EXTENDED REPORT

The rheumatoid arthritis articular damage score: first steps in developing a clinical index of long term damage in RA

T R Zijlstra, H J Bernelot Moens, M A S Bukhari

Ann Rheum Dis 2002;**61**:20–23

See end of article for authors' affiliations

Correspondence to: Dr T R Zijlstra, Medisch Spectrum Twente, Secretariaat Reumatologie, Postbus 50000, 7500 KA Enschede, The Netherlands; tr.zijlstra@worldonline.nl

Accepted 28 June 2000

Objective: To design and validate a clinical method for scoring irreversible long term articular damage in rheumatoid arthritis (RA).
Methods: The rheumatoid arthritis articular damage score (RAAD score) is based on examination of 35 large and small joints. Concise definitions were formulated to score each joint on a three point scale (0, no irreversible damage; 1, partially damaged; 2, severe damage, ankylosis, or prosthesis). The RAAD score was determined for 121 patients with RA with a large range of disease duration. Interobserver agreement was studied in 39 patients scored by three observers. Data on disease duration, Health Assessment Questionnaire, disease activity score, and Larsen score were collected for 121, 78, 47, and 45 patients, respectively.
Results: The RAAD score correlated well with the Larsen score (r_s=0.81) and disease duration (r_s=0.68) and (as intended) not with disease activity (r_s=0.10). Good interobserver agreement was found for total scores and individual joints. The wide range of RAAD scores for patients with the same disease duration suggested good discriminating power, especially after >10 years.
Conclusion: The RAAD score is a quick and feasible method for measuring the long term articular damage in large RA populations. It has good reliability and construct validity and deserves further study to assess its discriminant validity.

Disease status in rheumatoid arthritis (RA) can be expressed in terms of inflammatory activity or of damage. Indices of disease activity are reversible, whereas measures of damage should represent the irreversible results of disease activity over time. Although damage in RA can occur in skin or organs by amyloidosis or vasculitis, joint damage is the most prominent feature of disease outcome. Articular damage is generally assessed by radiographs, which may show the destruction of bone and cartilage. Several radiological damage scores have been developed, each having specific characteristics for reproducibility and sensitivity to change.[1] Despite their usefulness in studying disease progression, there are some drawbacks. Firstly, radiographs represent mainly osseous changes, whereas part of the articular damage in RA is in the soft tissues surrounding the bones. Secondly, methods for scoring radiographic damage concentrate on the hands and feet, whereas damage in larger joints may be of equal importance for a patient's functional ability. Thirdly, the cost of measuring radiographic damage makes these methods less suitable for studying large numbers of patients or for use in developing countries.

Plant *et al* found that a rheumatologist could predict the Larsen radiographic score by clinical examination with surprising accuracy in the small hand joints (though less so in the feet).[2] Kuper *et al* showed that radiographic damage in large joints was significantly related to the damage in hands and feet, a physical disability index, and cumulative disease activity.[3]

Observations like these suggest that it is possible to develop a score for irreversible articular damage, based on clinical examination of large and small joints, which may be useful for measuring long term damage in large patient groups. Such a score would be helpful in comparing the effects of different treatment strategies or the results in different rheumatology centres or in different countries, particularly after longer disease duration.

Attempts to design a clinical damage score in RA have been published before, but until now these have not been widely used. Symmons *et al* developed the OSRA, a simple measure of overall status.[4] In its section on damage, the number of destroyed large joints, and the need for splints, collar, special shoes or surgery on small joints are used as a measure of articular damage. Recently Cranney *et al* reported their deformity index,[5] a measure of limited joint motion and deformity, adapted from the joint alignment and motion scale[6] and the Escola Paulista de Medicina-range of motion scale[7]. However, the deformity index was not formally validated.

PATIENTS AND METHODS

We set out to design a score that is quick and easy to obtain, using information obtained by physical examination. It should measure irreversible damage, which implies that the score can only increase over time. Based on our clinical experience and common sense we formulated the rheumatoid arthritis articular damage (RAAD) score. In this method 35 joints or joint groups are scored on a three point scale (0, no irreversible damage; 1, partly damaged; 2, severe damage, ankylosis, or prosthesis). The definitions for scoring each joint are concise in order to make the method accessible to inexperienced assessors (table 1). The only tool needed is a goniometer, although most joints can be assessed without one. Metatarsophalangeal (MTP) joints of each foot are scored as a single joint.

