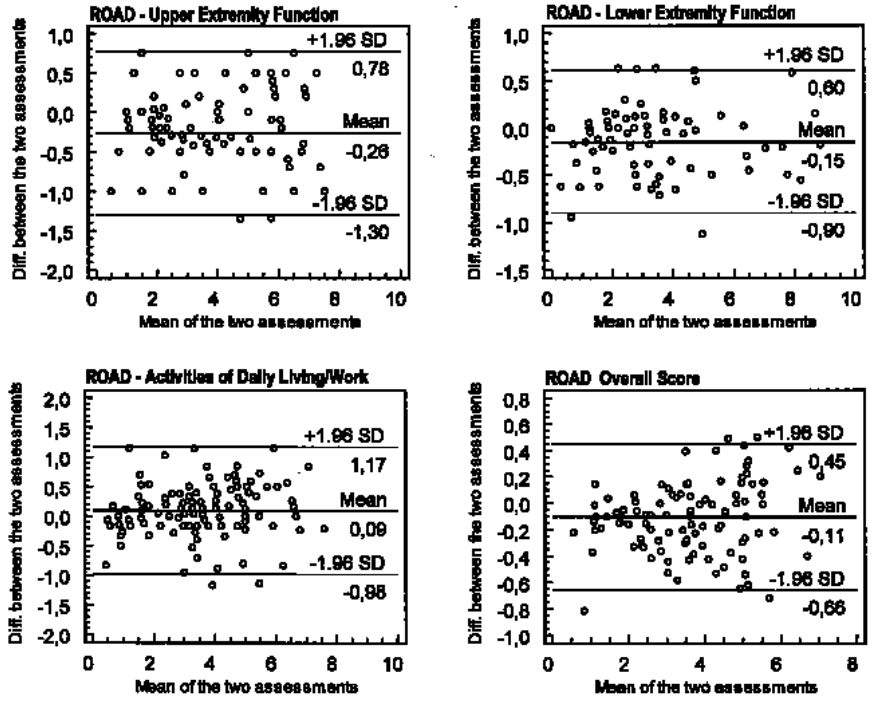

Fig. 3. Bland-Altman plots for the ROAD (upper extremity function, lower extremity function, activities of daily living/work, and overall score).

All three ROAD dimensions were significantly correlated with one another. The strongest correlation was between upper extremity function and activities of daily living/work (rho = 0.469; $p < 0.0001$). ROAD scales and the overall score were moderately to highly correlated with all clinical measures, including pain, morning stiffness, joint swelling, RAI, patient and physician disease activity assessment, ESR, CRP, and DAS (Table III), although there was no statistically significant correlation with ESR and CRP.

Reliability study

The reliability of the ROAD index was evaluated in 86 patients over a one-week period. Nine subjects were excluded due to an excessive time lapse (more than 10 days) between the test and retest. For the remaining 77 subjects, there was a mean of 6.5 days between the test and retest (± 2.1 days). The intra-class correlation coefficients of the three subscales (upper extremity function, lower extremity function and activity of daily living/work) were 0.927, 0.892, and 0.851, respectively. Bland and Altman plots of repeatability are given in Figure 3. For all subscales, 95% of the differences against the means were less than two standard deviations.

Internal and external responsiveness study

Table IV shows the responsiveness of the various instruments in the RAsample over a 6-month period. A significant improvement was seen for all the ROAD subscales, SF-36 PCS (all at $p < 0.0001$), HAQ, and PGA of functional disability (both at $p < 0.001$). The EULAR response criteria were also used as external criterion to calculate ES and SRM. The results indicate that the ROAD subscales were slightly more sensitive to perceived change in functional disability than those of the SF-36 PCS, HAQ, and the subjects' rating of global disability. The upper extremity function and activities of daily living/work dimensions of the ROAD appeared to be the most responsive (1.021 and 1.035 for ES; 0.722 and 0.721 for SRM, respectively). The calculation of

