Reliability of Measures of Disease Activity and Disease Damage in Rheumatoid Arthritis: Implications for Smallest Detectable Difference, Minimal Clinically Important Difference, and Analysis of Treatment Effects in Randomized Controlled Trials

MARISSA N.D. LASSERE, DÉSIRÉE van der HEIJDE, KENT R. JOHNSON, MAARTEN BOERS, and JOHN EDMONDS

ABSTRACT. We evaluate measurement properties of common rheumatoid arthritis (RA) assessments. Included are a comprehensive literature review and new data on the reliability and smallest detectable difference (SDD) for different classes of these measures. We found that certain common measures such as joint counts, pain, and patient global all had poor reliability and showed large SDD compared to multi-item measures of physical/psychological function or compared to radiographic measures. We discuss the implications of these findings on the use of composite endpoints such as the ACR20 or the EULAR responder index in RA clinical trials, particularly the introduction of misclassification bias that arises from differential measurement error. Finally, we consider generically how the concept of the SDD might or might not relate to the concept of the minimal clinically important difference. (J Rheumatol 2001;28:892-903)

Key Indexing Terms:
MINIMAL CLINICALLY IMPORTANT DIFFERENCE
RANDOMIZED CONTROLLED TRIALS
HEALTH STATUS
RELIABILITY
ARTICULAR ASSESSMENT
RADIOGRAPHS

> If assessment methods are insensitive, then no significant differences will be the outcome of the clinical trial, but if the errors of the method employed have not been properly ascertained then a significant difference may be found where none exist.
>
> — Bradford Hill, 1963

INTRODUCTION

Chronic symptomatic diseases are inherently complex to describe and study, and rheumatoid arthritis (RA) exemplifies this challenge. Recent work on clinical trial outcome measures[1-3], measurement variability [the smallest detectable difference (SDD) — see below], and definitions of the minimum clinically important difference (MCID) are all parts of this process. Many assessment measures have been assembled, but there has been little critical evaluation of this repertoire using modern measurement theory[4]. Reliability, in particular, has not been systematically investigated in the literature. Our article is a first step in assessing the measurement properties of common RA measures. It focuses on a literature review and new data on reliability for different classes of RA measures, and it speculates about implications of these data for MCID research, for certain composite measures, and for randomized controlled trial (RCT) design and analysis.

From the Department of Rheumatology, St. George Hospital, Sydney, Australia; Department of Rheumatology, University Hospital Maastricht, Maastricht, The Netherlands; Limburg University Center, Diepenbeek, Belgium; CDER, Food and Drug Administration, Rockville, Maryland, USA; and Department of Clinical Epidemiology, VU University Hospital, Amsterdam, The Netherlands.

M. Lassere, MB, BS, PhD, FAFPHM, FRACP, Staff Specialist in Rheumatology, Department of Rheumatology, St. George Hospital; D. van der Heijde, MD, PhD, Professor of Rheumatology, Department of Internal Medicine, Division of Rheumatology, University Hospital Maastricht, and Limburg University Center; K. Johnson, MD, Medical Officer, Center for Drug Evaluation and Research, FDA (Note: Views are not necessarily those of the FDA); M. Boers, MSc, MD, PhD, Rheumatologist, Professor of Clinical Epidemiology, Department of Clinical Epidemiology, VU University Hospital; J. Edmonds, MB, BS, FRACP, Professor of Rheumatology, Department of Rheumatology, St. George Hospital.

Address reprint requests to Dr. M. Lassere, Staff Specialist in Rheumatology, Department of Rheumatology, St. George Hospital, Kogarah, NSW 2217, Australia. E-mail: lasserem@sesahs.nsw.gov.au

Reliability and the SDD

Differences between individuals revealed by a measuring method should represent, to the greatest extent possible, real differences in the phenomenon of interest. Because all measurement entails error, we have defined the smallest detectable difference as that amount of difference for which anything smaller cannot be reliably distinguished from random error in the measurement. It reflects random variability, as seen in repeated measures by one particular observer over a time during which all other causes of variability can be assumed constant, for example, repeating a

swollen joint assessment repeated in the same patient within a one hour period.

The SDD is one of several descriptions of random variability, or, differently viewed, of the reliability or precision of measurements on a continuous scale. The SDD directly measures the amount of variation in the original scale units. It can be calculated using the standard error of the measurement or the limits of agreement method. The latter method has been used to calculate the SDD of radiological progression[5].

Directly measuring reliability is advantageous because it provides a useful yardstick of error in scale based units. Variability can also be expressed as a ratio of variances — a reliability coefficient (RC), which is an indirect measure of random measurement error. However, because it is unitless (scale independent), the RC can easily be used to compare reliability of measurements across different phenomena. Furthermore, they can be used to determine the effect of measurement error on treatment effects observed in RCT.

The SDD and the reliability coefficient are context-, patient-, and observer-specific[6]. Nonetheless, they harbor information fundamental to evaluating the use of measures generally. They offer a way to rank measures. Narrow SDD — those that extend only a small portion of the total scale — may indicate a more useful measurement than broad SDD. Finally, the SDD concept can be extended beyond individuals to populations by characterizing distributions of SDD in those populations.

The minimum clinically important difference (MCID)

The motivation for defining the minimum clinically important difference is to understand and ultimately to inform the dynamics of the therapeutic encounter of patient and clinician. Given the complexities of clinical medicine, it is not surprising that agreement on the definition of the MCID has been elusive. Three strategies in MCID research have been described[7]. One approach uses statistical descriptions of the population ("distribution based"), a second relies on experts ("opinion based"), and a third is based on sequential hypothesis formation and testing ("predictive/data driven based"). The distribution based model, in turn, has 3 methods of estimation: the effect size (ES), the standard error of measurement (SEM), and the limits of agreement (LOA) (see Table 1). The latter 2 are more intuitive for clinicians because the MCID remains in the original scale of the measure of interest.

Relation of the SDD to the MCID

How the SDD should or might relate to the MCID is unclear. In the absence of robust, predictive epidemiology to invoke in a data driven approach (as, for example, can be done in the case of blood pressure and serious vascular outcomes), one might propose that the SDD should at least constitute a provisional "lower bound" for the MCID. However, if such epidemiologic evidence did exist, then changes even smaller than the SDD, indeed any real change, may well be shown to be important on clinical and public health grounds, especially if it translated into a reduction of serious clinical outcomes. At this point in rheumatology we can proceed by investigating the SDD of commonly used measures in RA.

Reliability and its implications for the analysis of treatment effects in RCT

Variability of any cause will negatively affect the power of a RCT, so within-patient measurement variability is a cause for concern. Furthermore, more variability is introduced by use of a difference score (endpoint minus baseline) rather than a static score, because the measurement errors add together, while whatever is common to the 2 measures cancels out[8].

