Comparative Responsiveness of Four Elbow Scoring Instruments in Patients with Rheumatoid Arthritis

YTJE A. DE BOER, JOHANNA M.W. HAZES, PAUL C.A. WINIA, RONALD BRAND, and PIET M. ROZING

ABSTRACT. ***Objective.*** This prospective study investigated the comparative responsiveness to change of 4 different elbow scoring instruments: 2 Hospital for Special Surgery elbow assessment scales, the Mayo Clinic Elbow Performance Index, and the Elbow Functional Assessment (EFA) Scale.

Methods. A group of patients with rheumatoid arthritis (RA) (median age 60 yrs) undergoing either elbow arthroplasty (22 elbows) or synovectomy with radial head excision (3 elbows) were evaluated both before and after surgery (median 7 mo postoperatively). Changes in the scores obtained using the scales under study were calculated and analyzed. The patient's opinion of global perceived effect of the intervention was used as an external criterion to classify them as "improved" or "non-changed." Responsiveness was evaluated with 3 different statistical approaches: using paired t statistics (pre and postsurgery scores), effect size statistics (standardized response mean, effect size, and responsiveness ratios), and receiver operator characteristic curves. Minimal clinically important difference was estimated using patient satisfaction as the external criterion.

Results. Each of the elbow rating measures under study proved to be responsive to change when evaluating patients with RA undergoing elbow arthroplasty or synovectomy. The EFA scale had the highest power to detect a clinically meaningful difference and had the best discriminative ability to distinguish improved from no-change patients, as shown by all responsiveness statistics applied.

Conclusion. Using the EFA scale requires smaller sample sizes to achieve a fixed level of statistical power than the other scales we studied. (J Rheumatol 2001;28:2616–23)

Key Indexing Terms:
RESPONSIVENESS RHEUMATOID ARTHRITIS ELBOW SCORING SCALES

Over the last decades, the use of rating instruments has become widespread in clinical practice as they provide a practical tool to report the effectiveness of orthopedic interventions. Further, their use may allow comparison of different disease processes, patient populations, and types of treatment. For a long time, rating scales have often been used without formal testing of their measurement characteristics. Yet in recent years increasing emphasis has been placed on measurement theory in the evaluation of surgical orthopedic treatments, and there is a broad consensus that outcome scales should have established and proven reliability and validity before they are used as an outcome measure[1-8]. Nevertheless, properly designed reliability and validity studies are still needed for the majority of commonly employed scores in orthopedic surgery[1,7].

Review of the literature reveals a number of elbow scoring indices[9-12], each assigning their own numerical values on a variety of subscales like pain, range of motion (ROM), and function. These instruments are frequently used in clinical orthopedic research, both to measure cross sectional differences between patients or groups (discriminative purpose), e.g., to compare the outcomes of different surgical procedures, but especially to detect longitudinal change within individuals over time (evaluative purpose)[13,14].

The usefulness of evaluative instruments as an outcome measure depends on their sensitivity to the type and magnitude of changes that occur as a result of treatment[15-18]. This property is also termed sensitivity to clinically important change, or responsiveness[13,14,17,19,20]. Knowledge of the instrument's responsiveness to intervention effects is essential as it permits accurate estimation of sample size to assure adequate statistical power[19,21]. Instruments are not useful in evaluative studies when their power to detect a difference is too low, that is, requiring an unattainable sample size to achieve a fixed level of statistical power[14]. Accordingly, highly responsive instruments are crucial to decrease the number of subjects required in studies of the efficacy of interventions[15,21]. Despite its importance, this measurement attribute is least well tested and documented for most measures of outcome in rheumatic disease[17,20,23].

As conceptualized by Guyatt, *et al*[13,16,24], the necessary

From the Departments of Orthopaedic Surgery and Rheumatology, Leiden University Medical Center, Leiden, The Netherlands.

Supported by "Het Nationaal Reumafonds," the National Association Against Rheumatism of The Netherlands, grant NR 822.

Y.A. de Boer, MD, MSc, Department of Orthopaedic Surgery, Leiden University Medical Center; J.M.W. Hazes, MD, PhD, Rheumatologist, Department of Rheumatology, Leiden University Medical Center, currently Professor of Rheumatology, Erasmus University Medical Center, Rotterdam; W.P.C.A. Winia, MD, Department of Orthopaedic Surgery, Slotervaart Hospital, Amsterdam; R. Brand, PhD, Department of Medical Statistics; P.M. Rozing, MD, PhD, Professor of Orthopaedic Surgery, Leiden University Medical Center.

Address reprint requests to Dr. Y.A. de Boer, Department of Orthopaedic Surgery, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands. E-mail: ydboer@ortho.azl.nl

Submitted February 6, 2001; revision accepted June 14, 2001.

measurement properties of outcome instruments should be 2-fold; they have to be valid and also characterized by a high ratio of signal to noise. The latter can be summarized in a reliability coefficient for discriminative instruments and in a responsiveness index for evaluative instruments. As such, responsiveness should be considered as a separate concept, and not as another aspect of validity[13,24]. On the other hand, although reliability and responsiveness to change are different properties, they are related[25].