Abbreviations: DAS, disease activity score; HAQ, Health Assessment Questionnaire; MCP, metacarpophalangeal; MTP, metatarsophalangeal; PIP, proximal interphalangeal; RA, rheumatoid arthritis; RAAD, rheumatoid arthritis articular damage

Table 1 The RAAD score: definitions for scoring damage in individual joints. Contractures and other deformities should only be scored when they are expected to be irreversible without surgery

Joint type	Definitions for scoring irreversible articular damage	Max score
Cervical spine	1: severe limitation of motion, ankylosis, or known cervical subluxation	
	2: history of medullary compression or surgical fusion	2
Shoulder	1: external rotation <45° (anatomical limitation, not due to pain), or severe crepitus	
	2: ankylosis or prosthesis	4
Elbow	1: flexion contracture <30°	
	2: flexion contracture >30°, ankylosis, history of radial head resection, or prosthesis	4
Wrist	1: extension or flexion <30°, or volar/ulnar/radial shift	
	2: ankylosis, prosthesis or history of ulnar head resection	4
MCP*	1: ulnar deviation	
	2: subluxation, ankylosis, or prosthesis	20
PIP	1: flexion contracture	
	2: Swan neck or boutonnière deformity, ankylosis, or prosthesis	20
Hip	1: internal rotation <10°	
	2: prosthesis or Girdlestone pseudarthrosis	4
Knee	1: medial or lateral deviation >10° due to arthritis, or flexion contracture <20°	
	2: flexion contracture >20° or prosthesis	4
Ankle	1: fixed valgus deformity <20°	
	2: prosthesis, ankylosis, arthrodesis, or valgus deformity >20°	4
MTP	1: visible deformity due to arthritis	
	2: history of Kates-Kessel or other arthroplasty of the forefoot	4
Maximum total score		70

*After analysis of our results we suggest that MCP 1 should be scored "1" if it shows impairment of normal extension (see text).

Patients fulfilling the 1987 ACR criteria for RA[8] were selected from our outpatient clinic and rheumatology ward to obtain a sample with a wide range of disease duration and activity. Forty seven patients (17 male, 30 female) gave informed consent. Their mean age was 63 years (range 27–84), and mean disease duration was 16 years (range 1–48). In these patients, an RAAD score was determined by three observers on the same day: an experienced rheumatologist (HBM), a rheumatology trainee (TZ), and a rheumatology nurse specialist with little experience in joint examination, who had a brief training in using the score. Apart from the original score (RAAD-1), we also computed two alternative RAAD scores. In RAAD-2 only the number of damaged joints was counted. In RAAD-3 all joints were scored 1 or 2, and metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joints of each hand were scored as a single joint.

On the same day, the 28 joint counts for tenderness (T) and swelling (S) were done, patients completed a visual analogue scale for general wellbeing (G), and the erythrocyte sedimentation rate (ESR) was measured. From these variables the disease activity score (DAS28) was calculated, using the formula: $DAS28 = 0.56T + 0.28S + 0.70\ln ESR + 0.014G$.[9]

All patients completed a Dutch version of the Health Assessment Questionnaire (HAQ).[10] For each patient, recently taken radiographs of hands and feet were collected, or a new set of radiographs was obtained. These radiographs were scored using the Larsen method,[11] in which the first metatarsal joint was left out and the wrist was scored as one joint and multiplied by five (maximum score 190). All radiographs were scored by two observers with any initial disagreement finally agreed by consensus.

A random selection of 121 patients with RA of known disease duration, visiting our outpatient clinic, were scored by HBM using the RAAD score. In 78 of them, an HAQ score was obtained on the same day. These cross sectional data were used to get an impression of the RAAD scores over time.

For statistical analysis SPSS was used. To assess interobserver variability, RAAD scores of three observers were analysed with Spearman's rank correlation and Friedman's test. Spearman's rank correlation was also used to assess correlation between the RAAD score, DAS28, disease duration, and the Larsen score. κ Statistics[12] were used to assess the interobserver agreement of the RAAD score for individual joints.