The question of interest in an RCT is what response difference has been demonstrated between groups treated differently (the top level in the cube of discrimination[9]). For most measures (endpoints), the variability between patients will be greater than that within patients, so the response of interest is often smaller than differences seen between individuals. Thus, it becomes advantageous to adjust the end-of-trial score (result) by the baseline score. One also relies on randomization to balance, on average, baseline scores across treatment groups. The easiest, although not most efficient adjustment[10], is to calculate the difference score (end minus baseline). However, if the measure has a large within-patient

Table 1. Distribution based models of defining minimum clinically important difference (MCID).

Name	Method of Calculation	Units
Effect size	Mean change divided by SD at baseline.	Unitless
Standard error of measurement	SD of measurement multiplied by the square root of 1 minus the reliability coefficient of the measurement, i.e., $\sigma\sqrt{1-RC}$. Any one of several reliability coefficients can be used including the various intraclass correlation coefficients[20].	Original scale units
Limits of agreement	The CI for the difference between measurements on the same subject is estimated by $mean_{diff} - t_{0.05,\,n-1} \bullet (SD_{diff})$ and $mean_{diff} + t_{0.05,\,n-1} \bullet (SD_{diff})$ where $t_{0.05,\,n-1}$ is the critical point of the t distribution.	Original scale units

SD: standard deviation. CI: confidence interval.

variation over time, the advantage of a difference score over a simple end-of-trial score is lost[10]. In addition, the error of a difference score is increased (therefore its reliability decreased) over a component because component errors accumulate in the difference score[8,11]. This loss of reliability and its implication in the design and analysis of RCT is not well recognized.

Another, and perhaps more important, example of how within-patient measurement variability (or error) can cause problems for RCT analyses is the use of responder analyses such as the American College of Rheumatology (ACR) and EULAR improvement criteria[1,2,12], which require dichotomizing or trichotomizing the observed response of a continuous measure based on predefined cutpoints. Responder analyses are readily understood by clinicians; they want to know how many patients improved, rather than the average of some measure. However, this clinical attractiveness may be falsely reassuring for 2 reasons. (1) Information, and thus power, is lost when an intrinsically continuous measure is dichotomized; this is generally known. (2) Measurement error can bias both the size and direction of estimates of treatment effects in RCT when the measure is a dichotomized responder variable, if the measurement error systematically differs across treatment arms. The differential measurement effect across treatment arms causes unbiased estimates on a continuous scale to become biased estimates of the proportion of patients responding[13]. If there is no measurement error, as when death is the endpoint, the reliability coefficient is 1.0, and misclassification error shrinks to zero. However, most measures have error, consequently the reliability coefficient is less than 1.0 and nondifferential misclassification bias may occur.

METHODS

Field studies of reliability in RA

Joint examination. The interobserver agreement of joint examination measures (see Appendix I) by 2 rheumatologists was tested throughout the course of a longitudinal study of outcome in RA (Study A). The observers met prior to the study to briefly discuss the method of the examination and criteria for scoring, but no formal training session was held. All patients were examined in quick succession by each rheumatologist working independently and blinded to other joint recordings. Patients were examined at the same time and place on each occasion. The study population was 10 patients from the rheumatology clinic of a Sydney teaching hospital who met ACR criteria for RA. Patients were selected to reflect the spectrum of disease activity and damage.

Patient self-report measures. Study B evaluated the test-retest reliability of patient self-report physical function using the Health Assessment Questionnaire (HAQ) Disability Index[14] and patient pain and global assessment [measured on a horizontal 100 mm visual analog scale (VAS) with a score of 0 representing the best and 100 the worst situation] in 24 patients attending routine followup visits to their rheumatologist. Questionnaires were administered on day 1 and day 8.

Study C evaluated the test-retest reliability of patient self-report physical function using the HAQ Disability Index, and generic health status using the Medical Outcome Survey Short Form-36 (SF-36)[15,16] in 26 patients with RA. Questionnaires were administered on day 1 and day 2.

Radiographic measures. Forty-eight radiographs of the hands and wrists of patients with RA were scored by 2 rheumatologist observers (Study D) using the Scott/Larsen method (0–150)[17]. The 48 radiographs were scored in random order, independently and blindly, after a preliminary standardization session. A second set of 20 radiographs of the hands and wrists of patients with RA were scored by 2 rheumatologist observers (Study E) using the Scott/Larsen method (0–150) scored. The radiographs were read in random order, independently and blindly, after several training and standardization sessions. A third set of radiographs (135 patients; hands, wrists, feet) from a randomized controlled trial of early RA (COBRA study) were read paired and chronologically by 2 observers (Study F) using the van der Heijde modified Sharp method (0–448 scale)[18]. Scores at baseline and at 12 months were analyzed separately. Finally, a fourth set of radiographs (hands, wrists, feet) of 52 patients, selected to be representative of the spectrum of radiological progression, were obtained from a randomized controlled trial of early RA (COBRA study). They were read paired and chronologically by 2 observers (Study G) using the van der Heijde modified Sharp method (0–448 scale) and by another 2 observers using the Scott modified Larsen method (0–200). The scores at baseline and at 12 months were analyzed separately.

Data extraction from review of published studies

The Medline database (1966–1996) was searched for RA studies on joint examination measures and radiographic measures, and for studies on developing and testing of the HAQ and SF-36 questionnaires (English and non-English questionnaires). The references of all located articles were scanned for any unidentified articles. One reviewer (ML) critically appraised all papers and described details on reliability according to the statistical method used. If the data were provided, SDD (by limits of agreement) and fixed effects interclass correlation coefficient (ICC) were calculated.

Statistical analysis. Bland and Altman's 95% limits of agreement method was used to estimate the smallest detectable difference of radiological progression[5,19]. The mean ($mean_{diff}$) and standard deviation (SD_{diff}) of the paired differences of radiological progression between readers were calculated. Because radiological progression was calculated in all studies using the mean score of 2 readers, one must first divide the SD of the paired differences by the square root of 2. The confidence interval (CI) for the difference between measurements on the same subject is estimated by: $mean_{diff} - t_{0.05,\,n-1} \bullet (SD_{diff}/\sqrt{2})$ and $mean_{diff} + t_{0.05,\,n-1} \bullet (SD_{diff}/\sqrt{2})$ where $t_{0.05,\,n-1}$ is the critical point of the t distribution. This CI is the 95% limits of agreement as defined by Bland and Altman, and is the SDD of radiological progression for the mean score if the same 2 readers are used in all studies. However, if different readers are used across studies, then the 95% upper and lower limits of agreement in turn have their own 95% CI, and these wider limits must be used to determine the SDD of radiological progression for the mean score of *any* 2 readers. The 95% CI around the upper and lower limits were determined first by calculating the standard error of the limits of agreement, SE_{limits}, which is roughly $\sqrt{[(\{3*(SD_{diff}/\sqrt{2})\}^2)/n]}$. Therefore the 95% CI for the lower limit of agreement is: $mean_{diff} - t_{0.05,\,n-1} \bullet (SD_{diff}/\sqrt{2}) \pm t_{0.05,\,n-1\,(df)}(SE_{limits})$ and the 95% CI for the upper limit of agreement is: $mean_{diff} + t_{0.05,\,n-1} \bullet (SD_{diff}/\sqrt{2}) \pm t_{0.05,\,n-1\,(df)}(SE_{limits})$. Finally, to facilitate comparisons across measures, we provide the fixed effects ICC (Type 1.2)[20]; and for each SDD, the percentage of the maximum actual score (the range in that dataset, smallest to largest) and maximum possible score (full definition of the measure) were calculated.