Recently, the discriminative properties of various elbow scoring instruments were compared[8,26]. However, their usefulness as evaluative instruments remained to be shown. We compared the relative responsiveness of frequently used elbow rating instruments, as applied to patients with rheumatoid arthritis (RA) who were undergoing primary total elbow joint replacement or elbow synovectomy with radial head excision.

MATERIALS AND METHODS

Patients. Consecutive RA patients scheduled for primary total elbow joint replacement in the Leiden University Medical Center and for elbow synovectomy with radial head excision at the Slotervaart Hospital in Amsterdam were recruited for this prospective followup study. The study was approved by the institutional review boards of each hospital.

From November 1996 up to and including February 1998, 23 patients were included, 2 of whom underwent bilateral procedures. Bilateral procedures were performed in 2 different sessions, with a period of at least 5 months between them. All patients had to be free of serious cognitive impairments, had to meet the American Rheumatism Association 1987 criteria for definite or classic RA[27], and have preoperative radiographs of the elbow showing articular damage grade III, IV, or V according to the classification of Larsen, *et al*[28]. Patients were informed about the aim of the study and gave their consent. Their main symptoms included pain in the elbow and restricted movement. In 22 cases a Souter-Strathclyde total elbow prosthesis was inserted, in 3 patients a synovectomy of the elbow was performed.

All patients were evaluated preoperatively and at followup, using 4 different elbow scoring systems: the Hospital for Special Surgery (HSS) Total Elbow Scoring Systems[10,11], the Mayo Clinic (Mayo) Performance Index for the Elbow[12], and the Elbow Function Assessment (EFA) Scale[26].

Description of the elbow scoring instruments. The HSS Elbow Assessment Scale is a 100 point rating system composed of ordinal scores for pain when bending (15 points) and at rest (15 points), function and activity (20 points), ROM (28 points), strength (10 points), and deformity (12 points)[11].

A shortened version of this scale (HSS2), in which no physical examination is required, is divided into 3 sections, assessing pain on a 5 point ordinal subscale (50 points), function (30 points), and activity (20 points)[10].

The Mayo Clinic Performance Index for the Elbow[12] is made up of a 4 point ordinal subscale for pain (45 points), daily function (25 points), motion (20 points), and stability (10 points), resulting in a possible maximum of 100 points.

Designed as a functional elbow evaluation tool, the EFA Scale[26] concentrates on self-reported performance of elbow-specific activities of daily living (ADL; 35 points). Unlike other elbow assessment instruments, it assesses the degree of pain (30 points) on a visual analog scale (VAS). Measurements of range of elbow motion (35) complete the final score to a possible maximum of 100 points.

Questionnaires. The subjective items of the cited elbow scoring instruments were compiled in one self-assessment questionnaire, which was completed by the patients one day prior to the operation as well as at followup evaluation. In addition, on both occasions a Dutch equivalent of the revised Arthritis Impact Measurement Scales[29-31] (AIMS-2) was completed. It is a combination of health status scales that measures quality of life of individuals, which assesses physical, emotional, and social well being, designed for individuals with rheumatic diseases. A normalization procedure converts scores to a normal standard of 0–10, the higher the score, the more disabled the patient. All scores were corrected for comorbidity, following the original instructions.

At followup the questionnaire package also asked the patients whether the results of surgery differed from their expectations of it, and in what way. Further, the patient was asked to indicate the global perceived effect of the intervention on a 5 point ordinal scale with the following levels of measurement: much improved, slightly improved, no change, slightly worsened, much worsened. This measure of improvement was used as an external criterion for evaluating responsiveness. If the patient had indicated much improvement, we categorized the patient as "improved"; if slightly improved, no change, or slightly worsened we categorized as "non-changed." Additionally, patients were asked to point out their overall satisfaction with the result of elbow surgery on a horizontal unscaled 10 cm VAS. On this scale the left extremity represented "dissatisfied" (score 0) and the right extremity "very satisfied" (10 points). A score of 0 to 2.5 was considered "dissatisfied," from 2.5 to 5.0 "somewhat satisfied," a score of 5.0 to 7.5 represented "moderately satisfied" patients, and the category patients rated between 7.5 and 10 was considered "very satisfied."

Physical examination. A standard form composed of various objective variables concerning the upper extremity was completed prior to the operation by one single registrar in orthopedic surgery (YADB). Active ROM of both elbows, shoulders, and wrists were measured and registered in degrees, using a two arm 8-inch plastic goniometer and standard positioning[32], with zero degrees as the neutral starting position. Varus-valgus instability of the elbow was assessed in 45° elbow flexion, or at maximum extension if flexion deformity exceeded 45°, and rated as stable (no apparent laxity), moderate instability (< 10°) and gross instability (> 10°). Isometric muscle strength of elbow flexion and extension was measured by placing a manual dynamometer (Microfet, Hoggan, Health Industries Inc., Draper, OR, USA) just above the wrist joint, recording the mean value of 3 measurements.

Shoulder function was assessed using the Shoulder Function Assessment scale[33], an instrument that reliably measures shoulder function in patients with RA, producing an overall score out of 70.

