

ULRR

Outcome measures useful to physiotherapists working with patients with rheumatoid arthritis

Item Type	Article
Authors	Kennedy, Norelee;Stokes, Emma
Citation	British Journal of Therapy an Rehabilitation;9(9), pp. 357-364
Publisher	Mark Allen Publishing
Download date	2026-06-14 19:07:26
Item License	https://creativecommons.org/licenses/by-nc-sa/1.0/
Link to Item	https://hdl.handle.net/10344/2969

Outcome Measures useful to Physiotherapists working with patients with Rheumatoid Arthritis

Norelee Kennedy

Chartered Physiotherapist, Postgraduate Student,
School of Physiotherapy,
University of Dublin
Trinity College

Ms. Emma Stokes

Lecturer,
School of Physiotherapy,
University of Dublin
Trinity College

Address for correspondence:
School of Physiotherapy,
Trinity Centre for Health Sciences,
St. James's Hospital,
Dublin 8
Ireland.

Phone: +353 (0)1 6082110

Fax: + 353 (0) 1 4531915

e-mail: kennedyn@tcd.ie

ABSTRACT

This paper discusses outcome measures suitable for use by physiotherapists working with patients with rheumatoid arthritis. The measures have been successfully used in clinical practice, audit and research in the area of rheumatology and with other chronic diseases. Together, they provide comprehensive measurement of the disease. Both objective and patient-reported measures are included. The tools are as follows: grip strength, a visual analogue scale to measure pain, a timed walk test, the Arthritis Impact Measurement Scale 2, the Schedule for the Evaluation of Individual Quality of Life and the Quality of Life Index. The tools have been found to be easy to administer and are suitable for clinical use. They provide an effective means of evaluating treatment in both short and long term settings, making them ideal measures for a chronic disease.

INTRODUCTION

Rheumatoid Arthritis (RA) is a chronic, disabling disease that affects 1% of the adult Irish population. The number of people affected is similar in the United Kingdom and the United States. It is a commonly treated disease in both rheumatology units and general outpatient departments. These patients present a challenge to clinicians due to the relapsing nature of the disease, with intervention required at any point along the acute-chronic continuum. The aims of treatment are long term as well as focusing on immediate problems. Outcome measures used with this patient group should focus on the multifarious effects of the disease.

This paper suggests a number of outcome measures, which may be useful for physiotherapists working with people with rheumatoid arthritis. These measures will be discussed in relation to their measurement properties and their suitability for use in rheumatoid arthritis

Outcome measures in rheumatoid arthritis – what has been suggested to date?

Interest in the use of formal outcome measures in RA has developed under the auspices of the International League of Associations for Rheumatology (ILAR). OMERACT (Outcome measures in Rheumatoid Arthritis Clinical Trials) is a gathering of professionals interested in outcome measures in RA (Dequeker 1999). Recent work has included the incorporation of toxicity, generic health status and economic evaluation as well as outcome measures in osteoarthritis and osteoporosis. The output from an OMERACT Conference in 1993 resulted in the designation of eight endpoints to be used in research trials in RA (Table 1) (Tugwell and Boers 1993). Similarly, the American College of Rheumatology (ACR) (Felson et al 1993)

has recommended a core set of outcome measures to be utilised in trials involving patients with RA (Table 1).

These outcome measures are intended for medical trials and would not be entirely relevant to physiotherapy practice; indeed their use would perhaps preclude accurate measurement of change as a result of physiotherapy intervention. They are mentioned here because a number of themes emerge on analysis of the recommendations, namely that one measure may not be sufficient, the requirement for measurement within the context of the International Classification of Functioning, Disability and Health (ICF) (see below) levels and the incorporation of patient reported measures.

The WHO has agreed on a new classification system, the International Classification of Functioning, Disability and Health, the ICF (WHO 2001). The overall aim of this new classification is to provide a unified and standard language and framework for the description of health and health-related states. The new system defines health in two parts a) Functioning and Disability, and b) Contextual Factors.

Functioning and Disability is composed of two components: a) Body Functions and Structures and b) Activities and Participation, while the second part, Contextual Factors, contains two elements: Environment and Personal factors.

The outcome measures presented here will be mentioned also with respect to the definitions incorporated in the new classification.

Outcome Measures for Use in Current Research

As part of ongoing and previous (Kennedy et al 2001) research into the clinical and cost-effectiveness of in-patient and out-patient rehabilitation for patients with RA in Ireland, the authors employed and evaluated a variety of outcome measures considered to be appropriate for use in physiotherapy departments that are attended by patients with RA. Given the constraints of time in the clinical setting, the feedback

from patients, and the measurement properties of the measures, we suggest the following:

- 1. Visual Analogue Scale** to measure **Pain** intensity
- 2. Grip strength** using a mechanical hand held dynamometer
- 3. Timed “Up and Go” test**
- 4. The Arthritis Impact Measurement Scale 2 (AIMS2)**
- 5. Quality of life** may be measured by the **Quality of Life index** or by using the **Schedule for the Evaluation of the Individual Quality of Life (SEIQoL)**.