RESULTS

Tables 2 and 3 summarize the demographic and clinical characteristics of the patients with RA participating in these studies. Although differences in disease duration reflect different source populations (clinical practice vs early

Table 2. Characteristics of patients: joint examination and self-report measures.

Characteristic	Joint Examination Study A	Patient Self-report Clinical Study B	Patient Self-report Clinical Study C
No of patients	10	24	26
Male, %	20	12.5	23
Age, yrs, mean	65	61	56
Disease duration, yrs, mean	14	16	6
HAQ, mean	1.3	1.3	1.0
Patient global 100 mm VAS, mean	42	37	NA
Patient pain 100 mm VAS, mean	46	37	37
RF positive (ever), %	80	84	75
Erosive, %	80	79	40

VAS: visual analog scale. RF: rheumatoid factor.

Table 3. Characteristics of patients: radiographic measures.

Characteristic	Radiographs Study D	Radiographs Study E	Radiographs Study F	Radiographs Study G
No. of patients	48	20	147	52
Male, %	18	15	41	47
Age, yrs, mean	59	58	50	48
Disease duration, yrs, mean	11	12	5 mo	6 mo
HAQ, mean	1.3	1.2	1.4	1.4
Patient global 100 mm VAS, mean	41	39	50	51
Patient pain 100 mm VAS, mean	48	39	54	52
RF positive (E = ever, C = current), %	67 (E)	80 (E)	58 (C)	NA
Erosive, %	73	75	74	NA

VAS: visual analog scale. RF: rheumatoid factor.

Table 4. Interobserver reliability of joint examination: Study A (10 patients, 2 observers).

Measure	Scale	Mean	Maximum Score	ICC	Mean Difference	SDD 95% LOA	SDD Maximum Actual Score, %	SDD Maximum Possible Score, %
Tender 28	0–28	5.3	14	0.64	0.9	–6.9, 8.7	± 56	± 12
Tender 74	0–74	10.4	30	0.81	2.3	–8.8, 13.4	± 37	± 21
Swollen 28	0–28	5.1	19	0.52	1.0	–11.4, 13.4	± 65	± 43
Swollen 68	0–68	6.7	20	0.52	0.6	–12.7, 13.9	± 67	± 20

ICC: Interclass correlation coefficient. SDD: Smallest detectable difference. LOA: Limits of agreement.

disease RCT), the patient groups are generally similar on many demographic and clinical characteristics.

The statistical analysis of reliability of the joint examination is listed in Table 4, that for the self-report questionnaire measures in Table 5, and that for the radiographic status scores in Table 6. Overall, joint swelling, joint tenderness, VAS patient pain, and patient global assessment had very poor reliability and large SDD compared to the multi-item measures of physical and psychological function and radiographic measures.

Although the mean difference between joint examinations was small, the SDD were large and neither the tender nor swollen joint assessments achieved ICC greater than 0.9. Overall, the swollen joint count had poorer reliability than the tender joint count. Although joint examination signs may change as a result of repeated examination, steps were taken to limit this source of variability. The evaluation was limited to 2 examiners separated by only a 30 min interval. The test-retest reliability of the self-report measures varied in performance. In general, multi-item single dimension measures (e.g., test-retest HAQ, SF-36 physical function, SF-36 mental health dimensions) had

Table 5. Test-retest reliability of patient self-report clinical measures: Studies B (24 patients, day 1 vs day 8) and C (26 patients, day 1 vs day 2).

Measure	Scale	Mean	Maximum Score	ICC	Mean Difference	SDD 95% LOA	SDD Maximum Actual Score, %	SDD Maximum Possible Score, %
Study B								
HAQ	0–3	1.2	2.875	0.91	–0.05	–0.69 to 0.59	± 22	± 21
Pain	0–100	37	93	0.75	–4.8	–54 to 44	± 53	± 49
Patient Global	0–100	37	80	0.75	–4.7	–41 to 32	± 46	± 37
Study C								
HAQ	0–3	0.94	2.25	0.95	0.10	–0.29 to 0.48	± 17	± 13
SF36 PCS*	0–100	36	54	0.87	–1.4	–12.7 to 9.9	± 21	± 11
SF36 MCS*	0–100	55	65	0.70	1.8	–6.7 to 10.4	± 13	± 9
Physical function	0–100	55	95	0.84	–3.0	–29 to 24	± 28	± 27
Role physical	0–100	40	100	0.77	–1.0	–59 to 57	± 58	± 58
Pain	0–100	52	89	0.75	0.8	–26 to 28	± 30	± 27
General health	0–100	53	92	0.94	–0.8	–18 to 16	± 18	± 18
Social	0–100	68	100	0.74	–0.5	–38 to 37	± 38	± 38
Role mental	0–100	74	100	0.58	1.5	–66 to 70	± 68	± 68
Vitality	0–100	59	85	0.95	–0.4	–11 to 11	± 13	± 11
Mental health	0–100	76	100	0.93	–0.2	–12 to 12	± 12	± 12

For definitions see Table 4.

Table 6. Interobserver reliability of radiographic assessment: Studies D, E, F [at Baseline (1) and at 1 year followup (2)] and G [at Baseline (1) and at 1 year followup (2)].

Study	n	Scale	Mean	Maximum Score	ICC	Mean Difference	SDD 95% LOA	SDD Maximum Actual Score, %	SDD Maximum Possible Score, %
Scott/Larsen									
Study D	48	0–150	43	122	0.95	1.39	–19.8, 22.6	± 17	± 13
Study E	20	0–150	56	123	0.99	–1.45	–11.4. 8.5	± 8	± 7
Study G (1)	52	0–200	15	57	0.86	–0.27	–17.0, 16.5	± 30	± 8
Study G (2)	52	0–200	24	65	0.84	–1.06	–20.7, 18.6	± 30	± 10
van der Heijde/Sharp									
Study F (1)	147	0–448	8.5	60	0.90	0.4	–10.0 to 10.9	± 17	± 2
Study F (2)	135	0–448	17.6	106	0.92	–1.3	–17.7 to 15.1	± 15	± 4
Study G (1)	52	0–448	13	60	0.92	–2.1	–14.4 to 10.3	± 21	± 3
Study G (2)	52	0–448	27	97	0.92	–0.5	–19.2 to 21.5	± 21	± 4

For definitions see Table 4.

better reliability (larger ICC and smaller SDD) than single item measures using VAS (pain, patient global). Examination of the radiographic results clearly show how the conditions external to the measurement process influence the magnitude and direction, and thereby the variability, of the SDD. Studies F and G differed by one condition — the latter were radiographs that were selected to represent the spectrum of radiographic damage — from little to maximal damage — in patients with early disease using the van der Heijde/Sharp scoring method. Although there was no difference in the intraclass correlation coefficients (0.92 vs 0.90), the SDD varied from about 10 to 20. In Study G, the radiographs were scored using 2 methods, the Scott/Larsen and the van der Heijde/Sharp. Although the absolute SDD cannot be directly compared, the ICC and the SDD as a percentage of the maximal actual score for the 2 scoring systems show that the van der Heijde/Sharp method performs better on reliability than the Scott/Larsen scoring method (ICC 0.92 vs 0.86, 0.84, respectively; SDD as percentage of maximal actual score 21% vs 30%, respectively).