Followup evaluation. All patients included were seen in followup evaluation from June 1, 1997, to September 30, 1998. We combined most patient evaluations needed for this study with their regular appointments at the outpatient department with their orthopedic surgeon or rheumatologist. The average time from operation to followup was 7 months [standard deviation (SD) 4.2 mo; range 2–15]. The questionnaire package was mailed to the patients one week prior to followup evaluation and completed at home. Next, subjects were examined at the outpatient department by the same registrar who performed the preoperative examinations.

In one patient, who had a long history of RA with multiple joint arthroplasties, there was disturbed wound healing, which was followed by deep infection; completion of the questionnaires and followup measurements were performed after the treatment for infection and implant removal.

Statistical analysis. The collected data were entered on a computer, using a specialized data management application (Project Manager, MRDM Leiden). Analyses were performed using the SPSS statistical package for Windows (Release 7.5.2) and MedCalc (Version 5.0).

Aggregated preoperative and postoperative scores were computed for the operated side for each elbow rating scale. Subsequently, scores were stratified for the improved and the non-changed patient group, and differences between score levels were calculated by subtracting the preoperative score from that obtained at followup. Thus, a positive change indicates improvement. Next, the confidence intervals (95%) of the paired differences were calculated. Differences were judged to be statistically significant if $p < 0.05$.

A variety of strategies have been proposed to determine responsiveness. In the absence of a single standard[20], we selected 2 different strategies: paired t tests on difference scores and effect size statistics. Further, the discrimina-

tive ability of each scale was calculated using the receiver operating characteristic (ROC) method. Each method will be clarified below.

Paired t statistics comparisons and relative efficiency. An index of responsiveness can be calculated by comparing preoperative and postoperative values of improved patients, using paired t test analyses for within-patient changes[19,22,34]. The scale with the largest value of t statistics is judged to be the most responsive.

In addition, the relative efficiency to detect changes was calculated by comparing each instrument with the Hospital for Special Surgery scale[11]. The choice of the latter as standard was arbitrary. Relative efficiency was computed by squaring the ratio of appropriate t values[22], e.g.,

$$\text{relative efficiency (Mayo versus HSS)} = (t_{Mayo}/t_{HSS})^2$$

A relative efficiency > 1 means that this scale was a more efficient tool for measuring change than the HSS scale[11]. In addition, the instrument with highest relative efficiency has the highest power for a fixed sample size, or requires fewer patients to achieve a fixed level of statistical power[22].

Effect size statistics. Effect size statistics relate the magnitude of the change to the variability in the score[35]. Larger effect sizes indicate higher responsiveness. Several SD are used as a denominator for calculating effect sizes, as will be explicated below. The usual determination of effect size is calculated by taking the mean change for a single group and dividing it by the SD of baseline score of that group[19,36].

The standardized response mean (SRM) is a variant of the effect size. It is calculated as the mean change between followup and preoperative scores divided by the SD of these changes[35,37]. Absolute values of 0.2 are considered small, values of 0.5 are moderate, and those 0.8 or more represent large effects[35]. The values have direct implications for sample size determinations, since the ratio of sample sizes required to detect a given clinical effect is equal to the square of the ratio of the SRM[20].

A third index is the responsiveness ratio, which relates the minimal clinically important difference (MCID)[38] to the variability in stable subjects[17]. This concept was also illustrated in analogy with signal-to-noise ratios[16,17,24], in which the (smallest) meaningful clinical change one wishes to detect in a specific population of patients stands for the signal, and the within-subject variability unrelated to true clinical change represents the noise. As the magnitudes of the MCID for the scales under study are unknown, the mean change score for patients who were somewhat satisfied was used as an estimation for the MCID for improvement[21]. Accordingly, responsiveness rates were determined by calculating the ratio of the MCID to the intersubject variability of the change scores in non-changed patients.

Receiver operating characteristic method. ROC curves visualize the relation of true positive rate (sensitivity) on the y-axis against false-positive rate (100 – specificity) plotted on the x-axis for multiple cutoff points of a diagnostic test. Similarly, Deyo and Centor[34] advocated this method to qualify an instrument in its ability to discriminate between patients who have changed as a result of treatment (sensitivity) and those who have not benefited (specificity). Using the patient's opinion of global perceived effect of the intervention as a dichotomously external criterion of change, ROC curves can be constructed for various instrument cutpoints in change score. In this way, the ROC curves provide information on the sensitivity and specificity to discriminate between improved and non-changed patients. The area under the ROC curve indicates the probability that the instrument classifies improved patients correctly from randomly selected pairs of improved and non-changed patients[34], the larger the area, the better the scale. A line that runs diagonally across the figure from lower left to upper right will have an area of 0.5; this represents an instrument that does not discriminate. In addition, a ROC curve can provide an indication of which change score represents the optimal cutoff point to discriminate between improved and non-changed patients, by selecting the point closest to the upper left corner of the curve[19]. The corresponding value represents the optimal combination between sensitivity and specificity for deciding whether the instrument change score of an individual represents a true clinical improvement.

RESULTS

Characteristics of the study population. Demographic data of the patients are presented in Table 1. The median age at elbow surgery of the subjects was 60 years (range 40–79) and disease duration was between 8 and 66 years (median 24). Sixteen patients had the operation on the dominant side.

