Author: Lukas, Cédric
Title: Methodological aspects of outcome assessment in inflammatory rheumatic diseases
Date: 2012-05-31

Chapter 2

Automated measurement of joint space width in small joints of patients with rheumatoid arthritis.

Published in Lukas C, Sharp JT, Angwin J, Boers M, Duryea J, Hall JR, Kauffman JA, Landewé R, Langs G, Bernelot Moens HJ, Peloschek P, and van der Heijde D.

J Rheumatol 2008;35(7):1288-93.

ABSTRACT

Aim: Comparison of performances of 5 (semi)automated methods in measuring joint space width (JSW) in rheumatoid arthritis.

Methods: Change in JSW was determined by 5 methods of JSW measurement on 4 radiographs per patient from 107 patients included in the COBRA trial (comparing sulfasalazine alone or in combination with methotrexate and corticosteroids). For each method the number of patients with sufficient available results was assessed (efficiency). An independent repeated measurement was carried out on a random sample of 30 patients' baseline and 1 year- radiographs, to evaluate within-method reliability of change scores. Discriminatory ability (DA) of the measurement methods (between the 2 treatment arms) was compared, with the DA of the Sharp-van der Heijde score (SHS) and its two components (erosion- and JSW scores).

Results: The overall succes rate varied widely between methods. Applying the prior-chosen threshold of a minimum of 50% available joints with a change score per patient resulted in a succes rate >92% in 4/5 methods. Repeatibility of measurements was good for most methods (intraclass corelation coefficient ≥0.80 in 4/5 methods). Almost all measurements methods in 3 follow-up periods (12/14) showed a lower mean loss of JSW in patients from the intensive treatment group, although rarely statistically significant, confirming the known difference in structural damage. The JSW as measured by the (semi)automated systems often showed higher DA than the JSW score of SHS, but lower than the total SHS and erosion scores.

Conclusion: Although efficiency of the methods should be further improved, results already show good reliability and encouraging DA of most methods. Optimal information may be obtained with a combination of scoring of erosions and (semi)automated measurement of JSW.

INTRODUCTION

Rheumatoid arthritis (RA) is a chronic inflammatory disease leading to progressive destruction of joint structures, and thus functional disability. Besides inflammation (synovitis, acute phase reactants, pain), a major outcome of clinical trials in RA is radiographic progression. However, the current scoring methods, although widely applied, have several limitations such as limited generalizability due to the difficulty of standardizing scoring by different readers. The use of an ordinal scale is also a theoretical limitation to measurement accuracy, which could be improved by an assessment of damage on a continuous metric scale, provided that the latter is reproducible. In this context, a subcommittee within the OMERACT imaging committee was formed after OMERACT 6 to test reliability, sensitivity to change and feasibility of computer-based methods for measuring radiographic damage in the small joints of the hands and feet in patients with RA, and priority was given to assessment of joint space width [1]. Preliminary exercises with small numbers of patients were conducted and results, presented at OMERACT 7, were considered sufficiently encouraging to proceed with further studies including larger sets of images. In the present exercise, a metrological comparison between most currently available methods for (semi)automatically measuring joint space width was performed using principles outlined by the OMERACT filter. Radiographs from a randomized controlled trial with proven efficacy on radiographic outcomes of an intensive therapeutic intervention in early RA (COBRA [2]) were digitized and used to evaluate the different systems.

Automated measuring systems

Main characteristics of five computer-based measurement systems of joint space width that were evaluated in this exercise are summarized in Table 1. The tested methods differ with regard to many aspects: the number of assessed joints across the five systems ranges from 8 (only including the 4 metacarpophalangeal joints (MCPs) of each hand for the least comprehensive system) to 34 (including 4 MCPs, 4 proximal interphalangeal joints (PIPs), 4 wrist joints and 5 metatarsophalangeal joints (MTPs) of each hand and foot for the most comprehensive systems. The workflow of

the programs is also diverse: the localization of the joints on the digitized radiographs can be performed by an algorithm or they have to be cropped, rotated, or centered for the measurement procedure by a technician. The same actually applies for the measurement process itself, which can be performed completely automatically by algorithms or with variable amounts of user input (e.g. pre-segmentation of contours). Finally, the absolute values obtained can not be directly compared across methods, because they represent different entities (shortest or averaged distance, with different defined margins of the joint etc.).