We present them, for consideration, as a battery of tests for use by physiotherapists. Initially, a brief review of the properties of measurement will be given. Each outcome measure will then be discussed within the context of its measurement properties. With regard to the ICF classification, pain (WHO, 2001, p68) and grip strength (WHO, 2001, p96) are components of “Body Structures and Function” whereas the functional measures, such as AIMS 2 and the Timed ‘Up & Go’ would comprise descriptors included in the “Activities and Participation” component.

Measurement properties

To measure something is to quantify it and to determine the extent of it by comparison with a standard unit (Wade 1992).

Selection of an outcome measure should consider the level of measurement and the measurement properties of reliability, validity and sensitivity. Reliability pertains to the consistency or repeatability of measurements (Table 2). However, it should be noted that not all measurements are equally error free (Rothstein 2001). The critical issue is whether the error will affect the use of the measure. Reliability is measured using various statistics depending on the level of measurement (see Table 2). General

consensus on the acceptable range of reliability is a correlation (r) from 0.65 to 0.94. A more discussion of reliability can be found in McDowell and Newell (1996).

Validity refers to the appropriateness of the measure for the purpose it was intended (Table 3) and sensitivity refers to the ability of the measure to detect the relevant clinical change. Evidence of each of these properties should be sought for an outcome measure before deciding on its use. Further information on each of these can be obtained from Rothstein (2001), Streiner and Norman (1995), Wade (1992) and Richman et al (1980). The outcome measures suggested for use with rheumatoid arthritis will now be examined with respect to each of these properties.

For clarity, Table 4 outlines each outcome measure with respect to reliability, validity and sensitivity. Only available evidence for each component is listed in the section below.

OUTCOMES FOR USE WITH RHEUMATOID ARTHRITIS

1. Visual Analogue Scale to measure pain

Description

A Visual Analogue Scale is a 100 millimetre straight line used to measure the intensity of pain. The ends of the line define the extreme limits of pain, with the 0mm end of the scale represents 'no pain', while the 100mm end represents the 'worst pain ever experienced'. The line can be orientated vertically or horizontally. The test can measure pain at one point in time (absolute) or over time (comparative) (Downie et al 1978, Scott and Huskisson 1976, Dalton and McNaull 1998, Chapman et al 1985, Carlsson 1983).

Scoring

The measure is scored by counting the number of millimetres from 0mm to the patient score. A VAS is easy to administer and appears to be understood by all patients.

Attention should be paid to the reproducibility of the line as photocopying can alter the length of the line and thus invalidate results.

Reliability

a) Inter-rater reliability

Not applicable as scale is self-reported

b) Intra-rater reliability

Not applicable as scale is self-reported

c) Test-retest

Dixon and Bird (1981) investigated the ability of subjects to reproduce an existing mark on a VAS. While reproducibility was variable, it was noted that the most variability was demonstrated +/- 2cm from the midpoint.

Validity

a) Content

In asking a patient to describe the intensity of pain, we are explicitly directing them towards a measure of pain; hence there is no doubt about what domain is being measured, albeit a one-dimensional component of a multi-dimensional.

b) Construct

Pain is difficult to define and is multi-faceted: thus construct validity is hard to establish. However, it has been noted that higher scores on other measure of pain are linked with higher scores on VAS (Downie et al 1978).

c) Criterion – related

Downie et al (1978) demonstrated good correlation between a numerical rating scale ($r=0.9$), a simple descriptive scale ($r=0.78$), a vertical VAS ($r=0.88$) and the horizontal VAS.

Sensitivity

The VAS is a sensitive, measure of pain provided that the divisions used on the scale are uniform and consist of at least twenty divisions (Scott and Huskisson 1976)

Completion Time

This can take up to 5 minutes from explanation of task to completion by the patient

2. Grip strength

Grip Strength is most objectively measured using a dynamometer, such as the Jamar dynamometer.

Description

The Jamar dynamometer is an industry standard in use for over 35 years.

Techniques to aid in the standardisation of the procedure have been identified and listed (Owners Manual for Jamar Dynamometer).

Reliability

a) Inter-rater reliability

This was demonstrated by testing the correlation between two testers for testing of both hands. Correlation (r) between the two testers was calculated as 0.99 for testing of both hands (Pearson Product Moment Correlation Co-efficient (PPMCC)) demonstrating good inter-rater reliability (Mathiowetz, Weber, Volland and Kashman 1984).

b) Test-retest reliability

Again, the correlation of two separate observations of grip strength tests a week or less apart was calculated. The correlation (r) was 0.88 for a mean of three tests for the right hand and $r = 0.92$ for the mean of three tests for the left hand, as measured by the Pearson correlation coefficient (Mathiowetz, Weber, Volland and Kashman 1984).