By self-assessment of global perceived effect of the operation performed, 18 patients rated themselves as much improved (improved), and 6 patients as no change or slightly improved (non-changed) at time of followup. The latter group indicated that surgery did not meet their expectations: e.g., one patient could not reach her mouth, another could not extend the arm as far as he had expected to be able to do. As only one patient indicated herself as much worsened, analysis for detecting deterioration could not be performed. This patient was excluded from analysis of responsiveness, analogously to the definition of responsiveness by Guyatt, *et al*[17].

Clinical characteristics and scores on the elbow instruments. Table 2 provides information on the clinical characteristics and the mean aggregated scores on the Dutch Arthritis Impact Measurement Scale 2 (AIMS-2) and the various elbow rating instruments of the patients in the improved and non-changed group, as measured and calculated at both occasions.

Preoperatively, the non-changed group experienced less pain on the elbow but showed a lower arc of motion of the elbow, in comparison to the improved group. The preoperative scores on the Mayo Clinic and HSS scale[11] showed only small differences between the improved and non-changed group, but was evidently higher for the improved group on the HSS2 scale[10] and Elbow Function Assessment scale.

Postoperatively, the improved group established a distinctly higher elbow score on all instruments, and also a higher sagittal range of elbow motion, in comparison to the non-changed group. Moreover, the improved group showed a significant rise ($p < 0.05$) in range of elbow motion and improve-

Table 1. Patient characteristics (N = 25 elbows), expressed as median (range) unless stated otherwise.

Sex, male/female, number	6/19
Age at time of intervention, yrs	60 (40–79)
Disease characteristics	
Rheumatoid factor positive, %	84
Disease duration, yrs	24 (8–66)
ESR, mm/h	37 (9–106)
Right handed, %	92
Side of operation right/left, number	15/10
Operation at dominant side, %	64
Joint destruction operated side, Larsen score, % III/IV/V	16/44/40
Followup period, mo	7.1 (2–15)
Previous surgical procedures at the operated side	
Elbow*, number	7
Shoulder†, number	1
Wrist††, number	6

* Synovectomy/excision radial head. † Synovectomy. ‡ Synovectomy/arthrodesis/distal ulnar resection/arthroplasty.

Table 2. Mean clinical characteristics and mean scores on the elbow rating scales under study of the improved and non-changed patient group, preoperatively (Pre) and at followup (FU), and the mean paired change (postsurgery-presurgery) with the 95% confidence interval (CI) of the difference.

	Improved Patients (n = 18)				Non-Changed Patients (n = 6)			
	Pre	FU	Change	(95% Cl)	Pre	FU	Change	(95% Cl)
Dutch AIMS-2								
Physical dimension, points	4.0	3.8	–0.18	(–0.6, 0.2)	3.5	3.5	–0.03	(–1.6, 1.5)
Arm function, points	4.2	2.9	–1.1	(–2.1, –0.08)*	4.5	4.1	–0.4	(–3.3, –2.5)**
Pain dimension, points	6.6	5.7	–0.67	(–1.3, 0.02)	6.1	3.9	–2.2	(–3.1, –1.3)**
Elbow								
Sagittal arc of motion, degrees	93	104	12	(–22, –1)*	78	93	15	(–38, 8)
Score on Mayo scale, points	53	87	33	(26, 41)**	52	71	19	(– 2, 40)
Score on HSS scale, points	52	75	26	(19, 33)**	55	64	9	(–3, 20)
Score on HSS2 scale, points	33	74	41	(29, 53)**	46	53	7	(–8, 22)
Score on EFA scale, points	45	76	33	(27, 39)**	54	65	10	(–2, 23)
Ipsilateral shoulder								
Score on SFA scale (range 0–70)	43	43	0.7	(–2, 4)	50	52	2	(–10, 14)

*p < 0.05;** p < 0.01. HSS: Hospital for Special Surgery, EFA: Elbow Function Assessment, SFA: Shoulder Functional Assessment.

Table 3. Mean change scores [standard deviation (SD)] of the improved (Imp) and non-changed (Non) patient groups for the Hospital for Special Surgery (HSS[11] and HSS2[10]), Mayo Clinic[12], and Elbow Function Assessment (EFA[26]) elbow rating scales. Also, t values (relative efficiency) on differences between pre and post-surgery scores, several calculations of effect sizes, and the area under the receiver operating characteristic (ROC) curves are given for each scale.