Table 1. Main characteristics of the five methods.

Method	Assessed joints on each side	n joints	Joint location	Region of measurement	Measurement	Method for calculation of mean joint space width
A [5, 6]	MCP 2-5 PIP 2-5 MTP 1-5 4 Wrist joints	34	Manual: 3 points on each MCP, MTP, wrist joint	MCP: radian drawn from the perceived center of the arc PIP: full breadth of the joint	Automated	distance between proximal and distal joint margins on vertical lines
B [7, 8]	MCP 2-5 PIP 2-5 MTP 2-5	24	Automated		Automated	distance between proximal and distal joint margins on vertical lines
C [9, 10]	MCP 2-5 PIP 2-5 MTP 1-5	26	Manual in COBRA set: 1 point in PIP 2-5 1 point in MCP 2-5	Fixed number of mm (dependent on joint type) centered within joint	Automated	Averaged shortest distance at multiple locations
D [4, 11]	MCP 2-5	8	Automated	Central part of the joint	Automated	Averaged shortest distance at multiple locations
E [1, 12]	MCP 2-5 PIP 2-5 MTP 1-5 4 Wrist joints	34	Manual: 1 marker on medial and lateral margins of joints	60% of joint span, from medial to lateral sides	Automated	Averaged shortest distance at multiple locations

MCP: metacarpophalangeal joints; MTP: metatarsophalangeal joints; PIP: proximal interphalangeal joints; wrist joints: capitatum-naviculare; capitatum-lunatum; radius-naviculare; radius-lunatum.

PATIENTS AND METHODS

Radiographic films of wrists, hands and feet from 107 patients included in the COBRA trial [2] of whom all time points were available (baseline, 6, 12 and 18 months) were digitized at 20 pixels per millimeter (50 micron pixel size), at 8-bit gray scale. Digitization was performed centrally, and a copy of the resulting batch of pictures was sent to the developers of five methods for (semi)automated measurement of joint space width. This clinical trial compared radiographic outcomes of patients with early RA who had been randomly allocated to one of 2 treatment regimens (sulfasalazine alone *vs.* combination therapy of methotrexate, sulfasalazine and temporary high dose of corticosteroids).

Treatment arm, patient identity and time points were randomized and blinded to the person who applied the automated method. Because every method measures joint space width differently, we focused on change over time. Change between baseline and a follow-up time point in a single joint is called change in a joint-pair. For the feasibility aspect we first determined the percentage of joints in which measurements were successful. Thereafter, we calculated the proportion of patients that could be measured based on different thresholds – 10%, 20%, 30% etc of successfully measured joints. Intra-method reliability (reproducibility) of measurement was evaluated by calculation of intra-class correlation coefficient (ICC) (2-way mixed, absolute agreement) based on two independent readings of baseline- and 12 months- radiographs in a random sample of 30 patients, and comparing the change in measurement over this time period. Because a correct interpretation of an ICC ("relative agreement") would require comprehensive description of data distributions, we also present results of "absolute agreement" by means of smallest detectable change (SDC) values [3]. The discriminatory ability of the methods was assessed by a paired t-test comparing mean change of joint space width over 3 time periods (baseline-6 months, baseline-12 months and baseline-18 months) per treatment arm. Only patients with more than 50% evaluable joint-pairs as defined above were included in this latter analysis. Radiographs of month 18 were not measured by method A, due to time constraints. Because the different methods measured a different kind and number of small joints of hands and feet, which may theoretically jeopardise a fair comparison of the five systems, the discriminatory ability was also calculated when taking into account only the second to fifth MCP joints, because these were assessed by all methods. This discriminatory ability was also compared with the original scoring method which was applied in the COBRA trial (Sharp-van der Heijde method, mean score of 2 independent readers), with additional separate comparisons of erosion- and joint space narrowing scores. All analyses were carried out using SPSS 12 and Excel 2002.