Table IV. Acomparison of responsiveness indices for functional measures

	Mean change from baseline*		Effect size		Standardized response mean		ROC-AUC	
	Mean	Std. Error	Mean	Std. Error	Mean	Std. Error	Mean	Std. Error
ROAD - Overall Score	-1.970	0.160††	1.083	0.108	0.877	0.068	0.865	0.044
ROAD - Upper Extremity Function	-1.885	0.273††	1.021	0.128	0.722	0.090	0.751	0.043
ROAD - Lower Extremity Function	-1.947	0.264††	0.827	0.104	0.708	0.091	0.761	0.044
ROAD - Activities of Daily Living/Work	-2.078	0.218††	1.035	0.107	0.731	0.123	0.781	0.041
HAQ	-0.484	0.071†	0.914	0.119	0.794	0.091	0.774	0.041
SF36-PCS	16.154	2.047††	0.701	0.114	0.634	0.083	0.659	0.059
PGAfunctional disability	-1.537	0.222†	0.709	0.094	0.702	0.093	0.678	0.058

* Absolute mean changes with results of paired t tests. † $p < 0.001$; †† $p < 0.0001$. All mean changes are all in expected direction (improved).

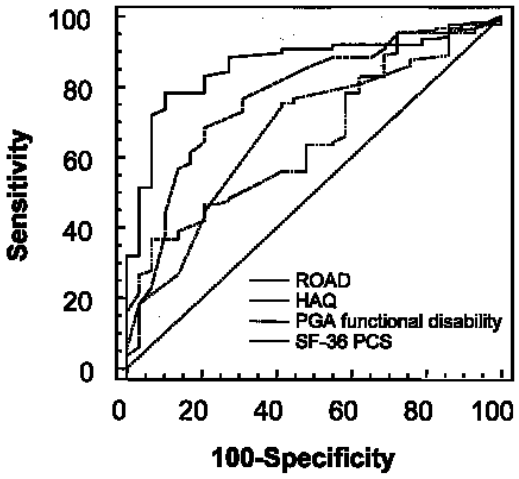

Fig. 4. ROC curves illustrating the relationship between sensitivity and complement of specificity (100-specificity) for the ROAD, HAQ, PGA functional disability, and SF-36 measures, using the EULAR definition of response or improvement as an external indicator. The area under the ROC curve (AUC) in this setting can be interpreted as the probability of correctly distinguishing patients with improvement from those without improvement.

SRM generally yielded somewhat smaller numbers, but did not change the interpretation of the data (Table IV).

Figure 4 presents the ROC curves generated for changes in the ROAD, HAQ, SF-36 PCS and PGAfor functional disability at the 6-month follow-up. True positive proportions (sensitivity) and false positive proportions (100-specificity) for the discrimination between clinical improvement and clinical stability are plotted for multiple cut-off points. The highest area under the curve was recorded for ROAD (0.865, with 95% C.I. from 0.801 to 0.914). For the HAQ the AUC was 0.774 (95% C.I. from 0.701 to 0.836), for the SF-36 PCS the AUC was 0.659 (95% C.I. from 0.580 to 0.732), and for the PGA of functional disability AUC was 0.678 (95% C.I. from 0.599 to 0.750). The difference between the ROAD and HAQ (differences between areas = 0.091± 0.045 with 95% C.I. from 0.002 to 0.180), ROAD and SF-36 PCS (differences between areas = 0.205± 0.061 with 95% C.I. from 0.085 to 0.326), and ROAD and PGA of functional disability (differences between areas = 0.187 ± 0.054 with 95% C.I. from 0.081 to 0.293) were all significant ($p = 0.045$, $p = 0.001$, and $p = 0.002$, respectively).

Discussion

Measurement of a patient's perception of health is regarded as a standard approach in clinical practice, in controlled clinical trials and longitudinal observational studies, as well as in other types of epidemiological research. Such measurements are performed in rheumatic diseases using a "rheumatology specific" questionnaire such as the HAQ to assess physical function, or a disability score and pain intensity score on a visual analogue scale (VAS), or a "general" questionnaire to capture information concerning several dimensions of health such as the SF-36 questionnaire.

Physical disability is the most powerful determinant of all severe long-term outcomes in RA, such as work disability, costs and premature mortality (15-17,57,58). Data from self-report patient questionnaires are recognized to be valid and reliable (2, 4, 8, 11). Questionnaires are valuable research tools but generally are not incorporated into routine medical care, mainly because of their length, the extra time needed to administer and complete the questionnaires, and complex and non-intuitive scoring systems (31, 32, 34).

