Multiple Computer-based Methods of Measuring Joint Space Width Can Discriminate Between Treatment Arms in the COBRA Trial — Update of an Ongoing OMERACT Project

JOHN T. SHARP, JANE ANGWIN, MAARTEN BOERS, JEFF DURYEA, AXEL FINCKH, JAMES R. HALL, JOOST A. KAUFFMAN, ROBERT LANDEWÉ, GEORG LANGS, CÉDRIC LUKAS, H.J. BERNELOT MOENS, PHILIPP PELOSCHEK, C. VIBEKE STRAND, and DÉSIRÉE van der HEIJDE

ABSTRACT. Previously reported data on 5 computer-based programs for measurement of joint space width focusing on discriminating ability and reproducibility are updated, showing new data. Four of 5 different programs for measuring joint space width were more discriminating than observer scoring for change in narrowing in the 12 months interval. Three of 4 programs were more discriminating than observer scoring for the 0–18 month interval. The program that failed to discriminate in the 0–12 month interval was not the same program that failed in the 0–18 month interval. The committee agreed at an interim meeting in November 2007 that an important goal for computer-based measurement programs is a 90% success rate in making measurements of joint pairs in followup studies. This means that the same joint must be measured in images of both timepoints in order to assess change over time in serial radiographs. None of the programs met this 90% threshold, but 3 programs achieved 85%–90% success rate. Intraclass correlation coefficients for assessing change in joint space width in individual joints were 0.98 or 0.99 for 4 programs. The smallest detectable change was < 0.2 mm for 4 of the 5 programs, representing 29%–36% of the change within the 99th percentile of measurements. (J Rheumatol 2009;36:1825–8; doi:10.3899/jrheum.090353)

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

A committee to examine and test the feasibility and reliability of computer-based measurements of features in hand and foot radiographs was established by OMERACT in 2002. Individuals known to be working on computer methods of measurement were invited to join, and 5 groups have actively participated[1-8]. Two groups, Ziekenhuis Groep Twente, Hengelo, and the Medical University of Vienna, Vienna, Austria, were collaborating with technical departments, and graduate students were contributing much or all of the programming skills. Two new groups have recently joined in the committee's efforts.

As its initial project, the committee undertook testing programs to measure joint space width[9,10]. There is agreement that a successful computer-based program for measur-

From the University of Washington, Division of Rheumatology, Department of Medicine, University of Washington, Seattle, WA, USA; GlaxoSmithKline, London, England; Department of Clinical Epidemiology and Biostatistics, Free University Medical Center, Amsterdam, The Netherlands; Brigham and Women's Hospital, Harvard University, Boston, MA, USA; Division of Rheumatology, University Hospital of Geneva, Geneva, Switzerland; University of Twente, Enschede; University Hospital Maastricht, Maastricht, The Netherlands; Graz University of Technology, Graz, Austria; University Hospital Lapeyronie, Montpellier, France; Rheumatology, Ziekenhuis Groep Twente, Hengelo, The Netherlands; Department of Radiology, University of Vienna, Vienna, Austria; Division of Immunology/Rheumatology, Stanford University, Palo Alto, CA, USA; and Leiden University Medical Center, Leiden, The Netherlands.

The lead author, John T. Sharp, died September 14, 2008. Professor emeritus and retired rheumatologist, he chaired the OMERACT subcommittee on automated joint measurement. The members of the subcommittee and co-authors of this article commemorate Prof. Sharp as an inspiring, scholarly, and ever-enthusiastic forerunner in the development of automated joint measurement, and intend to continue the scientific work on automated joint measurement in his spirit.

J.T. Sharp, MD, Professor, University of Washington, Division of Rheumatology, Department of Medicine, University of Washington; J. Angwin, GlaxoSmithKline; M. Boers, MSc, MD, PhD, Department of Clinical Epidemiology and Biostatistics, Free University Medical Center; J. Duryea, MD, Brigham and Women's Hospital, Harvard University; A. Finckh, MD, Professor, Division of Rheumatology, University Hospital of Geneva; J.R. Hall, Snoqualmie, WA, USA; J.A. Kauffman, PhD, University of Twente; R. Landewé, MD, University Hospital Maastricht; G. Langs, MSc, Graz University of Technology; C. Lukas, University Hospital Lapeyronie; H.J. Bernelot Moens, Rheumatology, Ziekenhuis Groep Twente; P. Peloschek, MD, Department of Radiology, University of Vienna; C.V. Strand, MD, Professor, Division of Immunology/Rheumatology; D. van der Heijde, Professor, Leiden University Medical Center.

