Computer Based Methods for Measurement of Joint Space Width: Update of an Ongoing OMERACT Project

JOHN T. SHARP, JANE ANGWIN, MAARTEN BOERS, JEFF DURYEA, GABRIELE von INGERSLEBEN, JAMES R. HALL, JOOST A. KAUFFMAN, ROBERT LANDEWÉ, GEORG LANGS, CÉDRIC LUKAS, JEAN-FRANCIS MAILLEFERT, HEIN J. BERNELOT MOENS, PHILIPP PELOSCHEK, VIBEKE STRAND, and DÉSIRÉE van der HEIJDE

ABSTRACT. Computer-based methods of measuring joint space width (JSW) could potentially have advantages over scoring joint space narrowing, with regard to increased standardization, sensitivity, and reproducibility. In an early exercise, 4 different methods showed good agreement on measured change in JSW over time in the small joints of the hands and feet. Despite differences in measurement values between methods, measurement of within-joint change over time showed no systematic differences. The within-method variation was small, with intra-operator variation being smaller than inter-operator variation. Although this initial study was limited in terms of the number of patients and timepoints (total 10), the number of joints was relatively high (340 joints), so the results were considered strong evidence supporting the validity of computer-based JSW measurements to continue the study of the potential value of JSW by comparison of measurements to manual scoring of joint space narrowing using the COBRA trial images. (J Rheumatol 2007;34:874–83)

Key Indexing Terms:
JOINT SPACE MEASUREMENT   COMPUTER-BASED
RELIABILITY   PRECISION   SENSITIVITY TO CHANGE

Introduction

A subcommittee within the OMERACT imaging committee was formed after OMERACT 6 to test reliability, sensitivity, and feasibility of computer-based methods for measuring radiographic damage in the small joints of the hands and feet in patients with rheumatoid arthritis (RA). Initial efforts concentrated on measurement of joint space width (JSW). A report described the rationale and objectives of the committee in detail[1]. Briefly, measurement of damage on a continuous metric scale would be preferable to scoring damage on an ordinal scale if the measurements were highly reproducible, especially if they were more sensitive to change. Further, the difficulty of standardizing scoring by different readers even within a single center is an obstacle to comparing results across studies, and even limits of agreement are specific to a given study and cannot be generalized to others[2]. Metric measurements could be a means of reducing this obstacle.

To test reliability, sensitivity, and discriminating ability, investigators participated in a series of exercises to measure JSW in the finger, wrist, and toe joints, the joints that are most regularly involved in RA. The results of measurements on radiographs of 2 patients with RA at 2 timepoints using 3 different computer programs were presented at OMERACT 7 and encouraged more studies. Developers of 4 different methods and 10 different readers participated in a second exercise, called exercise A, involving radiographs from 4 patients at a total of 10 different timepoints, 2 or more for each patient[3-10]. The study determined variation of repeated JSW measurements by the same operator, between operators, and between methods, and the extent to which any differences in JSW methods affected the measurement of JSW change over time. The results of the analysis of reliability of repeated measurement and agreement between methods, which are reported here, were considered sufficiently encouraging for the committee to agree that a more comprehensive study employing a

From the University of Washington School of Medicine, Seattle, Washington, USA; GlaxoSmithKline, Middlesex, England; VU University Medical Center, Amsterdam, The Netherlands; Brigham and Women's Hospital, Harvard University, Boston, Massachusetts, USA; Synarc, San Francisco, California, USA; Snoqualmie, Washington, USA; University of Twente, Enschede, The Netherlands; University Hospital Maastricht, Maastricht, The Netherlands; Graz University of Technology, Graz, Austria; Dijon University Hospital, University of Burgundy, Dijon, France; University Medical Center, Utrecht, The Netherlands; University of Vienna, Vienna, Austria; and Stanford University, Palo Alto, California, USA.

J.T. Sharp, MD, Affiliate Professor of Medicine, University of Washington; J. Angwin, MA, Respiratory and Inflammation Centre of Excellence in Drug Discovery, GlaxoSmithKline R&D, UK; M. Boers, MSc, MD, PhD, Professor of Clinical Epidemiology, VU University Medical Center; J. Duryea, PhD, Assistant Professor, Brigham and Women's Hospital, Harvard University; J.A. Kauffman, Signals and Systems Group, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente; R. Landewé, MD, PhD, Associate Professor of Rheumatology, University Hospital Maastricht; C. Lukas, MD, Rheumatologist, University Hospital Maastricht; J.F. Maillefert, MD, PhD, Associate Professor of Rheumatology, University of Burgundy; H.J. Bernelot Moens, MD, Rheumatologist, Ziekenhuis Groep Twente, Hengelo, The Netherlands; G. Langs, Institute for Computer Graphics and Vision, Graz University of Technology; P. Peloschek, MD, Radiologist, Medical University of Vienna; D.M.F.M. van der Heijde, MD, PhD, Professor of Rheumatology, University Hospital Maastricht; G. von Ingersleben, Radiologist, Synarc; J.R. Hall, Software Engineer, Snoqualmie, WA, USA.

Address reprint requests to Dr. J.T. Sharp, 8387 N.E. Sumanee Place, Bainbridge Island, WA 98110, USA. E-mail: johntsharp@comcast.net.

trial set of images was the appropriate next test of the methods. The investigators in the COBRA trial[11] agreed to make the images, which had recently been digitized, available for this study, referred to as exercise B. Analysis of COBRA results is incomplete, but the available preliminary data are included in this report.