The disability assessment component of the ROAD index assesses a patient's level of functional ability and includes questions concerning fine movements of the upper extremities, locomotor activities of the lower extremities, and activities that involve both upper and lower extremities (22). The twelve items were extrapolated from a reduction process conducted on an initial pool of 122 questions (22). Eight items are very similar to items in the HAQ; one explores working capacity, one the capacity to stand, and two the fine movements of the hands. Assessment of the functional capacity of the upper extremity using a specific instrument is not usually included for clinical evaluation in early RA. Although large joint involvement is an early phenomenon in patients with RAand is strongly associated with physical disability, a strong correlation with small joint involvement in clinical practice has been demonstrated. (59-61).

In our experience, patients find the 12 items of the ROAD index easy to complete within a few minutes (22), indi-

cating its suitability for use in observational studies, but also in randomized controlled trials, with 3 separate scales – one for upper extremity functioning (5 items), one for lower extremity function (4 items) and one for activities that involve both upper and lower extremities (3 items). All items refer to the preceding week and are scored on a 5-point Likert-scale ranging from 0 (without any difficulty) to 4 (unable to do) (22). During the development of the ROAD index, a total score was calculated in addition to the three subscores, because of the assumption that one summary score might be easier for clinicians. However, when comparing the information derived from the overall score (which told whether the patient had any problems whatsoever) to the information derived from the three subscores, which gave a much more detailed picture as to what areas were causing difficulties, we decided that the use of three subscores was preferable.

Our findings suggest that the ROAD index is a valid, reliable, and responsive instrument for the quick and simple assessment of functional disability in early RA patients. The validation of an instrument is an on-going process and testing validity arises not from a single experiment, but from a series of converging experiments. Assessing validity means measuring the extent to which a technique measures that which it is intended to measure (36, 37). If ROAD is a valid measure of functional disability, one would expect its values to be modestly correlated with biological parameters such as the ESR or CRP, but more highly correlated with the patients'subjective perceptions of their disabilities, for example with the HAQ and SF-36 PCS, in accordance with the suggestion of Liang and Jette (46). To fully establish construct validity, the investigator should also demonstrate which variables are not correlated with the construct of interest.

In this study, the pattern hypothesized *a priori* of higher correlations to subscales with a high capacity to measure physical health and the majority of the clinical measures of disease activity (e.g. swollen and tender joints) and weak or absent correlations with laboratory tests was found, confirming that the ROAD index measures the suggested constructs. In this report, as in others (55, 56), a strong association was found between the ROAD subscale scores and the patient's emotional state on the SF-36 Mental Component Scale. The ROAD index is sensitive to psychosocial factors, as are self-report instruments in general, as such factors contribute to the pain and physical impairment reported by patients. If, however, a patient's emotional state markedly influences pain and physical health status perception, the resultant random measurement error would restrict the validity of the ROAD or other self-report questionnaires to only relatively large studies.

Another way of assessing content validity is to study the floor and ceiling effects of each dimension. A ceiling effect makes it impossible to measure improvement, while a floor effect makes it impossible to measure deterioration. The ROAD had negligible floor and ceiling effects in patients with early RA, indicating its potential for measuring improvement over time. The HAQ appears to have a floor effect (5, 62). This would imply that the instrument is less sensitive in detecting improvement at the lower end of the scale. A report by van den Ende *et al.* (63) concluded that HAQ is not an appropriate instrument to detect changes in physical impairment due to short-term physical exercise therapy, a finding confirmed by other authors (64). The acceptability of the ROAD index was in general very good, with no disturbing questions, few confusing items, a very low percentage of missing data for items and scales, and a short duration of administration of the questionnaire. This reinforces its expected validity (face validity).

Previously, we confirmed that the ROAD subscales were internally consistent, with a Cronbach's coefficient alpha of 0.882 for the upper extremity function subscale, 0.838 for the lower extremity function subscale, and 0.811 for the activities of daily living/work subscale (22). In this report, reliability was tested by examining the stability of the instrument scores in patients reporting 'no change' in their condition over one week. The ICCs of the three subscales of upper extremity function, lower extremity function, and activities of daily living/work were 0.92, 0.89, and 0.85, respectively.