Validity

a) Content

A dynamometer directly measures grip strength therefore the Jamar dynamometer has good content validity.

Evidence of its validity has been established (Mathiowetz, Weber, Volland and Kashman 1984). To do this, a number of known weights were hung from the dynamometer. Accuracy of measurement for the dynamometer was established to be +/- 3%.

Average Completion times:

The test takes approximately 5 minutes to complete.

3. The Timed “Up & Go” test

This test was developed by Podsiadlo and Richardson (1991) as a modified version of the “Get Up & Go” test (Mathias, Nayak and Isaacs 1986).

Description

The Timed “Up & Go” test is a quick test of basic mobility. It requires no special equipment (except for a rehabilitation chair and a stop-watch); thus it is a simple test to perform.

Although originally designed for use with older patients, it has been used as an outcome measure by McMeeken et al (1999) investigating the effect of knee extensor and flexor muscle training in individuals with RA. Following specific muscle training using an isokinetic dynamometer for 17 patients with RA, faster Timed “Up & Go” test times, greater peak speed (from the isokinetic dynamometer), less pain and better quality of life were recorded. It has also been used to measure the effects of a period of rehabilitation in a rheumatic diseases unit (Kennedy et al 2001).

A standardised procedure is available for the test (Podsiadlo and Richardson 1991).

Scoring

The time taken to complete the test for each patient is recorded in seconds/milliseconds.

Reliability

a) Inter-rater reliability

To test the interrater reliability, twenty-two patients were rated by three raters in random order at different times in the same day while performing the test. The correlation (r) (intraclass correlation coefficient (ICC)) was 0.99, indicating high interrater reliability (Podsiadlo and Richardson 1991). As sample size was small here, a high correlation would be necessary to demonstrate good reliability.

b) Intra-rater reliability

Twenty patients were observed by the same rater on two consecutive visits to the day hospital (physiotherapists $n = 20$, physicians $n = 10$). The correlation (r) (ICC) was 0.99, again indicating good intra-rater reliability (Podsiadlo and Richardson 1991). Again, due to the small sample size, it is necessary to have a high ICC value to demonstrate good inter-rater reliability.

c) Internal Consistency

Not applicable

Validity

a) Criterion – related

Podsiadlo and Richardson (1991) hypothesised that the timed “Up and Go” test would correlate to patients balance, walk speed and functional capacity. Thus its relationship to the Berg Balance Scale, gait speed and the Barthel Index of Activities of Daily Living was established. When compared to the Berg Balance Scale the correlation r (Pearson correlation coefficient) was -0.72 , for Gait Speed $r = -0.55$ and for the Barthel Index, $r = -0.51$, indicating good criterion-related validity.

Average completion times are under 20 seconds for patients who are independent in basic transfers.

4. Arthritis Impact Measurement Scale 2 (AIMS2)

The AIMS was developed in 1980 by Meenan (1982). A revised edition, the AIMS2 was developed in 1992 (Meenan et al 1992). This was a more comprehensive and sensitive version of the original.

Description:

The AIMS2 is a self-administered evaluation instrument to measure patient outcome in the rheumatic diseases. It is a 78-item questionnaire containing scales such as Mobility Level, Walking and Bending, Hand and Finger Function, Arm Function, Self-Care Tasks, Household Tasks, Pain, Social Activity, Level of Tension and Mood.

Reliability

a) Inter-rater reliability

Not applicable as questionnaire is self-report

b) Test-retest

A score of $r =$ or > 0.8 for all components of the scale except work ($r =0.78$) as measured by ICC of scores from two administrations of the instrument separated by two to three weeks in a sample of 45 patients over a 2 week period was demonstrated by Meenan et al (1992).

c) Internal Consistency

The reliability of all 12 scales of the questionnaire, as measured by the internal consistency (Cronbach's alpha) of their component items, were r between 0.72 – 0.91 for the RA group (n=299), (Meenan et al 1992).

Validity

a) Content

This was established by Meenan et al upon development of the questionnaire (1982) by giving the questionnaire to 335 rheumatologists to analyse from the point of view of clarity and relevance of the chosen items to the arthritis.

b) Construct

Meenan (1992) examined the validity of the AIMS2 and demonstrated that when patients assigned an area as a problem, then this area was associated with a poorer AIMS2 score in that area. Potts and Brandt (1987) provide further evidence of the validity of the original scale. The responses of one hundred and twenty patients who scored the AIMS were examined to determine if a correlation existed between their scores and the importance rating that they apportioned to each section such as pain, anxiety, depression, dexterity and physical activity. The results obtained had a significance level of $p=0.01$, providing evidence for the validity of the pain, anxiety, depression, dexterity and physical activity subscales of the AIMS. However, the authors conceded that the subscale on household activity might not be appropriate for male respondents. Finally, Hughes et al (1991) adapted the AIMS for use with frail, elderly respondents. Scores were obtained from four hundred and thirty eight elderly patients (mean age 76 years). This revised version was interviewer administered. A generic and an arthritis specific score were obtained to help identify if a co-morbid condition was implicated. Ninety percent of subjects had osteoarthritis, whereas just three percent had RA. The authors concluded that due to the chronic, disabling nature of RA, the investigator-administered GERI-AIMS may be appropriate in frail, elderly patients.

c) Criterion – related

This was not established, as the authors of the questionnaire were unable to locate a suitable standard with which it could be compared.