METHODS

Radiographs and joints. Hand and foot radiographs were digitized at 20 pixels per millimeter (50 micron pixel size) for both exercises and at 12-bit gray-scale for exercise A and 8-bit gray-scale for exercise B. Radiographs were blinded to time order. Three different computer-based methods (A[9], B[7,8], and E[3,4]) measured 34 joints of each patient at each timepoint including 4 proximal interphalangeal (PIP), 4 metacarpophalangeal (MCP), and 4 wrist joints in each hand and 5 metatarsophalangeal (MTP) joints in each foot (Figure 1, Appendix). One method has not been extended to measure wrist joints (D[5,6]) and another method in development measured only MCP joints(F[11]). Exercise A included 4 patients, 3 with 2 timepoints and one with 4. Exercise B included 107 patients and 428 timepoints, baseline, 6 months, 1 year, and last available radiograph.

Measurement methods

Brief outlines of computer methods are given in the Appendix. The methods differ in a number of details. First, in locating the joint and joint margins, method D uses neural networks, although the method has not yet been successfully adapted to measure wrist joint spaces. Method B requires the operator to place one point on the PIP and another point on the MCP joints of digits II–V and uses these locations to find the joint margins[7,8]. In method E, the operator locates MCP joints by placing 3 points on each metacarpal head. MTP and wrist joints are treated similarly; PIP joints are oriented vertically and located manually[3,4]. In method A, the operator locates each joint by placing a marker on the medial and lateral extent of the joint space (Hall JR, Sharp JT, unpublished data). Method F requires the operator to create a region of interest (ROI) that includes the hand. Bones with joints are detected automatically without additional operator input[10]. The region of measurement varies. Programs B and D measure a fixed number of millimeters depending on joint type (PIP, MCP, MTP, wrist joint) and centered within the joint. Program E uses a radian drawn from the perceived center of the metacarpal head as the region for measuring MCP joints. For PIP joints, the full breadth of the joint is selected. In method A, 60% of the joint span, medial to lateral sides, is selected for measurement, the joint span having been delineated earlier. In all methods, the location of joint margins is based on the density along lines crossing the joint space.

Calculation of the mean distance across the joint space also varies. The shortest distance across the joint space is found and averaged at multiple evenly spaced locations in methods A and B. Program D orients digits vertically and measures the distance across the joint space on multiple vertical lines before averaging. Method E uses a similar procedure for PIP joints, but measures MCP joint space along radial lines extending from the approximate center of the metacarpal head. In method F the highest gray level in the concave base of the proximal phalanx is defined as the distal joint margin. The minimal distance across the joint space to the metacarpal head is used to calculate the final measurements.

A

B

Figure 1. A. Joints measured in each hand. B. Joints measured in each foot.

Additional information can be derived from measurement of each individual joint. The standard deviation of the measurements within a joint indicates the symmetry and variability within that joint. Minima and maxima indicate the range of measurements and provide additional information on symmetry. The correlation coefficient between the measurements and their location can be combined with the slope of the measurements plotted against location to give an indication of the extent of asymmetry.

Statistical analysis

JSW and annualized change in JSW measurements over time were compared using Stata (Stata Corp., College Station, TX, USA), SPSS (SPSS Inc., Chicago, IL, USA), and Excel (Microsoft Corp., Redmond, WA, USA). Determination of the smallest detectable difference (SDD) used the method of Lassere, *et al*[12] based on the limits of agreement analysis of Bland and Altman[13]) or by analysis of variance. Analysis of the COBRA data was carried out using SPSS.

RESULTS

The overall success rates for measurements in the COBRA set (exercise B) are shown in Table 1. Not all methods were able to measure all joints; failure rates for wrist joints were 25% and 100% for methods B and D, respectively, 15% for MTP joints by method A and 2% for MCP joints by method E. Method F at present has been trained only for measurement of MCP joints, accounting for its low success rate. Methods B and D failed to read any wrist joints in the COBRA set.

Results of the initial exercises indicate the probable utility of metric JSW measurements in clinical trials and longterm studies in RA. The mean measurements and the mean annualized change in measurements in exercise A are shown for each method for all joints and separately for PIP, MCP, wrist, and MTP joints in Table 2. The different methods demonstrate consistent ranking in mean JSW magnitude across all joint groups.

There is considerable variation in JSW for all joints and even within each joint group, as reflected in the standard deviation. This represents differences between the patients, the timepoints, and individual joints within each group; for example, the second and fifth PIP joints. Although measured JSW differ, the standard deviations of JSW show remarkable consistency across methods within each joint area. This indicates that the variation in JSW between patients and between joints is similar for all methods, but care should be taken in interpreting the wrist, MTP, and overall results due to the different numbers of joints successfully measured by the different methods. The number of patients (4 with 10 timepoints in exercise A) and, hence, the number of change measurements per individual joint (for example the left PIP digit 4) is too small to evaluate within-joint variability or within-joint discriminative ability. Measurements of the rate of change of

Table 1. Change in measurement comparing treatment groups for all available data (exercise B: monotherapy was sulfasalazine, COBRA was combination therapy with sulfasalazine, methotrexate, and prednisolone.