The results which we report here indicate that all the ROAD subscale scores improved significantly due to intervention and the ROAD effect sizes ranged from 0.827 to 1.035, indicating that the ROAD is a valid measure of change over time. The calculation of the SRM generally yielded somewhat smaller numbers, but did not change the interpretation of the data. One drawback of this study was that no placebo group was included as a control for responsiveness, and it is possible that the use of an open label treatment may have the effect of increasing pre- and post-differences, particularly in the subjective assessment.

In conclusion, the ROAD questionnaire demonstrated somewhat stronger correlations than any of the other instruments studied. The ROAD index proved to be more responsive to meaningful clinical change than the HAQ, SF-36 PCS or the patient's rating of global disability in early RA. Of course, future research is needed to compare the ROAD index with other recent instruments shown to have acceptable psychometric properties, such as the HAQ-II (14). Also, it will be important to evaluate how useful the ROAD index is relative to instruments that are more specific to body regions (i.e., upper extremity, lower extremity, etc.).

References

1. FELSON DT, ANDERSON JJ, BOERS A *et al*.: The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. *Arthritis Rheum* 1993; 36: 729-40.
2. RUSSAK SM, CROFT JD, FURST DE *et al*.: The use of rheumatoid arthritis health-related quality of life patient questionnaires in clinical practice:lessons learned. *Arthritis Care Res* 2003; 49: 574-84.
3. PINCUS T, STRAND V, KOCH G *et al*.: An index of the three core data set patient questionnaire measures distinguishes efficacy of active treatment from placebo as effectively as the American College of Rheumatology 20% response criteria (ACR20) or the disease activity score (DAS) in a rheumatoid arthritis clinical trial. *Arthritis Rheum* 2003; 48: 625-30.
4. PINCUS T, SOKKA T, KAVANAUGH A:

Quantitative documentation of benefit/risk of new therapies for rheumatoid arthritis: patient questionnaires as an optimal measure in standard care. *Clin Exp Rheumatol* 2004; 22 (Suppl. 35): S26-S33.

