

**Functional Status and Health-related Quality
of Life Assessment in Patients with
Rheumatoid Arthritis**

Dr Graeme Hawthorne

Centre for Health Program Evaluation
Department of General Practice and Public Health
The University of Melbourne

Dr Rachelle Buchbinder

Melbourne Rheumatology Group
Cabrini Medical Centre
Department of Epidemiology and Preventive Medicine
Monash University

Dr John Defina

Sirius Research
Melbourne, Australia

CENTRE PROFILE

The Centre for Health Program Evaluation (CHPE) is a research and teaching organisation established in 1990 to:

- undertake academic and applied research into health programs, health systems and current policy issues;
- develop appropriate evaluation methodologies; and
- promote the teaching of health economics and health program evaluation, in order to increase the supply of trained specialists and to improve the level of understanding in the health community.

The Centre comprises two independent research units, the Health Economics Unit (HEU) which is part of the Faculty of Business and Economics at Monash University, and the Program Evaluation Unit (PEU) which is part of the Department of General Practice and Public Health at The University of Melbourne. The two units undertake their own individual work programs as well as collaborative research and teaching activities.

PUBLICATIONS

The views expressed in Centre publications are those of the author(s) and do not necessarily reflect the views of the Centre or its sponsors. Readers of publications are encouraged to contact the author(s) with comments, criticisms and suggestions.

A list of the Centre's papers is provided inside the back cover. Further information and copies of the papers may be obtained by contacting:

The Co-ordinator
Centre for Health Program Evaluation
PO Box 477
West Heidelberg Vic 3081, Australia
Telephone + 61 3 9496 4433/4434 **Facsimile** + 61 3 9496 4424
E-mail CHPE@BusEco.monash.edu.au
Web Address <http://ariel.unimelb.edu.au/chpe/>

ACKNOWLEDGMENTS

This project was made possible by a project grant from the Victorian Health Promotion Foundation.

The Health Economics Unit of the CHPE is supported by Monash University.

The Program Evaluation Unit of the CHPE is supported by The University of Melbourne.

Both units obtain supplementary funding through national competitive grants and contract research.

The research described in this paper is made possible through the support of these bodies.

AUTHOR ACKNOWLEDGMENTS

We would like to acknowledge the support given to this project by the project participants, who gave of their time so willingly. Ethics approval was provided by the Cabrini Hospital Ethics Committee.

We would like to thank Ms Mahshid Nasserri for her valuable work in database construction, and the preliminary analyses.

Dr Hawthorne's position at The University of Melbourne is funded by the Victorian Consortium for Public Health.

The project was funded by Wyeth Pharmaceuticals.

Table of Contents

1	Introduction	1
1.1	Study objectives.....	3
2	Research Method	3
2.1	Research method.....	3
2.2	Study instruments.....	4
2.3	Data analysis.....	6
3	Findings	6
3.1	Participant details.....	6
3.2	Instrument data.....	7
3.3	Validity and reliability of instruments.....	10
4	Discussion	12
5	Conclusion	15
	References	17

List of Tables

Table 1	Participant details (n=139)*.....	20
Table 2	Participants' health status.....	21
Table 3	Baseline HAQ for all respondents by demographic characteristics and health status.....	23
Table 4	Baseline EQ5D Utility scores for all respondents by demographic characteristics and health status.....	25
Table 5	Baseline patient global assessment of disease severity for all respondents by demographic characteristics and health status.....	27
Table 6	Baseline AQoL for all respondents by demographic characteristics and health status.....	29
Table 7	Baseline SF-36 summary scales by demographic characteristics and health status.....	31
Table 8	Correlations between the instruments at baseline.....	32
Table 9	Internal consistency as a measure of reliability.....	36
Table 10	Baseline and retest correlations.....	37
Table 11	The HAQ questionnaire data.....	42
Table 12	Validation of the AQoL, EQ5D and SF-36 by levels of arthritis disability (HAQ recorded).....	43

List of Figures

Figure 1	Year with arthritis.....	22
Figure 2	HAQ scores.....	22
Figure 3	EQ5D utility scores.....	24
Figure 4	Patient global assessment of arthritis symptom severity.....	26
Figure 5	AQoL scores.....	28
Figure 6A	SF-36 Mental health summary scale.....	30
Figure 6B	SF-36 Physical health summary scale.....	30
Figure 7A	HAQ vs. AQoL.....	33
Figure 7B	HAQ vs. EQ5D Utility.....	33
Figure 7C	EQ5D VAS.....	34
Figure 7D	HAQ vs. SF-36: PCS.....	34
Figure 7E	HAQ vs. SF-36: MCS.....	35
Figure 8A	Test-retest scatterplot: patient global.....	38
Figure 8B	Test-retest scatterplot: AQoL.....	38
Figure 8C	Test-retest scatterplot: EQ5D Utility.....	39
Figure 8D	Test-retest scatterplot: EQ5D VAS.....	39
Figure 8E	Test-retest scatterplot: HAQ.....	40
Figure 8F	Test-retest scatterplot: SF-36 PCS.....	40
Figure 8G	Test-retest scatterplot: SF-36: MCS.....	41

1 Introduction

This study was undertaken to investigate the relationship between functional ability in patients with rheumatoid arthritis (RA) and their level of health-related quality of life (HRQoL). Information on HRQoL is particularly useful in evaluating health outcomes as collecting self-report HRQoL is the only direct way of obtaining the patients' perspective. As such it may be useful in evaluating models of care for RA.

HRQoL can be defined and measured in a number of ways.

Firstly, the HRQoL measure may focus on aspects of health status that are specific to the area of primary interest, for example a specific disease, population of patients or problem. The rationale for this approach is the purported increased responsiveness of the disease-specific instrument that may result from including only important aspects of HRQoL that are relevant to the disease and patients being studied (Guyatt, Feeny et al. 1993). For example, one of the most widely used disease-specific arthritis tools is the Health Assessment Questionnaire (HAQ) (Fries, Spitz et al. 1980). This considers aspects of physical function that are deemed to be important in daily life; such as dressing, eating, mobility or the amount of pain a person is in. The HAQ has been shown to be responsive in RA clinical trials. The disadvantages of disease-specific instruments are: (a) they cannot be used to compare the effectiveness of interventions across different diseases; and (b) mostly they do not provide preference estimates precluding their use in economic evaluation.

A more inclusive definition of HRQoL extends to how well the patient is coping physically, socially and mentally. This infers that HRQoL provides a broad picture of how patients are feeling in general about their overall life. Multidimensional health status scales or health profiles attempt to measure the salient aspects of HRQoL, such as functional status, symptoms, mood and sense of well-being. For example the SF-36 comprises eight scales including physical function, role physical, bodily pain, general health, vitality, social function, role emotional, and mental health (Ware, Kosinski et al. 1994). Like disease-specific HRQoL measures, multidimensional health status scales are psychometric measures that assign a "value" for a particular health state according to the item response chosen. These are usually weighted (McDowell and Newell 1987). Generic profile instruments, however may not be responsive to changes in specific conditions. Although they allow comparability between interventions for different diseases, they do not capture preferences for health states, which prevents their use in cost-utility analysis (Suarez-Almazor 2000).

HRQoL measures incorporating the value that a person places upon a particular health state are called 'utility' or 'preference' measures (Guillemin 1999). Utilities are usually defined as preferences for health states; they may be elicited using one of five standard methods: visual analog rating (VAS) scale¹, standard gamble (SG), time-trade off (TTO), person-trade off (PTO) or magnitude estimation (ME)². In addition the evaluation may be of disaggregated health states

¹ Although VAS-techniques are widely used, they fail to meet the requirements for preference elicitation since there is no trade-off (between life-length and life-quality) or uncertainty (in the outcome) in the evaluation. Thus evidence of equal reliability and validity between VAS and the preference techniques is insufficient to establish the appropriateness of VAS for the elicitation of preferences.

² For a discussion of these different methods the reader is referred to Drummond (Drummond, O'Brien et al. 1998).

(as in MAU-type descriptive systems) or through the use of vignettes describing complex health states. The literature suggests that the various possible combinations do not produce directly comparable evaluations. In order to use utilities to produce QALYs (quality adjusted life years), three conditions must be met to satisfy the underlying theory: (a) the weights must be based on preferences; (b) these preferences must be anchored on a 0.00–1.00 scale where 0.00=death and 1.00=good health (note that this allows health states worse than death, eg. in the case of a person committing suicide he/she has made the evaluation that it is preferable to be dead than alive); and (c) the preferences should form an interval scale.

Utility scores reflect both the health status and the value of that health status. The obtained utilities can be used to calculate QALYs; these are simply the utility score multiplied by the time that a person spends in that health state. Where QALYs are calculated for several interventions for different diseases, the cost-per-QALY gained can be computed. Normally the intervention with the lowest cost-per-QALY is the preferred intervention because it provides the best health outcomes for the lowest cost.

Regarding whose values should be used in utility instruments, it is generally accepted that for economic evaluations providing information to assist with health policy decisions and the allocation of resources, the preference weights should reflect the values of the public (Gold 1996). For example, the *Assessment of Quality of Life* (AQoL) instrument utility weights were elicited from a representative sample of the Australian population, using the TTO procedure (Hawthorne, Richardson et al. 2000).

Recommendations of a core set of outcomes to be used in clinical trials of RA were prepared in 1993 by the American College of Rheumatology (ACR) and an international conference committee (Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT)) (Felson, Anderson et al. 1993; Tugwell and Boers 1993). These were formed after review of the evidence from previous trials that specific measures meet required criteria for evaluative instruments. The proposed measures included tender and swollen joint counts, patient pain assessment, patient and physician global assessments of efficacy, patient-assessed physical function (a “disease-specific” measure of HRQoL) and level of an acute phase reactant. None of these directly captures the patient’s perspective of their overall HRQoL.

Subsequently, it was proposed that a generic HRQoL instrument be included in the core set as well (Wells, Boers et al. 1999). This is becoming increasingly important as financial restraints on health care have made it necessary to compare the cost effectiveness of different interventions for different diseases. Concern about the validity of generic measures of HRQoL as outcome measures in rheumatology led to the creation of an international task force (the OMERACT/ILAR Task Force on Generic Quality of Life) specifically set up to evaluate them further (Wells, Boers et al. 1999).

Given the paucity of validation studies, we set out to provide some additional data concerning the validity and reliability of both generic profile and utility HRQoL outcome measures in RA. Our hypothesis was that the study results would support the use of HRQoL outcome measures as being able to capture the patients’ perspective in clinical trials of arthritis treatments. This is important in developing a more patient relevant perspective on clinical trial outcomes. The results of the study will be of value in establishing the medical importance of different therapies to patients themselves, and in evaluating the likely cost-utilities of those therapies.

1.1 Study objectives

The overall objectives were to:

- Establish the level of correlation between functional status (ie. how well the patient is functioning with respect to activities of daily living) and the overall level of a patient's HRQoL. This was to be achieved through the direct comparison of several different functional status, health status and utility-based quality of life measures in a cross-sectional survey of a cohort of RA patients with stable disease attending an Australian community-based private rheumatology practice. It was expected that these patients would be a representative sample of RA sufferers with a wide spectrum of disease severity.
- Provide additional data on the validity and reliability of the different instruments in this setting.

2 Research Method

2.1 Research method

Design

Data were collected via a mailed battery of instruments to 150 patients with RA receiving routine medical care.

Test-retest reliability of instruments was assessed in 51 cases randomly selected for follow-up at 2 weeks. Again, mail administration of the instrument battery was used.

A convenience sample of 50 patients attended the Cabrini Hospital rooms to complete the Problem Elicitation Technique (PET), as this requires a trained interviewer for administration (see below). Consecutive patients who returned their questionnaires and were due for review were approached to attend.

Participants

Participants were recruited from the community-based private rheumatology practices of the clinical investigator (RB). A selected sample of 200 RA patients was identified on the basis of the following inclusion criteria, with RA severity across the spectrum from mild to severe levels of disability:

- current patients attending Cabrini Medical Centre or the Dandenong rooms of the clinical investigators (defined as attended within the previous six months);
- age > 18 years;
- ability to read and speak English adequately;

- able to attend Cabrini Medical Centre (applies to PET subset of patients).

The withdrawal criteria were:

- desire to withdraw from participation in the survey at any time;
- urgent need for surgery or other extensive medical diagnosis or treatment in the 2 weeks in between the initial and retest questionnaires for the test-retest sub-sample;
- otherwise considered advisable in the opinion of the clinical investigator.