	No.	Monotherapy	COBRA	t test	p
Δ (Baseline – 6 mo)					
Joint space score	107	1.580 (3.208)	0.947 (1.929)	1.253	0.213
Erosion score	107	5.130 (5.405)	2.105 (3.323)	3.429	0.001
Total score	107	6.710 (7.147)	3.053 (4.460)	3.124	0.002
Angwin (method E)	106	–0.068 (0.084)	–0.031 (0.073)	–2.466	0.015
Duryea (D)	26	–0.041 (0.067)	0.029 (0.108)	–1.901	0.069
Kauffman (B)	98	–0.024 (0.041)	0.001 (0.067)	–2.251	0.027
Peloschek (F)	35	–0.072 (0.135)	–0.062 (0.146)	–0.198	0.844
Sharp (A)	101	–0.076 (0.148)	–0.027 (0.149)	–1.639	0.104
Δ (Baseline – 12 mo)					
Joint space score	107	3.122 (5.234)	2.605 (4.682)	0.537	0.592
Erosion score	107	7.878 (8.381)	4.061 (5.925)	2.666	0.009
Total score	107	11.000 (11.618)	6.667 (9.493)	2.113	0.037
Angwin (method E)	107	–0.088 (0.110)	–0.054 (0.094)	–1.740	0.084
Duryea (D)	23	0.051 (0.296)	–0.012 (0.078)	0.715	0.482
Kauffman (B)	99	–0.024 (0.054)	–0.012 (0.065)	–0.927	0.356
Peloschek (F)	33	–0.021 (0.283)	–0.061 (0.165)	–0.507	0.616
Sharp (A)	103	–0.096 (0.151)	–0.028 (0.154)	–2.245	0.027
Δ (Baseline – last radiograph)					
Joint space score	107	4.827 (6.905)	4.026 (6.780)	0.601	0.549
Erosion score	107	10.918 (11.125)	6.439 (8.569)	2.294	0.024
Total score	107	15.745 (16.144)	10.465 (14.32)	1.784	0.077
Angwin (method E)		NA	NA	NA	NA
Duryea (D)	25	–0.083 (0.097)	0.038 (0.228)	–1.580	0.127
Kauffman (B)	100	–0.037 (0.075)	–0.029 (0.078)	–0.521	0.603
Peloschek (F)	37	–0.082 (0.215)	–0.042 (0.199)	–0.579	0.566
Sharp (A)	103	–0.111 (0.203)	–0.055 (0.176)	–1.520	0.133

Table 2. Mean joint space width and rate of change for PIP, MCP, wrist joints, and MTP in millimeters (exercise A).

Method	n	Mean JSW (SD)	n	Mean Rate of Change (SD)
All joints				
Sharp (method A)	319	1.55 (0.53)*	118	–0.08 (0.33)**
Kauffman (B)	340	1.42 (0.46)	136	–0.08 (0.27)
Duryea (D)***	240	1.28 (0.49)	96	–0.08 (0.20)
Angwin (E)	338	1.77 (0.56)	135	–0.09 (0.24)
PIP				
Sharp (method A)	78	0.90 (0.19)	31	–0.12 (0.23)
Kauffman (B)	80	0.79 (0.16)	32	–0.04 (0.11)
Duryea (D)	80	0.78 (0.19)	32	–0.06 (0.17)
Angwin (E)	80	1.10 (0.24)	32	–0.09 (0.11)
MCP				
Sharp (method A)	80	1.41 (0.35)	32	0.00 (0.49)
Kauffman (B)	80	1.39 (0.31)	32	–0.01 (0.19)
Duryea (D)	80	1.31 (0.32)	32	–0.01 (0.17)
Angwin (E)	78	1.59 (0.30)	31	–0.06 (0.14)
Wrist joints*				
Sharp (method A)	76	1.89 (0.38)	29	–0.11 (0.30)
Kauffman (B)	80	1.61 (0.26)	32	–0.17 (0.43)
Angwin (E)	80	2.12 (0.44)	32	–0.04 (0.33)
MTP				
Sharp (method A)	85	1.98 (0.32)	26	–0.09 (0.23)
Kauffman (B)	100	1.80 (0.25)	40	–0.11 (0.25)
Duryea (D)***	80	1.76 (0.33)	32	–0.18 (0.23)
Angwin (E)	100	2.16 (0.39)	40	–0.15 (0.29)

* Mean JSW (SD) for all patients, all timepoints, and all relevant measured joints. ** Mean annualized rate of JSW change (SD) between the first 2 timepoints for all patients and all relevant measured joints. *** No wrist joints were measured.

JSW show no systematic differences between methods, and again, standard deviations show remarkable consistency across methods within each joint area (Table 2, last column).

These results are presented graphically in Figures 2 and 3. Each point on the graph in Figure 2 is the JSW for a single joint for one timepoint per patient determined by one method, plotted against the averaged measurements of the same joint over all methods (340 sets of points). Where repeated measurements were recorded, the first was used. In Figure 3, each point represents the change in JSW between 2 timepoints for a single joint by one method, plotted against the average change in measurements for all readers of the same joint (204 sets of points). Systematic differences are seen between methods in Figure 2, but not in Figure 3. Again, the data demonstrate differences between methods in absolute measurements, but with very similar measurement of change in serial radiographs.

Not all programs completed intra-operator and inter-operator repeated measurements. Method B as used in exercise A, but not in COBRA, was the most fully automated method. Repeated measurements were not tested for this method, assuming that operator input had little effect on the final measurement. Developers of the less automated methods, A and E, completed intra-operator repeat measurements, and developers of methods A, D, and E completed repeated measurements by different operators. As expected, agreement within methods is better than between methods. The SDD for duplicate measurements by the same reader within methods varied from 0.124 to 0.260 mm, and the mean of differences was tight around zero, varying from –0.001 to 0.009 mm. Analysis of the inter-reader data reveals greater variability for methods A and E, suggesting that further automation of some steps, or more operator training, might improve performance for these methods. The SDD for these 2 methods ranged between 0.156 and 0.406 mm for method E and between 0.376 and 0.599 mm for method A. The inter-reader variability for method D is in the same range as the intra-reader data, probably reflecting its greater automation. The absolute values of the SDD should not be compared across methods, because the wrist measurements were not carried out by 2 methods and these measurements are the most variable.