RESULTS

Feasibility of the five methods is presented in Figure 1. If only 50% measurable joint-pairs per patient was required, all methods except method D provided

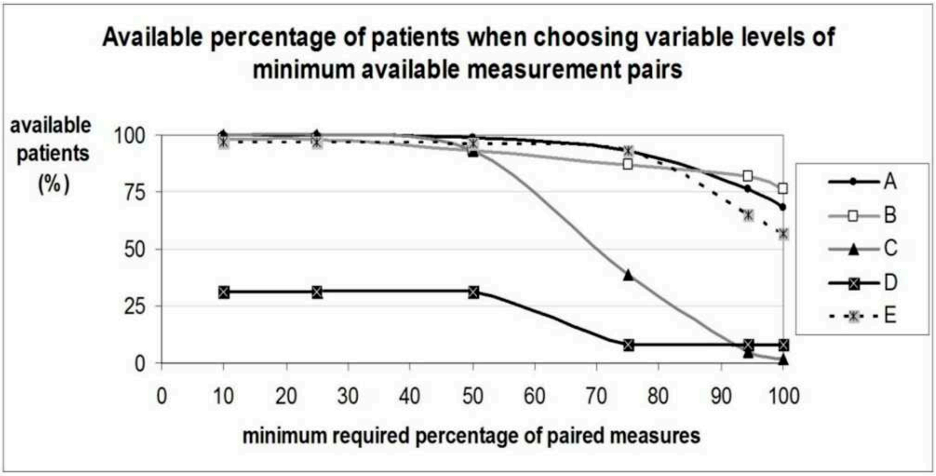

Figure 1. Efficiency of the five methods (available proportion of patients *-Y axis-* depending on various required successful reading rates *-Xaxis-* for baseline-to-12 months time-period). Calculations are based on intentionally measured joints, as described by the respective methods.

evaluable data. If a higher requirement (75% evaluable joint-pairs per patient) was applied, only 3 methods (A,B and E) provided evaluable data. The yield of methods A, B and E only gradually decreased by increasing requirements, but the yield of method C dropped sharply if more than 50% of evaluable joint pairs was required. Method C seemed to have a pattern of randomly missing joints that could not be evaluated (e.g. third right MCP, second left PIP etc.), occurring in many patients, whereas methods A, B, and E were mainly unable to read an entire radiograph but in a lower frequency.

Table 2. Intra-method reliability (independent repeated measurements) in assessing change in joint space over the baseline to 12 months- period on two independent readings in a random sample of 30 patients, per method (compared values are mean change per patient, with patients with less than 50% joints available being excluded).

		Based on all measured joints			Based on separated joint groups (MCPs, PIPs, MTPs, Wrists)		
		Valid cases	ICC	SDC†	Range of valid cases	Range of ICCs	Range of SDCs†
Method	A	100%	0.98	21%	100%	0.96 - 0.98	20% - 30%
	B	100%	0.96	41%	100%	0.78 – 0.97	22% - 43%
	C	96.7%	0.80	57%	42.4% - 96.7%	0.60 – 0.80	41% - 79%
	D	23.3%	0.81	59%	23.3%*	0.81*	59%*
	E	96.7%	0.41	71%	86.7% - 96.7%	0.24 – 0.54	51% - 78%

†expressed as percentage of 99^{th} observed percentile in values of change in the respective method; *only MCPs are measured in method D

ICC: Intraclass Correlation Coefficient; SDC: Smallest Detectable Change

Intra-reader reliability (repeatability) of the methods is summarised in Table 2. In general, repeatability was very high for most methods: Four of the 5 methods achieved an ICC over 0.8 (considered the cut-off for "good reliability"). Interestingly, when data were analysed based on joint groups (eg. only MCP joints), reliability of the measures appeared to be highly dependent on the region that was measured. In particular, the PIP joints (when measured) consistently resulted in the worst reliability for each of the methods, while MCP and MTP joints achieved most reliable measurements. The SDCs were expressed as the percentage of the 99^{th} observed percentile in values of change in the respective method in order to have a comparable measure across methods. The SDCs ranged from 21% to 71%, with most of the methods around 50%.