The literature search located very few studies that included data on the SDD or data that could be used to calculate SDD for these measures. The results of the search

Table 7. Field studies of reliability: literature review of joint examination.

Measure	ICC Intra	ICC Inter	SDD 95% LOA Intra	SDD 95% LOA Inter	Other Analyses (CV, P)
Lansbury[54]					CV 27%
Ritchie Index[31]			–9, 12	–31, 25	P 0.94
Ritchie Index[55]	0.70				
ARA[55]	0.67				
28 TJC[34]		0.48			
14 Swollen JC[34]		0.15			
Ritchie Index (binary)[56]		0.83			
Ritchie Index[56]		0.85			
ARA 68 (Tender)[57]					Observer Effect –9.2 to 4.5 > mean joint count
Ritchie Index ARA[58]					CV 35–47%
Active joint[59]	0.95		–8, 8		
Ritchie Index[32]			–7, 9	–10, 14	
Ritchie Index[33]	0.70	0.52		–10, 10	
ARA[33]	0.87	0.48		–13, 13	
Ritchie Index[60]		RAI 0.81			
Lansbury		Lansbury 0.68			Variance: patients 74% examiners 2%, residual 22%

ICC Intra/inter: Intraclass/Interclass correlation coefficient. SDD: smallest detectable difference. LOA: limit of agreement. CV: Coefficient variation. P: Pearson correlation.

Table 8. Field studies of reliability: literature review of self-report measures.

	Test-retest ICC	Correlation Pearson (P) Rank (R)	Other Analyses
HAQ and MHAQ[61]		HAQ0.78 (P) MHAQ 0.91 (P)	
Swedish HAQ[62]		0.91 (R)	
Dutch HAQ[63]		0.89 (R)	Standard error for single observation = 0.26
Portuguese HAQ[64]	0.91		
French HAQ[65]	0.96		
Italian HAQ[66]		0.99 (R)	
Spanish HAQ[67]		0.89 (R)	
Spanish HAQ[68]		0.89 (P)	
VAS Pain[69]		Illiterate 0.71 (P), literate 0.94 (P)	
VAS Pain[70]	0.91		

ICC: Interclass correlation coefficient. HAQ: Health Assessment Questionnaire. M: modified HAQ. VAS: visual analog scale.

are given in Tables 7, 8, and 9. There were no studies on the HAQ that would allow calculation of SDD.

DISCUSSION

Historical perspective

A cursory look at these results is sobering. Clinical measures assumed for years to be robust, in the sense of having unquestioned "face validity" (such as joint counts and VAS global), are shown here to be much more "unstable," much more noisy, than one would have suspected. Nonetheless, they have been used historically as primary endpoints in RA trials. Thus, it is incumbent on RA methodologists to carefully reassess the effect of endpoint selection in clinical trials. However, before discussing how these SDD data may

Table 9. Field studies of reliability: literature review of radiographic measures.

Scoring Method	Kappa (K) Correlations Intraclass (ICC) Pearson (P) Rank (R) Intra	Kappa (K), Correlations Intraclass (ICC) Pearson (P) Rank (R) Inter	SDD 95% LOA Intra Inter
Larsen[71]	0.95 (P)		
Genant[41]	0.77–0.93 (P)		
Sharp (modified)[72]	0.93 (P)	0.84–0.92 (P)	
Larsen[42]	0.94–0.97 (P)	0.93–0.95 (P)	
Amos[42]	0.78–0.94 (P)	0.72–0.77 (P)	
Scott[73]	0.77–0.93 (P)		
Larsen			
Sharp 1971[74]	Random effects ICC 0.95	Random effects ICC 0.26	± 15
	> 0.9 (P)	0.74–0.96 (P)	± 70
Sharp (modified)[72]	0.93 (P)	0.84–0.92 (P)	
10 methods[75]	> 0.76 (P)	Fixed effects ICC 0.41–0.78	
Kaye[76]	0.91–0.97 (P)	0.91–0.96 (P)	
Kaye (simplified)[77]	0.91–0.97 (P)	0.90–0.98 (P)	
Sharp hands, wrists, feet[78]	0.97 (P)	0.94 (P)	
Larsen hands, wrists, feet[38]			± 8 ± 11
Sharp[79]	Sharp 0.94 (P)		
Larsen	Larsen 0.93 (P)		
Sharp[80]	JSN 0.86 (P) Erosions 0.91 (P)	JSN 0.86 (P) Erosions 0.89 (P)	
Sharp[43]	Sharp JSN kappa 0.58	Sharp JSN kappa 0.50	
Larsen	Sharp Erosion kappa 0.48 Larsen kappa 0.53 Sharp 0.96 (P) Larsen 0.92 (P)	Sharp Erosion kappa 0.39 Larsen kappa 0.35 Sharp 0.90–0.97 (P) Larsen 0.91–0.99 (P)	
Sharp[40]	Fixed effects ICC 0.97		± 30
Larsen	Fixed effects ICC 0.88		± 25
Scott/Larsen[81]	0.89–0.90 (P)		
Larsen[39]	0.95 (P)		± 6.6
van der Heijde/Sharp[82]	0.94–0.99 (P)	0.80–0.92	
Larsen[83]		0.94 (P)	
Wassenberg[84]	Fixed effects ICC 0.91		

SDD: smallest detectable difference. LOA: limit of agreement.

affect use of RA measures in RCT, we briefly review the history of various classes of RA measures. One dominant research question is "What measure best predicts future outcomes?" This is the basis of the evidence based, predictive approach to measurement research[7]. This is not to disparage face validity approaches to measurement, but face validity is more relevant to selecting outcomes than determining predictors of outcomes.

The multifaceted metrology in RA arose over numerous decades, with early work driven by the high prevalence of poor outcomes, the anecdotal reports of remission with injectable gold, and the dawn of medical statistics, then by early UK controlled trials in the 1940s and 1950s. As a consequence the history of measurement in RA is a rich and varied one. Lassere[21] offers a comprehensive review of this topic. The period of disillusionment following early enthusiasm with corticosteroids, and the delay until 1961[22] of definitive evidence that gold was efficacious, may have contributed to a clinical milieu emphasizing careful observation and description. Clinicians were concerned that inflammation in RA, appearing responsive to therapy, may be pathophysiologically dissociated from cartilage and bone destruction, leading to measures reflecting this dichotomy between "disease activity" and "disease damage."