Address correspondence to Dr. R. Landewé, Maastricht University Medical Center, Department of Internal Medicine, Subdivision of Rheumatology, PO Box 5800, 6202AZ Maastricht, The Netherlands. E-mail: r.landewe@mumc.nl

ing joint space width in metric units must demonstrate equal or greater sensitivity to change than observer scoring in discriminating between treatment arms in clinical trials. To this end, 5 developers have measured joint space width in the COBRA trial image set to compare computer measurements with observer scoring[10].

Since the data were presented at OMERACT 8, one program that had not yet completed measurements on the full COBRA set has now done so, and one other program has been extensively revised[9,10]. The additional data are included in Table 1 (also available from: www.omeract.org and originally published in[9,10]). The data demonstrate that computer-based measurements are more discriminatory than observer scoring in 11 of 14 comparisons. It should be noted that data were taken into consideration only if the program was able to successfully measure at least 50% of the required joint pairs per patient. This prerequisite explains why the numbers of patients that were assessed differed across methods.

The committee has agreed that data should be complete in at least 90% of paired measurements in order to use automated scoring in studies. This implies successful measurements of joint space width for the same joints at both timepoints in serial radiographs, a requirement necessary for calculating change in joint space width as an outcome measure.

The committee has identified multiple causes of measurement failure, which are listed in Table 2. In order to get some insight into the practical reasons for measurement failure, reasons were recorded in the reevaluation of the entire set of 107 cases using the revised Sharp program, which measures a set of 34 individual joints. Overall, 2.9% of all single-joint measurements failed for a reason. Only 0.2% of single-joint measurements failed due to the inability of the computer program to find the joint margins. The criterion of at least 90% successful joint assessment per patient (≥ 31/34 successful joints) was met in 92% of cases if only one successful timepoint was required, and in 87% of the cases if 2 successful timepoints (for assessing change) were required. This result is comparable to results of the 2 best performing programs in the previous report[10].

Although our committee has not proposed a standard for reliability, intraclass correlation coefficients (ICC) give an indication of relative agreement, and the smallest detectable changes (SDC) give an indication about absolute agreement of the measurement programs. In order to test reproducibility of the 5 computer programs, the complete assessment was repeated in a set of 30 selected paired cases, all belonging to the COBRA dataset. This set included radiographs

Table 1. Discriminatory ability of 5 computer programs in comparison with semiquantitative observer scoring in measuring change in joint space width in the monotherapy group and the COBRA therapy group of the COBRA trial, taking all measured joints into consideration. Published in part in a previous OMERACT report[10]. Since then, Duryea has completed the measurement by his program, one program (method D in[10]) was left out due to measurement failure, and a revised Sharp program was added. Data presented here include the mean change over all measured joints, but the type and number of joints measured per program differ.

	No. Patients Included*	Monotherapy, mm	COBRA Therapy, mm	t Test	p
Change between baseline and 6 mo					
Angwin method	105	–0.068 (0.084)	–0.031 (0.073)	–2.466	0.015
Duryea method**	104	–0.062 (0.077)	–0.011 (0.096)	–2.959	0.004
Kauffman-Moens method**	98	–0.024 (0.041)	0.001 (0.067)	–2.251	0.027
Sharp-Hall method	102	–0.076 (0.148)	–0.028 (0.148)	–1.639	0.104
Sharp revised method	93	–0.033 (0.143)	0.005 (0.118)	–1.153	0.252
Observed scoring (joint space only)	107	1.580 (3.208)	0.947 (1.929)	1.253	0.213
Change between baseline and 12 mo					
Angwin method	107	–0.088 (0.110)	–0.054 (0.094)	–1.743	0.084
Duryea method**	99	–0.035 (0.168)	–0.038 (0.091)	0.094	0.925
Kauffman-Moens method**	99	–0.024 (0.054)	–0.012 (0.065)	–0.927	0.356
Sharp-Hall method	104	–0.096 (0.151)	–0.029 (0.152)	–2.239	0.027
Sharp revised method	100	–0.059 (0.132)	–0.012 (0.111)	–2.184	0.031
Observed scoring (joint space only)	107	3.122 (5.234)	2.605 (4.682)	0.537	0.592
Change between baseline and 18 mo					
Angwin method	NA	NA	NA	NA	NA
Duryea method**	103	–0.071 (0.100)	–0.032 (0.157)	–1.493	0.139
Kauffman-Moens method**	100	–0.037 (0.075)	–0.029 (0.078)	–0.521	0.603
Sharp-Hall method	104	–0.111 (0.203)	–0.054 (0.175)	–1.548	0.125
Sharp revised method	97	–0.067 (0.170)	–0.017 (0.134)	–1.605	0.112
Observed scoring (joint space only)	107	4.827 (6.905)	4.026 (6.780)	0.601	0.549