If comparison is limited to the MCP joint measurements, the discriminatory ability of the five methods (the difference in change in MCP joint space width between treatment groups) is shown in tables 3a, 3b and 3c for all three evaluated time intervals. Only patients with at least 50% of evaluable joint-pairs were included in these analyses. This implies that a direct comparison of discriminatory ability across methods can not be made, because different patients and joints were used depending on the various reading efficiencies. Again, one can get an impression of accuracy of each method in assessing the MCPs by comparing the values given in second column of the tables: compared to the other methods, method D had a consistently lower number of evaluable patients for all time periods. Although no significant difference was found when focusing on MCPs only, the great majority of the comparisons (12/14) showed a treatment effect in the "expected direction", that is reflecting a lower loss of joint space width in the patients from the intensively treated group, as compared to the monotherapy group. On the other hand, when using all available data (and thus all measured joints for each method), results dramatically improved in methods measuring a substantial number of joints (methods A, B, C and E), especially for the baseline to 6 months time-period (Tables 4a, 4b and 4c). Additionally, the more comprehensive systems were able to better discriminate between treatments than the "default" Sharp-van der Heijde score for change in joint space narrowing using the average of 2 independent readers' scores. The erosion score and the aggregated erosion and joint space narrowing score of the SvdH method, however, outperformed the (semi)automated methods with respect to discrimination.

Table 3. Discriminatory ability of the five methods in measuring mean change of joint space width on **last 4 metacarpo-phalangeal joints** of both hands over the evaluated time intervals.

3a. baseline-to-6 months time interval

Method	n*	Mean change (SD)		t-test	p-value
		monotherapy	COBRA		
A	105	-0.059 (0.095)	-0.045 (0.104)	-0.693	0.490
B	104	-0.055 (0.110)	-0.035 (0.122)	-0.884	0.379
C	80	-0.036 (0.078)	-0.013 (0.075)	-1.326	0.189
D	35	-0.072 (0.135)	-0.062 (0.146)	-0.198	0.844
E	102	-0.086 (0.272)	-0.087 (0.313)	-0.008	0.993

SD: standard deviation *n refers to the number of available patients for analysis with a 50%-available joint-pairs threshold

3b. baseline-to-12 months time interval

Method	n*	Mean change (SD)		t-test	p-value
		monotherapy	COBRA		
A	106	-0.074 (0.120)	-0.051 (0.124)	-1.003	0.369
B	98	-0.064 (0.127)	-0.035 (0.127)	-1.136	0.259
C	82	-0.034 (0.087)	-0.031 (0.082)	-0.185	0.853
D	33	-0.021 (0.283)	-0.061 (0.165)	-0.507	0.616
E	104	-0.107 (0.308)	-0.060 (0.403)	-0.657	0.512

SD: standard deviation *n refers to the number of available patients for analysis with a 50%-available joint-pairs threshold

3c. baseline-to-18 months time interval

Method	n*	Mean change (SD)		t-test	p-value
		monotherapy	COBRA		
A	N/A	N/A	N/A	N/A	N/A
B	103	-0.067 (0.133)	-0.055 (0.132)	-0.445	0.657
C	80	-0.079 (0.132)	-0.041 (0.102)	-1.451	0.151
D	37	-0.082 (0.215)	-0.042 (0.199)	-0.579	0.566
E	104	-0.143 (0.349)	-0.060 (0.368)	-1.164	0.247

SD: standard deviation *n refers to the number of available patients for analysis with a 50%-available joint-pairs threshold

Table 4. Discriminatory ability of the five methods in measuring mean change of joint space width on **all measured joints** over the evaluated time intervals.

4a. baseline-to-6 months time interval

Method	n*	Mean change (standard deviation)		t-test	p-value
		monotherapy	COBRA		
A	105	-0.068 (0.084)	-0.031 (0.073)	-2.466	**0.015**
B	104	-0.062 (0.077)	-0.011 (0.096)	-2.959	**0.004**
C	98	-0.024 (0.041)	0.001 (0.067)	-2.251	**0.027**
D	35	-0.072 (0.135)	-0.062 (0.146)	-0.198	0.844
E	102	-0.076 (0.148)	-0.028 (0.148)	-1.639	0.104
Erosion Score	107	5.130 (5.405)	2.105 (3.323)	3.429	**0.001**
Joint space narrowing score	**107**	**1.580 (3.208)**	**0.947 (1.929)**	**1.253**	0.213
Total Score	107	6.710 (7.147)	3.053 (4.460)	3.124	**0.002**