With some exceptions, the reliability of the measures of articular disease, the questionnaire measures, and radiographic assessment in this study were similar to results available in the published literature. However, the interpretation of these results and any conclusions drawn differ, depending on what is considered acceptable reliability. What error, or what degree of reliability, is acceptable depends not on statistical considerations but on the context

of the measurement. A given measure, instrument, or method may have acceptable reliability in one context but not in another. A measure may have acceptable reliability for research, but unacceptable reliability for clinical practice. An instrument with acceptable reliability for cross sectional research may be unacceptable for longitudinal research[23]. In general, a measure for individual clinical decision-making requires a higher standard of reliability than a measure for decisions on groups, where the objective is to compare mean responses across groups. In the latter case, the increased sample size can overcome poor reliability, at least to some extent. Therefore, ultimately, the acceptability of reliability depends as much on the purpose of the instrument and how it will influence subsequent decisions as it does on the size and sources of measurement error.

Literature review

Many studies did not attempt to evaluate the reliability of the joint examination method or index in their hands[24-29], or reported results only as percentage agreement[30]. Only one paper used a variant of Bland and Altman's 95% limits of agreement to evaluate the reliability of the clinical articular assessment. The other SDD entries in Table 7 were recalculated using data from the original publications.

The interobserver 95% limits of agreement for the Ritchie articular index was –31 to 25 in Ritchie's original paper[31], –10 to 14 in a later study by Lewis, *et al*[32], and 10 in a very small study of 4 patients by Thompson, *et al*[33]. In Ritchie's study the examiners were trained but uncalibrated rheumatologists, whereas in Lewis's study 2 metrologists underwent simultaneous training and calibration. Ritchie herself considered that differences of more than 20 between 2 observers on an individual patient were important and wrote "this clearly renders the index invalid if observations are made on the same patient by different clinicians."[31]

Few published studies have evaluated the reliability of joint swelling and none could be reanalyzed using the limits of agreement method. Hansen, *et al*[34] reported the interobserver random effects ICC for a 14 joint swelling count as 0.14. Although Bellamy, *et al*[28] reported a reliability coefficient of 0.98 for the American Rheumatism Association graded 66 joint swelling index, the method of calculating the reliability coefficient as reported in the paper was incorrect (G. Wells, coauthor, personal communication). A revised fixed effects ICC (~0.74) was recalculated using the components of variance analysis provided in Bellamy's report[21]. One explanation for the discrepant ICC results is study design [Hansen, *et al* 0.14, this study (Study A) 0.52, and Bellamy, *et al* 0.74]. Hansen, *et al* evaluated 10 patients, as we did, but Hansen used 5 examiners, whereas we only had 2, and better ICC are expected with fewer examiners. Bellamy, *et al* evaluated 6 patients with 6 examiners, but the examiners had undergone considerable calibration.

Much more has been published on the reliability of radiographic scoring methods than for clinical articular assessments. This is not surprising given the difficulty of clinical studies compared to studies of radiographs. Real patients are needed for the former, evaluated by multiple observers, within short time periods, whereas radiographs can be read anytime by multiple observers. However, some early studies only reported the results as percentage agreement[35,36] or the residual variance (using analysis of variance methods of analysis) was not provided in the result[37].

About one-half of studies evaluating the reliability of their scoring method attempted a comprehensive assessment of both intra and interobserver reliability. However, few used an optimal method of analysis. O'Sullivan, *et al*[38] reported an intraobserver 95% limits of agreement for the Larsen score (hands, wrists, and feet, 0–210) as 8 and an interobserver agreement of 11. The observers had undergone considerable training. Ruckman, *et al*[39] reported an even smaller intraobserver agreement of 7, whereas Guth, *et al*[40] intraobserver agreement was 25 (Larsen score 0–150). The discrepancy between Ruckmann, *et al* and Guth's results is easily explained. Ruckmann, *et al* evaluated 24 patients, with 21 having a Larsen score of less than 10, so there was a predominance of very low scores, whereas the Guth, *et al* assessed 71 patients, with 66 having a Larsen score of 6–75, thus a much broader range of scores. These results confirm our findings that conditions external to the measurement process influence the magnitude and direction of measurement error, and that the SDD will be small if the abnormalities that are being measured are themselves negligible.

We located 2 reports that directly compared status scores with difference scores in the literature review. In both reports the reliability of the difference score (albeit employing a Pearson or rank correlation coefficient) was less than the measure's status score[41-43], a finding consistent with the theory of measurement[11].

Reliability and implications for the ACR20 and EULAR response criteria

The ACR20 and the EULAR response criteria of disease activity differ greatly in their development, selection, weighting, and component aggregation. A look at their calculation reveals stages at which random measurement error, the SDD, may have an effect. The implications of SDD in the use of these measures for clinical trials and epidemiology have not been studied.

The ACR20 was designed to reflect a "clinically significant" change in its 7 components, using a Delphi approach to define this change as 20%[1,44,45]. Thus, each component is first assessed for responder or nonresponder status by the 20% change test. This process bisects the population of interest (the distribution), and the "cutpoint" used has implications for introduction of measurement error (see below). These 7 outcomes, then, are surveyed to arrive at the overall responder/nonresponder decision. Thus, the population is

again bisected, now by each patient's overall ACR20 test. In addition to these heuristic considerations, the ACR derivation process was data driven. Multiple differently structured composites, varying both the cutpoint for clinical significance and the cutpoint for the number and distribution of components, were systematically studied for their ability to separate the pooled placebo from the pooled treatment response in a traditional DMARD database[1].

In contrast, the EULAR Response Index was developed and structured quite differently. The process here differed from explicit Delphi approaches. The EULAR Response Index was developed from the Disease Activity Score (DAS). The DAS employed an implicit "expert opinion or experiential based approach"[6,7]. The selection and weighting (the formula) of the pooled components were analytically derived from information implicit in the clinical decisions of rheumatologists in clinical practice. The statistical method of discriminant analysis was used to discriminate between low and high disease activity as judged by real treatment decisions[46]. Unlike the ACR20 the components of the DAS (and later the EULAR Response Index) remained as status scores rather than difference scores. Subsequently the DAS was validated using a "hypothesis-testing/predictive" approach[7]. Thereafter, the DAS, like the ACR20, was cast in a categorical (trichotomous) form, the EULAR Response Index using an RCT database[12]. So the DAS is first aggregated, then invoked as a responder decision, whereas the ACR20 first invokes a responder decision for all components, then is aggregated. So all else being equal, one would expect the DAS to engender less random measurement error than the ACR20.