For the purposes of psychometric analysis the study sample was calculated to include 150 cases; it was assumed that there could be up to a 25% refusal rate, hence the oversampling to 200 potential subjects in case more than 150 subjects were needed to obtain a satisfactory level of response.

The first 150 subjects in the sample of 200 patients were sent a mailed questionnaire together with a covering letter explaining the purpose of the study and consent form. There were 139 out of 150 patients who completed the questionnaire (response rate = 92.7%). This was considered a sufficient response rate and no further patients were approached.

No patients withdrew from the study and all patients who were approached to attend for administration of the PET completed PET questionnaires.

2.2 Study instruments

The outcome measures and instruments used in this study are summarised and described below:

Measure :	Arthritis severity	Disease-specific health status	Overall health status	Utility of health status
Instrument :	Patient global assessment of severity of symptoms	HAQ Disability Index (fixed item) PET (patient preference)	SF-36 EQ-5D feeling thermometer	EQ-5D utility AQoL

Patient global impression of disease severity

An 11-point rating scale (0 – 10) referred to as the *Patient Global Impression of Severity of Arthritis* was used to determine the respondent’s global assessment of the extent to which arthritis affects them.

The scale range was 0 representing an asymptomatic state to 10 representing severe symptoms. The higher the score the worse the effect of arthritis; i.e. the lower the score the better the health state.

Health Assessment Questionnaire (HAQ)

The HAQ is an arthritis-specific HRQoL measure (Fries, Spitz et al. 1980) that has had wide applicability in RA studies. It focuses on those aspects of a patient’s medical condition that especially involve physical function and pain and comprised two components – the HAQ Disability

Index and HAQ Pain. The HAQ Disability Index is comprised of eight weighted scales³ measuring a person's ability to dress, arise from a chair or bed, eat, walk, perform basic toileting (hygiene), reach, grip and perform normal activities. Each scale, referred to as a component, is scored on a 4-point scale, from 0 = no difficulty in performing the task through to 3 = unable to do the task. Each component scale is then weighted by the use made of a mechanical aid or assistance given by a helper.

The eight component scales are summed and the sum divided by eight to provide an aggregate score between 0 and 3, where 0 = no difficulty and 3 = unable to perform (Fries, Spitz et al. 1980). Thus a high HAQ Disability Index score indicates loss of functional capacity and the lower the HAQ score the better the functional state.

The HAQ Pain scale measures pain on a horizontal VAS, with “no pain” at one end (scored 0) and “severe pain” at the other (scored 10).

Problem Elicitation Technique (PET)

The PET is a disease-specific patient preference HRQoL measure which was developed from the *McMaster Arthritis Patient Preference Disability Questionnaire (MACTAR)* (Tugwell, Bombadier et al. 1987). In contrast to conventional questionnaires, the PET asks individuals to identify their own problems related to the disease that they would most like to see improve as a result of therapy. Patients are asked about the level of difficulty, severity or frequency, as well as the importance of each identified problem. By focusing only on those disabilities with the potential to change and judged to be important by the patient, the hypothesis is that “noise” created by including items not relevant to the specific patient will be reduced, thereby enhancing the responsiveness to detect change (Buchbinder, Bombardier et al. 1995). The PET requires a trained interviewer for completion. It has been used in RA clinical trials and has established responsiveness (Buchbinder 1991). A report on the PET will be issued separately.

SF-36

The *SF-36* is a generic multi-dimensional health status measure that comprises eight scales measuring Physical Function, Role Physical, Bodily Pain, General Health, Vitality, Social Function, Role Emotional, and Mental Health. These scales are collapsed into two summary scales, the Physical Component Scale (PCS) and the Mental Health Component Scale (MCS). The scores on the eight scales are transformed into 0–100 point scales, where 100 represents the best health state. The PCS and MCS are presented as T-scores, with a mean of 50 and a standard deviation of ± 10 (Ware, Kosinski et al. 1994).

The SF-36 has been widely used for many health conditions, including RA. This enables results from the present survey to be compared with published data on the SF-36. These data include those from the National Health Survey; including values for persons suffering from RA (Hill, Parsons et al. 1999).

This report only presents results at the PCS and MCS levels due to sampling limitations.

³ The scales are actually ‘items’: each comprises a stem and responses.

EQ5D (EuroQol) Utility and Feeling Thermometer

The *EQ5D* — formerly the *EuroQoL* — is a utility instrument consisting of 5 items, each of which has 3 ordinal response levels. In addition, it includes a visual analogue self-rating scale, the Feeling Thermometer (EuroQol Group 1990).

The utility items measure Mobility, Self-care, Usual Activities, Pain/Discomfort and Anxiety/Depression. Where these are reported separately, the EQ5D can be used as a health profile instrument. More usually it is used to calculate utility values (EQ5D Utility). The utility weights were obtained from a representative sample of the UK population, using the TTO. The utilities are computed using a regression model in which each item level is considered. The range of scores provides for values from –0.59 to 0.00 (death) to 1.00 (full health) (Dolan, Gudex et al. 1995). The higher the score, the better the HRQoL.

The visual analogue self-rating scale, the *EQ5D Feeling Thermometer* is a global health status measure. Patients indicate how good or bad their own health is on a 0–100 vertical VAS, where 0 is worst imaginable health state and 100 is best imaginable health state.

AQoL

The *Assessment of Quality of Life* (AQoL) is a newly developed HRQoL utility instrument. It was designed to overcome some of the limitations of earlier instruments (Hawthorne and Richardson 1995). The AQoL descriptive system comprises 12 items in 5 dimensions. Item responses are all ordinal scales with four levels per item. The dimensions are Illness, Independent Living, Social Relationships, Physical Senses and Psychological Wellbeing (Hawthorne, Richardson et al. 1999). The utility weights were derived from an Australian population sample using TTO. During the calculation of the utility index, the Illness dimension score is not used. A multiplicative function is used to combine the remaining four dimensions into the utility index. The range of scores is between –0.04 (health states worse than death) to 0.00 (death) to 1.00 (full health) (Hawthorne, Richardson et al. 2000). The higher the score the better the HRQoL.

2.3 Data analysis

Data analyses were conducted in SPSS (SPSS 1998).

Due to the number of analyses undertaken and the restricted number of respondents, to control for Type I errors the significance level was set at $\alpha = 0.01$.

3 Findings

The findings are presented in three parts.

3.1 Participant details

Participant details are shown in Table 1. This provides details of participants' gender, age, marital status, education level and work status. As shown in the table, there were no significant differences between the baseline completers and the test-retest completers on any of the demographic measures. The data show that 80% of participants were female, that the mean age was 58 years, that 70% of participants were married, that 97% had completed high school and that 33% were retired.

Table 2 shows the self-reported health status of participants. Based on the EQ5D's VAS scale, the mean rating of health was 68/100; and consistent with this 37% of respondents rated their health on the SF-36's general health scale as being 'good'; only 35% rated their health status as 'fair' or 'poor'.

On average, participants reported that they had suffered from arthritis for 10.4 years, although there were extreme differences in time periods ranging from just 0.5 through to 44 years. The data are shown graphically in Figure 1, which reveals a skewed distribution. The average pain severity was 4.1 on the HAQ Pain scale, where the higher numbers indicated greater pain.

Finally, 81% of participants reported they did not receive any form of assistance; 19% reported receiving assistance from spouses, children, friends, cleaner/home helpers, etc. The details are given in the notes to Table 2.

When these data were examined by baseline and test-retest completers no significant differences were reported for any of the measures in Table 2.

3.2 Instrument data

Details of the instrument data are given in Tables 3, 4, 5 and 6, where the data are broken down by the demographic variables described in the previous section.

HAQ

Figure 2 shows the distribution of HAQ scores, which are clearly skewed; most respondents obtained less severe scores. Table 3 shows HAQ scores broken down by the various demographic and health status variables. There were significant differences by:

- gender: women obtained higher HAQ scores indicating worse functioning ability;
- the EQ5D VAS health rating: lower VAS scores were associated with higher HAQ scores, indicating that as health declined functional ability also declined;
- the SF-36 general health question: those reporting excellent/very good health obtained the lowest HAQ scores, whereas those reporting fair/poor health obtained the highest HAQ scores;
- reported pain level: the more severe the pain the greater the loss of function as measured by the HAQ;
- having assistance: those with assistance obtained significantly worse (higher) HAQ scores.

EQ5D

Details of the EQ5D are shown in Table 4. The distribution of scores provided in Figure 3 shows a marked skew towards the lower boundary of the EQ5D, and that most respondents obtained scores in the range 0.60–0.80.

When analysed by the demographic and health status variables, it was revealed that significantly different scores were obtained on:

- the EQ5D VAS health rating: lower VAS scores were associated with lower EQ5D scores, indicating that as health declined HRQoL also declined;
- the SF-36 general health question: those reporting excellent/very good health obtained the highest EQ5D scores, whereas those reporting fair/poor health obtained the lowest EQ5D scores;
- reported pain level: the more severe the pain the greater the loss of HRQoL as measured by the EQ5D;
- having assistance: those with assistance obtained significantly lower EQ5D scores.

Patient global impression of disease severity

Figure 4 shows the data distribution for patient global impression of disease severity. This reveals that the data were normally distributed, unlike those for any of the other instruments. The details of the patient global impression of disease severity are given in Table 5.

The only variables for which there were significant differences were:

- the EQ5D VAS health rating: lower VAS scores were associated with higher patient global scores, indicating that as health declined respondents' impression of the severity of arthritis increased;
- the SF-36 general health question: those reporting excellent/very good health obtained the lowest patient global scores, whereas those reporting fair/poor health obtained the highest patient global scores;
- reported pain level: the more severe the pain the greater the reported severity of arthritis as shown by higher patient global scores;
- having assistance: those with assistance obtained significantly higher patient global scores.

AQoL

The distribution of AQoL utility scores is shown in Figure 5. This suggests the data were bimodal, with one cohort of respondents with a high HRQoL clustered around 0.90, and a second cohort reporting a much poorer HRQoL clustering around a value of 0.50.

As shown in Table 6, there were significant differences in AQoL scores by:

- gender: women obtained significantly lower scores when compared with men;

-
- the EQ5D VAS health rating: lower VAS scores were associated with significantly lower AQoL scores, indicating that as health declined so did HRQoL;
 - the SF-36 general health question: those reporting excellent/very good health obtained the highest utility scores, whereas those reporting fair/poor health obtained the lowest AQoL scores;
 - years with arthritis: there was a modest correlation between the length of time since the onset of arthritis and AQoL scores. The longer the time with arthritis the lower the reported HRQoL;
 - reported pain level: the more severe the pain the greater the reported loss of HRQoL as shown by lower AQoL scores;
 - having assistance: those with assistance obtained significantly lower AQoL scores.

SF-36

Figures 6A & 6B show the data distributions for the two SF-36 summary component scales, the MCS and PCS. These reveal that scores on the MCS were negatively skewed, and that there was a convergence of scores in the range 50–65. A different pattern was evident for the PCS, where scores were fairly evenly spread around the range 20–50.

Table 7 shows the breakdown and analysis of SF-36 MCS and PCS scores. For the scales significant differences were reported for:

- the EQ5D VAS health rating: lower VAS scores were associated with significantly lower PCS and MCS scores, indicating that as perceived health declined so did both physical and mental health status;
- the SF-36 general health question: those reporting excellent/very good health obtained the highest physical and mental health status scores, whereas those reporting fair/poor health obtained the lowest physical and mental health status scores;
- years with arthritis: there was a modest correlation between the length of time since the onset of arthritis and physical health status scores. The longer the time with arthritis the lower the reported health status. There was no significant association between years with arthritis and mental health status;
- reported pain level: the more severe the pain the greater the reported loss of physical and mental health status as shown by lower PCS and MCS scores;
- having assistance: those with assistance obtained significantly lower physical health status scores. There was no significant association between having assistance and mental health status.

3.3 Validity and reliability of instruments

Examination of the validity and reliability of the different instruments was undertaken through correlation, reliability and test-retest analyses.

Correlation analysis between the instruments

Table 8 shows the correlations between the different measures. The correlation between the patient global impression of disease severity and the HAQ was 0.50; suggesting these two measures of arthritis are measuring different things, albeit related.