Sensitivity

Anderson et al (1989) investigated the sensitivity of the AIMS2 to the short-term clinical changes in arthritis. Having analysed data from three clinical trials, substantial improvements in physical function, psychological status and pain and overall arthritis impact were detected. These changes occurred at initial assessment, at four weeks and/or at eight weeks, indicating the sensitivity of the AIMS2 to detect change.

Meenan et al (1984) also reported that the AIMS was sensitive to clinically meaningful drug-induced change in patients with rheumatoid arthritis.

Average completion times are approximately twenty minutes.

Quality of life

Information provided by quality of life instruments can be useful in determining prognosis (Bell, Bombadier and Tugwell 1990). Measures of quality of life have been developed out of a need to define the impact of health care interventions on individuals. The choice of quality of life measure depends on whether a disease specific or a generic measure is required. Commonly used generic tools include the Sickness Impact Profile, the Nottingham Health Questionnaire, the McMaster Health Index and the SF-36. It is becoming increasingly common to use both a disease-specific and a generic measure.

6. The Schedule for the Evaluation of Individual Quality of Life (SEIQoL)

The SEIQoL (O'Boyle 1994) is a quality of life measure that evaluates individual quality of life. As each individual's life experiences are unique, a need exists for

individual assessment and rating of personal quality of life. The development of the questionnaire was based on a number of basic propositions. Firstly, QoL is individual and secondly, that a person's judgement of their overall QoL is made based on all the domains or areas of that person's life important to them.

Description

Patients are asked to name five important cues (areas) in their lives, to rate the current status of each cue and their importance relative to each other.

Scoring

An index is calculated which can be analysed using parametric statistics. Frequently elicited cues include family, relationships, health, finances, living conditions, work, social, religious or spiritual life.

Scores range from 0 –100, with a higher score indicating superior quality of life.

Reliability

Evidence for the reliability of the questionnaire could only be found with respect to internal reliability. Internal reliability relates to consistency amongst individuals with respect to their judgements of their quality of life. Correlation scores r ranged from 0.49 to 0.74 (O'Boyle 1994). More research is needed to determine the reliability of this tool due to the wide range of reliability scores obtained.

Validity

Evidence for the validity of the questionnaire could only be found with respect to internal validity. Internal validity was explained by the author as reflecting the proportion of the variance in the overall QoL judgement. This was measured by the square of r (r^2) of scores from various studies involving the SEIQoL. Scores ranged from 0.62 to 0.79 (O'Boyle 1994). However, more work is necessary to demonstrate evidence of construct and criterion referenced validity.

Average completion times are 10 to 20 minutes

7. The Quality of Life Index

This index was developed in 1981 (Spitzer et al 1981). Originally developed for cancer patients, it is suitable for use with patients with a chronic physical disease such as rheumatoid arthritis.

Description

The measure consists of 5 subscales: Activity, Daily Living, Health, Support and Outlook. For each heading there are three possible choice answers

Scoring

Scores for each question are summed to determine the overall index, ranging from 1 – 10, with a high score again indicating satisfaction with quality of life.

Reliability:

a) Inter rater reliability

This was assessed by two Canadian physicians who independently assessed 64 physicians within seven days using the instrument, resulting in an interrater correlation coefficient (r) of 0.81 (as measured by Spearman's rank correlation coefficient)(Spitzer et al 1981).

b) Internal consistency

This was determined by calculating Cronbach's α to test if the items in the questionnaire are similar to the underlying constructs or theories that relate to quality of life. The correlation was calculated as $r = 0.775$, indicating good internal consistency (Spitzer et al 1981).

Validity:

a) Content

A pre set *a priori* criterion of 51% for each of the two independent panels that reviewed the index was set. The first panel consisted of 34 lay people composed of healthy people, people with cancer and other chronic diseases and patient's relatives. The second panel, an expert panel, consisted of 34 health professionals and methodologists. Each panel had to respond in the affirmative for each question by a majority vote (51%). This was achieved resulting in high content validity (Spitzer et al 1981).

b) Construct

This was examined by testing a unidimensional scale and a comprehensive multidimensional scale as well as the QoL index on a sample of well and sick patients. Agreement was strong and statistically significant at 0.005 and 0.001 levels for the panel of sick patients (Spitzer et al 1981). A statistically significant result of $p = 0.005$ or 0.001 would provide strong evidence for the construct validity of the tool.