* No. of patients in which at least 50% of the attempted joint measurements per patient were successful.
** Duryea and Kauffman-Moens did not measure wrist joints. NA: not assessed.

Table 2. Potential reasons for measurement failure in the assessment of joint space width by computer programs.

Inappropriate imaging technique	Joint not included in the image
	Image quality (minimal dynamic range)
Patient-related	Subluxation
	Flexion contracture
	Severe asymmetry of the joint
	Osteoarthritis
	Destruction of the joint
Program-related	Inability to locate the joint
	Inability to locate the joint margins

taken at baseline and radiographs at 18 months' followup. Part of these data have been published in an aggregated manner[10]. In Table 3 the results of 4 of the 5 previous computer programs plus the results of the revised Sharp method are summarized more comprehensively, so that the reliability per program per joint group can be investigated.

Expectedly, ICC for status scores were higher than ICC for change scores for all programs. If all joints were taken into account, 4 of the 5 programs yielded ICC for status scores between 0.97 and 0.99. ICC for change scores were above 0.80 for the same 4 of the 5 methods, reflecting acceptable relative agreement. In contrast, however, the

Table 3. Relative and absolute agreement for 5 computer programs with respect to measuring joint space width and change in joint space width, as measured in a set of 30 paired images of the COBRA trial that were assessed twice by all methods.

	Relative Agreement		Absolute Agreement
	ICC Status Measurements (% valid cases)	ICC Change Measurements (% valid cases)	SDC (% of 99-percentile change)*
Angwin method			
MCP	0.98 (100)	0.93 (100)	0.07 (31)
MTP	0.96 (98)	0.93 (97)	0.12 (25)
PIP	0.98 (98)	0.89 (98)	0.06 (38)
Wrist	NA	NA	NA
All joints	0.99 (98)	0.92 (98)	0.09 (36)
Duryea method			
MCP	0.98 (100)	0.98 (100)	0.11 (31)
MTP	0.99 (97)	0.94 (97)	0.11 (20)
PIP	0.95 (100)	0.70 (100)	0.15 (34)
Wrist	NA	NA	NA
All joints	0.99 (95)	0.88 (97)	0.12 (29)
Kauffman-Moens method			
MCP	0.82 (74)	0.57 (59)	0.15 (36)
MTP	0.91 (81)	0.76 (72)	0.20 (43)
PIP	0.77 (86)	0.31 (78)	0.08 (20)
Wrist	NA	NA	NA
All joints	0.97 (80)	0.71 (70)	0.15 (35)
Sharp-Hall method			
MCP	0.23 (97)	0.09 (95)	1.10 (57)
MTP	0.62 (96)	0.08 (92)	0.79 (58)
PIP	0.22 (95)	0.06 (93)	0.98 (62)
Wrist	0.83 (84)	0.43 (73)	0.52 (70)
All joints	0.62 (93)	0.12 (88)	0.89 (41)
Sharp-revised method			
MCP	0.96 (99)	0.81 (99)	0.13 (45)
MTP	0.97 (98)	0.88 (97)	0.17 (31)
PIP	0.94 (98)	0.67 (97)	0.12 (48)
Wrist	0.89 (97)	0.82 (95)	0.31 (42)
All joints	0.98 (98)	0.83 (97)	0.19 (36)

* The SDC was expressed as the percentage of the change value representing the 99-percentile of the set of change values measured by the program. ICC: intraclass correlation coefficient; SDC: smallest detectable change; MCP: metacarpophalangeal; MTP: metatarsophalangeal; PIP: proximal interphalangeal. NA: not available.