Data from the initial analysis of the COBRA trial, exercise B, permits comparison of the sensitivity of measurements to scoring, but must be evaluated in light of the proportion of joints that were successfully measured by each method, shown in column 2 of Table 1. The results of t tests and p values are shown for each comparison of average change in measurement between patients treated with sulfasalazine only versus combination therapy, measuring change between baseline and 6 months, baseline and 12 months, and baseline to last radiograph only if the number of successfully measured joint pairs was 50%, considering only those joints in which measurement was attempted. This analysis ignores the fact

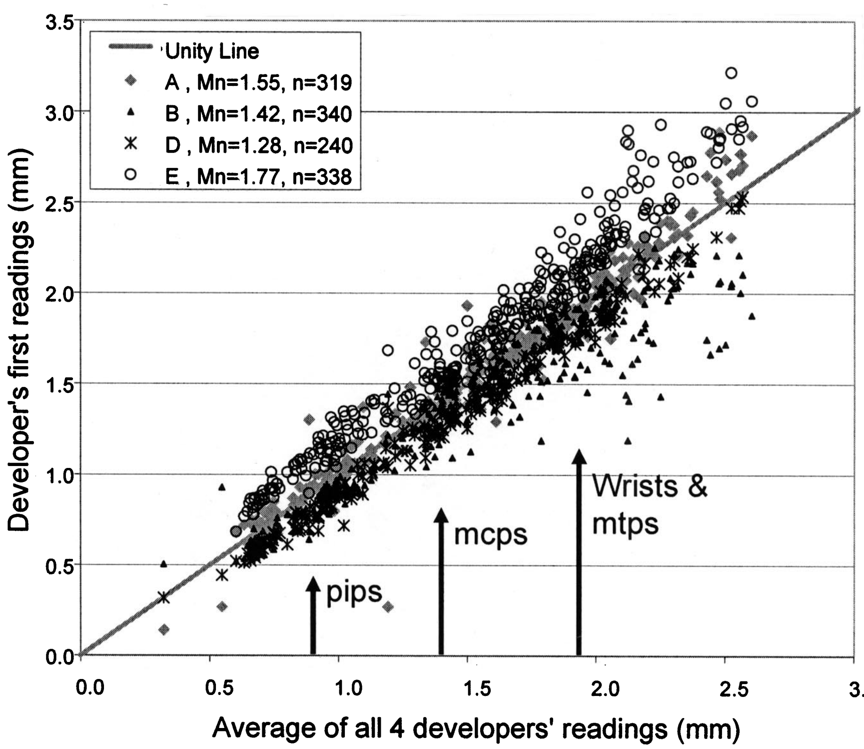

Figure 2. Comparing methods: JSW measurements (mm); each point on the graph is the JSW for a single joint for one timepoint per patient determined by one method, plotted against the averaged measurements of the same joint over all methods (340 sets of points). The developer's first recording is shown. Vertical lines indicate average JSW measurements for PIP, MCP, wrists, and MTP joints (over all patients, all timepoints, and all joints of this type).

that one investigator did not have sufficient time in his schedule to complete the measurements and 3 investigators did not make any measurements of the wrist joints. With these points in mind, loss of joint space was more frequently detected by measurement than by scoring joint space narrowing using 4 of 5 methods for the 0 to 6 month and 0 to 12 month intervals and by 2 of 4 methods for the 0 to last radiograph interval. Not surprisingly, measurements were less sensitive to change than erosion and total score by the manual method.

DISCUSSION

Exercise A was focused on determination of reproducibility of computer based methods for measurement of the average distance between proximal and distal joint edges in radiographic images of the fingers, toes, and wrist joints. JSW measurement of the COBRA images in exercise B was carried out specifically to compare measurements to scores of joint space narrowing regarding sensitivity to change. Neither within-reader nor within-method repeated measurements were completed for all images in either exercise. Considering the duplicate measurements that were completed in exercise A, intra-reader and intra-method agreement indicate the error term is within a fraction of a millimeter. As expected, intra-method agreement was better than comparisons between methods. One program that required more operator input (method A) exhibited more intra- and inter-operator variability than more fully automated programs. Method D showed particularly good inter-operator agreement, as expected since there is very little operator input, but results by this method do not include wrist measurements, which vary more than finger and toe measurements.

Although JSW measurements showed systematic differences between methods, change measured in serial radiographs showed good agreement, with no observable systematic differences. JSW changes in exercise A (Table 3) were averaged for all measured joints per patient per method, and show excellent agreement considering that each set of results (i.e., each row) is for a single patient, and not all methods measured all joints. Three of these patients had definite progression by scoring and one (Patient 95, Table 3) was non-progressing. The JSW changes are in agreement with scoring. A trend of decreasing radiographic progression over time was shown by all methods for Patient 203. Although not proven, this could represent a treatment effect, since disease modifying therapy was started after the baseline radiograph was obtained and, provided there was a lag in the treatment effect, this may not have influenced the progression rate until later. The joint margins in the radiographs of Patient 203 were depicted particularly clearly, and this may be a factor in the particularly high level of agreement in these results. The 3 other patients had less well distinguished joint margins, and the level of agreement for these patients is very encouraging.