Average completion time: 3 to 5 minutes

DISCUSSION

While the importance of employing carefully constructed and relevant instruments or scales to measure intervention is now well-recognised in rehabilitation, the decision-making paradigms used by therapists in choosing the outcome measurements are often hampered by the fact that studies on the measurement properties are published in diverse sources. In addition the relevance of the measurement properties of the

instrument are often not grounded clearly in language that is relevant to the 'measurer' and the 'measured'.

If a method of measuring outcome is not reliable, it is pointless to employ it, since the change in the patients' score might be due to errors within the measure itself or errors on the part of the therapist using it. Inter-rater reliability considers inter-observer variability. Hence it relates to the reproducibility of test results obtained between one tester and another. In an instrument or scale that is patient-reported, this is not a relevant domain to investigate, however, with tests of strength and timed tests, this type of reliability is essential. The inter-rater reliability of the methods of measuring grip strength (Jamar dynamometer) and speed during a functional task (Timed Up & Go) have been investigated, with reported correlation co-efficients of $r=0.97$ and $r=0.99$ respectively. Similarly the measurements of quality of life, using the Quality of Life index will yield consistency between different assessors enhancing the use of the index in daily patient management.

Intra-rater reliability considers the reproducibility of a measure and how it may be subject to errors or changes that occur due to a lack of accuracy in repeating the test within the individual tester. Test-retest reliability, defined clearly by Feinstein (1987) 'requires examination of two different versions of the same index', however the term is often used inter-changeably with intra-rater reliability. The Timed "Up and Go" test demonstrates good reliability over different assessments for the same tester, indicating its usefulness and stability as a repeated measure to document actual change in reported pain. However, the evidence for 'test-retest'/intra-rater reliability was established using just 8 patients: thus further investigation may be warranted to ascertain greater accuracy. Strong evidence for this type of reliability is available for

grip strength also ($r = 0.80$) and for the AIMS2 ($r = 0.80$). Repeated testing of different forms of the scales has not been reported.

A final aspect of reliability is internal consistency - the manner in which the different components of a scale relate to each other, how the individual items in a multi-dimensional scale form a unit. It is not relevant for uni-dimensional measures. A high level of agreement suggests that individual components relate to one another in a way that suggests overall contribution to one domain. If the level of agreement is too high, one component might be irrelevantly included and should have been omitted in its development. Evidence of internal consistency was found for the AIMS2. Thus there is high correlation between different divisions of the AIMS2, contributing to its usefulness as a measurement tool in arthritis. Both quality of life measures have internal consistency (QoL index: $r = 0.775$ and the SEIQoL: $r = 0.49$ to 0.74).

And so to validity, which is the process of examining the measurement with a view to considering its content, the groups to which it may most appropriately pertain and the extent of the interpretation that may be applied to its output (McDowell & Newell, 1986). Criterion validity is concerned with the extent to which a measurement relates to the “gold” standard measurement in the area. This is usually the same phenomenon. In examining criterion validity, the relationship of the measurement to another, taken at the same point in time, refers to concurrent validity and its ability to relate to a measure taken in the future refers to its predictive validity. If no ‘gold standard’ exists for comparison, and in comparing it with other existing scales or instruments, a relationship is noted, this may be referred to as convergent validity. Evidence of this is available for the VAS and its correlation to the Verbal Rating Scale (VRS) another tool utilised in pain measurement. The VAS correlates well to the VRS ($r = 0.70$) indicating the ability of the VAS to measure pain intensity. The Timed “Up and Go”

test has also been correlated to other standards in the area. When correlated with the Berg Balance Scale (convergent) the correlation r was = - 0.72, with gait speed (criterion) $r = -0.55$ and for the Barthel Index, $r = -0.51$, indicating the various components of criterion validity. New measurement tools are born out of a need to measure different dimensions of health. To this end, exact correlation with another standard tool in the area would not be expected. Correlations of below $r = 0.65$ can thus be acceptable in view of differing needs. Construct validity is concerned with the scores or results obtained using the new measure and underlying theories upon which the measure is based. For example, in a sample of people with established gait disorders, one might expect a lower score on an instrument measuring gait, than in a 'normal' sample. If this hypothesis is tested and found to be true, construct validity is supported. As pain is multi-dimensional, it is difficult to establish construct validity for the VAS. No explicit evidence for the construct validity of grip strength or for the Timed "Up and Go" test was found to be reported. The evidence for the construct validity of the AIMS2 is available, indicating its correlation to its underlying theory that health encompasses physical, psychological and social states. Similarly, the QoL index was examined for its correlation to its underlying constructs and results were found to be significant. Content validity reflects the adequacy of the tool in reflecting the aims outlined in the original definition of the measure. It is often quite apparent from the scale, but can be supported by user consensus, often carried out as part of the preliminary development process. The VAS demonstrates content validity since it measures self-reported pain, for which it was intended. Similarly, the Jamar dynamometer measures grip strength accurately, allowing for its suitability to the clinical setting. Grip strength is frequently decreased in the rheumatoid patient, resulting in decreased hand function. Thus its inclusion in assessing the outcome of

treatment allows for accurate measurement of intervention and the displaying of any resulting improvement.