The highest correlation was between the AqoL and the HAQ ($r = -0.76$). Other high correlations were between the SF-36 PCS and the HAQ ($r = -0.74$); the EQ5D utility and the patient global ($r = -0.73$); and the AqoL with the SF-36 PCS ($r = -0.72$). The lowest correlations were between the two SF-36 scales ($r = 0.12$, NS); and between the SF-36 MCS and the HAQ ($r = -0.29$) and also with the patient global ($r = -0.29$).

Scatterplots of the HAQ versus the AqoL, EQ5D utility, EQ5D VAS, SF-36 MCS and SF-36 PCS are given in Figures 7A–E.

Reliability analysis

The internal consistency of each of the instruments was examined, using the baseline data and again at follow-up. For all Cronbach α analyses weighted items were used, based on instrument developers' algorithms. The results are presented in Table 9. This shows that the Cronbach α s at baseline were satisfactory⁴ for the AqoL, EQ5D Utility and HAQ. The Cronbach α s were unsatisfactory for both SF-36 scales. At follow-up, the Cronbach α for the HAQ was satisfactory, those for the other instruments were unsatisfactory.

Test-retest analysis

Table 10 shows the Spearman test-retest correlations⁵. The correlations for the AqoL, HAQ, SF-36 MCS and SF-36 PCS were all within the accepted range for test-retest reliability. The values for patient global and the EQ5D were outside the acceptable range.

⁴ The range of values generally recognised as implying acceptable internal consistency is between 0.70–0.90, although there is controversy over the precise level which would be unacceptable. Nunnally (1967) stated that values as low as 0.50 were acceptable; by 1978 Nunnally was arguing for 0.70 (Pedhazur and Schmelkin 1991). Low values indicate poor correlations, suggesting unreliable measurement (ie. that the different items are measuring different things, and that there is a high proportion of unique variance associated with each item) and very high values suggest redundancy in the measurement (ie. that the different items are really measuring the same thing and that there is a very little variance associated with each item) (Cortina 1993).

⁵ Most authors refer to reliability being demonstrated with Pearson correlations, and argue that the accepted values for demonstrating reliability are test-retest correlations of >0.80 . For identical forms Anastasi argued the correlation should be about 0.90 or higher (Anastasi 1976). In the present study, because of the skewed data distributions, Pearson correlations produce spuriously high estimates. Spearman correlations were therefore used.

Scatterplots of the baseline and retest data are shown in Figures 8A–G.

In-depth analysis of the HAQ

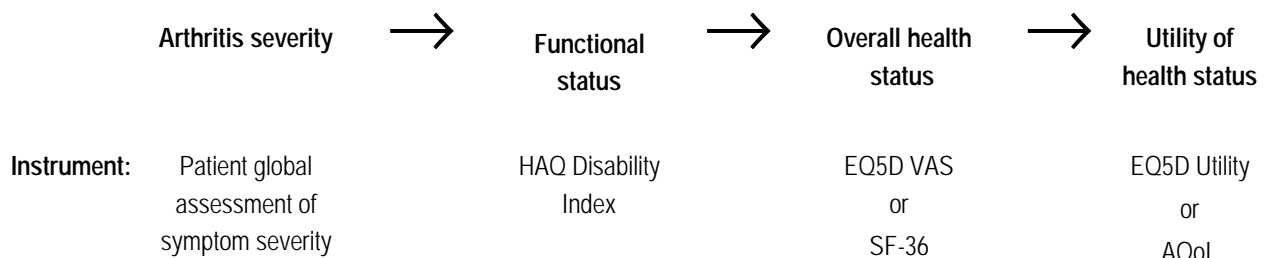
Table 11 provides detailed test-retest data for the components of the HAQ, based on examining each item or scale using the standard measure of agreement between two observers, kappa. In this analysis, the test and retest are treated as the observers. Also shown in the table the percentage of cases where there was complete agreement on the test-retests was from 73% (reaching) through to 93% (dressing).

The table reveals that the range of kappas⁶ was from 0.55 (gripping) through to 0.89 (dressing). Dressing (0.89) was the only item showing excellent agreement; while gripping (0.55) was the only item with moderate agreement. All other items obtained kappas indicating good agreement between the test and retest.

As these data would suggest, test-retest reliability for the HAQ was excellent ($r_s = 0.93$).

Predicting health-related quality of life

We hypothesized a model (presented below) regarding the relationship between severity of arthritis symptoms, functional status, overall health status and the utility of that health status. Although the analyses above suggest there is moderate correlation between arthritis severity and functional status, the patient global assessment of severity was not particularly reliable when compared with the HAQ as demonstrated by the very ordinary patient global test-retest correlation (0.66)⁷. On the other hand the HAQ test-retest correlation was excellent at 0.93. The HAQ has been widely demonstrated to be a valid measure of the impact of arthritis and we examined whether it can be used to predict the impact of RA on overall health status and the value of that health state according to the model:



Using the HAQ as the predictor, regression analyses showed that:

⁶ Unweighted kappas were used to avoid any intra-responder confounding; essentially this has resulted in a conservative analysis. The accepted values of kappa are: ≤0.20 = poor; 0.21–0.40 = fair; 0.41–0.60 = moderate; 0.61–0.80 = good; and 0.81–1.00 = excellent (Landis and Koch 1977).

⁷ The test-retest was performed 2 weeks apart. It is possible that the disease severity changed over this period although patients were identified as being stable. Alternatively, patients' global impression of the severity of their symptoms of arthritis may reasonably be assumed to fluctuate day-to-day.

- 27% of the EQ5D VAS scores were explained ($r^2 = 0.27$, $F = 51.11$, $df = 1;136$, $p < 0.01$, $\beta = -14.78$; 95%CI: -18.86 , -10.69). The regression equation for predicting EQ5D VAS scores: $S_{VAS} = 81.48 + (-14.78 * HAQ)$.
- 54% of the variation in the SF-36 PCS scores was explained ($r^2 = 0.54$, $F = 151.93$, $df = 1;129$, $p < 0.01$, $\beta = -11.70$; 95%CI: -13.57 , -9.82). The regression equation for predicting SF-36 PCS scores: $S_{PCS} = 48.19 + (-11.70 * HAQ)$.
- 8% of the variation in the SF-36 MCS scores was explained ($r^2 = 0.08$, $F = 11.60$, $df = 1;129$, $p < 0.01$, $\beta = -3.99$; 95%CI: -6.30 , -1.67). The regression equation for predicting SF-36 MCS scores: $S_{MCS} = 53.08 + (-3.99 * HAQ)$.
- 35% of the variation in EQ5D Utility scores was explained ($r^2 = 0.35$, $F = 74.05$, $df = 1;135$, $p < 0.01$, $\beta = -0.19$; 95%CI: -0.24 , -0.15). The regression equation for predicting EQ5D scores: $S_{EQ5D} = 0.86 + (-0.20 * HAQ)$.
- 57% of the variation in AQoL scores was explained ($r^2 = 0.57$, $F = 180.72$, $df = 1;134$, $p < 0.01$, $\beta = -0.23$; 95%CI: -0.26 , -0.19). The regression equation for predicting AQoL scores: $S_{AQoL} = 0.85 + (-0.23 * HAQ)$.

The construct validity of the four generic HRQoL instruments (SF-36, EQ5D VAS, EQ5D Utility and AQoL) was also assessed by mapping scores against recoded HAQ scores. The construct tested was that as HAQ scores increased, there would be ordinal monotonic decreases in AQoL, EQ5D Utility, EQ5D VAS and SF-36 scores. HAQ scores were subdivided into four categories based on even distances within category; the categories were 0.00–0.74 = no difficulty, 0.75–1.49 = some difficulty, 1.50–2.24 = much difficulty, and 2.25–3.00 = unable to do. These categories were based on the original HAQ item responses (see Table 11). The results are given in Table 12.

For the AQoL, EQ5D Utility and EQ5D VAS the scores ordinally decreased as arthritis disability increased. For the SF-36 the results were less certain. For the PCS, measuring physical health status, the scores ordinally decreased as expected. However, for the MCS, measuring mental health status, the mean score for those reporting 'much difficulty' on the HAQ was higher than for those reporting 'some difficulty'.

The intervals for the AQoL were even: 24% between the HAQ 'no difficulty' and 'some difficulty', 27% between 'some difficulty' and 'much difficulty' and 27% between 'much difficulty' and 'unable to do'. For the EQ5D Utility they were 24%, 16% and 25% respectively. The intervals were 17%, 13% and 16% for the EQ5D VAS, 27%, 22% and 17% for the SF-36 PCS, and 12%, -2% and 10% for the SF-36 MCS.

4 Discussion

This two-stage study, comprising a cross-sectional mail survey of a selected sample of 139 of 150 current RA patients, and a follow-up of 51 respondents randomly chosen from the base cohort, was designed to investigate the validity and reliability of several self-reported measures in

arthritis intervention evaluation. Given the research design, the study was essentially an 'observational' or 'pre-experimental' design (Colton 1974).

Although a sample of RA patients was identified from computerized records, there may have been some selection bias in identifying the sample itself, due to ensuring that a broad spread of disease severity was present in the sample for the initial mail-out. It may also be noted that the respondents were all volunteers. However because the demographic and clinical characteristics of the participants are comparable to RA patients in general, this suggests the results may be generalisable to RA patients treated in community care in Australia. There was a spectrum of mild, moderate and severe disease as evident by the distribution of scores for both patient global impression of current symptom severity and functional status. The latter was skewed towards those with mild to moderate disease again suggesting the respondents may be comparable to community-based rheumatology practice. Patients were receiving usual care and their condition was not expected to change over the 2 weeks between the baseline and follow-up data collection, but the respondents may have had other treatments unknown to the researchers. In addition we relied upon the integrity of the respondents to self-complete as honestly as possible. The findings are subject to all the associated limitations (Cook and Campbell 1979).

The HAQ proved to be a reliable and valid measure of overall functional status in RA. It performed well in terms of discriminating between respondents, the scores varying as expected when calibrated against those on the EQ5D VAS, the SF-36 general health question and the reported pain level (Table 3). The internal consistency of the HAQ was satisfactory at both baseline and follow-up, and the test-retest HAQ correlation was excellent (Table 10). Apart from questioning the use of the HAQ in relation to the test-retest kappa for the item "gripping" (Table 11), our findings were in keeping with the results of previous studies (Fries, Spitz et al. 1980; Fries, Spitz et al. 1982; Bombardier, Ware et al. 1986; Buchbinder, Bombardier et al. 1995).

When the HAQ was used as a predictor of HRQoL in terms of overall health status and utility, the proportion of explained variance varied — and not entirely in the hypothesized directions. Given the hypothesized model between functional status, overall health status and utility it was expected that there would be greater explained variance levels with the more proximal indicators (health status) than with the more distal indicators (utility). The results showed that the proportion of explained variance was highest for the AQL and the SF-36 PCS scale (57% and 54% respectively; Table 12). This would suggest that the functional disutility being identified by the HAQ more closely measured physical losses rather than general health or mental health losses (as measured by the EQ5D VAS (an assessment of global health) or the SF-36 MCS (which measures mental health). When the content of the HAQ is examined (Table 11), this interpretation is consistent: the items are all about physical activities of daily living: dressing, rising, eating, walking, hygiene, reaching, gripping and general activities. There is no compelling reason why losses in these areas of a person's health would necessarily imply their mental health or overall global health need be poor (although this may usually be the case).

Regarding the measurement of overall health status among those suffering RA, the two instruments used in this study were the EQ5D VAS and the SF-36. The EQ5D VAS is a single item indicator which correlated moderately with the HAQ at baseline; moderate correlations with the SF-36 scales were also noted at baseline (Table 8). The reliability of the EQ5D VAS as measured by test-retest was just satisfactory whereas for the SF-36 PCS and MCS it was slightly better (Table 10). The correlation at baseline between the EQ5D VAS and the SF-36 PCS was

good, whereas it was moderate for the SF-36 MCS (Table 8). Regarding the more reliable and valid measure, it was thus difficult to choose: on the one hand the EQ5D VAS is a single index, whereas the SF-36 summary scales are combinations of measures; on the other hand the EQ5D VAS exhibited properties which were almost equivalent to those exhibited by the SF-36 PCS scale.