Radiographic evaluation of joint damage is useful in evaluating status in relation to other disease features, in describing longterm outcome, and in measuring progression of disease in longitudinal studies and in therapeutic trials. The results of our exercise A and the preliminary results of exercise B (the

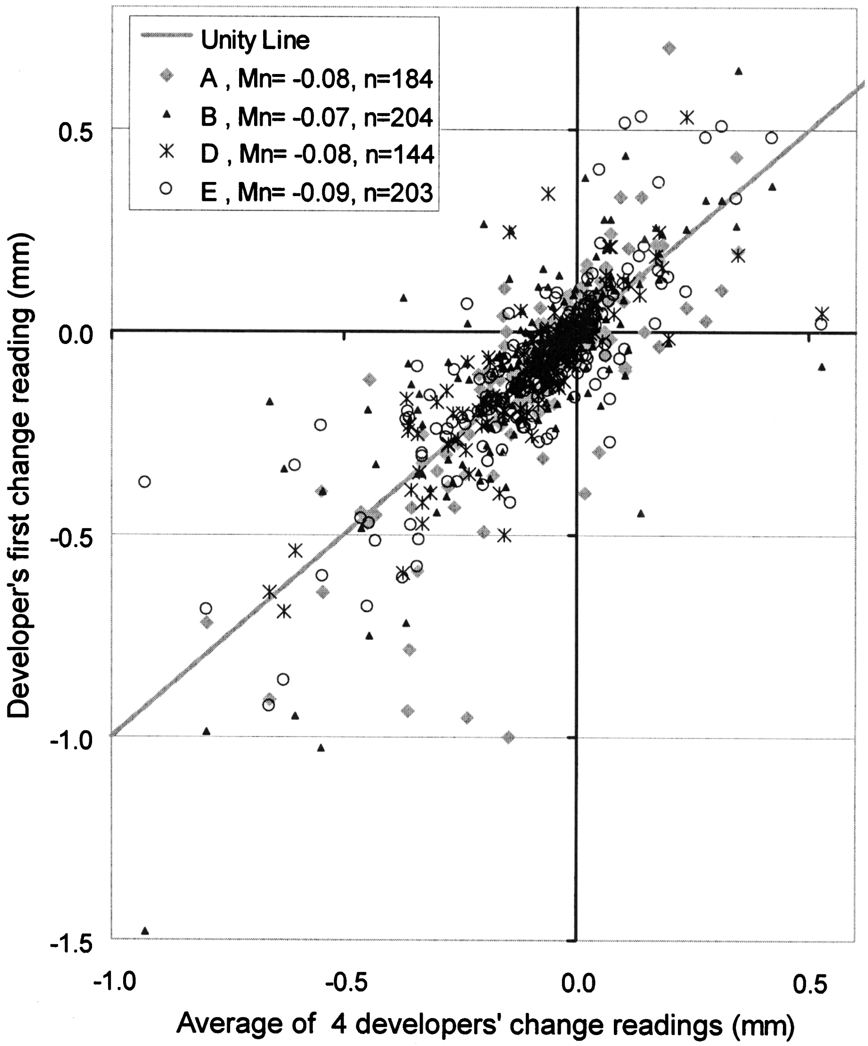

Figure 3. Comparing methods: JSW change (mm); each point on the graph represents the change in JSW between 2 timepoints for a single joint by one method plotted against the average change in measurements for all readers of the same joint (204 sets of points).

COBRA data) indicate that computer based measurements are likely to be useful both in status evaluation and in measurement of change in serial radiographs for clinical trials and studies describing the course of the disease. Although there is a difference in absolute measurements between methods, the similarity of change measurements indicates that serial studies would be reproducible if measurements were all made by the same method. And since change is the focus of longitudinal studies, provided the same method was used for measurements in all studies compared, comparability across studies

Table 3. Within-patient annualized JSW change over time (mm)* (exercise A).

Patient	Days from Baseline	Method A Sharp	Method B Kauffman	Method D Duryea	Method E Angwin	Average
95	352	–0.008	–0.003	–0.032	0.031	–0.003
136	219	–0.043	–0.024	–0.070	–0.046	–0.045
156	239	–0.105	–0.121	–0.054	–0.145	–0.106
203	185	–0.162	–0.188	–0.175	–0.192	–0.179
203	364	–0.127	–0.073	–0.115	–0.149	–0.116
203	798	–0.047	–0.021	–0.026	–0.048	–0.035

* The change in annualized JSW measurement, averaged over all measured joints per patient per time interval. Patient 203 has 4 radiographs included in the exercise. The annualized change measured from baseline is shown for each timepoint.

should be valid. This should be a powerful tool in conducting longterm studies, particularly when comparison of treatment regimens is of interest.

One note of caution: establishing normal values for the healthy population will be difficult. Considerable variability in JSW occurs among healthy males of similar age[4] and among RA patients across sexes in joints considered normal on inspection by experts[3]. Narrowing of joints is a common feature of disease, but there is extreme variability in otherwise apparently normal joints. Age, genetic makeup, regular activity, and perhaps other constitutional factors influence JSW in "normal" individuals. Hence, a "normal" standard has not been defined and large population based studies will be required to define "normal" measurements for each individual joint stratified for sex, age, and other factors.

Feasibility is a multifaceted issue. For some of the programs studied here the time for measurement of 34 joints in both hands, wrists, and feet is greater than 10 minutes per patient per timepoint. This is considerably longer than the average time spent scoring the same number of joints. Although some methods are considerably shorter and it is anticipated that there will be further improvement in time performance with further refinement of the computer programs under development, it is not yet certain that measurements will be achieved in the same timeframe as scoring. So as the methods now stand there is a tradeoff between greater sensitivity and longer time required to collect the data. However, in therapeutic trials the burden of a time penalty for using computer based joint space measurements will be offset by greater precision and reliability, resulting in significant reduction in number of patients required to demonstrate benefit, adding additional benefit to the greater sensitivity and reliability of the measurements. A further benefit of computer based measurements is the expectation that less training will be necessary for obtaining more reliable results. Although this is a central issue in adopting measurements in clinical trials, it is premature to attempt to answer the question definitively until more data are available on reliability and precision and more advanced versions of the programs have been written and tested. Even so, successful measurement of more than 90% of selected joints by 3 of the programs participating in the measurements on COBRA images is considered encouraging enough to warrant testing in parallel with scoring in a prospective study.