Evidence of the content validity of the Timed “Up and Go” test was not found. It clearly examines a timed task involving two functional activities – gait speed and sit to stand and as such is a suitable inclusion in the measurement of any disease resulting in impaired lower limb function. The AIMS2 also has content validity as it contains various items relevant to the arthritic patient such as pain, mobility, hand function, walking and bending as well as social activities and anxiety and depression. All of these areas, and more, can be affected in the patient with a rheumatic disease. Similarly, the QoL index displays content validity by measuring the various aspects of quality of life relevant to patients with a chronic disease such as activity, daily living, health, support and outlook.

The final property of a measurement tool is sensitivity, which refers to the ability of the tool to detect relevant clinical change. This is important in the clinical setting, as certain outcomes of treatment may indeed be small but relevant if they produce improved function. Liang et al (1995) suggest that it is this measurement property that is least well studied. This is consistent with the information we present in this paper. The AIMS2 is documented as a sensitive tool and has shown that it can reflect changes following intervention that may be relevant to the rheumatoid patient. All of the methods of measuring outcome described in this paper have been successfully employed in clinical research, as part of an audit of clinical practice in a rheumatology unit (Kennedy et al 2001) and are employed in an ongoing research trial. While the measures have been described separately, their interrelationship has also been previously considered and yielded results that support the inclusion of an eclectic battery of tools to measure outcome in Rheumatoid Arthritis. The relevance

of the measurement properties to each outcome measure and to the clinical setting will now be discussed. Pain, grip strength and walking time have all demonstrated a relationship to self-reported functional ability (Spiegel et al, 1987; Ward and Leigh, 1993, Vliet Vlieland et al, 1996; Nordenskiold and Grimby, 1997) but they have not shown the ability to relate to disease markers such as Erythrocyte Sedimentation Rate and C-Reactive Protein. Borstlap et al (1993) investigated the overlap between clinical and laboratory tests and a Dutch quality of life measure. Results indicated that laboratory and clinical measures (Erythrocyte Sedimentation Rate, C-Reactive Protein, grip strength and functional classification) were significantly correlated ($p < 0.001$) with physical dimensions of the quality of life measure. van der Heide et al (1993) noted that self-reported measures of physical ability reflect current disease activity whereas the number of assistive devices/aids used reflects disease duration, and suggest the importance of differentiating between *difficulty in performance* and *duration of disease*.

Advantages associated with these suggested measures include their low cost, sensitivity, ease of administration, evidence of validity and reliability for each measure and the existence of established inter-relationships. The SEIQoL usually takes between 5 and 15 minutes to complete. This may limit its use in the clinical setting due to time constraints. It does allow for exploration of individual issues with patients and may allow for the consideration of cognitive issues such as coping strategies and perceptions of how illness impacts on a range of aspects pertinent to particularities of specific patients. Scharloo et al (1999) note that after controlling for disease duration and severity, these issues are relevant to health outcomes in RA.

Due to the frequent involvement of the finger joints and wrists in the disease, hand function is affected in patients with RA. The undertaking of everyday tasks can be

severely limited. Dressing, feeding and other self-care tasks are negatively affected. The impingement on hand function can be assessed by measuring grip strength and by using hand function tests such as the Jebsen Taylor Hand Function test (Jebsen et al, 1969). However, with regard to clinical assessment, grip strength may be more acceptable measure to physiotherapists, as many hand function tests can be time consuming and in addition are often completed by the occupational therapy department, which would thus duplicate assessments.

CONCLUSION

The use of outcome measures is imperative to determine the effectiveness of treatment. This is particularly relevant with a chronic disease such as rheumatoid arthritis. The outcome measures presented here have been found to be valid and reliable and suitable for use in clinical and research settings. By measuring at the levels of impairment, disability and handicap they provide a comprehensive assessment of the disease process and allow for the accurate evaluation of intervention.