5. WOLFE F: Which HAQ is best? A comparison of the HAQ, MHAQ and RA-HAQ, a difficult 8 item HAQ /DHAQ, and a rescored 20 item HAQ (HAQ20): Analyses in 2491 rheumatoid arthritis patients following leflunomide initiation. *J Rheumatol* 2001; 28: 982-9.
6. WOLFE F, LASSERE M, VAN DER HEIJDE D *et al*.: Preliminary core set of domains and reporting requirements for longitudinal observational studies in rheumatology. *J Rheumatol* 1999; 26: 484-9.
7. GUYATT GH, VAN ZANTEN SJOV, FEENY DH, PATRICK DI: Measurement quality of life in clinical trials: a taxonomy and review. *Can Med Assoc* 1989; 140: 1441-8.
8. FRANCHIGNONI F, SALAFFI F: Generic and specific measures for outcome assessment in orthopaedic and rheumatologic rehabilitation. *In* BARAT M and FRANCHIGNONI F (Eds.): *Assessment in Physical Medicine and Rehabilitation - Advances in Rehabilitation*. 2004; 16: 45-77.
9. SCOTTDL, GARROOD T: Quality of life measures: use and abuse. *Bailliere's Clin Rheumatol* 2000; 14: 663-87.
10. FRIES JF, SPITZ PW, KRAINES RG, HOLMAN HR: Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980; 23: 137-45.
11. PINCUS T, SUMMEY JA, SORACI SA, WALLSTON KA, HUMMON NP: Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. *Arthritis Rheum* 1983; 26: 1346-53.
12. PINCUS T, SWEARINGEN C, WOLFE F: Toward a multidimensional Health Assessment Questionnaire (MDHAQ): Assessment of advanced activities of daily living and psychological status in the patient-friendly health assessment questionnaire format. *Arthritis Rheum* 1999; 42: 2220-30.
13. TENNANT A, RYSER L, STUKI G, WOLFE F: An 8-item international version of the HAQ: the inter-HAQ. *Arthritis Rheum* 1999; 42 (Suppl): S74 (abstract).
14. WOLFE F, MICHAUD K, PINCUS T: Development and validation of the Health Assessment Questionnaire. II. A revised version of the Health Assessment Questionnaire. *Arthritis Rheum* 2004; 50: 3296-305.
15. SOKKA T, PINCUS T: Markers for work disability in rheumatoid arthritis. *J Rheumatol* 2001; 28: 1718-22.
16. SOKKA T, HÄKKINEN A, KRISHNAN E, HANNONEN P: Similar prediction of mortality by the health assessment questionnaire in patients with rheumatoid arthritis and the general population. *Ann Rheum Dis* 2004; 63: 494-7.
17. WOLFE F, ZWILLICH SH: The long-term outcomes of rheumatoid arthritis: A 23-year prospective, longitudinal study of total joint replacement and its predictors in 1600 patients with rheumatoid arthritis. *Arthritis Rheum* 1998; 41: 1072-82.
18. WOLFE F: A reappraisal of HAQ disability in rheumatoid arthritis. *Arthritis Rheum* 2000; 43: 2751-61.
19. KVIEN TK, UHLIGH T: The Oslo with arthritis registries. *Clin Exp Rheumatol* 2003: 21 (Suppl. 31): S118-122
20. PINCUS T, SOKKA T: Uniform databases in early arthritis: Specific measures to complement classification criteria and indices of clinical change. *Clin Exp Rheumatol* 2003; 21 (Suppl. 31): S79-S88.
21. MCGONAGLE D, CONAGHAN PG, O'CONNOR P *et al*.: The relationship between synovitis and bone changes in early untreated rheumatoid arthritis: a controlled magnetic resonance imaging study. *Arthritis Rheum* 1999; 42: 1706-11.
22. SALAFFI F, BAZZICHI L, STANCATI A *et al*.: Development of a functional disability measurement tool to assess early arthritis: the Recent-Onset Arthritis Disability (ROAD) questionnaire. *Clin Exp Rheumatol* 2005; 23: 628-36.
23. ARNETT FC, EDWORTHY SM, BLOCH DA *et al*.: The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum* 1988; 31: 315-24.
24. KORPELA M, LAASONEN L, HANNONEN P *et al*.: FIN-RAco trial group. Retardation of joint damage in patients with early rheumatoid arthritis by initial aggressive treatment with disease-modifying antirheumatic drugs: five-year experience from the FIN-RACo study. *Arthritis Rheum* 2004; 50: 2072-81.
25. RITCHIE DM, BOYLE JA, MCINNES JM *et al*.: Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. *Q J Med* 1968; 37: 393-406.
26. VAN GESTEL AM, PREVOO MLL, VAN'T HOF MA, VAN RIJSWIJK MH, VAN DE PUTTE LBA, VAN RIEL PLCM: The development and validation of the European League Against Rheumatism response criteria in rheumatoid arthritis. A comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism criteria. *Arthritis Rheum* 1996; 39: 34-40.
27. VAN DER HEIJDE DMFM, VAN'T HOF MA, VAN RIELPLCM *et al*.: Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. *Ann Rheum Dis* 1990; 49: 916-20.
28. VAN DER HEIJDE DMFM, VAN'T HOF MA, VAN RIELPLCM, VAN LEEUVEN MA, VAN RIJSWIJK MH, VAN DE PUTTE LBA: A comparison of validity of single variables and composite indices for measuring disease activity in rheumatoid arthritis. *Ann Rheum Dis* 1992; 51: 177-81.
29. VAN DER HEIJDE DMFM, VAN'THOF MA, VAN RIELPLCM, VANDE PUTTE LBA: Development of a disease activity score based on judgment in clinical practice by rheumatologists. *J Rheumatol* 1993; 20: 579-81.
30. SCOTT DL, PANAYI GS, VAN RIEL PLCM, SMOLEN J, VAN DE PUTTE LBA, and the CONSENSUS STUDY GROUP OF THE EUROPEAN WORKSHOP FOR RHEUMATOLOGY: Disease activity in rheumatoid arthritis: preliminary report of the Consensus Study group of the European Workshop for Rheumatology Research. *Clin Exp Rheumatol* 1992; 10: 521-5.
31. 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: 473-83.
32. APOLONE G, MOSCONI P: The Italian SF-36 Health Survey: translation, validation and norming. *J Clin Epidemiol* 1998; 51: 1025-36.
33. RANZA R, MARCHESONI A, CALORI G *et al*.: The Italian version of the functional disability index of the Health Assessment Questionnaire. A reliable instrument for multicenter studies on rheumatoid arthritis. *Clin Exp Rheumatol* 1993; 11: 123-8.
34. WARE J, KOSINSKI M, BAYLISS M, ROGERS WH, RACZEC 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; 4: AS264-79.
35. KIRSHNER B, GUYATT GH: Methodological framework for assessing health indices. *J Chron Dis* 1985; 1: 27-36.
36. OXMAN AD, GUYATT GH: Validation of an index of the quality of review articles. *J Clin Epidemiol* 1991; 44: 1271-8.
37. STREINER DL. NORMAN GR: *Health Measurement Scales: A Practical Guide to Their Development and Use.* 2nd ed. New York, Oxford University Press, 1996.
38. JAESCHKE R, SINGER J, GUYATT GH: Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clin Trials* 1989; 10: 407-15.
39. SALAFFI F, STANCATI A, SILVESTRI CA, CIAPETTI A, GRASSI W: Minimal clinically important changes in chronic musculoskeletal pain intensity measured on a numerical rating scale. *Eur J Pain* 2004; 8: 283-91.
40. KOSINSKI M, ZHAO SZ, DEDHIYA S, OSTERHAUS JT, WARE JE, JR: Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. *Arthritis Rheum* 2000; 43: 1478-87.
41. ANGST F, AESCHLIMANN A, MICHEL BA, STUKI G: Minimal clinically important rehabilitation effects in patients with osteoarthritis of the lower extremities. *J Rheumatol* 2002; 29: 131-8.
42. EHRICH EW, DAVIES GM, WATSON DJ, BOLOGNESE JA, SEIDENBERG BC, BELLAMY N: Minimal perceptible clinical improvement with the Western Ontario and McMaster Universities Osteoarthritis Index Questionnaire and global assessment in patients with osteoarthritis. *J Rheumatol* 2000; 27: 2635-41.
43. DEYO RA, DIEHR P, PATRICK DL: Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clin Trials* 1991; 12: 142S-58S.
44. BLAND JM, ALTMAN DG: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1: 307-10.
45. DEBRUIN AF, DIEDERIKS JPM, DEWILLE LP, STEVENS FCJ, PHILIPSEN H: Assessing the responsiveness of a functional status mea-