Instrument Mean change on Scale Score (SD)	t Value*(RE)	Effect Size Calculations SRM	 ES	 RR (MCID)	ROC Area
HSS					
Imp 26 (13.9)	7.9 (1.00)	1.86	1.47	1.12 (12.0)	0.83
Non 9 (10.8)	1.9	0.79	0.63		
HSS2					
Imp 41 (24.1)	7.2 (0.83)	1.70	1.55	0.94 (13.0)	0.86
Non 7 (13.9)	1.2	0.50	0.42		
Mayo					
Imp 33 (15.2)	9.3 (1.39)	2.19	1.89	0.76 (15.0)	0.7
Non 19 (19.9)	2.4	0.97	1.70		
EFA					
Imp 33 (12.6)	11.1 (1.97)	2.61	2.07	1.77 (21.5)	0.91
Non 10 (11.9)	2.1	0.87	0.94		

*Paired t test on differences between preoperative and postoperative scores.
RE: relative efficiency, the squared ratio of the paired t-statistic on difference score on the scales compared to that of the HSS score (see text). SRM: standardized response mean, calculated as the mean change between postoperative and preoperative scores divided by the SD of changes in score[35]. ES: effect size, calculated as the mean change divided by the SD of the preoperative mean in that group[36]. RR: responsiveness ratio of the minimal clinically important difference (MCID) to the SD of change score in non-changed patients[17].

ment on the elbow scoring systems ($p < 0.01$), this in contrast to the non-changed group, in which these increases were not significant. Pain, both as recorded on VAS and on ordinal scales, had significantly decreased in the improved group, while no significant pain reduction was found in the non-changed group. Strikingly, the latter group showed pain reduction on the general pain subdimension of the Dutch AIMS-2.

Far higher differences were found between the preoperative and postoperative scores on the various elbow rating instruments within the improved group (Table 3). The mean change of scores in the improved patients ranged from 26 for the HSS scale[11] up to and including 41 points for the HSS 2[10] scale. In the non-changed group mean changes varied from 7 points on the HSS 2 scale[10] up to and including 19 points on the Mayo scale.

Minimal clinically important difference. Comparing the order of mean change scores on the instruments for the 4 categories of satisfaction we found an increasing change score with higher levels of satisfaction for all instruments. That is, the greatest change in scores was found in patients who were most satisfied with the operative result. The differences in scale scores between the dissatisfied patient and the somewhat satisfied patients were 20 (Hospital for Special Surgery[11]), 29 (Hospital for Special Surgery 2 scale[10]), 10 (Mayo), and 32 (Elbow Function Assessment) points. The MCID ranged from a value of 12 points for the HSS scale[11] to 22 points for the Elbow

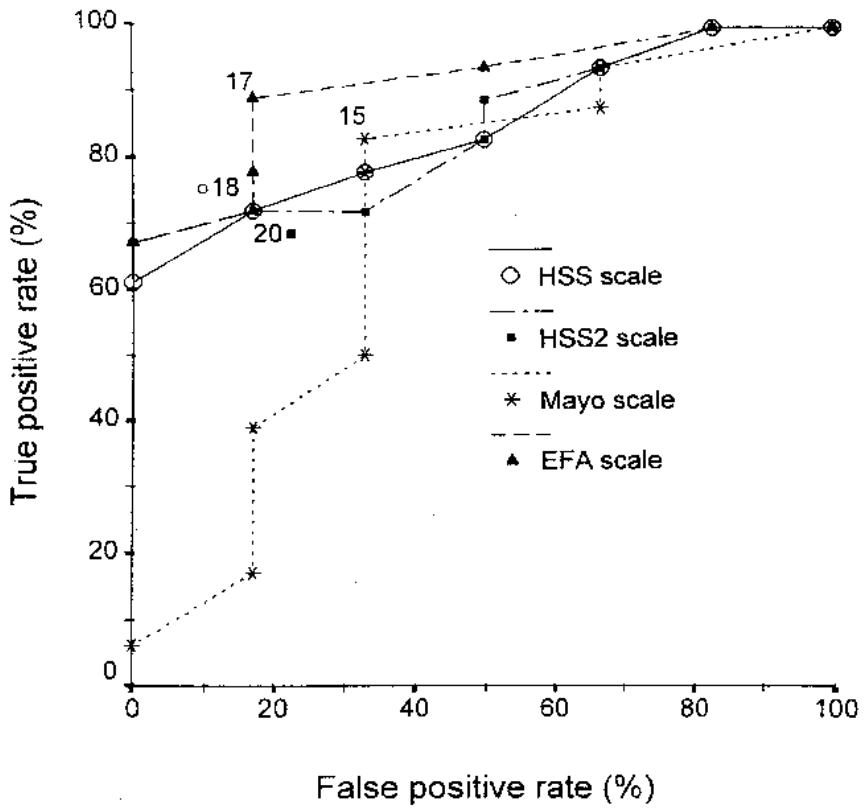

Figure 1. Receiver operating characteristic curves for the Hospital for Special Surgery (HSS), Mayo, and Elbow Functional Assessment (EFA) scales comparing improved and non-changed patients at a mean of 7 months of followup after elbow surgery. The indicated values represent the optimal instrument change score for distinguishing between improved and non-changed patients.

Function Assessment scale (Table 3), using patient satisfaction as the external criterion.

Responsiveness. Paired t statistics. The results of the paired t test analyses for the comparison between preoperative and postoperative scores of the scales under study indicate that each scale was able to detect statistically significant changes in the improved group (Table 3). The lowest value was achieved with the HSS 2 scale[10] (t = 7.2; relative efficiency = 0.83), and the highest with the Elbow Function Assessment scale (t = 11.1; relative efficiency = 1.54). The changes on all scales between the pre- and postsurgery scores in the non-changed group were not statistically significant.