KEY POINTS

1. The use of outcome measures is necessary to determine clinical effectiveness
2. Outcome measures used should aim to evaluate within the context of the International Classification of Functioning, Disability and Health (ICF)
3. A comprehensive range of outcome measures are presented which encompass all aspects of a chronic disease such as rheumatoid arthritis
4. These measures can be routinely incorporated into routine clinical practice
5. Use of these outcome measures can facilitate accurate evaluation of treatment for rheumatoid arthritis

REFERENCES

- Anderson J, Firschein H and Meenan R (1989) Sensitivity of a Health Status Measure to short-Term Clinical Changes in Arthritis. *Arthritis and Rheumatism* 32: 844 – 850
- Bell M, Bombadier C and Tugwell P (1990) Measurement of Functional Status, Quality of Life, and Utility in Rheumatoid Arthritis. *Arthritis and Rheumatism* 1: 591-601
- Borstlap M, van de Laar M, Zant J and van der Korst J (1993) Components of health: an analysis in rheumatoid arthritis using quality of life questionnaires and clinical and laboratory variables. *Annals of Rheumatic Diseases* 52: 650 – 654
- Carlsson A M (1983) Assessment of Chronic Pain Aspects of the reliability and validity of the Visual Analogue Scale. *Pain* 16: 87-101
- Chapman C, Casey K, Dubner R, Foley K, Gracely R and Reading A (1985) Pain Measurement: an overview. *Pain* 22: 1-31
- Cole B, Einch E, Gowland C and Mayo N (1995) *Physical Rehabilitation Outcome Measures* 1st edn. Williams and Wilkins, Baltimore
- Dalton J and McNaull F (1998) A call for standardizing the clinical rating of pain intensity using a 0 to 10 rating scale. *Cancer Nursing* 21: 46-49
- Dequeker J (1999) ILAR and OMERACT. *Journal of Rheumatology* 26: 201-203
- Devlin J, Gough A, Huissoon A, Perkins P, Holder R, Reece R, Arthur V and Emery P (1997) The acute phase and function in early rheumatoid arthritis. C-Reactive Protein levels correlate with functional outcome. *Journal of Rheumatology* 24:9-12

Downie W, Leatham P, Rhind V, Wright V, Branco J and Anderson J (1978) Studies with pain rating scales *Annals of Rheumatic Diseases* 37:378-381

Feinstein A (1987) *Clinimetrics* Yale University Press New Haven and London

Felson D, Anderson J, Boers M, Bombadier C, Chernoff M, Fried B, Furst D, Goldsmith C, Kieszak S, Lightfoot R, Paulus H, Tugwell P, Weinblatt M, Widmark R, Williams J and Wolfe F (1993). The American College of Rheumatology Preliminary Core Set of Disease Activity Measures of Rheumatoid Arthritis Clinical Trials. *Arthritis and Rheumatism* 36: (6)729 - 739

Hughes S, Edelman P, Chang R, Singer R and Schuette P (1991) The GERI-AIMS. Reliability and Validity of the Arthritis Impact Measurement Scales Adapted for Elderly Respondents. *Arthritis and Rheumatism* 34:856-865

Jebsen R, Taylor N, Trieschmann R, Trotter M and Howard L (1969). An objective and standardised test of hand function. *Archives of Physical Medicine and Rehabilitation* 50:311-319

Kennedy N, Sayers K, Cassidy C, Ward S, Bresnihan B, Fitzgerald, O and Stokes E (2001) The effect of a period of rehabilitation in a rheumatic diseases unit on patients with rheumatoid arthritis. *British Journal of Therapy and Rehabilitation*. 8: 29 – 36

Liang M (1995) Evaluating Measurement Properties. *Journal of Rheumatology* 22:1191-1192

Mathias S, Nayak USL, Isaacs B (1986) Balance in the elderly patient: The “Get Up & Go” test. *Archives of Physical Medicine and Rehabilitation*. 67: 387 - 389

Mathiowetz V, Weber K, Volland G, Kashman N (1984) Reliability and Validity of Grip and Pinch Strength Evaluations. *Journal of Hand Surgery [Britain]* 9A: 22-6

McMeeken J, Stillman B, Story I and Kent P (1999) The effects of knee extensor and flexor muscle training on the timed-up-and-go test in individuals with rheumatoid arthritis. *Physiotherapy Research International* 4: 55 - 67

Meenan R (1982) The AIMS Approach to Health Status Measurement: Conceptual Background and Measurement Properties. *Journal of Rheumatology* 9: 785 –788

Meenan R, Anderson J and Kazis L (1984). Outcome assessments in clinical trials: evidence for the sensitivity of a health status measure. *Arthritis and Rheumatism* 27: 1344-1352

Meenan R, Mason J, Anderson J, Guccione A and Kazis L. AIMS2 (1992) The Content and Properties of a Revised and Expanded Arthritis Impact Measurement Scales Health Status Questionnaire. *Arthritis and Rheumatism* 35:1 –10

Nordenskiold U and Grimby G (1997) Assessments of disability in women with rheumatoid arthritis in relation to grip force and pain. *Disability and Rehabilitation* 19: 13-19

O’Boyle CA (1994) The Schedule for the Evaluation of the Individual Quality of Life (SEIQoL). *International Journal of Mental Health* 23:3-23

Podsiadlo D and Richardson S (1991) The Timed “Up & Go”: a test of Basic Functional Mobility for Frail Elderly Persons. *Journal of the American Geriatric Society* 39: 142 – 148