Table 2 Results in 47 patients, showing means, standard deviations, minimum and maximum values

Variable	Mean	SD	Min	Max
Age (years)	63.3	13.1	27	84
Disease duration (years)	16	12	1	48
Disease activity score	5.0	1.4	2.0	8.2
HAQ	1.1	0.8	0.0	2.9
RAAD observer 1 (n=47)	10.3	11.7	0	51
RAAD observer 2 (n=41)	11.2	12.7	0	51
RAAD observer 3 (n=45)	11.3	12.3	0	53
Larsen score (n=45)	68	44	1	185

RESULTS

The RAAD score appeared to be easily applicable. After a short learning period, it took about two minutes for each patient, depending on the amount of damage.

Table 2 shows the mean scores in 47 patients. All patients were assessed by observer 1, 41 by observer 2, and 45 by observer 3. Thirty nine patients were seen by all three observers. In 45 patients radiographs of hands and feet were available.

Figure 1 shows the relation between the RAAD scores of two different observers. Spearman's r_s for rheumatologist *v* rheumatology nurse was 0.89. For rheumatologist *v* trainee it was 0.90, for trainee *v* rheumatology nurse 0.95. Different ways of computing the RAAD score (RAAD-2 and -3) showed similar degrees of correlation. Because correlation is not the same as agreement, we also compared the RAAD scores of three observers. Using Friedman's non-parametric test for multiple related samples, we found no statistically significant difference between observers.

Table 3 shows the κ values for interobserver agreement in individual joints. There was moderate to good agreement for most joints, although agreement for MCP 1 and the PIP joints

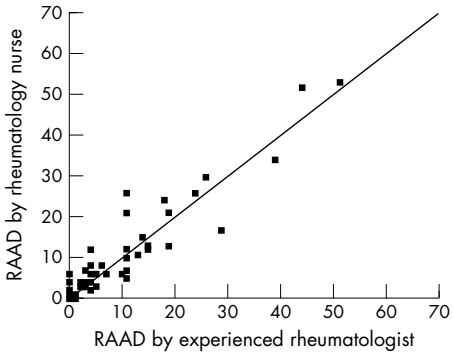

Figure 1 Scatterplot showing correlation of RAAD scores by two different observers. The diagonal line indicates perfect agreement. For these observers, Spearman's r_s=0.89 (n=45, p<0.01).

was less favourable. κ Values for the ankle joint could not be computed for observer 2 because he never scored a value of 1 for this joint.

In table 4 correlation between the RAAD score and other measures is shown using Spearman's r_s. The RAAD score correlated well with the Larsen score. There was no significant correlation between the RAAD score and DAS28, whereas HAQ and DAS28 did correlate (Spearman's r_s=0.42, p=0.003).

Figure 2 shows data on disease duration *v* RAAD score for 121 patients assessed by observer 1. There was a wide range of RAAD scores for patients with the same disease duration, reflecting the large variation in disease outcome in RA. Spearman's rank correlation for RAAD-1 with disease duration was 0.68, for RAAD-2 was 0.67, and for RAAD-3 was 0.69 (p<0.001). This indicates that alternative methods of computing the score did not influence correlation with disease duration in this group of patients.

DISCUSSION

Tugwell and Bombardier described the quality of scoring methods in terms of feasibility, reliability, validity, and responsiveness.[13] The RAAD score is a cheap and quick method of assessing damage. We think it is feasible for studying long term damage in large groups of patients—for instance, in comparing outcome between hospitals or countries, or long term (>5 years) treatment strategies.

We tested interobserver variability and found little difference between the results of experienced and inexperienced observers. Interobserver variability was low for the total score as well as for most individual joints. The rather low level of agreement for the first MCP joint may have been caused by an inadequate definition: ulnar deviation was proposed for all MCP joints. For MCP 1, impairment of normal extension would be more appropriate as an intermediate (grade I) damage score. In the PIP joints, grade I damage may be difficult to assess because of pre-existing osteoarthrosis, and should perhaps be omitted. In the ankle joint, a damage score of 1 or 2 occurred only in a minority of patients. We believe this reflects the actual low occurrence of damage in this joint.

In some joints, particularly the cervical spine, shoulder, and hip, it may be difficult to distinguish irreversible damage from reversible impairment due to inflammation. If, for instance, shoulder movement is impaired, the observer has to decide if this is a fixed impairment or one that might improve after a corticosteroid injection.