Effect size statistics. For all instruments, SRM were large for the improved group (Table 3), ranging from 1.70 (Hospital for Special Surgery 2 scale[11]) to 2.61 (Elbow Function Assessment), indicating that the scales were responsive to the improvement experienced in this group. It was striking that the SRM for the non-changed group were moderate to large. This means that there was score improvement of some patients who did not experience global improvement by surgery.

As expected, the effect size statistics in the improved group are higher than those in the non-changed group (Table 3). With the exception of the Mayo scale, the differences between the effect size in the improved and non-changed group allow the conclusion that the instruments can discriminate between the improved and the non-changed patients. For the non-changed group, effect size was the lowest for the HSS 2 scale[10], indicating that this scale would be the most specific to detect change. The responsiveness rates to surgery, calculated as the ratio of MCID to the SD of the change score in the non-changed group (Table 3), ranged from 0.76 for the Mayo scale to 1.77 for the Elbow Function Assessment scale. Given the various responsiveness ratios, it was possible to calculate the necessary sample sizes[17]. To achieve a statistical power of 90% for paired groups ($\alpha = 0.05$), we would need 17 patients per group using the HSS scale[11], 24 patients using the HSS 2 scale[10], and 37 patients per group using the Mayo scale. Using the Elbow Function Assessment scale would only require 7 patients in each group.

Generally, the effect sizes of the Elbow Function Assessment scale were superior to the other scales, as manifested by all responsiveness statistics applied. However, with respect to the HSS and the Mayo scales, both the SRM and effect size calculations revealed another rank order of responsiveness than the responsiveness ratio value. The former values indicated that the Mayo scale was more responsive to change than the HSS scales, while according to the latter the Mayo scale would be the least responsive.

Receiver operating characteristic curves. Figure 1 presents the ROC curves for changes as a result of surgical intervention of the rheumatoid elbow on 4 elbow rating instruments, using the patient global perceived effect as the external criterion for discrimination between improved and non-changed. For each instrument the curve was to the left above the diagonal, showing some discriminative ability. However, the shape of both HSS curves and the Elbow Function Assessment curve, and the areas under these curves (Table 3), showed that these scales discriminated more reliably between patients who experienced improvement and those who did not change than the Mayo scale did.

At the point on the ROC curve closest to the left upper angle, the sensitivity and the specificity of the Elbow Function Assessment scale were 89 and 83%, respectively, to distinguish improved from non-changed patients, with a corresponding change score of 17 points and a positive predictive value of 94%. Both HSS scales reached 72 and 83%, with a change score of 18 points for the HSS scale[18] and 20 points for the HSS 2 scale[10] and a positive predictive value of 93% for either scale. The sensitivity and specificity for the Mayo scale were 83 and 67%, at an optimal cutoff score of 15 points, with a positive predictive value of 88%.

DISCUSSION

We studied the evaluative qualities of 4 elbow scoring scales in a group of patients with RA being treated with total elbow arthroplasty and synovectomy. To ensure the usefulness of elbow rating scales for clinical purposes, it is crucial that such measures of outcome are sensitive to the type and magnitude of treatment changes that occur with arthroplasty and synovectomy. The present investigation made it clear that each of the elbow rating measures under study proved to be sensitive to change. However, while the Elbow Function Assessment scale showed the highest responsiveness on all responsiveness statistics applied, the rank order of responsiveness for both Hospital for Special Surgery scales and the Mayo scales was not consistent when comparing the results of the t values, SRM, or effect size calculations on the one hand and the responsiveness ratio on the other. This may be related to some disadvantages of the statistical techniques used.

An important disadvantage of using the largest paired t statistics as a value of responsiveness is that this method does not account well for the score variability that might have occurred in the apparently non-changed subjects[19].

Concerning the effect size calculations, the SRM and effect sizes compare the magnitude of change to the SD of change[19,36] or the SD of the baseline scores[36], respectively. A shortcoming of the effect size is that a test may be assumed to be responsive if the differences of the scores before and after an intervention have a large mean and a small SD[39]. Moreover, as the effect size does not incorporate the response variance, it is thought to be unsuitable to comparing the response means of various instruments[40]. A potential disadvantage of both the effect sizes and the SRM is that the score changes in the numerator may overestimate treatment effects. This is because some change (score improvement) is often observed even in non-changed patients[19], as confirmed by our results (Table 3). In addition, the differences in SRM and effect sizes among instruments in part reflect differences in weights given to the subscales and in coverage of the scale dimensions. The latter might explain the small differences that were found in effect size calculations between both HSS scales, and also the high SRM when we compare them to other instruments reported in the literature[21,41].

In studies on responsiveness, responsiveness rates are calculated by using both the MCID and the mean change score in improved patients as the numerator. A lot of investigators use the latter[42], since there is no single best method for estimating the MCID. We decided to calculate responsiveness rates by using the estimated MCID as a numerator, as it is likely to suppose that the mean change score in improved patients would be larger than the smallest change that might be considered clinically relevant, and hence would result in an overestimation of responsiveness ratio. It should be realized that the MCID will depend on the population of patients[43] and the criterion being used as an external indicator of minimal clinical importance. Like others[21], we chose patient satisfaction as a criterion to estimate MCID because satisfaction is a relevant variable in elective surgical procedures, although we realize that we may not have established with certainty the single best estimate of the MCID of each scale. Also, we are aware that the level of satisfaction can depend on the degree of happiness of both patient and surgeon as well as their attitude to each other[44].