sure: the Sickness Impact Profile versus the SIP68. *J Clin Epidemiol* 1997; 50: 529-40.
46. LIANG MH, JETTE AM: Measuring functional ability in chronic arthritis: a critical review. *Arthritis Rheum* 1981; 24: 80-6.
47. SALAFFI F, STANCATI A, CAROTTI C: Responsiveness of health status measures and utility-based methods in patients with rheumatoid arthritis. *Clin Rheumatol* 2002; 21: 478-87.
48. GUYATT G, WALTER S, NORMAN G: Measuring change over time: assessing the usefulness of evaluative instruments. *J Chronic Dis* 1987; 40: 171-8.
49. KAZIS L, ANDERSON J, MEENAN R: Effect sizes for interpreting changes in health status. *Med Care* 1989; 27: S178-9.
50. KATZ JN, LARSON MG, PHILIPS CB, FOSSEL AH, LIANG MH: Comparative measurement sensitivity of short and longer health status instruments. *Med Care* 1992; 30: 917-25.
51. DEYO RA, CENTOR RM: Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test peformance. *J Chronic Dis* 1986: 39: 897-906.
52. COHEN J: *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, Laurence Erlbaum, 1988.
53. HANLEY JA, MCNEIL BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143: 29-36.
54. HANLEY JA, MCNEIL BJ: A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148: 839-43.
55. SALAFFI F, PIVA S, BARRECAC, CACACE E, CIANCIO G, LEARDINI G, on behalf of the GONARTHROSIS AND QUALITY OF LIFE (GOQUOLA) STUDY GROUP: Validation of an Italian version of the arthritis impact measurement scales 2 (ITALIAN-AIMS2) for patients with osteoarthritis of the knee. *Rheumatology* 2000; 39: 720-6.
56. SALAFFI F, LEARDINI G, CANESI B *et al*.: Reliability and validity of the Western Ontario and McMaster Universities (WOMAC) Osteoarthritis Index in Italian patients with osteoarthritis of the knee. *Osteoarthritis Car - tilage* 2003; 11: 551-60.
57. SOKKA T, KRISHNAN E, HÄKKINEN A, HANNONEN P: Functional disability in rheumatoid arthritis patients compared with a community population in Finland. *Arthritis Rheum* 2003; 48: 59-63.
58. PINCUS T, BROOKS RH, CALLAHAN LF: Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. *Ann Intern Med* 1994; 120: 26-34.
59. KUPER HH, VAN LEEUWEN MA, VAN RIEL PL *et al*.: Radiographic damage in large joints in early rheumatoid arthritis: relationship with radiographic damage in hands and feet, disease activity, and physical disability. *Br J Rheumatol* 1997; 36: 855-60.
60. DROSSAERS-BAKKER KW, KROON HM, ZWINDERMAN AH, BREEDVELD FC, HAZES JM: Radiographic damage of large joints in long-term rheumatoid arthritis and its relation to function. *Rheumatology* 2000; 39: 998-1003.
61. SCOTTDL, SMITH C, KINGSLEY G: Joint damage and disability in rheumatoid arthritis: an updated systematic review. *Clin Exp Rheumatol* 2003; 21: S20-7.
62. STUKI G, STUKI S, BRUHLMANN P, MICHELBA: Ceiling effects of the Health Assessment Questionnaire (HAQ) and its modified version in some ambulatory rheumatoid arthritis patients. *Ann Rheum Dis* 1995; 54: 461-5.
63. VAN DEN ENDE CH, BREEDVELD FC, DIJKMANS BA, HAZES JM: The limited value of the Health Assessment Questionnaire as an outcome measure in short term exercise trials. *J Rheumatol* 1997, 24: 1972-7.
64. BREMANDER ABI, PETERSSON IF, ROOS EM: Validation of the Rheumatoid and Arthritis Outcome Score (RAOS) for lower extremity. *Health Quality of Life Outcomes* 2003; 1: 55.