With respect to the two HRQoL utility measures, the EQ5D and the AQL, both showed evidence of internal consistency, although the AQL Cronbach α was higher suggesting the items were more closely related than those in the EQ5D. This was particularly evident on the retest, where the EQ5D Cronbach α was unsatisfactorily low (Table 9). Regarding the test-retest correlations (Table 10), the performance of the EQ5D utility as measured by this test was unsatisfactory, falling below the accepted standards for reliability (Anastasi 1976). This finding was surprising given that the EQ5D comprises just 5 items each with 3 levels whereas the AQL utility score involves 12 items with 4 levels. Under these circumstances it would be expected that the test-retest correlations for the EQ5D would be higher than those of the AQL. This suggested that the AQL would be the preferred instrument.

Apart from correlation analyses and the basic psychometric tests described above, the construct validity of the four generic HRQoL instruments measures was examined based upon the argument that as functional status worsened (increase in HAQ score), health status would also decline. This was the case for the AQL, EQ5D Utility, EQ5D VAS and SF-36 PCS and provides *prima facie* evidence that the AQL, EQ5D Utility, EQ5D VAS and SF-36 PCS are sufficiently discriminating to be used as outcome measures in arthritis programs for capturing patient perspectives. The evidence for the SF-36 MCS is more equivocal: there was evidence of inconsistency in the MCS as arthritis disability increased. The mean score for those reporting 'much difficulty' on the HAQ was higher than for those reporting 'some difficulty'. This is inconsistent: it seems highly unlikely that those reporting greater difficulty dressing, rising, eating, walking, with personal hygiene, reaching, gripping or performing their usual activities would have better mental health when compared with those with fewer task difficulties.

Several studies have performed empirical analyses of the EQ5D in RA. Hurst *et al* assessed the clinimetric properties of the EQ5D in a sample of RA patients stratified according to functional class. They found it valid, responsive to change and sufficiently reliable for group comparisons. They also found that some patients with severe longstanding disease had health states that attracted utility values below zero, ie. from a societal perspective they were regarded as being in states 'worse than death' (Hurst, Jobanputra *et al.* 1994; Hurst, Kind *et al.* 1997). Wolfe and Hawley investigated the properties of the EQ5D in a postal survey of 1372 rheumatic disease patients, including 537 with RA (Wolfe and Hawley 1997). They described a similar distribution of EQ5D utility scores to that found in our study, with particular sparsity between 0.25 and 0.5. This could be attributed to substantial problems in the scaling of individual item responses that force scale compression in the mid-ranges.

A study in Sweden by Kobelt *et al.* examined the correlation between EQ-5D utility values and HAQ score in a cross-sectional outpatient survey. The study showed the mean utility value for each of six health states for different levels of RA (defined by intervals of HAQ score) that the authors used in their Markov economic model (Kobelt, Eberhardt *et al.* 1999). We did not have access to the full Kobelt *et al.* dataset, but as an approximation, we used the midpoint of the HAQ score in defining each health state, along with the mean utility value that was associated with that

health state. We performed a simple linear trend/regression analysis on the data and found that the equation was:

$$S_{EQ5D} = 0.77 + (-0.17*HAQ)$$

This crude approximation closely resembles those derived from the correlation between HAQ score and the EQ5D and the EQ5D Feeling Thermometer in our own work. This therefore suggests a measure of external validation of our own findings.

The EQ5D has also been assessed in a study assessing the potential implications that utility ratings, done by different groups using various elicitation techniques, would have on hypothetical cost-utility analyses of interventions for RA (Suarez-Almazor 2000). In this study, patients' direct utility ratings of their own health using VAS, TTO and SG methods were compared to the indirect utility scores calculated from their EQ5D profile, using the same weights as in our study. The EQ5D utility scores were significantly different to and lower than all other evaluations, with TTO and SG giving the most similar results. Peeters *et al* examined the validity of utilities derived by rating scale and SG, as HRQoL measures in RA in the setting of a clinical trial examining the value of recombinant human erythropoietin. Rating scale utilities were reported to be sensitive to change and correlated well with standard disease activity measures as well as changes in these measures (Peeters, Jongen-Lavrencic *et al.* 1999).

In terms of predicting health status, using stepwise regression Hurst *et al* showed that EQ5D utility values and EQ5D VAS were explained best as a function of pain, disability, disease activity and mood ($r^2 \approx 70\%$), although other variables (side-effects, years of education) were required to explain the visual analogue scores (Hurst, Kind *et al.* 1997). Others have shown that other factors such as coping strategies and illness perceptions also contribute to health outcome in RA (Scharloo, Kaptein *et al.* 1999).

5 Conclusion

Subject to the limitations associated with 'pre-experimental' designs and the small sample size, the findings from this study support the validity and reliability of the HAQ as an outcome measure for use with rheumatoid arthritis interventions. The HAQ was found to be sensitive, valid and reliable.

Turning to the measures of health status, the EQ5D VAS performed very creditably, although its reliability was just satisfactory. The SF-36 provided some interesting findings. The PCS scale performed very creditably, whereas the MCS scale showed some inconsistencies; it should be borne in mind that there is no necessary reason why mental health must be affected by arthritis, although it would seem likely there is a relationship. Given that the SF-36 also provides eight sub-scales covering Physical Function, Role Physical, Bodily Pain, Vitality, General Health, Social Function, Role Emotion and Mental Health⁸ it would be the preferred instrument except in those situations where brevity was at a premium.

Comparing the clinimetric properties of the two utility measures, the AQoL showed greater reliability as measured by test-retest; it possessed higher internal consistency estimates at both test and retest; and it was also the more sensitive of the two utility instruments when compared with the HAQ. However because different methods of eliciting and analysing utilities can have

⁸ These were not reported in this study due to sampling limitations.

substantial implications for economic evaluations of interventions for RA, further research needs to be done assessing the validity and reliability of utility measures in RA.

Overall, the results from this study suggest that the measures used can be used with confidence as outcome measures in studies of RA. Given the study limitations, however, this work needs to be replicated in a further study. Ideally, such a study would involve random sampling (for generalisability), clinical indicators of severity (to ensure participants covered the spectrum of rheumatoid arthritis enabling indicator scores to be mapped against clinical criteria), test-retest on a large enough sample (for greater confidence in the reliability estimates), and interview completion (for verification).

References

- Anastasi A 1976, *Psychological Testing*, New York: Macmillan Publishing.
- Bombardier C, Ware J, Russell I et al 1986, 'Auranofin therapy and quality of life in patients with rheumatoid arthritis: results of a multicentre trial', *American Journal of Medicine*, 36: 565-587.
- Buchbinder R 1991, *Critique of the "Problem Elicitation Technique"*, Wellesley, Wellesley Hospital.
- Buchbinder R, Bombardier C, Yeung M et al 1995, 'Which outcome measures should be used in rheumatoid arthritis clinical trials? Clinical and quality-of-life measures' responsiveness to treatment in a randomized controlled trial', *Arthritis Rheumatism*, 38: 1568-1580.
- Colton T 1974, *Statistics in Medicine*, Boston, Little, Brown and Company.
- Cook T and Campbell D 1979, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Boston, Houghton Mifflin.
- Cortina J 1993, 'What is coefficient alpha? Examination of theory and applications', *Journal of Applied Psychology*, 78: 98-104.
- Dolan P, Gudex C, Kind P et al 1995, *Social Tariff for EUROQoL: Results from a UK General Population Survey*, York, Centre for Health Economics, University of York.
- Drummond M, O'Brien B, Stoddart G et al 1998, *Methods for the Economic Evaluation of Health Care Programmes: Oxford*, Oxford University Press.
- EuroQol Group 1990, 'EuroQol: an new facility for measurement of health-related quality of life', *Health Policy*, 16: 199-208.
- Felson D, Anderson J, Boers M et al 1993, 'The American College of Rheumatology preliminary set of core disease activity measures for rheumatoid arthritis clinical trials', *Arthritis Rheumatism*, 36: 729-740.
- Fries J, Spitz P, Kraines G et al 1980, 'Measurement of patient outcome in arthritis', *Arthritis Rheumatism*, 23: 137-145.
- Fries J, Spitz P and Young D 1982, 'The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales', *Journal of Rheumatology*, 9: 789-793.

-
- Gold M 1996, 'Panel on cost-effectiveness in health and medicine', *Medical Care*, 34S: DS197-199.
- Guillemin F 1999, The value of utility: assumptions underlying preferences and quality adjusted life years, *Journal of Rheumatology*, 26: 1861-1863.
- Guyatt G, Feeny D and Patrick D 1993, 'Measuring health-related quality of life', *Annals of Internal Medicine*, 118: 622-629.
- Hawthorne G and Richardson J 1995, 'An Australian MAU/QALY Instrument: Rationale and Preliminary Results', Proceedings Health Outcomes and Quality of Life Measurement Conference, Canberra.
- Hawthorne G, Richardson J, Day N et al 2000, Construction and Utility Scaling of the Assessment of Quality of Life (AQoL) Instrument, Melbourne, Centre for Health Program Evaluation.
- Hawthorne G, Richardson J and Osborne R 1999, 'The Assessment of Quality of Life (AQoL) Instrument: a psychometric measure of health related quality of life', *Quality of Life Research*, 8: 209-224.
- Hill C, Parsons J, Taylor A, et al 1999, 'Health related quality of life in a population sample with arthritis', *Journal of Rheumatology*, 26: 2029-2035.
- Hurst N, Jobanputra P, Hunter M et al 1994, 'Validity of Euroqol--a generic health status instrument--in patients with rheumatoid arthritis', Economic and Health Outcomes Research Group, *British Journal of Rheumatology*, 33: 655-662.
- Hurst N, Kind P, Ruta D et al 1997, 'Measuring health-related quality of life in rheumatoid arthritis: validity, responsiveness and reliability of EuroQol (EQ5D)', *British Journal of Rheumatology*, 36: 551-559.
- Kobelt G, Eberhardt K, Jonsson L et al 1999, 'Economic consequences of the progression of rheumatoid arthritis in Sweden', *Arthritis and Rheumatism*, 42: 347-356.
- Landis J and Koch G 1977, 'The measurement of observer agreement for categorical data', *Biometrics*, 33: 150-174.
- McDowell I and Newell C Eds. 1987, *Measuring health: a guide to rating scales and questionnaires*, New York, Oxford University Press.
- Pedhazur E and Schmelkin L 1991, *Measurement, Design and Analysis: An Integrated Approach*, Hillsdale, Lawrence Erlbaum.

Peeters H, Jongen-Lavrencic M, Bakker C et al 1999, 'Recombinant human erythropoietin improves health-related quality of life in patients with rheumatoid arthritis and anaemia of chronic disease: utility measures correlate strongly with disease activity measures', *Rheumatology International*, 18: 201-206.

Scharloo M, Kaptein A, Weinman J et al 1999, 'Predicting functional status in patients with rheumatoid arthritis', *Journal of Rheumatology*, 26: 1686-1693.

SPSS 1998, *SPSS for Windows*, Chicago, SPSS Inc.

Suarez-Almazor M 2000, 'Ratings of arthritis health states by patient, physicians and the general public. Implications for cost-utility analyses', OMERACT Conference, Toulouse.

Tugwell P and Boers M 1993, 'Development consensus on preliminary core efficacy endpoints for rheumatoid arthritis clinical trials', *Journal of Rheumatology*, 20: 555-556.

Tugwell P, Bombadier C, Buchanan W et al 1987, 'The MACTAR Patient Preference Disability Questionnaire - an individualized functional priority approach for assessing improvement in physical disability in clinical trials in rheumatoid arthritis', *Journal of Rheumatology*, 14: 446-451.

Ware J, Kosinski M and Keller S 1994, *SF-36 Physical and Mental Health Summary Scales: A User's Manual*, Boston, The Health Institute, New England Medical Centre.

Wells G, Boers M, Shea B et al 1999, 'Sensitivity to change of generic quality of life instruments in patients with rheumatoid arthritis: preliminary findings in the generic health OMERACT study', *Journal of Rheumatology*, 26: 217-221.

Wolfe F and Hawley D 1997, 'Measurement of the quality of life in rheumatic disorders using the EuroQol', *British Journal of Rheumatology*, 36: 786-793.