As argued earlier, responsiveness rates are prone to bias since the numerator and denominator are based on different samples, presuming that the variance in the patients who don't change is about equal to the variance of the subjects who do change[20,25].

Next to sensitivity, the specificity of a scale to change is also important, since instruments may reflect changes that have no clinical relevance[34]. In this respect, the ROC method has the advantage of visualizing not only the sensitivity of an instrument, but also the specificity of its score changes[19]. In addition, with this method one can identify the optimal cutoff points in change scores, although it should be noted that all cutoff points are essentially arbitrary and depend on the purpose of a study. The ROC method indicated that the Mayo grading system discriminated less accurately than the Elbow Function Assessment and the HSS scales between patients who experienced improvement and those who did not change.

Although AIMS scales were shown to be sensitive to clinical changes in health status secondary to total hip and knee joint replacement[22,37,40], this seemed not to be the case in the current patient group after elbow joint replacement (Table 2). A decrease in disease activity in the non-changed group rather than the operation result itself may have caused the improve-

ment of this group on the pain subdimension. Still, generic health measures are useful as they permit comparability across conditions and populations[16]. This may be of importance especially in determining the clinical result of total elbow arthroplasty, as concluded in a study of Weiland, *et al*[45], showing that the most important overall factor of success was the patient's baseline health status.

We must acknowledge possible limitations associated with the data presented in this report. Generally, it should be realized that statistical techniques measuring responsiveness to any change may not capture well its responsiveness to clinically meaningful or important change. For any method of calculating responsiveness, the choice of external criteria for change is disputable. We used the patient's own global perceived effect to select subgroups of improved and non-changed patients for estimating responsiveness. There may be concern that it is not reasonable to expect patients to assess change themselves, as validity may be compromised by the patient's mood and expectations[46]. However, the importance of the patient's perception of treatment benefit for orthopedic procedures is increasingly recognized[1,3,5] as perceived pain relief and functional improvement are relevant to the patients themselves. Like others, we used patient satisfaction as the external measure of MCID, because satisfaction reflects both perceived change and preference for this change[21]. Since our external criterion cannot be regarded as a gold standard, more comparisons of elbow scales against several external criteria may be needed. Consideration should also be given to the limited sample size, especially that of the non-changed group. Finally, our comparisons are based on outcomes of elbow arthroplasty and synovectomy in patients with RA, and may not be generalizable to other patient groups or interventions. Despite these recognized shortcomings, this study does appear to have merit and relevance in demonstrating responsiveness to change of elbow assessment scales in patients with RA.

In conclusion, each of the elbow rating measures under study proved to be responsive to change, as experienced in patients with RA undergoing either elbow arthroplasty or synovectomy. The Elbow Function Assessment scale had the highest power to detect clinically meaningful difference and had the best discriminative ability to distinguish improved from non-changed patients, as revealed by all responsiveness statistics applied. Therefore, using the EFA scale will require smaller sample sizes to achieve a fixed level of statistical power than the other scales under study.