SDC varied from 29% to 41% of the measured range (99th percentile), which indicates considerable residual measurement error.

DISCUSSION

Theoretically, measuring change in joint space width in metric units should be more reproducible than semiquantitative observer scoring, since a well constructed computer program may reduce operator/observer influence to near zero. Greater reproducibility may translate into greater sensitivity to change over time. The studies done by this committee support this contention. In addition, metric units make more sense to most people than van der Heijde, Larsen, or Sharp units. For someone not regularly using score data, a statement that a joint decreases width by 0.1 or 0.5 mm is more readily visualized than a joint narrowing score increasing by 1 or 2 units.

Are computer-based programs for measuring joint space width ready for use in clinical trials? The committee members believe they are, provided the conditions for digitizing the images meet the high standards employed in the full COBRA trial, in which images were digitized at 50-micron pixel size. Under these conditions an acceptable success rate was initially obtained by 3 computer programs. The other 2 programs were unable to measure a sufficient number of paired joints to be useful in clinical trials. The 50-micron pixel size (20 pixels/mm) is a higher resolution than is regularly employed by many clinical research organizations, which use 100-micron pixel size (10 pixels/mm) in clinical trials. Unless future image resolution is standardized at the higher resolution, computer programs need to be validated for use on images recorded at the 100-micron pixel size. The issue of digitization resolution needs to be resolved before measurements with the current instruments can be recommended for clinical trials employing resolution lower than 50-micron pixel size.

Additionally, since COBRA included only patients with early disease and consequently with little baseline damage, the programs need to be validated in patients with more extensive baseline damage. Theoretically, this may lead to lower success rates per patient, since measurement failure is more common in damaged joints.

Although the high standard of a 90% success rate in measuring both images in paired image sets was not reached by any of the programs, the majority of failures were due to image problems or structural abnormalities that precluded measurement. These factors may also influence observer scoring. No record was available as to how the readers who scored the radiographs handled these structural problems, other than subluxation, which is treated as joint space narrowing in the Sharp-van der Heijde method. Methods for imputation of data missing for technical reasons should be carefully considered. Multiple imputation based on mixed-effects analysis, as suggested recently by Baron, *et al*, may offer the smallest bias[11].

Most of the other issues that can affect the success rate relate to image quality. Undoubtedly, many of these issues have been ignored too long. The failure of radiographers to include all structures, such as the wrist, little finger, and little toe, is an inexcusable lapse in meeting appropriate standards for investigational studies. From personal experience we can say that such failures are common in clinical trials. Poor dynamic range of images is frequently due to using cheap film. OMERACT should take a strong and positive stand on this issue.

In the future the subcommittee on computer-based measurements is planning studies to evaluate the effect of image resolution of 50- versus 100-micron pixel size on measurements and to examine whether recording the image at 8- or 16-bit gray scale has an effect on computer-based measurements. As indicated above, further evaluations need to be done in patients with more extensive damage. Beyond that we will begin studies on measuring erosions.

REFERENCES

1. 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.
2. 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.
3. Finckh A, de Pablo P, Katz JN, et al. Performance of an automated computer-based scoring method to assess joint space narrowing in RA. A longitudinal study. Arthritis Rheum 2006;54:1444-50.
4. 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.
5. Kauffman JA, Slump CH, Moens HJB. Segmentation of hand radiographs with multilevel connected active appearance models. Proc SPIE Medical Imaging 2005;5747:1571-81.
6. Kauffman JA, Slump CH, Moens HJB. Detection of joint space narrowing in hand radiographs. Medical Imaging: Image Processing 2006;6:144-9.
7. 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.
8. 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 2002;113 Suppl:37-8.
9. Sharp JT, Angwin J, Boers M, et al. Computer based methods for measurement of joint space width: Update of an ongoing OMERACT project. J Rheumatol 2007;34:874-83.
10. Lukas C, Sharp JT, Angwin J, et al. Automated measurement of joint space width in small joints of patients with rheumatoid arthritis. J Rheumatol 2008;35:1288-93.
11. Baron G, Ravaud P, Samson A, Giradeau B. Missing data in randomized controlled trials of rheumatoid arthritis with radiographic outcomes: A simulation study. Arthritis Rheum 2008;59:25-31.