Tables and Figures

Table 1 Participant details (n=139)*

		<i>Baseline sample</i>		<i>Test-retest sample</i>		<i>All participants</i>		<i>Statistics</i>
		<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	
Gender	Male	18	21%	9	18%	27	20%	$\chi^2 = 0.15, p = 0.70, NS$
	Female	67	79%	40	82%	107	80%	
Age (years)	Mean (sd)	57.7 (12.1)		59.2 (13.1)		58.3 (12.5)		F = 0.43, p = 0.51, NS
	Minimum–Maximum	22–83						
Marital status	Single	7	8%	6	12%	13	9%	$\chi^2 = 1.69, p = 0.64, NS$
	Married	60	69%	37	73%	97	70%	
	Divorced/Separated	10	12%	3	6%	13	9%	
	Widowed	10	12%	5	10%	15	11%	
Education level	Primary (a)	1	1%	4	8%	5	4%	$\chi^2 = 2.61, p = 0.11, NS$
	High (a)	49	59%	33	66%	82	62%	
	Tertiary	33	40%	13	26%	46	35%	
Working status	Fulltime	24	32%	8	20%	32	28%	$\chi^2 = 5.63, p = 0.13, NS$
	Parttime	10	13%	4	10%	14	12%	
	Houseworker (a)	13	13%	15	15%	28	24%	
	Houseworker + help (a)	2	3%	1	3%	3	3%	
	Retired	26	35%	12	30%	38	33%	

Notes:

* Some missing values for individual variables

a = Categories combined during data analysis due to small numbers

Table 2 **Participants' health status**

		<i>Baseline sample</i>		<i>Test-retest sample</i>		<i>All participants</i>		<i>Statistics</i>
		<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	
Health rating (a)	Mean (sd)	67.6 (20.1)		69.2 (23.7)		68.2 (21.4)		F = 0.16, p = 0.69 (b)
	Minimum – Maximum	0.0–100.0						
Health status (c)	Excellent (d)	4	5%	2	4%	6	4%	$\chi^2 = 0.34, p = 0.84$ (f)
	Very good (d)	19	22%	14	28%	33	24%	
	Good	33	38%	18	35%	51	37%	
	Fair (e)	25	29%	33	36%	38	28%	
	Poor (e)	5	6%	4	8%	9	7%	
Arthritis years	Mean (sd)	10.2 (8.8)		10.8 (9.5)		10.4 (9.1)		U = 1761.5, p = 0.63 (g)
	Minimum – Maximum	0.5 – 44.0						
Pain suffered (h)	Mean (sd)	4.1 (2.8)		3.9 (2.8)		4.1 (2.8)		F = 0.42, p = 0.62 (b)
	Minimum – Maximum	0.0 – 10.0						
Assistance	None reported	72	81%	41	80%	113	81%	$\chi^2 < 0.01, p = 0.94$ (f)
	Assistance received (i)	17	19%	10	20%	27	19%	

Notes:

a = Visual analog scale (VAS) from the EQ5D. Scale range: 0.0–100.0 where a higher score indicates better health

b = ANOVA

c = Q1 from the SF-36. Scale range shown in table, where a lower score indicates better health

d = Categories combined during data analysis due to small numbers

e = Categories combined during data analysis due to small numbers

f = Chi-square

g = Mann-Whitney U-test

h = Q47 from the HAQ; scale range: 0.0 – 10.0 where a higher score indicates greater pain

i = Refers to help from spouse (8 cases), child (2), friend (1), cleaner/home help (11), child care (1), gardener (1), council support (2) and allied health professional (1)

Figure 1 Year with arthritis

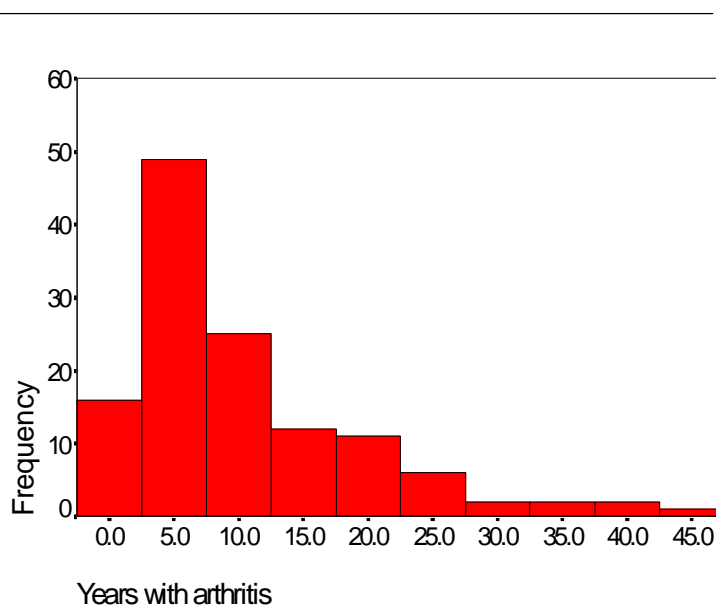


Figure 2 HAQ scores

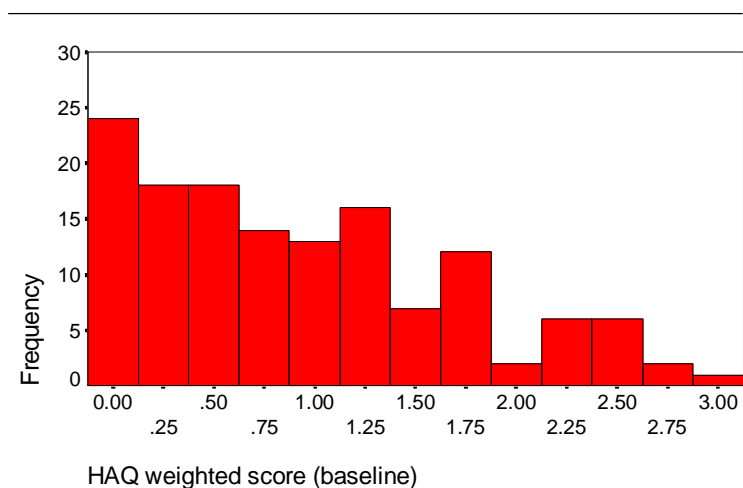


Table 3 Baseline HAQ for all respondents by demographic characteristics and health status

		HAQ		Statistics (a)
		Median	IQR	
Gender	Male	0.25	0.75	$\chi^2 = 11.10, p < 0.01^*$
	Female	0.88	1.13	
Age		$r = 0.19$ (b)		$p = 0.02, NS$
Marital status	Single	0.75	0.94	$\chi^2 = 4.29, p = 0.23, NS$
	Married	0.63	1.00	
	Divorced/Separated	1.50	1.56	
	Widowed	0.63	0.75	
Education level	Primary/High	1.00	1.25	$\chi^2 = 2.54, p = 0.11, NS$
	Tertiary	0.50	1.00	
Working status	Fulltime	0.38	1.09	$\chi^2 = 4.55, p = 0.21, NS$
	Parttime	0.63	0.78	
	Houseworker	0.63	0.88	
	Retired	1.00	1.38	
Health rating (b)		$r = -0.52$		$p < 0.01^*$
Health status	Excellent/Very good	0.25	0.63	$\chi^2 = 36.9, p < 0.01^*$
	Good	0.63	1.13	
	Fair/Poor	1.38	1.25	
Arthritis years		$r = 0.27$ (b)		$p = 0.02, NS$
Pain suffered		$r = 0.49$ (b)		$p < 0.01^*$
Assistance	None reported	0.50	0.88	$\chi^2 = 39.96, p < 0.01^*$
	Assistance received	1.75	1.13	

Notes:

For description of variables see Tables 1 & 2

a = Kruskal-Wallis 1-Way ANOVA unless otherwise stated

b = EQ5D VAS; Pearson correlation

Figure 3 EQ5D utility scores

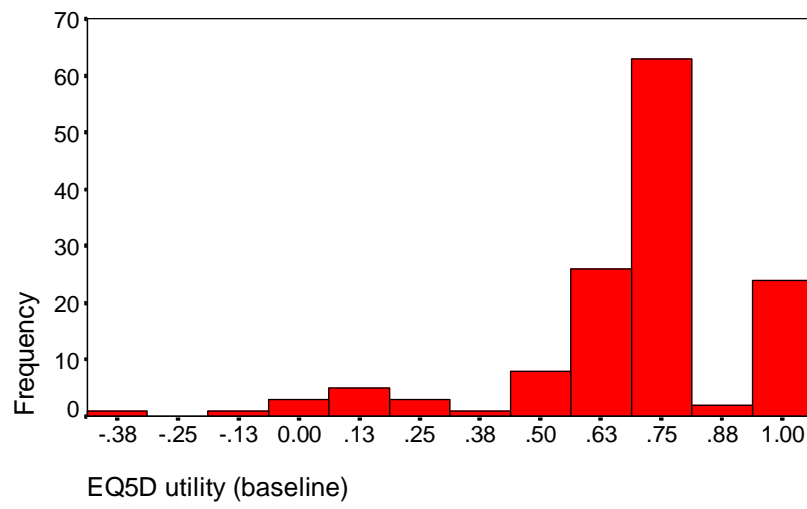


Table 4 Baseline EQ5D Utility scores for all respondents by demographic characteristics and health status

		<i>EQ5D</i>		<i>Statistics (a)</i>
		<i>Median</i>	<i>IQR</i>	
Gender	Male	0.76	0.11	$\chi^2 = 3.84, p = 0.05, NS$
	Female	0.69	0.21	
Age		$r = -0.09$ (b)		$p = 0.32, NS$
Marital status	Single	0.78	0.16	$\chi^2 = 0.02, p = 0.57, NS$
	Married	0.69	0.18	
	Divorced/Separated	0.62	0.31	
	Widowed	0.73	0.21	
Education level	Primary/High	0.69	0.21	$\chi^2 = 0.19, p = 0.67, NS$
	Tertiary	0.69	0.18	
Working status	Fulltime	0.73	0.11	$\chi^2 = 1.94, p = 0.51, NS$
	Parttime	0.69	0.18	
	Houseworker	0.76	0.26	
	Retired	0.69	0.21	
Health rating (b)		$r = -0.70$		$p < 0.01^*$
Health status	Excellent/Very good	0.80	0.27	$\chi^2 = 52.78, p < 0.01^*$
	Good	0.73	0.11	
	Fair/Poor	0.59	0.27	
Arthritis years		$r = 0.10$ (b)		$p = 0.29, NS$
Pain suffered		$r = 0.66$ (b)		$p < 0.01^*$
Assistance	None reported	0.73	0.18	$\chi^2 = 13.88, p < 0.01^*$
	Assistance received	0.60	0.38	

Notes:

For description of variables see Tables 1 & 2

a = Kruskal-Wallis 1-Way ANOVA unless otherwise stated

b = EQ5D VAS; Pearson correlation

Figure 4 Patient global assessment of arthritis symptom severity

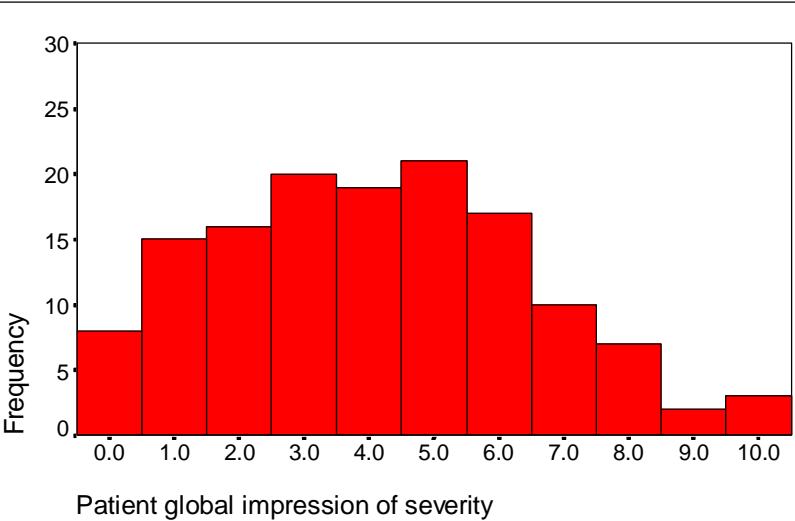


Table 5 Baseline patient global assessment of disease severity for all respondents by demographic characteristics and health status