Clinical assessment of cervical instability may be difficult. If from earlier radiographs a patient is already known to have significant cervical instability, we do not object to using this information in the RAAD score. However, no new radiographs should be taken for this purpose.

More detailed definitions or more extensive training and consensus meetings may reduce interobserver variability. However, both strategies make the method less easily applicable for wide scale use. Therefore, we prefer to keep definitions as simple as possible.

From the rather high level of interobserver agreement in our sample, we do not expect much intraobserver variability to

Table 3 Value of weighted κ and percentage of joints in which observers agreed, for separate joints in 39 patients. Left and right side of the body were summed. Agreement is considered poor if κ<0.20, fair if 0.21–0.40, moderate if 0.41–0.60, good if 0.61–0.80, very good if 0.81–1.00

	Observer 1 *v* 2		Observer 1 *v* 3		Observer 2 *v* 3	
Joint type	κ	% Agreed	κ	% Agreed	κ	% Agreed
Cervical spine	0.60	87	0.54	85	0.80	92
Shoulder	0.83	92	0.64	82	0.72	87
Elbow	0.72	82	0.72	82	0.85	90
Wrist	0.59	67	0.66	70	0.56	67
MCP 1	0.49	85	0.30	86	0.22	83
MCP 2	0.75	85	0.57	77	0.64	79
MCP 3	0.70	85	0.59	81	0.75	87
MCP 4	0.61	85	0.51	79	0.83	92
MCP 5	0.63	86	0.50	79	0.66	85
PIP 1	0.49	92	0.37	86	0.42	88
PIP 2	0.49	82	0.51	81	0.39	81
PIP 3	0.46	77	0.46	77	0.36	77
PIP 4	0.39	82	0.36	83	0.45	87
PIP 5	0.75	94	0.39	87	0.31	85
Hip	0.69	87	0.82	94	0.85	94
Knee	0.79	94	0.76	92	0.76	91
Ankle	–	96	0.15	85	–	87
MTP	0.79	82	0.78	79	0.78	79

Table 4 Spearman's rank correlation of RAAD score with disease duration, radiographic damage (Larsen), disease activity (DAS), and functional capacity (HAQ)

	RAAD versus:			
	Disease duration	Larsen	DAS	HAQ
Number of patients	121	45	47	78
Spearman's r_s	0.68 (p<0.01)	0.81 (p<0.01)	0.10 (p=0.51)	0.50 (p<0.01)

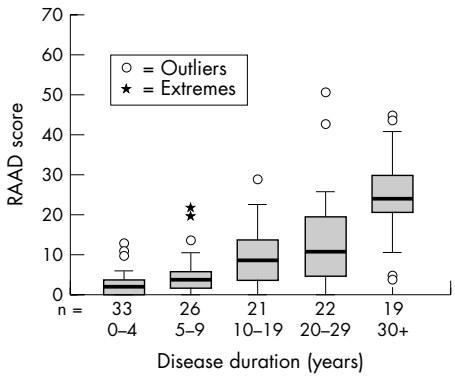

Figure 2 Box plot showing results of RAAD score *v* disease duration for 121 patients scored by observer 1. Boxes with horizontal lines represent interquartile range and median. Outliers and extremes are indicated separately. Note that the boxes cover various ranges of disease duration. The numbers of patients are displayed on the x axis.

occur. However, further study should focus on this aspect, especially since variation in swelling and inflammation within a single patient may influence the score in some joints.

Five aspects of the validity of outcome measures in RA can be distinguished: face validity, content validity, construct validity, criterion validity, and discriminant validity. Face validity means credibility. We believe that our clinical definitions are a sensible way of describing articular damage and have good face validity, but we welcome any suggestions for improvement.

Content validity deals with the question of whether a measure covers all aspects of the subject. We think that assessing clinical damage in large and small joints will render better content validity than assessing radiographic changes in small joints only.

To assess construct validity (that is, does this method correspond to theoretical concepts in articular damage?) we studied the correlation of the RAAD score with a number of other variables. As we expected, there is a positive correlation with disease duration and HAQ score (convergent validity) and no correlation with actual disease activity (divergent validity).