Further reliability measurements were made within-image and these do not take into account differences due to repeated imaging, such as change in position of hands and angle of x-ray beam. The development of automated, computer based measurements of joint damage is a continuing process. At present, experience with measurement of JSW appears to be fulfilling its anticipated reliability and potential value. Future development of computer based measurements by the cooperating program managers will encourage those whose programs are not fully automated to modify computer methods to incorporate features that reduce operator dependence without sacrificing accuracy of measurements. For those programs that are more fully automated but successful in less than a satisfactory proportion of cases, it is appropriate to question whether any operator-dependent steps would improve the success rate and improve reliability without penalizing time requirement.

ACKNOWLEDGMENT

The committee is indebted to the investigators in the Western Consortium of Rheumatologists for permission to use images in exercise 2 and the organizing investigators of the COBRA study for permission to use the COBRA images. GlaxoSmithKline supported digitization of the COBRA images.

REFERENCES

1. Sharp JT, van der Heijde D, Angwin J, et al. Measurement of joint space width and erosion size. J Rheumatol 2005;32:2456-61.
2. Sharp JT, Wolfe F, Boers M, et al. Variability of precision in scoring radiographic abnormalities in rheumatoid arthritis by experienced readers. J Rheumatol 2004;31:1062-72.
3. Angwin J, Lloyd A, Heald G, Nepom G, Binks M, James MF. Radiographic hand joint space width assessed by computer is a sensitive measure of change in early rheumatoid arthritis. J Rheumatol 2004;31:1050-61.
4. Angwin J, Heald G, Lloyd A, Howland K, Davy M, James MF. Reliability and sensitivity of joint space measurements in hand radiographs using computerized image analysis. J Rheumatol 2001;28:1825-36.
5. Duryea J, Li J, Peterfy CG, Gordon C, Genant HK. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. Med Phys 2000;27:580-91.
6. Duryea J, Jiang Y, Zakharevich M, Genant HK. Neural network based algorithm to quantify joint space width in joints of the hand for arthritis assessment. Med Phys 2000;27:1185-94.
7. Kauffman JA, Slump CH, Moens HJB. Segmentation of hand radiographs with multilevel connected active appearance models. Medical Imaging: Image Processing 2005;5747-182. Proceedings of SPIE.
8. Kauffman JA, Slump CH, Moens HJB. Detection of joint space narrowing in hand radiographs. Medical Imaging: Image Processing 2006;6144-149. Proceedings of SPIE.
9. Sharp JT, Gardner JC, Bennett EM. Computer-based methods for measuring joint space and estimating erosion volume in the finger and wrist joints of patients with rheumatoid arthritis. Arthritis Rheum 2000;43:1378-86.
10. Peloschek P, Bogl K, Robinson S, Lomoschitz F, Graninger W, Kainberger F. Computer-assisted radiologic quantification of hand and foot changes in rheumatoid arthritis. Wien Med Wochenschrift Suppl 2002;113:37-8.
11. Boers M, Verhoeven AC, Markusse HM, et al. Randomised comparison of step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309-18.
12. Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999;26:731-9.
13. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.

APPENDIX — Methods A, B, D, E, F

Method of J. Hall in collaboration with J.T. Sharp (Method A); program written by J. Hall, Snoqualmie, WA, in collaboration with J.T. Sharp, MD, Bainbridge Island, WA.

Automatic Joint Space Measurement: Development of Automated Method for Joint Space Measurement

Figure i

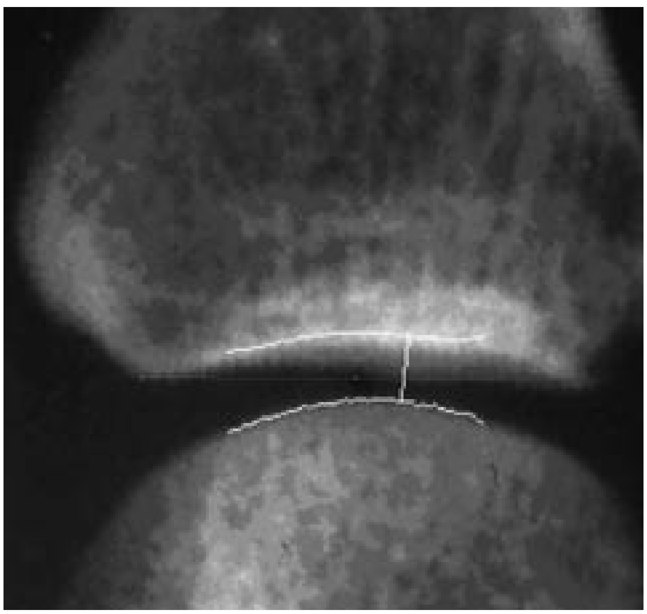

Figure ii

Radiographic images of hands and feet representing multiple timepoints were digitized and provided for the analysis. The digital image files were organized in folders and processed in batches. Using the program's interface the operator draws a line identifying the joint margin of each PIP, MCP, MTP, or wrist joint to be measured (Figure i). The operator identifies the span of each joint by placing markers on the radial and ulnar sides of each joint to be measured. The program then crops, rotates, and prepares an image file for each joint. Then each joint is measured automatically and the results are presented to the operator for review (Figure ii). The minimum and mean joint space results are saved in Excel™ spreadsheets and XML files along with supporting information.