		<i>Patient Global</i>		<i>Statistics (a)</i>
		<i>Mean</i>	<i>sd</i>	
Gender	Male	3.6	2.5	F = 0.96, p = 0.33, NS
	Female	4.1	2.3	
Age		$r = -0.03$ (b)		p = 0.69, NS
Marital status	Single	4.7	2.5	F = 1.22, p = 0.31, NS
	Married	4.1	2.4	
	Divorced/Separated	4.3	2.6	
	Widowed	3.1	1.7	
Education level	Primary/High	4.2	2.4	F = 0.57, p = 0.45, NS
	Tertiary	3.8	2.3	
Working status	Fulltime	3.7	2.2	F = 1.09, p = 0.36, NS
	Parttime	4.8	2.9	
	Houseworker	3.5	2.1	
	Retired	3.8	2.4	
Health rating (b)		$r = -0.62$		p < 0.01*
Health status	Excellent/Very good	2.5	1.9	F = 31.94, p < 0.01*
	Good	3.6	2.1	
	Fair/Poor	5.8	2.0	
Arthritis years		$r = 0.14$ (b)		p = 0.11, NS
Pain suffered		$r = 0.73$ (b)		p < 0.01*
Assistance	None reported	3.8	2.3	F = 7.30, p < 0.01*
	Assistance received	5.2	2.5	

Notes:

For description of variables see Tables 1 & 2

a = ANOVA unless otherwise stated

b = EQ5D VAS; Pearson correlation

Figure 5 AQL scores

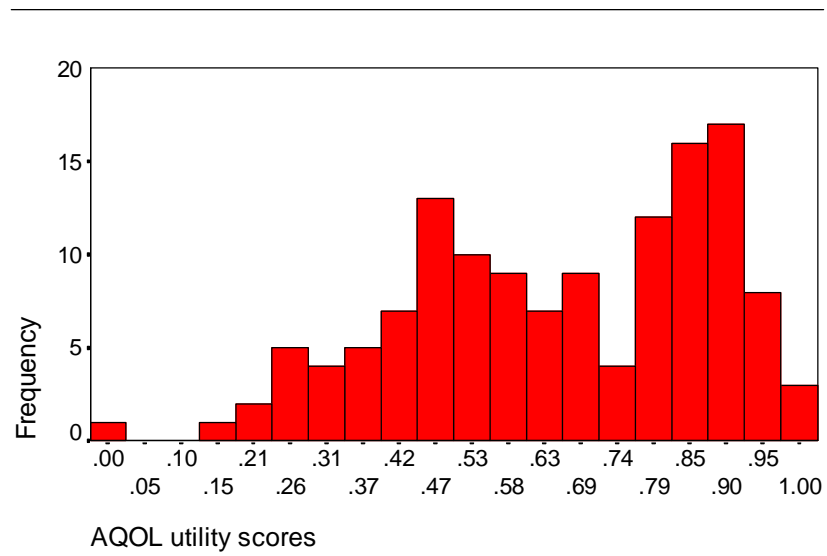


Table 6 Baseline AQoL for all respondents by demographic characteristics and health status

		AQoL		Statistics (a)
		Median	IQR	
Gender	Male	0.84	0.15	$\chi^2 = 8.41, p < 0.01^*$
	Female	0.62	0.38	
Age		$r = -0.19$ (b)		$p = 0.03, NS$
Marital status	Single	0.65	0.26	$\chi^2 = 3.18, p = 0.36, NS$
	Married	0.70	0.38	
	Divorced/Separated	0.46	0.40	
	Widowed	0.71	0.38	
Education level	Primary/High	0.65	0.38	$\chi^2 = 0.63, p = 0.43, NS$
	Tertiary	0.69	0.34	
Working status	Fulltime	0.78	0.33	$\chi^2 = 2.71, p = 0.44, NS$
	Parttime	0.84	0.27	
	Houseworker	0.76	0.29	
	Retired	0.67	0.39	
Health rating (b)		$r = 0.72$		$p < 0.01^*$
Health status	Excellent/Very good	0.87	0.22	$\chi^2 = 44.98, p < 0.01^*$
	Good	0.76	0.32	
	Fair/Poor	0.48	0.28	
Arthritis years		$r = -0.23$ (b)		$p = 0.01^*$
Pain suffered		$r = 0.56$ (b)		$p < 0.01^*$
Assistance	None reported	0.76	0.33	$\chi^2 = 28.13, p < 0.01^*$
	Assistance received	0.43	0.26	

Notes:

For description of variables see Tables 1 & 2

a = Kruskal-Wallis 1-Way ANOVA unless otherwise stated

b = EQ5D VAS; Pearson correlation

Figure 6A SF-36 Mental health summary scale

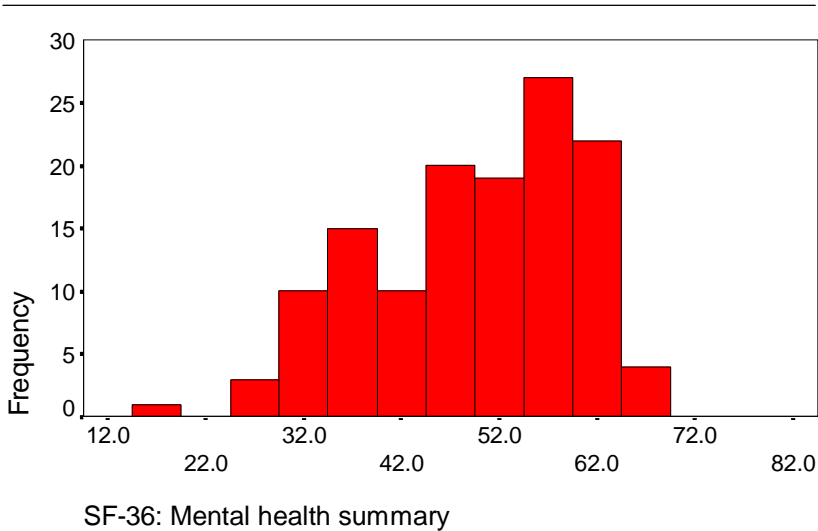


Figure 6B SF-36 Physical health summary scale

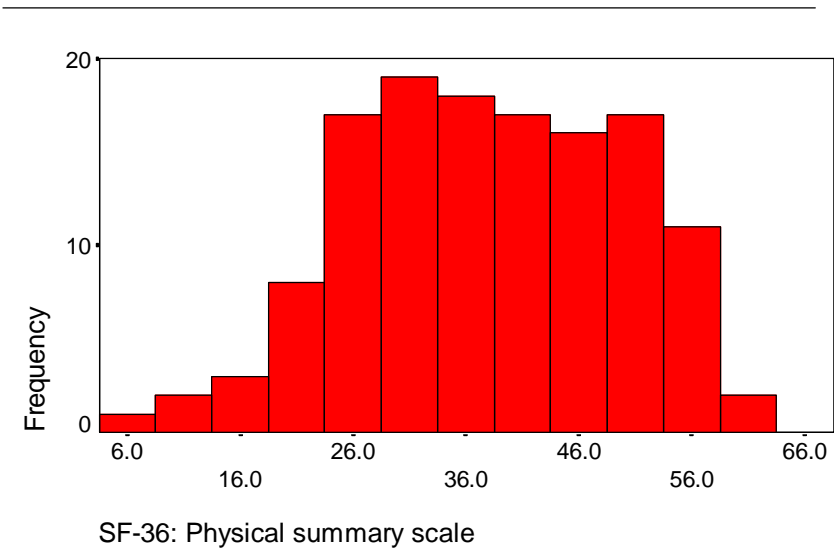


Table 7 Baseline SF-36 summary scales by demographic characteristics and health status

		<i>PCS</i>		<i>MCS</i>	
		<i>Median</i>	<i>IQR</i>	<i>Median</i>	<i>IQR</i>
Gender	Male	45.3	21.7	55.5	12.6
	Female	36.7	17.4	51.6	19.8
		$\chi^2 = 2.27, p = 0.13, NS$		$\chi^2 = 2.51, p = 0.11, NS$	
Age		$r = -0.11, p = 0.20, NS$		$r = 0.03, p = 0.73, NS$	
Marital status	Single	41.9	12.4	53.7	9.1
	Married	36.5	19.6	53.7	17.7
	Divorced/Separated	29.5	17.3	48.6	19.0
	Widowed	44.2	16.5	47.8	15.1
		$\chi^2 = 4.81, p = 0.19, NS$		$\chi^2 = 3.32, p = 0.34, NS$	
Education level	Primary/High	34.6	21.6	52.8	20.0
	Tertiary	42.2	13.5	50.9	11.1
		$\chi^2 = 1.27, p = 0.26, NS$		$\chi^2 = 0.26, p = 0.61, NS$	
Working status	Fulltime	40.7	18.4	52.2	12.3
	Parttime	36.9	20.0	55.6	18.9
	Houseworker	42.3	16.4	53.1	15.1
	Retired	37.7	20.3	48.8	18.6
		$\chi^2 = 2.98, p = 0.40, NS$		$\chi^2 = 3.43, p = 0.33, NS$	
Health rating (a)		$r = 0.68, p < 0.01^*$		$r = 0.43, p < 0.01^*$	
Health status	Excellent/Very good	47.4	11.2	55.6	12.2
	Good	40.6	15.2	54.5	11.5
	Fair/Poor	28.0	8.5	41.4	17.8
		$\chi^2 = 60.98, p < 0.01^*$		$\chi^2 = 24.06, p < 0.01^*$	
Arthritis years		$r = -0.23, p = 0.01^*$		$r = 0.02, p = 0.83, NS$	
Pain suffered		$r = -0.64, p < 0.01^*$		$r = -0.26, p < 0.01^*$	
Assistance	None reported	40.7	18.6	52.7	16.7
	Assistance received	27.0	13.3	48.1	22.6
		$\chi^2 = 21.37, p < 0.01^*$		$\chi^2 = 1.01, p = 0.32, NS$	

Notes:

For a description of the variables, see Tables 1 & 2

Statistics are Kruskal-Wallis 1-way ANOVA unless otherwise stated

a = EQ5D VAS; Pearson correlation

Table 8 Correlations between the instruments at baseline

	<i>Patient Global</i>	<i>AQoL</i>	<i>EQ5D Utility</i>	<i>EQ5D VAS</i>	<i>HAQ</i>	<i>SF-36:MCS</i>
<i>AQoL</i>	-0.55*					
<i>EQ5D Utility</i>	-0.73*	+0.68*				
<i>EQ5D VAS</i>	-0.62*	+0.72*	+0.70*			
<i>HAQ</i>	+0.50*	-0.76*	-0.60*	-0.52*		
<i>SF-36: MCS</i>	-0.29*	+0.42*	+0.35*	+0.43*	-0.29*	
<i>SF-36: PCS</i>	-0.62*	+0.72*	+0.62*	+0.68*	-0.74*	+0.12, NS

Notes:

* = Significant correlation, $p < 0.01$

N. cases: 139

Pearson correlations

Figure 7A HAQ vs. AqoL

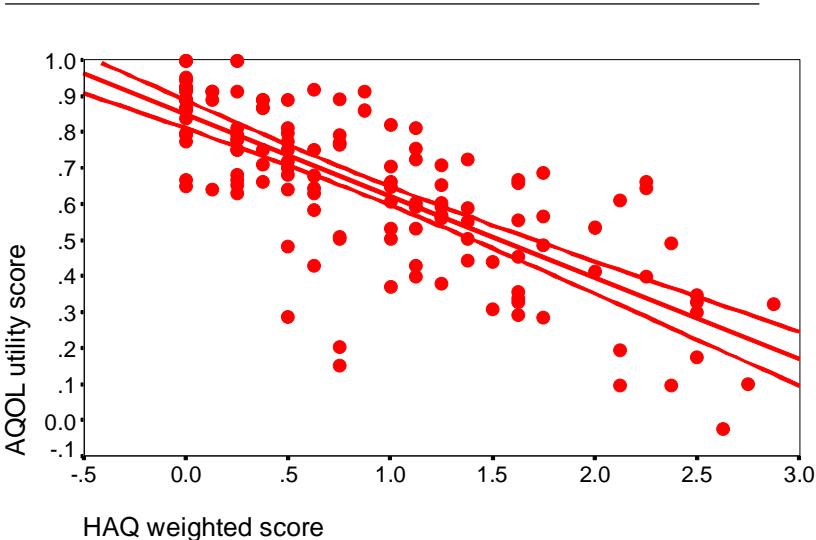


Figure 7B HAQ vs. EQ5D Utility

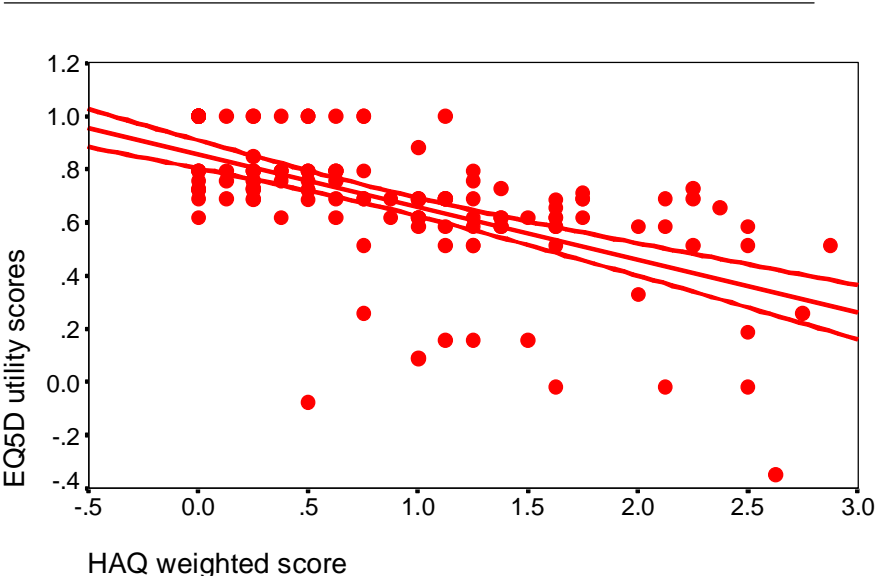


Figure 7C HAQ vs. EQ5D VAS

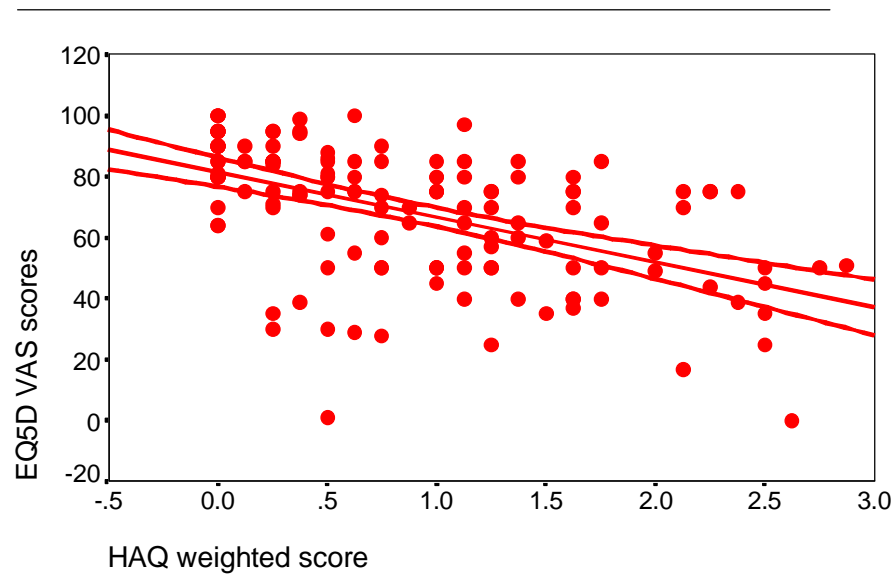


Figure 7D HAQ vs. SF-36: PCS

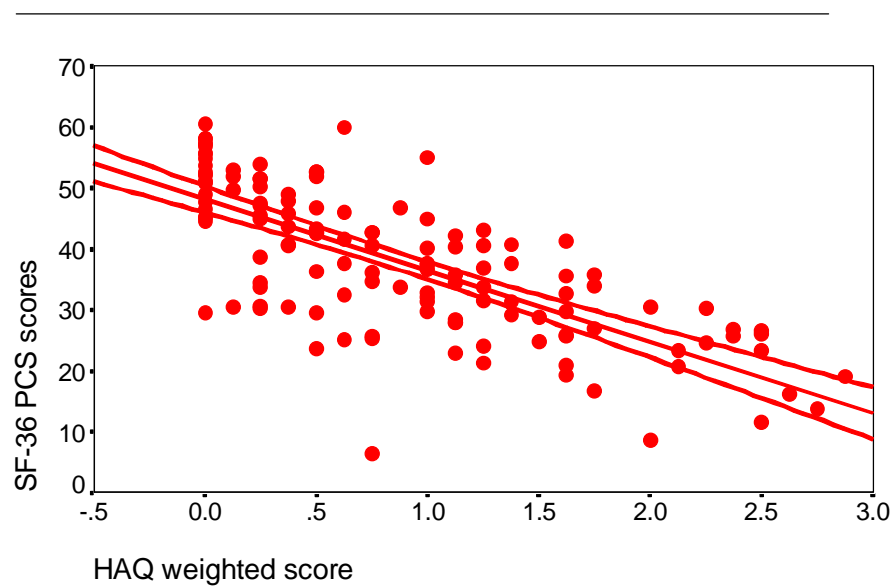


Figure 7E HAQ vs. SF-36: MCS

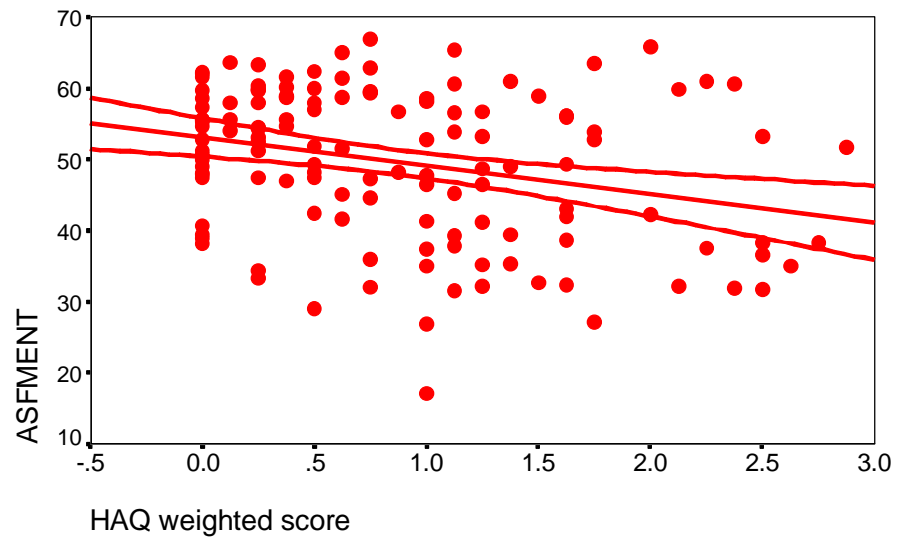


Table 9 Internal consistency as a measure of reliability

	<i>N. scales in instrument (a)</i>	<i>Cronbach a</i>	
		<i>Baseline</i>	<i>Retest</i>
Patient Global	1	N/A	N/A
AQoL	4	0.71	0.61
EQ5D	5 (c)	0.67	0.54
EQ5D VAS	1	N/A	N/A
HAQ	8	0.92	0.90
SF-36: MCS	8	0.36	0.38
SF-36: PCS	8	0.45	0.47

Notes:

All instruments employ additive models, except for the AQoL which has a multiplicative model

a = For each instrument, Cronbach α assessed by examination of weighted sub-scales (i.e. components, dimensions, scales or values as specified by the instrument developers)

c = Based on the five weighted items

Table 10 **Baseline and retest correlations**

		<i>Baseline</i>						
		<i>Patient Global</i>	<i>AQoL</i>	<i>EQ5D</i>	<i>EQ5D VAS</i>	<i>HAQ</i>	<i>SF-36: MCS</i>	<i>SF-36: PCS</i>
<i>Retest</i>	<i>Patient Global</i>	0.66						
	<i>AQoL</i>		0.87					
	<i>EQ5D</i>			0.74				
	<i>EQ VAS</i>				0.81			
	<i>HAQ</i>					0.93		
	<i>SF-36:MCS</i>						0.87	
	<i>SF-36:PCS</i>							0.86

Notes:

N. cases: 51, data missing from 1 case

Time between administrations: 2 weeks

Spearman correlations; in all cases $p < 0.01$.