Potts M and Brandt K (1987) Evidence of the reliability and Validity of the Arthritis Impact Measurement Scales. *Arthritis and Rheumatism*. 30: 93 – 97

Richman J, Makrides L and Prince B (1980) Measurement procedures in research *Physiotherapy Canada* 32(5):253-259

Rothstein J (2002) Sick and tired of reliability *Physical Therapy* (2):editorial

Scharloo M, Kaptein AA, Weinman JA, Hazes JMW, Breedveld FC, Rooijmans HGM (1999) Predicting functional status in patients with rheumatoid arthritis. *Journal of Rheumatology* 26:1686-1693

Scott J and Huskisson E.C. (1976) Graphic representation of pain. *Pain* 2: 175-184

Spiegel J , Paulus H, Ward N, Spiegel T, Leake B and Kane R (1987) What are we measuring? An examination of Walk Time and Grip Strength. *Journal of Rheumatology* 14:80 –86

SpitzerW, Bobson A, Hall J, Chesterman E, Levi J, Shepherd R, Battista R and Catchlove B (1981) Measuring the Quality of Life of Cancer Patients. *Journal of Chronic Diseases* 34: 585 – 597

Tugwell P and Boers M (1993) OMERACT Conference on Outcome Measures for use in Rheumatoid Arthritis Trials: a Conclusion. *Journal of Rheumatology* 20: 590-591

van der Heide A, Jacobs JWB, van Albada-Kuipers GA, Kraaimaat FW, Greenen R, Bijlsma JWJ (1993) Self report functional disability scores and the use of devices: two distinct aspects of physical function in rheumatoid arthritis. *Annals of Rheumatic Diseases* 52: 497-502

Vliet Vlieland TPM, Van der Wijk TP, Jolie IMM, Zwinderman AH, Hazes JMW (1996) Determinants of hand function in patients with rheumatoid arthritis. *Journal of Rheumatology* 23:835-840.

Wade DT (1992). *Measurement in Neurological Rehabilitation*. Oxford University Press. Oxford.

Ward MM and Leigh JP (1993) The relative importance of pain and functional disability to patients with rheumatoid arthritis. *Journal of Rheumatology* 20:1494-1499.

World Health Organisation (2002). *International Classification of Functioning, Disability and Health*. World Health Organisation, Geneva

Table 1: Outcome measures in Rheumatoid Arthritis Clinical Trials (OMERACT) and American College of Rheumatology (ACR) recommendations for outcome measures in Rheumatoid Arthritis trials

OMERACT	ACR
Acute Phase Reactants	One Acute Phase Reactant
Tender joint count	Tender joint count
Swollen joint count	Swollen joint count
Pain	Patient's assessment of pain
Patient global assessment	Patient's global assessment of disease activity
Physician global assessment	Physician's global assessment of disease activity
Disability	Patient's assessment of physical function
Radiographic studies of joints (for studies of one year or longer)	

Table 2: Reliability – description of each type

Type	Inter-rater	Intra-rater	Internal Consistency	Test-retest
Description	The degree of agreement between different raters using the same test	The degree of agreement between one rater for each repetition of the same test	The overall consistency of the measure among the different items of a multiple response instrument	Repeatability of measures over time*
Measurement Method	Intraclass correlation (ICC) (for ratio scales) or less commonly, Pearson correlation coefficient and Spearman rank correlation coefficient		Cronbach's coefficient alpha	

* Test –retest and intra-rater reliability both refer to the repeatability of a measure over time. The reliability quoted by a test developer or analyser should be interpreted according to the method used to calculate it, that is does the method refer to the *tester* (intra-rater) or to the *test* (test-retest).

Table 3: Types of Validity

Type	Content	Construct	Criterion
Description	The adequacy of the sampling questions in reflecting the aims of the index that were specified in the conceptual definition of its scope (McDowell and Newell)	The degree to which the scores obtained concur with the underlying theories related to the content of the measure. It can be divided into Convergent and Discriminant	The extent to which the measure concurs with “gold-standards” in the same area. It can be of two types Predictive and Concurrent

Table 4: Availability of evidence for each outcome measure with respect to measurement properties

	Pain	Grip	TUG	AIMS2	SEIQoL	QoL
<u>Reliability</u>						
Inter-rater	Not applicable	Not available	Yes	Not applicable	Not available	Yes
Intra-rater	Not applicable	Yes	Yes	No	Not available	Not available
Test-retest	Yes	Yes	Not available	Yes	Not available	Not available
Internal Consistency	Not available	Not available	Not applicable	Yes	Yes	Yes
<u>Validity</u>						
Content	Yes	Yes	Not available	Yes	Not available	Yes
Construct	No	Not available	Not available	Yes	Yes	Yes
Criterion-related	Yes	Not available	Yes	Not available	Not available	Not available
<u>Sensitivity</u>	Yes	Not available	Not available	Yes	Not available	Not available