To complete the task the program prepares a small 8-bit black and white image rotated and cropped so that the joint is centered and horizontal. A measurement rectangle representing 60% of the joint span defines the edge detection search area. The image is smoothed by temporarily applying a Gaussian blur. Seven vertical lines equally spaced inside the measurement rectangle define pixel density profiles. By examining the peaks and valleys in each profile the program locates the shoulder of the distal edge and the top of the proximal bone. After 7 sample points are identified along the top and bottom edges, a polynomial curve is applied along the sample points to approximate each edge (Figure iii). Next, edge points are computed by examining the pixels along each polynomial.

Given the Y location arrays for both the distal and proximal edges prepared in the previous step, the minimum and average distance between the 2 curves is calculated and plotted so the operator can review them. When things went wrong, the search points were adjusted manually and the process of finding the edges was repeated. The operator reviewed each solution and accepted them based on his visual appraisal of the edge fit.

Joint Image Analysis is a plug-in application that runs with ImageJ, a public domain image processing program supported by the National Institutes of Health. It is written in Java and runs on Windows™ or Macintosh™.

Method of J. Kauffman and H. Bernelot Moens (Method B).

The program measures JSW of the MCP joints 2–5 and PIP joints 2–5 and MTP joints 1–5. This program is being developed in The Netherlands at the University of Twente in collaboration with Hospital Group Twente. Financial support is provided by the Dutch Arthritis Association.

The current state of the program enables automatic detection of the joints in a hand radiograph (Kauffman JA, Slump CH, Bernelot Moens HJ. Segmentation of hand radiographs by using multi-level connected active appearance models. In: Fitzpatrick JM, Reinhardt JM, editors. Medical imaging 2005: image processing. Proceedings of the SPIE 2005;5747:1571-81). Nevertheless, initialization was done manually for the COBRA data set, since the image quality was not sufficient for a large portion of the data.

Manual initialization was done by an operator indicating several points of interest. For each hand, 8 points where indicated: the proximal and distal end of the proximal phalanges 2–5 (Figure iv). The same was done for the feet, with 10 points for all 5 proximal phalanges.

The initialization points determine the approximate locations of the joints and the medial axes of phalanges. The angles of the medial axes provide information about the orientation of the joints. The joint margins are detected with a statistical shape model, which means that only plausible margin shapes are detected (Kauffman JA, Slump CH, Bernelot Moens HJ. Detection of joint space narrowing in hand radiographs. In: Reinhardt JM, Pluim JPW, editors. Medical imaging 2006: image processing. Proceedings of the SPIE

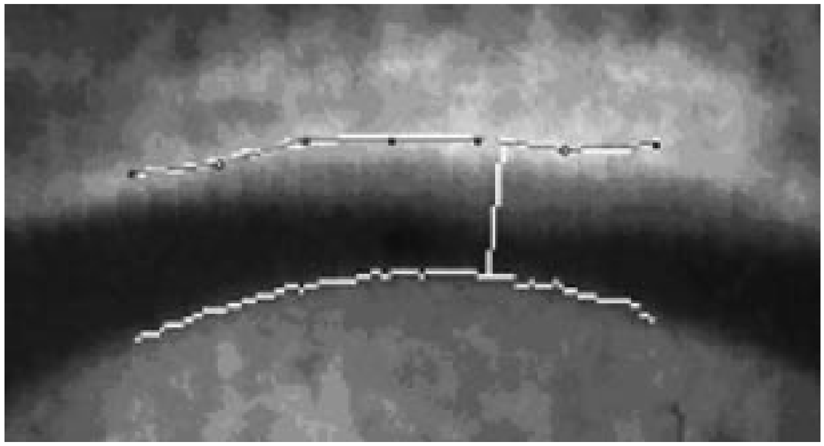

Figure iii

Figure iv. Initialization points.

2006;6144:1332-42). Both the pixel intensities and their gradients are used to find the margins. The properties of this model have been determined from a different set of 50 training samples, in which the joint margins had been delineated manually by a trained person.

After detection of the joint margins, the JSW is calculated within a region of 6 mm centered on the medial axis of the proximal phalanx (Figure 5). By calculating the point-line distance from the proximal margin to the distal margin, the measured JSW is not affected by the orientation of the joint. The results can be displayed graphically or exported to a table for MS Excel.

Figure v. Measurement results.

Method of J. Duryea (Method D).

In order to characterize a joint radiographically it is necessary to first determine its location. Manual joint space identification is very time consuming, particularly if the number of images is large. Therefore an automated rule-based algorithm was developed to determine joint space locations and the approximate orientation of the digits on digitized radiographs of the hand (Duryea J, Jiang Y, Countryman P, Genant HK. Automated algorithm for the identification of joint space and phalanx margin locations on digitized hand radiographs. Med Physics 1999;25:453-61). The locations of the DIP, PIP, and MCP joints were identified on Digits 2, 3, 4, and 5 (Figure vi).

Cropped images at each joint location were rotated to align the joint with the vertical axis of the image. An artificial neural network (ANN) based algorithm (Figure vii) was used to delineate the joint margins (Duryea J, Jiang Y, Zakharevich M, Genant HK. Neural network based algorithm to quantify joint space width in joints of the hand for arthritis assessment. Med Phys 2000;27:1185-94). Two lines tangent to the edge of the margins in the distal portion of the joint were used as landmarks to establish a coordinate system. By definition the lines are at x = 0 and x = 1.0. A central measurement region

Figure vi. The relevant joints.