However, there are other considerations. One could argue that as the EULAR Response Index combines change and achieved state into one criterion that has 3 levels, that this combination of 2 different domains may increase random measurement error. Further, the reliability of the composite depends on the reliabilities of the component measures, and these differ for the ACR20 and the EULAR Response Index. Measures that have poorer reliability, such as the patient global measured on a VAS, are a required component of the DAS formulation (DAS4) used to develop and test the EULAR response criteria, whereas patient global assessment may be excluded from the calculation of a particular patient's ACR20. Furthermore, the ACR20 gains in discriminatory power to detect differences by using 7 items and the 3/5 weighting of patient–specific responses, compared to the EULAR Responder Index, which keeps all items fixed.

Ranking the factors that influence the magnitude or predict the direction of the reliability of the ACR20 and EULAR Response Indices is not known and cannot be known without prospective field testing. Can the SDD shed light on useful characteristics of these instruments in prospective field testing? The interpretation of the single SDD for the DAS is straightforward. However, the effect of component SDD, combined into an ACR20, is less straightforward. One method of analysis is to calculate the kappa at any point within the derivation of the ACR20 — to assess how often the 20% percentage change threshold is reversed for each component — and then overall for each individual patient. Another method of analysis that more closely approximates the SDD approach to the ACR20 is to keep the percentage change from baseline for each of the patient's components in its original continuous scale.

Measures of functional disability, multidimensional health status, and health related quality of life

The ontology of function-disability measures is not as complex as that for the ACR20, so understanding how measurement error and its random component affect their use is easier. The impetus for developing function-disability measures such as the HAQ and the AIMS[47,48] was to facilitate clinic based and research based quantitation of function in RA. What can the patient do? Not do? These become the operative questions, not, what were his complaints or physical findings. Health related quality of life (QOL) measures are a more recent development. Although problematic in certain fundamental ways[49] they legitimately intend to capture the patient's preferences and judgments, rather than signs, symptoms, or functions. The question is not, what can one do, but what does one believe (one should be able to do). SDD for these measures are easily calculated, as we have demonstrated. Determining the MCID for these measures is more difficult, although not, in principle, insurmountable.

Radiographic measures

The Larsen and the Sharp scoring methods, the 2 major radiographic measures[50], arose in the 1960s and 1970s and have been modified and improved[17,18]. To have content validity[35], radiographic measures need to capture all the elements of structural degradation seen in advancing RA. The Sharp measure, in general, has better clinimetric properties (reliability, validity, and responsiveness to change) than the Larsen because it has more intervals, more items, and it separately scores the elements of joint space narrowing and erosions[51]. Use of a partial score, for example the erosion component of the Sharp score alone, undermines this rationale.

Level of response may affect measurement error — implications for RCT design

A common assumption in trial methodology is that measurement error is constant at different levels of response. In a clinical trial this means across treatment arms. This contention, however, has not been systematically investigated, and as we have seen in the work reported here, many factors, including the level of response in a particular arm of a clinical trial, may alter the measurement error. Additionally, the degree of this alteration may also vary

depending on what endpoint was chosen. What is critical for trial design and analysis is that if measurement error is not constant it ceases to be only error but now also introduces bias.

A bias of this type has been attributed to the choice of change scores or percentage change from baseline as endpoints[52]. Recently, Oppenheimer has shown that a bias also results from the choice of dichotomous rather than continuous endpoints[13]. Obviously it would be useful if both the direction and magnitude of this bias could be predicted. This may be possible. Bias, an underestimation or overestimation of the result, appears to be a function of 4 conditions: the number of change scores used in the endpoint, the cutpoint selected (i.e., how the distribution is bisected), the reliability of the endpoint itself, and the effect size seen in the RCT[13]. In any RCT, if the matrix of conditions is known, then one could predict both the direction and the magnitude of bias. Although the ACR20 and the EULAR response criteria may be robust to problems of bias and efficiency, further analyses of these measures should be carried out in light of these findings.

Finally, an important practical point for the research agenda is evident. Use of any cutpoint, as in an MCID or other responder/nonresponder measures, may have disadvantages because this dichotomy will cause loss of power and may introduce bias. We would argue that such dichotomies should not totally replace analyses of group means, although group means are neither intuitive nor attractive to the naïve clinician. Further, the SDD probably should have no necessary relation to the MCID, because it is a distribution based approach, based in formal statistics only, and we would assert that predictive/data based methodologies should underpin the MCID.

What then is the value of the SDD? It is another way of describing the error of measurement. It is more understandable because it uses the same scale as the measure, and it is truly quantitative. It makes sense to calculate the SDD, as this is the minimum we can assess apart from measurement error for an individual patient. Finally, the SDD can serve as a comparator against MCID derived from other research, particularly if the MCID is found to be smaller than the SDD.

Future research

Measurement error can be seen as the first step in determining where the minimum clinically important difference may be found. In the end, the relevance of a difference will have to be determined by how it relates to differences in longterm prognosis: remission, work disability, death, and burden of disease. Future research should be directed at clarifying such relations.

APPENDIX

Joint examination. Tenderness. Seventy-four joints were examined for joint tenderness, 66 by direct palpation and 8 joints or joint groups (cervical spine, lumbosacral spine, hip joints, subtalar and midtarsal joints) by pain on motion. Joint tenderness was defined as the presence (or degree) of a patient's discomfort when firm pressure (defined as blanching of the nail bed of the examiner's thumb) was applied directly on the joint line or in some cases over normal synovial reflections. Tenderness was recorded on a binary scale where 0 = no tenderness, 1 = presence of tenderness. Similarly, joint pain on motion was recorded as 0 = no pain on motion, and 1 = pain on motion.

Swelling. Sixty-eight joints were examined for joint swelling; the cervical spine, thoracolumbar spine, hip joints, and subtalar joints were excluded. Joint swelling was defined as soft tissue swelling either due to synovial thickening or to the presence of an effusion. Joint swelling was recorded on a binary scale where 0 = no swelling, 1 = presence of swelling. These definitions and methods used to ascertain joint tenderness, pain on motion, and joint swelling were recommended by the American Rheumatism Association[53]. The results are also given for the shortened form of joint assessment[28] comprising 28 joints.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, et al. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727-35.
2. van Gestel AM, Prevoo ML, van't hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of EULAR response criteria for rheumatoid arthrititis. Arthritis Rheum 1996;39:34-40.
3. Food and Drug Administration. Guidance for industry: Clinical development programs for drugs, device and biological products for the treatment of rheumatoid arthritis. US Department of Health and Human Services. Washington, DC: FDA 1999; www.fda./gov/cder/guidance/index.htm
4. Lassere M. Newer techniques in osteoarthritis and rheumatoid arthritis: clinical measures. APLAR J Rheumatol 1998;2:135-8.
5. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1998;26:731-9.
6. Lassere M, van der Heijde D, Johnson K, et al. Robustness and generalizability of the smallest detectable difference in radiological progression. J Rheumatol 2001;28:911-3.
7. Lassere M, van der Heijde D, Johnson K. Foundations of the MCID for imaging. J Rheumatol 2001;28:890-1.
8. Willett JB. Questions and answers in the measurement of change. In: Rothkopf EZ, editor. Review of research in education. Washington: American Educational Research Association; 1988:345-422.
9. Beaton D, Bombardier C, Katz J, Wright J, OMERACT MCID Working Group. Looking for important change/differences in studies of responsiveness. J Rheumatol 2001;28:400-5.
10. Norman GR, Streiner DL. Biostatistics. The bare essentials. St Louis: Mosby-Year Book; 1994.
11. Thorndike RM, Cunningham GK, Thorndike RL, et al. Measurement and evaluation in psychology and education. New York: Macmillan; 1991.
12. van Gestel AM, Anderson JJ, van Riel P, et al. ACR and EULAR improvement criteria have comparable validity in RA trials. J Rheumatol 1999;26:705-11.
13. Oppenheimer L, Kher U. The impact of measurement error on the comparison of two treatments using a responder analysis. Stat Med 1999;18:2177-88.
14. Fries JF, Spitz PW, Young DY. The dimensions of health outcomes: the Health Assessment Questionnaire, disability and pain scales. J Rheumatol 1982;9:789-93.
15. Ware JE Jr. SF-36 Physical and Mental Health Summary Scales: a user's manual. Boston: The Health Institute, New England Medical Center; 1994.
16. Ware JE Jr. SF-36 Health Survey: Manual and interpretation guide.