Figure 8A Test-retest scatterplot: patient global

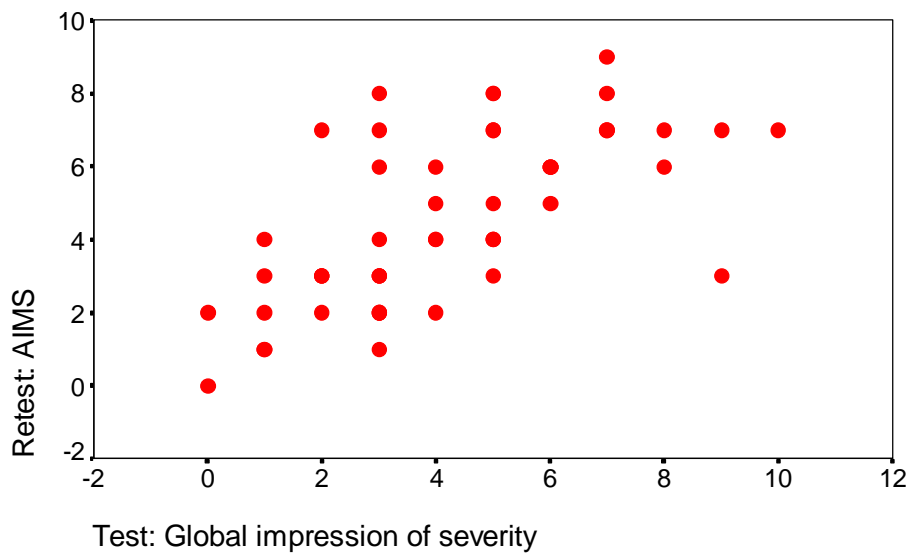


Figure 8B Test-retest scatterplot: AQoL

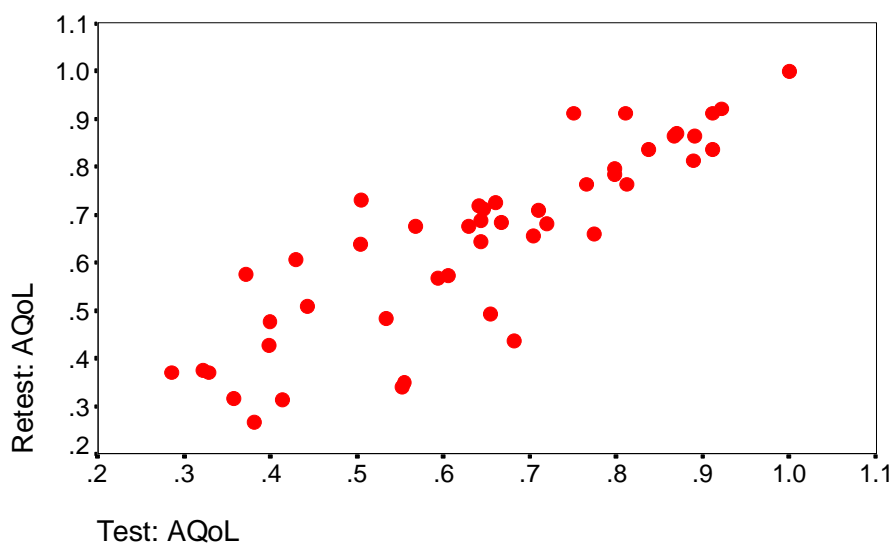


Figure 8C Test-retest scatterplot: EQ5D Utility

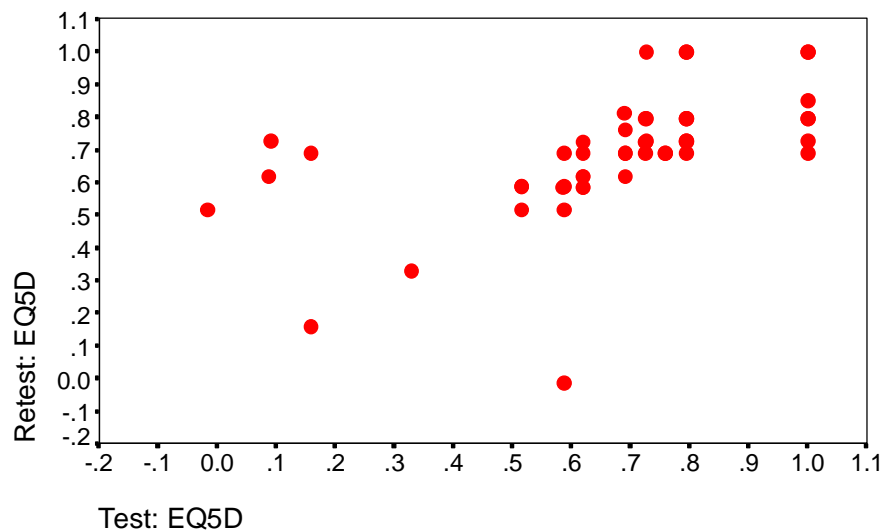


Figure 8D Test-retest scatterplot: EQ5D VAS

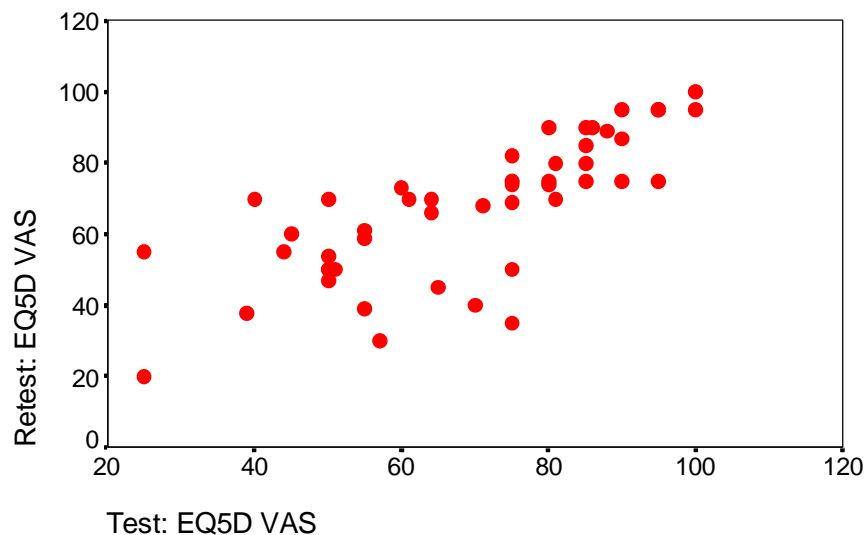


Figure 8E Test-retest scatterplot: HAQ

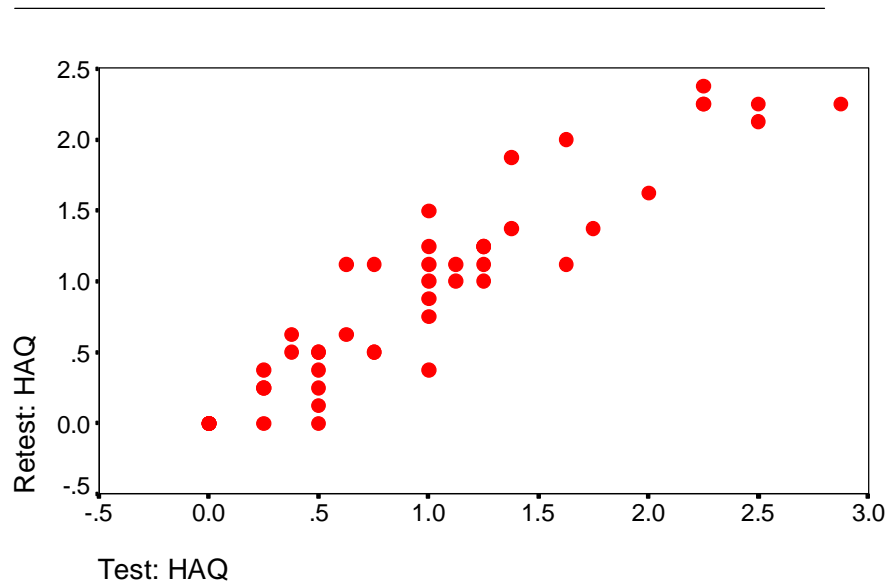


Figure 8F Test-retest scatterplot: SF-36 PCS

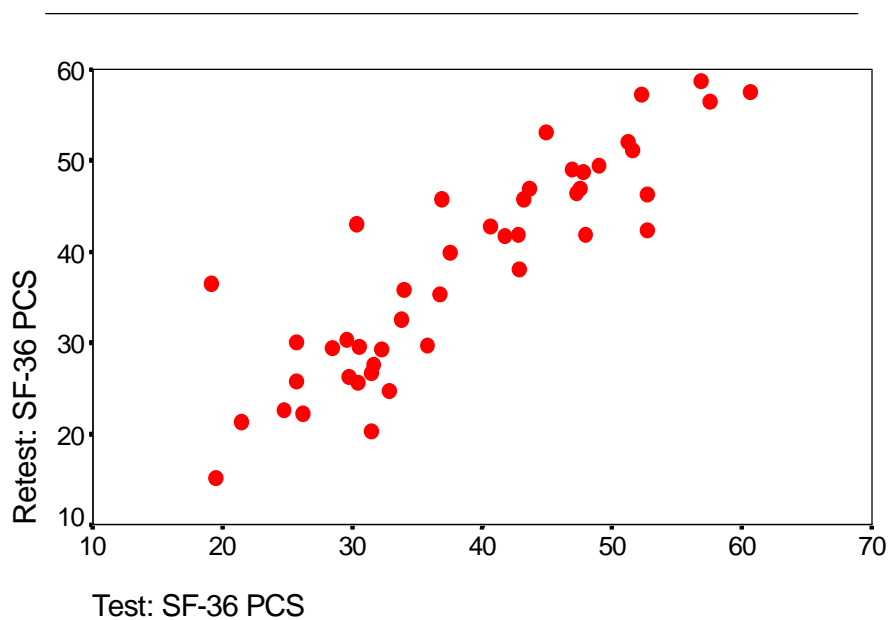


Figure 8G Test-retest scatterplot: SF-36: MCS

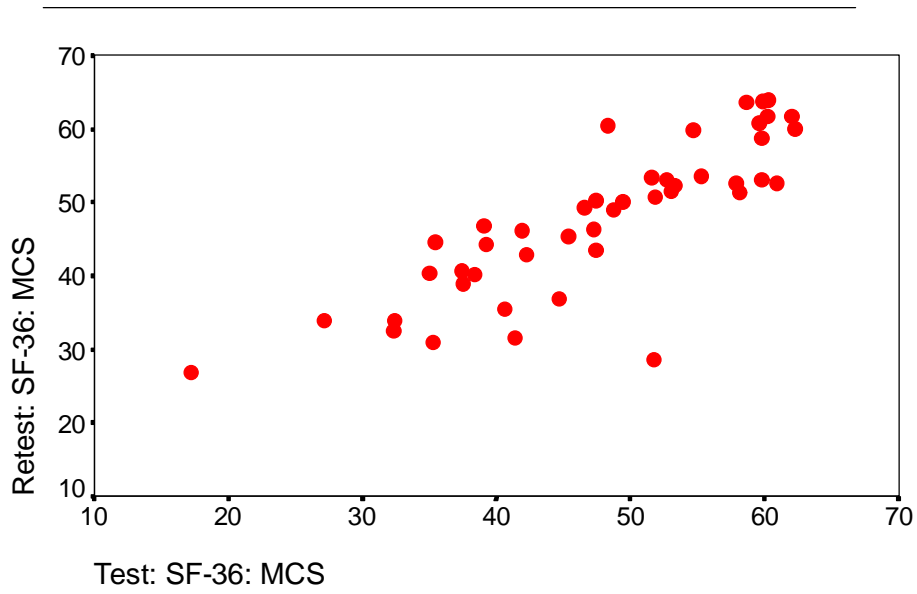


Table 11 The HAQ questionnaire data

		<i>All participants:</i>		<i>Test-retest sample</i>				<i>Statistics</i>
		<i>baseline</i>		<i>Test (baseline)</i>		<i>Re-test</i>		
		<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	<i>N</i>	<i>%</i>	
Dressing	No problem	65	49%	23	52%	22	50%	Agreement: 93% (a) k = 0.89, p < 0.01* (b)
	Some difficulty	36	27%	12	27%	12	27%	
	Much difficulty/Assisted	28	21%	6	14%	5	11%	
	Unable to do alone	5	4%	4	9%	4	9%	
Rising	No problem	71	54%	19	46%	23	56%	Agreement: 88% k = 0.79, p < 0.01*
	Some difficulty	40	30%	17	42%	14	34%	
	Much difficulty/Assisted	21	16%	5	12%	4	10%	
	Unable to do alone	0	0%	0	0%	0	0%	
Eating	No problem	63	47%	20	47%	23	54%	Agreement: 79% k = 0.68, p < 0.01*
	Some difficulty	38	28%	12	28%	10	23%	
	Much difficulty/Assisted	21	16%	8	19%	8	19%	
	Unable to do alone	12	9%	3	7%	2	5%	
Walking	No problem	83	61%	29	62%	31	66%	Agreement: 89% k = 0.80, p < 0.01*
	Some difficulty	36	26%	12	26%	10	21%	
	Much difficulty/Assisted	13	10%	4	9%	5	11%	
	Unable to do alone	5	4%	2	4%	1	2%	
Hygiene	No problem	66	47%	20	40%	21	42%	Agreement: 78% k = 0.69, p < 0.01*
	Some difficulty	23	16%	13	26%	12	24%	
	Much difficulty/Assisted	22	16%	8	16%	5	10%	
	Unable to do alone	25	18%	9	18%	12	24%	
Reaching	No problem	55	41%	16	33%	17	35%	Agreement: 73% k = 0.64, p < 0.01*
	Some difficulty	24	18%	11	22%	13	27%	
	Much difficulty/Assisted	34	26%	13	27%	9	18%	
	Unable to do alone	20	15%	9	18%	10	20%	
Gripping	No problem	40	30%	11	23%	13	28%	Agreement: 77% k = 0.55, p < 0.01*
	Some difficulty	15	11%	4	9%	4	9%	
	Much difficulty/Assisted	72	54%	31	66%	29	62%	
	Unable to do alone	6	5%	1	2%	1	2%	
Activities	No problem	55	41%	21	43%	21	43%	Agreement: 80% k = 0.71, p < 0.01*
	Some difficulty	35	26%	14	29%	11	22%	
	Much difficulty/Assisted	28	21%	7	14%	9	18%	
	Unable to do alone	15	11%	7	14%	8	16%	
HAQ score	Median			0.75		0.94		z = -1.44, p = 0.15, NS (c)
	(IQR)			(0.25–1.38)		(0.25–1.28)		

Notes:

- a = Agreement refers to the percentage of cases obtaining the same score at baseline and retest
- b = Kappa
- c = Wilcoxon Matched-Pairs Signed-Ranks Test

Table 12 Validation of the AQL, EQ5D and SF-36 by levels of arthritis disability (HAQ recoded)

		<i>HAQ levels (a)</i>				<i>Statistics (b)</i>
		<i>No difficulty</i>	<i>Some difficulty</i>	<i>Much difficulty</i>	<i>Unable to do</i>	
N. cases		66	41	19	11	
AQoL	Utility scores	0.79 (0.14)	0.60 (0.17)	0.44 (0.17)	0.32 (0.21)	F = 46.54, df = 3;132, p < 0.01*
EQ5D	Utility scores	0.82 (0.17)	0.62 (0.21)	0.52 (0.23)	0.39 (0.34)	F = 21.60, df = 3;133, p < 0.01*
EQ5D	VAS scores	77.92 (19.75)	64.67 (16.81)	56.16 (18.55)	47.00 (21.95)	F = 13.78, df = 3;134, p < 0.01*
SF-36	PCS (c)	45.99 (9.18)	34.48 (8.49)	26.78 (7.88)	22.25 (6.09)	F = 42.81, df = 3;127, p < 0.01*
	MCS (d)	52.89 (8.06)	46.80 (11.72)	47.93 (11.92)	43.27 (11.14)	F = 4.85, df = 3;127, p < 0.01*

Notes:

Values in cells are means and standard deviations

a = HAQ recoded into equal intervals:

0.00–0.74 = No difficulty, 0.75–1.49 = Some difficulty, 1.50–2.24 = Much difficulty, and 2.25–3.00 = Unable to do.

b = ANOVA

c = PCS: Physical health summary scale

d = MCS: Mental health summary scale