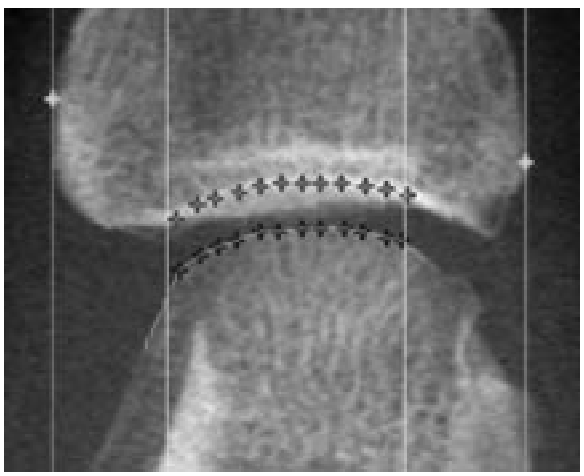

Figure vii. Output of ANN based joint delineation software.

between x = 0.25 and x = 0.75 was defined (Figure vii). The software then calculated the JSW as the average distance between the delineated margins in the central measurement region. The software can also determine JSW(x) for 0.25 > x > 0.75. For the OMERACT study we provided a measure of JSW (x = 0.5).

Method of M.F. James, G. Heald, J.H. Shorter, and J. Angwin (Method E).
The JSW measuring software was developed at GlaxoSmithKline R&D as described (Angwin J, Heald G, Lloyd A, Howland K, Davy M, James MF. Reliability and sensitivity of joint space measurements in hand radiographs using computerized image analysis. J Rheumatol 2001;28:1825-36; Angwin J, Lloyd A, Heald G, Nepom GT, Binks MH, James MF. Radiographic hand joint space width assessed by computer is a sensitive measure of change in early rheumatoid arthritis. J Rheumatol 2004;31:1050-61) and is based on software developed by James, *et al* (James MF, Heald G, Shorter JH, Turner RA. Joint space measurement in hand radiographs using computerized image analysis. Arthritis Rheum 1995;38:891-901).

MCP joints are located in the digitized image by positioning 3 user-input points along the curved metacarpal head, with one point identifying the midpoint of the measurement arc. The remaining measurement process is automatic. The 3 points are used to calculate the approximate center of the metacarpal head, from which ~180 radial lines of pixels traversing the joint space are sampled. To allow for MCP of differing size, the mean joint space is measured within a boundary arc ± 0.5 radian from the midpoint of the metacarpal head. The metacarpal margin is identified from a position along each line at the point of maximum slope in radiographic density. The proximal phalanx margin is defined on the same radial line at the point of peak radiographic density across the joint space. A variation of this routine is used for MTP and wrist measurements.

PIP joints are viewed horizontally and approximately located in the digitized image using a rectangular region of interest determined by the user. The same features of slope and peak are used to define the joint margins, but in contrast to MCP joints, the PIP joint margins are defined by sampling parallel, not radial, lines oriented vertically across the joint. The outer limits of the proximal phalangeal articular margin either are detected automatically or are limited by the operator when selecting the region of interest.

In each of the MCP and PIP processes, after an initial pass through the data, the radiographic joint margins are precisely located using a Gaussian function in a tracking procedure. The Gaussian function optimally combines the conflicting criteria of insensitivity to noise and accuracy (Canny JF. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 1986;8:679-98; Torre V, Poggio T. On edge detection. IEEE Trans Pattern Anal Mach Intell 1986;8:147-63) and contributes substantially to the success of the software in identifying valid anatomical margins irrespective of image digitization and noise. Having located the radiographic joint margins on all lines, the mean JSW is calculated as their linear separation averaged over the defined joint breadth (averaging ~180 values). Values are stored automatically in a spreadsheet.

Monitoring program performance. The computer program is not fully automatic. Serial radiographs are measured together but without knowledge of sequence. This allows the user to ensure that the same portion of each joint is selected at all timepoints.

Validation of computerized mean JSW measurements. Computer measurements of mean JSW were validated against manual measurements by 2 observers each using 2 different methods and repeated 3 times (Angwin J, Heald G, Lloyd A, Howland K, Davy M, James MF. Reliability and sensitivity of joint space measurements in hand radiographs using computerized image analysis. J Rheumatol 2001;28:1825-36). In the first manual method, the radiograph of a joint was examined under a microscope using an eyepiece with a graduated reticle. In the second method, the radiograph was digitized and standard distance routines were used to determine the length of lines placed manually across the joint space. In both cases, the agreement was within 0.01 mm, with tighter standard deviations for computer-repeated measurements (0.003 mm) than for the manual methods (0.020 mm).

Method of P.L. Peloschek and G. Langs (Method F).
AAMIR–RAQuantify. This study was partially supported by the Austrian Science Fund (FWF, grant no. P17083-N04).
The joint space width measurement program AAMIR–RAQuantify is accessible on the Internet by multiple users at different locations and does not require manual interaction on the measured joints.

Selection of area of joint to measure. The joint space is defined by the cortex of the metacarpal head and the concave, highest gray-level gradient in the base of the proximal phalanx as a reproducible substitute for the true distal joint margin.

Identifying the edges of the joint margin. The measurements are confined to the central part of the joint, which was defined by excluding the peripheral convex parts of the base of the proximal phalanx.

Calculation of distance across the joint space. The joint margin on the proximal phalanx serves as a reference contour, from which the minimum distance to the metacarpal head at a minimum of 80 points is determined. The mean of these is the final measurement.