Boston: The Health Institute, New England Medical Center; 1993.
17. Edmonds J, Saudan A, Lassere M, Scott D. Introduction to reading radiographs by the Scott modification of the Larsen method. J Rheumatol 1999;26:740-2.
18. van der Heijde D. Introduction to reading radiographs by the van der Heijde modification of the Sharp method. J Rheumatol 1999;26:743-5.
19. Bland M, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1:307-10.
20. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420-8.
21. Lassere MND. Dissertation. Quantile reference curves of outcome in rheumatoid arthritis. Sydney, Australia; 1998.
22. Research Sub-committee of the Empire Rheumatism Council. Gold therapy in rheumatoid arthritis. Final report of a multi-centre controlled trial. Ann Rheum Dis 1961;20:315-34.
23. Healy MJR. Measuring measuring errors. Stat Med 1989; 8:893-906.
24. Williams HJ, Ward JR, Reading JC, et al. Low-dose D-penicillamine therapy in rheumatoid arthritis. A controlled, double-blind clinical trial. Arthritis Rheum 1983;26:581-92.
25. Egger MJ, Huth DA, Ward JR, Reading JC, Williams HJ. Reduced joint count indices in the evaluation of rheumatoid arthritis. Arthritis Rheum 1985;28:613-9.
26. Fuchs HA, Callahan LF, Kaye JJ, Brooks RH, Nance E, Pincus T. Radiographic and joint count findings of the hand in rheumatoid arthritis: related and unrelated findings. Arthritis Rheum 1988;31:44-51.
27. Fuchs HA, Brooks RH, Pincus T, Callahan LF. A simplified twenty-eight-joint quantitative articular index in rheumatoid arthritis. Arthritis Rheum 1989;32:531-7.
28. Prevoo ML, van Riel PL, van't Hof M, et al. Validity and reliability of joint indices. A longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol 1993;32:589-94.
29. Smolen J, Breedveld F, Eberl G, et al. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. Arthritis Rheum 1995;38:38-43.
30. The Co-operating Clinics Committee of the American Rheumatism Association. A seven day variability study of 499 patients with peripheral rheumatoid arthritis. Arthritis Rheum 1965;8:302-34.
31. Ritchie DM, Boyle J, McInnes J, et al. Clinical studies with an articular index for the assessment of joint tenderness in patients with rheumatoid arthritis. Q J Med 1968;147 (New Series XXXVII):393-406.
32. Lewis P, O'Sullivan MM, Rumfield W, Coles E, Jessop JD. Significant changes in Ritchie scores. Br J Rheumatol 1988; 27:32-6.
33. Thompson PW, Hart LE, Goldsmith CH, Spector TD, Bell MJ, Ramsden MF. Comparison of four articular indices for use in clinical trials in rheumatoid arthritis: patient, order and observer variation. J Rheumatol 1991;18:661-5.
34. Hansen TM, Keiding S, Lauritzen SL, Manthorpe R, Sorensen SF, Wiik A. Clinical assessment of disease activity in rheumatoid arthritis. Scand J Rheumatol 1979;8:101-5.
35. Larsen A. Radiological grading of rheumatoid arthritis, an interobserver study. Scand J Rheumatol 1973;2:136-8.
36. Larsen A, Edgren J, Harju E, Laasonen L, Reitamo T. Interobserver variation in the evaluation of radiologic changes of rheumatoid arthritis. Scand J Rheumatol 1979;8:109-12.
37. Wassenberg S, Rau R. Problems of evaluating radiographic findings in rheumatoid arthritis using different methods of radiographic scoring: examples of difficult cases and a study design to develop an improved scoring method. J Rheumatol 1995;22:1990-7.
38. O'Sullivan MM, Lewis PA, Newcombe RG, et al. Precision of Larsen grading of radiographs in assessing progression of rheumatoid arthritis in individual patients. Ann Rheum Dis 1990;49:286-9.
39. Ruckmann A, Ehle B, Trampisch HJ. How to evaluate measuring methods in the case of non-defined external validity. J Rheumatol 1995;22:1998-2000.
40. Guth A, Coste J, Chagnon S, Lacombe P, Paolaggi JB. Reliability of three methods of radiologic assessment in patients with rheumatoid arthritis. Invest Radiol 1995;30:181-5.
41. Genant HK. Methods of assessing radiographic change in rheumatoid arthritis. Am J Med 1983;75:35-47.
42. Grindulis KA, Scott DL, Struthers GR. The assessment of radiological changes in the hands and wrists in rheumatoid arthritis. Rheumatol Int 1983;3:39-42.
43. Plant MJ, Saklatvala J, Borg AA, Jones PW, Dawes PT. Measurement and prediction of radiological progression in early rheumatoid arthritis. J Rheumatol 1994;21:1808-13.
44. Goldsmith CH, Boers M, Bombardier C, Tugwell P, and the OMERACT Committee. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. J Rheumatol 1993;20:561-5.
45. Felson DT, Anderson JJ, Boers M, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
46. Van der Heijde D, van't Hof M, van de Putte L, et al. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. Ann Rheum Dis 1990;49:916-20.
47. Meenan R, Gertman P, Mason J. Measuring health status in arthritis: The Arthritis Impact Measurement Scales. Arthritis Rheum 1980;23:146-52.
48. Meenan R. The AIMS approach to health status measurement. Conceptual background and measurement properties. J Rheumatol 1982;9:785-8.
49. Johnson KR. Cost-effectiveness analysis: assessing the assumptions. J Rheumatol 2000;27:1565-6.
50. Sharp JT, Lidsky MD, Collins LC, Moreland J. Methods of scoring the progression of radiologic changes in rheumatoid arthritis. Correlation of radiologic, clinical and laboratory abnormalities. Arthritis Rheum 1971;14:706-20.
51. Lassere M. Pooled metaanalysis of radiological progression: comparison of Sharp and Larsen methods. J Rheumatol 2000;27:269-75.
52. Kaiser L. Adjusting for baseline: change or percentage change? Stat Med 1989;8:1183-90.
53. American Rheumatism Association. Dictionary of the rheumatic diseases. Vol 1. Signs and symptoms. New York: Contact Associates International; 1982.
54. Lansbury J, Baier H, McCracken S. Statistical study of variation in systemic and articular indexes. Arthritis Rheum 1962;5:445-56.
55. Eberl D, Fasching V, Rahlfs I, Wolf R. Repeatability and objectivity of various measurements in rheumatoid arthritis. Arthritis Rheum 1976;19:1278-86.
56. Hart LE, Tugwell P, Buchanan WW, Norman GR, Grace EM, Southwell D. Grading of tenderness as a source of interrater error in the Ritchie articular index. J Rheumatol 1985;12:716-7.
57. Klinkhoff A, Bellamy N, Bombardier C, et al. An experiment in reducing interobserver variability of the examination for joint tenderness. J Rheumatol 1988;15:492-4.
58. Thompson PW. A comparison of articular indices in rheumatoid arthritis. Br J Rheumatol 1986;25:98-9.
59. Thompson P, Kirwan J, Currey H. A comparison of the ability of 28 articular indices to detect an induced flare of joint inflammation in