REFERENCES

1. Amadio PC. Editorial. Outcome measurements — more questions; some answers. J Bone Joint Surg 1993;75A:1583-4.
2. Beaton D, Richards RR. Assessing the reliability and responsiveness of 5 shoulder questionnaires. J Shoulder Elbow Surg 1988;7:565-72.
3. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perception of patients about shoulder surgery. J Bone Joint Surg 1996; 78B:593-600.
4. Keller RB. Measuring outcomes [editorial]. J Orthop Res 1996;14:171-2.
5. Levine DW, Simmons BP, Koris MJ, et al. A self-administered questionnaire for the assessment of severity of symptoms and functional status in carpal tunnel syndrome. J Bone Joint Surg 1993;75A:1585-92.
6. L'Insalata JC, Warren RF, Cohen SB, Altchek DW, Peterson MGE. A self-administered questionnaire for assessment of symptoms and function of the shoulder. J Bone Joint Surg 1997;79A:738-48.
7. Sun Y, Stürmer T, Günther KP, Brenner H. Reliability and validity of clinical outcome measurements of osteoarthritis of the hip and knee — a review of the literature. Clin Rheumatol 1997;16:185-98.
8. Turchin DC, Beaton DE, Richards RR. Validity of observer-based aggregate scoring systems as descriptors of elbow pain, function, and disability. J Bone Joint Surg 1998;80A:154-62.
9. Figgie MP, Inglis AE, Mow CS, Figgie HE III. Total elbow arthroplasty for complete ankylosis of the elbow. J Bone Joint Surg 1989;71A:513-20.
10. Figgie MP, Inglis AE, Mow CS, Wolfe SW, Sculco TP, Figgie HE III. Results of reconstruction for failed total elbow arthroplasty. Clin Orthop 1990;253:123-32.
11. Inglis AE, Pellicci PM. Total elbow replacement. J Bone Joint Surg 1980;62A:1252-8.
12. Morrey BF, Adams RA. Semiconstrained arthroplasty for the treatment of rheumatoid arthritis of the elbow. J Bone Joint Surg 1992;74A:479-90.
13. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurements: a clarification. J Clin Epidemiol 1989;42:403-8.
14. Kirshner B, Guyatt GA. Methodological framework for assessing health indices. J Chron Dis 1985;38:27-36.
15. Fortin PR, Stucki G, Katz JN. Measuring relevant change: an emerging challenge in rheumatologic clinical trials [editorial]. Arthritis Rheum 1995;38:1027-30.
16. Guyatt GH. A taxonomy of health status instruments. J Rheumatol 1995;22:1188-90.
17. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chron Dis 1987;40:171-8.
18. Hilliquin P, Menkes C-J. Rheumatoid arthritis — valuation and management: early and established disease. In: Klippel JH, Dieppe PH, editors. Rheumatology. London: Mosby Year Book Europe; 1994:3.13.1-3.13.14.
19. Deyo RA, Diehr P, Patrick DL. Reproducibility and responsiveness of health status measures — statistics and strategies for evaluation. Controlled Clin Trials 1991;12 Suppl:142S-58S.
20. Liang MH. Evaluating measurement responsiveness. J Rheumatol 1995;22:1191-2.
21. Stucki G, Liang MH, Fossel AH, Katz JN. Relative responsiveness of condition-specific and generic health status measures in degenerative lumbar spinal stenosis. J Clin Epidemiol 1995;48:1369-8.
22. Liang MH, Larson MG, Cullen KE, Schwarz JA. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542-7.
23. Liang MH, Jette AM. Measuring functional disability in chronic arthritis. Arthritis Rheum 1981;24:80-6.
24. Guyatt GH, Kirshner B, Jaeschke R. Measuring health status: what are the necessary measurement properties? J Clin Epidemiol 1992;45:1341-5.
25. Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol 1989;42:1097-105.
26. De Boer YA, van den Ende CHM, Eygendaal D, Jolie IMM, Hazes JMW, Rozing PM. Clinical reliability and validity of elbow functional assessment in rheumatoid arthritis. J Rheumatol 1999;26:1909-17.

27. Arnett FC, Edworthy SM, Bloch DA, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315-24.
28. 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.
29. Meenan RF, Gertman PM, Mason JH. Measuring health status in arthritis — The Arthritis Impact Measurement Scales. Arthritis Rheum 1980;23:146-52.
30. Meenan RF, Gertman PM, Mason JH, Dunaif R. The Arthritis Impact Measurement Scales: further investigation of a health status measure. Arthritis Rheum 1982;25:1048-53.
31. Riemsma RP, Taal E, Rasker JJ, Houtman PM, Van Paassen HC, Wiegman O. Evaluation of a Dutch version of the AIMS2 for patients with rheumatoid arthritis. Br J Rheumatol 1996;35:755-60.
32. American Academy of Orthopaedic Surgeons. Joint motion. Method of measuring and recording. Reprinted by The British Orthopaedic Association. Edinburgh and London: E.S. Livingstone; 1966.
33. Van den Ende CHM, Rozing PM, Dijkmans BAC, Verhoef JAC, Voogt-van der Harst EM, Hazes JMW. Assessment of shoulder function in rheumatoid arthritis. J Rheumatol 1996;23:2043-8.
34. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: An analogy to diagnostic test performance. J Chron Dis 1986;39:897-906.
35. Cohen J. Statistical power analysis for the behavioural sciences. New York: Academic Press Inc.; 1977:1-27.
36. Kazis LE, Anderson JJ, Meenan RF. Effect size for interpreting changes in health status. Med Care 1989;27:S178-89.
37. Liang MH, Fossel AH, Larson MG. Comparisons of five health status instruments for orthopaedic evaluation. Med Care 1990;28:632-642.
38. Jaeschke R, Singer J, Guyatt GH. Measurement of health status — Ascertaining the minimal clinically important difference. Controlled Clin Trials 1989;10:407-15.
39. Kreibich DN, Vaz M, Bourne RB, et al. What is the best way of assessing outcome after total knee replacement? Clin Orthop 1996;331:221-5.
40. Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. Med Care 1992;30:917-25.
41. Ruperto N, Ravelli A, Falcini F, et al. Responsiveness of outcome measures in juvenile chronic arthritis. Rheumatology 1999; 38:176-80.
42. Van der Windt DAWM, Van der Heijden JMG, de Winter AF, Koes BW, Deville W, Bouter LM. The responsiveness of the Shoulder Disability Questionnaire. Ann Rheum Dis 1998;57:82-87.
43. Prasad K. The Glasgow Coma Scale: a critical appraisal of its clinimetric properties. J Clin Epidemiol 1996;49:755-63.
44. Morris RW. A statistical study of papers in the Journal of Bone and Joint Surgery 1984. J Bone Joint Surg 1988;7B:242-6.
45. Weiland AJ, Weiss A-PC, Wills RP, Moore JR. Capitellocondylar total elbow replacement. A long-term follow-up study. J Bone Joint Surg 1989;71-A:217-22.
46. McFarlane AC, Brooks PM. Determinants of disability in rheumatoid arthritis. Br J Rheumatol 1988;27:7-14.