Appendix: Italian and English versions of the ROAD questionnaire

Le risposte alle domande - *Answers to questions*

0 = Sì, senza difficoltà – *Without any difficulty*
1 = Possibile, con lieve difficoltà - *With slight difficulty*
2 = Possibile, con qualche difficoltà - *With some difficulty*
3 = Possibile, con molta difficoltà - *With great difficulty*
4 = Impossibile - *Unable to do*

Vorremmo che Lei rispondesse alle seguenti domande relative alle Sue normali attività svolte nel corso dell'ultima settimana. - *Please, answer the following questions regarding your usual activities over the past week*

F1 - Funzionalità arti superiori - *Upper Extremity Function*
E'in grado di: - *Are you able to:*
1. Chiudere completamente la mano a pugno? - *Close your hand completely?*
2. Accettare una stretta di mano? - *Accept a hand shake?*
3. Abbottonarsi gli abiti? - *Do up buttons?*
4. Svitare un coperchio di un barattolo già aperto in precedenza? - *Open jars which have been previously opened?*
5. Raggiungere e afferrare un oggetto del peso di circa due chili posto sopra la Sua testa? - *Reach up and take down a 2 Kg object from above your head?*

F2 - Funzionalità arti inferiori - *Lower Extremity Function*
6. Stare in piedi in posizione eretta? – *Stand up?*
7. Camminare su un terreno piano? - *Walk on flat ground?*
8. Salire un piano di scale (esempio 5 gradini)? - *Climb up five steps or stairs?*
9. Salire e scendere dalla macchina? - *Get into and out of a car?*

F3 – Attività della vita quotidiana /lavorativa - *Activities of daily living/work*
10. Lavare ed asciugare tutto il corpo? - *Wash and dry your body?*
11. Fare attività vigorose quali trasportare oggetti o borse pesanti? - *Run errands and shop?*
12. Svolgere un lavoro retribuito o attività domestiche? – *Are you still able to do housework or/and your paid job?*

ROAD SCALE CONTENTS AND SCORING.

Scale	Number of items	Raw score range	Normalization
1. Upper extremity function	5	0-20	*S x 0.5
2. Lower extremity function	4	0-16	S x 0.625
3. Activity of daily living/work	3	0-12	S x 0.833

*S=Added raw score values.