Criterion validity (does the score correlate with the gold standard?) is difficult to assess because there is no gold standard for articular damage. We chose the Larsen score as a substitute and found good correlation with our damage score. A simple damaged joint count (RAAD-2) lacks information on severity of damage and offered no substantial time saving. We also studied a more or less weighted score (RAAD-3), because the PIP and MCP joints seem to be overweighted in RAAD-1. In our study group this modification did not change the properties of the score significantly, but we are currently collecting data on larger numbers of patients to study this item properly. For the time being, we recommend the original RAAD score, not its modifications.

We developed the RAAD for measuring long term damage. Assessing its discriminant validity (responsiveness or sensitivity to change) is best done in a prospective design, but this would take at least 10 years. Instead, one could compare RAAD scores from subsets of patients with RA in whom a difference in outcome (for example, rheumatoid factor positive and rheumatoid factor negative patients) is expected.

CONCLUSIONS

We developed the RAAD score as a clinical method for scoring long term articular damage in large groups of patients with RA. It is easy to perform and showed good interobserver reliability, even when used by an inexperienced observer. It correlates well with Larsen scores and with disease duration but not with disease activity, demonstrating its criterion and construct validity. Before recommending its use for research or follow up of patients with RA, its inter- and intraobserver variability and discriminant validity need to be assessed.

ACKNOWLEDGMENTS

Part of this study was performed while TR Zijlstra visited the ARC Epidemiology Unit in Manchester. He wishes to thank Professor D Symmons and other staff members for their hospitality and useful advice.

This study was supported by a Novartis Rheumatology Grant.

Authors' affiliations

T R Zijlstra, H J Bernelot Moens, Department of Rheumatology, Medisch Spectrum Twente, Enschede, The Netherlands
M A S Bukhari, ARC Epidemiology Unit, The University of Manchester, United Kingdom

REFERENCES

1 **Van der Heijde DMFM**. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillieres Clin Rheumatol 1996;10:435–53.
2 **Plant MJ**, Saklatvala J, Jones PW, Dawes PT. Prediction of radiographic damage in hands and feet in rheumatoid arthritis by clinical evaluation. Clin Rheumatol 1994;13:487–91.
3 **Kuper HH**, van Leeuwen MA, van Riel PLCM, Prevoo ML, Houtman PM, Lolkema WF, *et al*. Radiographic damage in large joints in early rheumatoid arthritis: relationship with radiographic damage in hands and feet, disease activity, and physical disability. Br J Rheumatol 1997;36:855–60.
4 **Symmons DPM**, Hassell AB, Gunatillaka KAN, Jones PJ, Schollum J, Dawes PT. Development and preliminary assessment of a simple measure of overall status in rheumatoid arthritis (OSRA) for routine clinical use. Q J Med 1995;88:429–37.
5 **Cranney A**, Goldstein R, Ba' Pham, Karsh J. A measure of limited joint motion and deformity correlates with HLA-DRB1 and DQB1 alleles in patients with rheumatoid arthritis. Ann Rheum Dis 1999;58:703–8.
6 **Spiegel TM**, Spiegel JN, Paulus HE. The joint alignment and motion scale: a simple measure of joint deformity in patients with rheumatoid arthritis. J Rheumatol 1987;14:887–92.
7 **Ferraz MB**, Oliviera LM, Araujo PMP, Atra E, Walter SD. EPM-ROM scale: an evaluative instrument to be used in rheumatoid arthritis trials. Clin Exp Rheumatol 1990;8:491–4.
8 **Arnett FC**, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, *et al*. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
9 **Prevoo MLL**, van 't Hof MA, Kuper HH, van Leeuwen MA, van de Putte LBA, van Riel PLCM. Modified disease activity scores that include 28-joint counts. Arthritis Rheum 1995;38:44–8.
10 **Van der Heijde DMFM**, van Riel PLCM, van de Putte LBA. Sensitivity of a Dutch health assessment questionnaire in a trial comparing hydroxychloroquine vs sulphasalazine. Scand J Rheumatol 1990;19:407–12.
11 **Larsen A**, Dale K, Eek M. Radiographic evaluation of rheumatoid arthritis and related conditions by standard reference films. Acta Radiol Diagn 1977;18:481–91.
12 **Cohen J**. A co-efficient of agreement for nominal scales. Educational and Psychological Measurement 1960;20:37–47.
13 **Tugwell P**, Bombardier C. Methodological framework for developing and selecting end points in clinical trials. J Rheumatol 1982;9:758–62.