rheumatoid arthritis. Br J Rheumatol 1988;27:375-80.
60. Bellamy N, Anastassiades TP, Buchanan WW, et al. Rheumatoid arthritis antirheumatic drug trials. I. Effects of standardization procedures on observer dependent outcome measures. J Rheumatol 1991;18:1893-900.
61. Pincus T, Summey J, Soraci S, Wallston K, Hummon N. Assessment of patient satisfaction in activities of daily living using a modified Stanford Health Assessment Questionnaire. Arthritis Rheum 1983;26:1346-53.
62. Ekdahl C, Eberhardt K, Andersson SI, Svensson B. Assessing disability in patients with rheumatoid arthritis. Use of a Swedish version of the Stanford Health Assessment Questionnaire. Scand J Rheumatol 1988;17:263-71.
63. van der Heijde D, van Riel PL, van de Putte LB. Sensitivity of a Dutch Health Assessment Questionnaire in a trial comparing hydroxychloroquine vs sulphasalazine. Scand J Rheumatol 1990;19:407-12.
64. Ferraz MB, Oliveira LM, Araujo PM, Atra E, Tugwell P. Crosscultural reliability of the physical ability dimension of the Health Assessment Questionnaire. J Rheumatol 1990;17:813-7.
65. Guillemin F, Briancon S, Pourel J. Validity and discriminant ability of the HAQ Functional Index in early rheumatoid arthritis. Disabil Rehabil 1992;14:71-7.
66. Ranza R, Marchesoni A, Calori G, et al. The Italian version of the Functional Disability Index of the Health Assessment Questionnaire. A reliable instrument for multicenter studies on rheumatoid arthritis. Clin Exp Rheumatol 1993;11:123-8.
67. Cardiel MH, Abello Banfi M, Ruiz Mercado R, Alarcon Segovia D. How to measure health status in rheumatoid arthritis in non-English speaking patients: validation of a Spanish version of the Health Assessment Questionnaire Disability Index (Spanish HAQ-DI). Clin Exp Rheumatol 1993;11:117-21.
68. Estevevives J, Batllegualda E, Reig A, et al. Spanish version of the Health Assessment Questionnaire: reliability, validity and transcultural equivalency. J Rheumatol 1993;20:2116-22.
69. Ferraz MB, Quaresma MR, Aquino LRI, Atra E, Tugwell P, Goldsmith CH. Reliability of pain scales in the assessment of literate and illiterate patients with rheumatoid arthritis. J Rheumatol 1990;17:1022-4.
70. Pincus T, Wolfe F, Callahan LF. Introduction: updating a reassessment of traditional paradigms concerning rheumatoid arthritis. In: Wolfe F, Pincus T, editors. Rheumatoid arthritis: pathogenesis, assessment, outcome and treatment. New York: Marcel Dekker; 1994:1-74.
71. Dawes PT, Fowler PD, Clarke S, Fisher J, Lawton A, Shadworth MF. Rheumatoid arthritis: treatment which controls the C-reactive protein and erythrocyte sedimentation rate reduces radiological progression. Br J Rheumatol 1982;25:44-9.
72. Pullar T, Hunter JCHA. Does second line therapy affect the radiological progression of rheumatoid arthritis? Ann Rheumatic Dis 1984;43:18-23.
73. Scott D, Coulton B, Bacon P, Popert A. Methods of x-ray assessment in rheumatoid arthritis. A re-evaluation. Br J Rheumatol 1985;24:31-9.
74. Sharp JT, Bluhm GB, Brook A, et al. Reproducibility of multiple-observer scoring of radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis. Arthritis Rheum 1985; 28:16-24.
75. Fries JF, Bloch DA, Sharp JT, et al. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum 1986;29:1-9.
76. Nance E, Kaye J, Callahan L, et al. Observer variation in quantitative assessment of rheumatoid arthritis: Part I. Scoring erosions and joint space narrowing. Invest Radiol 1986;21:922-7.
77. Kaye J, Nance E, Callahan L, et al. Observer variation in quantitative assessment of rheumatoid arthritis. II. A simplified scoring system. Invest Radiol 1987;22:41-6.
78. van der Heijde D, van Riel PL, Nuver Zwart IH, Gribnau FW, van de Putte L. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;1:1036-8.
79. Cuchacovich M, Couret M, Peray P, Gatica H, Sany J. Precision of the Larsen and the Sharp methods of assessing radiologic change in patients with rheumatoid arthritis. Arthritis Rheum 1992;35:736-9.
80. Salaffi F, Ferraccioli G, Peroni M, Carotti M, Bartoli E, Cervini C. Progression of erosion and joint space narrowing scores in rheumatoid arthritis assessed by nonlinear models. J Rheumatol 1994;21:1626-30.
81. Scott DL, Houssien DA, Laasonen L. Proposed modification to Larsen's scoring methods for hand and wrist radiographs. Br J Rheumatol 1995;34:56.
82. Van der Heijde D, van Leeuwen M, van Riel P, van de Putte L. Radiographic progression on radiographs of hands and feet during the first 3 years of rheumatoid arthritis measured according to Sharp's method (van der Heijde modification). J Rheumatol 1995;22:1792-6.
83. Fex E, Johnson U, Johnson K, Eberhardt K. Development of radiographic damage during the first 5-6 years of rheumatoid arthritis. A prospective follow up study of a Swedish cohort. Br J Rheumatol 1996;35:1106-15.
84. Bolten W, Brocal D, Sangha O, Kaser R. Reduced radiographic joint counts in rheumatoid arthritis [abstract]. Arthritis Rheum 1997;40 Suppl 9:S289.