4b. baseline-to-12 months time interval

Method	n*	Mean change (standard deviation)		t-test	p-value
		monotherapy	COBRA		
A	107	-0.088 (0.110)	-0.054 (0.094)	-1.743	0.084
B	99	-0.035 (0.168)	-0.038 (0.091)	0.094	0.925
C	99	-0.024 (0.054)	-0.012 (0.065)	-0.927	0.356
D	33	-0.021 (0.283)	-0.061 (0.165)	-0.507	0.616
E	104	-0.096 (0.151)	-0.029 (0.152)	-2.239	**0.027**
Erosion Score	107	7.878 (8.381)	4.061 (5.925)	2.666	**0.009**
Joint space narrowing score	**107**	**3.122 (5.234)**	**2.605 (4.682)**	**0.537**	0.592
Total Score	107	11.000 (11.618)	6.667 (9.493)	2.113	**0.037**

4c. baseline-to-18 months time interval

Method	n*	Mean change (standard deviation)		t-test	p-value
		monotherapy	COBRA		
A	N/A	N/A	N/A	N/A	N/A
B	103	-0.071 (0.100)	-0.032 (0.157)	-1.493	0.139
C	100	-0.037 (0.075)	-0.029 (0.078)	-0.521	0.603
D	37	-0.082 (0.215)	-0.042 (0.199)	-0.579	0.566
E	104	-0.111 (0.203)	-0.054 (0.175)	-1.548	0.125
Erosion Score	107	10.918 (11.125)	6.439 (8.569)	2.294	**0.024**
Joint space narrowing score	**107**	**4.827 (6.905)**	**4.026 (6.780)**	**0.601**	0.549
Total Score	107	15.745 (16.144)	10.465 (14.321)	1.784	0.077

*n refers to the number of available patients for analysis with a 50%-available joint-pairs threshold.

DISCUSSION

Applying available automated methods for measuring joint space width on a complete set of radiographs from a clinical trial in RA allowed us to compare "real-life" performance of a set of computer-based tools in terms of feasibility, effectiveness, reliability and discriminatory ability.

A single set of radiographs was used to evaluate performances of the automated methods of joint space measurement, and thus a direct comparison of the different methods was possible for some aspects as long as the missing values were limited (especially when focusing on classical metrological features such as reliability and discrimination). Many aspects of the different systems are definitely distinct. There are important differences between the methods with regard to the time needed to measure a radiograph, which is due to different software, and differences in the time needed for technician's interventions. While the fully automated systems virtually instantaneously localise and measure the joint of interest, the semi-automated methods require a substantial amount of operator's time: an approximation of the total time required (to obtain measurements in one patient, i.e. from radiographies of 2 hands and 2 feet when applicable), was 15, 5, 2.5, 7 and 22 minutes for methods A, B, C, D and E respectively. The operator's intervention also introduces a possibly important source of systematic error in semi-automated methods. For example if the margins of a joint are judged and marked in different ways from one operator to another or by one operator making 2 measurements at different times, this may compromise reliability and generalizability. As a consequence, the interpretation of results is operator-dependent to some extent, which is similar to the scoring systems such as the Sharp-van der Heijde score.

The SDCs help to interpret how important a small difference in ICC (e.g. 0.96 vs. 0.98) is when an absolute reliability cut-off is derived from the same data (e.g. the SDC dropping from 41% of the observed maximal change to 21%). The SDC is based on the limits of agreement between two measurements for a certain method and gives a cut-off level of the method above which a change can be seen as a real change beyond measurement error.

Clearly, the efficiency of the methods in providing successful measurements for all joints that were presented for measurement needs to be improved before the computer-based methods can be more widely used. Excluding too many patients from the radiographic analysis of a clinical trial due to unevaluable joints per patient is unacceptable, in that it can cause both a loss of important information and may yield potentially biased results (systematic selection of joints). Even if we applied a very lenient cut-off level for evaluability (more than 50% of joints with available paired measurements) a surprisingly low proportion of patients remained in the analysis with one of the methods, and a somewhat more stringent requirement also would have hindered a second system. As a comparison, the proportion of missing scores that was observed in the same trial for joint space narrowing score (Sharp-van der Heijde score) at a joint level was very low (<1%), while the most effective automated method (B) had a total of 4.4% of PIP and MCP joints considered "not measurable" in our set of radiographs (wrist joints were not measurable by method B). For radiographs produced with an imaging protocol consistent with the training set however (i.e. beam geometry, hand positioning), 7.4% of MCP joints were "not measurable" with method D [4].

However, despite these apparently low performances, even a modest requirement of 50% of comparable joints over time resulted in a consistently accurate discriminatory ability of most methods. The seemingly disappointing p-values should be compared with those from the scoring of joint spaces, namely Sharp-van der Heijde scoring system: The discriminatory ability of the scoring method is largely determined by the erosion score, while the joint space narrowing score contributes in a limited manner. Although this phenomenon is not specific for the COBRA trial, further evaluation in other trials will provide more insight in the real merits of the measurement methods.

There are limitations to this study, as it is an ad hoc comparison of systems which are at different stages of pre-clinical development, and results should therefore be interpreted carefully. Preliminary exclusions of anatomical regions or of sets of images may stem from decisions in software engineering rather than intrinsic limitations of particular methods. An additional limitation that should be mentioned is that adjustments of existing algorithms of the automated methods to characteristics of the image set was almost completely precluded.

In view of these limitations, the remarkably better discrimination in most automated methods should be regarded as a promising sign of efficiency in accurately assessing change in joint space width over time. This superiority of the semi-automated methods over the scoring with regard to their discriminatory abilities might be a consequence of the use of a continuous scale in the computer-based measurements of joint space width, while the Sharp-van der Heijde method (like all scoring systems) apply an ordinal scale, the latter being inherently less sensitive to change. In addition, a fully automated method for assessing all radiographic abnormalities in RA will require not only joint space width, but also erosion measurement. However, although such projects are under development, there is still a long way to go before they become available as a routine, and the general reliability and sensitivity of scoring makes the task especially difficult. An important consideration in this regard is to determine from a clinical point of view what is the optimal trade-off between automated measurement and scoring. One could imagine that optimal discriminatory ability will be obtained with a combination of scoring erosions and automated measurement of joint space width by computer-driven techniques.

REFERENCES

1. Sharp JT, van der Heijde D, Angwin J et al. Measurement of joint space width and erosion size. J Rheumatol. 2005 Dec;32(12):2456-61.
2. Boers M, Verhoeven AC, Markusse HM et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet. 1997 Aug 2;350(9074):309-18.
3. Lukas C, Braun J, van der Heijde D et al. Scoring inflammatory activity of the spine by magnetic resonance imaging in ankylosing spondylitis: a multireader experiment. J Rheumatol. 2007 Apr;34(4):862-70.
4. Peloschek P, Langs G, Weber M et al. An automatic model-based system for joint space measurements on hand radiographs: initial experience. Radiology. 2007 Dec;245(3):855-62.
5. Angwin J, Heald G, Lloyd A et al. Reliability and sensitivity of joint space measurements in hand radiographs using computerized image analysis. J Rheumatol. 2001 Aug;28(8):1825-36.
6. Angwin J, Lloyd A, Heald G et al. Radiographic hand joint space width assessed by computer is a sensitive measure of change in early rheumatoid arthritis. J Rheumatol. 2004 Jun;31(6):1050-61.
7. Duryea J, Li J, Peterfy CG, Gordon C et al. Trainable rule-based algorithm for the measurement of joint space width in digital radiographic images of the knee. Med Phys. 2000 Mar;27(3):580-91.
8. Duryea J, Jiang Y, Zakharevich M et al. Neural network based algorithm to quantify joint space width in joints of the hand for arthritis assessment. Med Phys. 2000 May;27(5):1185-94.
9. Kauffman JA, Slump CH, Moens HJB. Segmentation of hand radiographs with multilevel connected active appearance models. Medical Imaging: Image Processing. 2005;5747-182 SPIE.
10. Kauffman JA, Slump CH, Moens HJB. Detection of joint space narrowing in hand radiographs. Medical Imaging: Image Processing. 2006;6144-149. 2006. SPIE.
11. Peloschek P, Bogl K, Robinson S et al. [Computer-assisted radiologic quantification of hand and foot changes in rheumatoid arthritis]. Wien Med Wochenschr Suppl. 2002(113):37-8.
12. 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 Jun;43(6):1378-86.