13

RELIABILITY, SENSITIVITY TO CHANGE AND FEASIBILITY OF THREE RADIOGRAPHIC SCORING METHODS FOR HAND OSTEOARTHRITIS

J. Bijsterbosch, I.K. Haugen, C. Malines, E. Maheu, F.R. Rosendaal, I. Watt, F. Berenbaum, T.K. Kvien, D.M. van der Heijde, T.W.J. Huizinga, M. Kloppenburg

Ann Rheum Dis 2011;70(8):1465-7

ABSTRACT

Objective. To compare the reliability, sensitivity to change and feasibility of three radiographic scoring methods for hand osteoarthritis (OA).

Methods. Baseline, 2-year and 6-year hand radiographs of 90 patients with hand OA were read in triplicate in chronological order by three readers from different European centres using the OARSI atlas (OARSI), Kellgren-Lawrence grading scale (KL) and Verbruggen-Veys anatomical phase score (VV). Reliability was determined using intraclass correlation coefficients and smallest detectable change (SDC). Sensitivity to change was assessed by the proportion of progression above the SDC. Feasibility was reflected by the mean performance time.

Results. Intra- and interreader reliability was similar across methods. Interreader SDCs (% maximum score) for KL, OARSI and VV were 2.9 (3.2), 4.1 (2.9) and 2.7 (1.8) over 2 years and 3.8 (4.1), 4.6 (3.3) and 4.0 (2.5) over 6 years. KL detected a slightly higher proportion of progression. There were differences between readers, despite methods to enhance consistency. The mean performance time (SD, minutes) for KL, OARSI and VV was 4.3 (2.5), 9.3 (6.0) and 2.8 (1.5), respectively.

Conclusion. Methods had comparable reliability and sensitivity to change. Global methods were fastest to perform. For multicentre trials using a central reading centre and multiple readers may minimise interreader variation.

INTRODUCTION

Despite the high prevalence and health impact of hand osteoarthritis (OA), no structure modifying treatments exist.[1,2] The development of these treatments implies the need for reliable and sensitive outcome measures.[3] Structural damage is considered a primary outcome, with serial radiographs as recommended outcome measure. Various radiographic scoring methods exist to assess severity and progression of structural damage.[4-10] They differ with respect to the number of hand joints scored, the use of a global score as opposed to grading of individual radiographic features, the radiographic features scored and the grading of features. There is no consensus on the preferred method, but owing to these differences the choice for a method may depend on the study objective.

Only one previous study has compared scoring methods for hand OA, which was over a relatively short period of 1 year.[11] In order to gain further insight in the clinimetric properties of available scoring methods, we assessed the reliability, sensitivity to change and feasibility of three radiographic scoring methods for the assessment of hand OA over a period of 2 and 6 years.

PATIENTS AND METHODS

Study design and patient population

Patients were participants of the Genetics ARthrosis and Progression (GARP) study comprising 192 Caucasian sibling pairs with symptomatic OA at multiple sites in the hand or in at least two of the following sites: hand, knee, hip or spine. Patients were evaluated at baseline and some of them after 2 and 6 years. Details on the recruitment and selection have been published elsewhere.[12] The study was approved by the medical ethics committee.

Patients were eligible for the present study if they had hand OA defined by the American College of Rheumatology criteria for clinical hand OA[13] or if structural abnormalities were present and if baseline, 2-year and 6-year radiographs were available. From this group a sample of 90 patients was included to ensure variability in baseline and progression scores based on a previous study.[14] See appendix 1 for more information on inclusion and sampling.

Radiographs and scoring methods

Standardised hand radiographs (dorsal-volar) were obtained at baseline and follow-up by a single radiographer.

With the Kellgren-Lawrence grading scale (KL)[6,10], a global score, the distal interphalangeal (DIP), proximal interphalangeal (PIP), interphalangeal thumb (IP-1), metacarpophalangeal (MCP) and first carpometacarpal (CMC-1) joints were graded 0-4 as described in the atlas (0=no OA; 1=doubtful OA; 2=definite minimal OA; 3=moderate OA; 4=severe OA). Total scores range from 0 to 120.

Using the OARSI atlas (OARSI)[4] individual radiographic features were graded. Osteophytes (0-3), joint space narrowing (JSN) (0-3), subchondral erosions (0-1), sclerosis (0-1) and malalignment (0-1) were assessed in the DIP, PIP, IP-1 and CMC-1

joints. Pseudowidening (0-1) was assessed in the DIP joints and cysts (0-1) were assessed in the PIP and CMC-1 joints. Total scores range from 0 to 198.

The Verbruggen-Veys anatomical phase score (VV)[9] comprises five phases with a numerical value representing the evolution of hand OA: N=normal joint; S=stationary OA with osteophytes and JSN; J=complete loss of joint space in the whole or part of the joint; E=subchondral erosion; R=remodelling of subchondral plate. The DIP, PIP, IP-1 and MCP joints were assessed. This score ranges from 0 to 218.4.

Reading procedures

Radiographs of all time points were read simultaneously in chronological order blinded for patient characteristics by three readers (JB, IKH, CM) from three European centres independently. Readers attended a training session before starting the study. A standard set of radiographs with scores was available for individual practice.

For assessment of intrareader reliability a random sample of 40 sets of radiographs was rescored with each method.

To randomise patients as well as methods a random number was assigned to each possible patient-scoring method combination, resulting in 390 combinations ((90 sets + 40 sets for intrareader reliability)x3 methods). To avoid mistakes and confusion because of frequent switching between methods, we grouped scoring methods per 10 sets of radiographs.

Statistical analysis

To evaluate intra- and interreader reliability for status scores, intraclass correlation coefficients (ICCs) were estimated. For change scores measurement error due to intrareader and interreader variability was assessed by estimating the smallest detectable change (SDC).[15] Sensitivity to change was assessed by the percentage of progression above the SDC. This analysis was done for all joints together and for separate joint groups (DIP/PIP, MCP and CMC-1 joints). Feasibility was determined by the mean scoring time of three time points for all readers together. The relationship between radiographic scores and performance time was assessed using linear regression analysis.

RESULTS

At baseline the mean age was 60.2 years and 70 patients (78%) were female. The observed status and change scores are shown in appendix 2. There were differences between readers, especially for change scores.

Intrareader and interreader ICCs for status scores were high with little difference between methods (table 1, appendix 2 for separate joint groups). For change scores the intrareader SDCs were good, with reader 3 showing higher SDCs than the other readers (table 2). Over both follow-up periods the method with the best reliability varied between readers. Interreader SDCs were lowest for VV, although differences from the other methods were small (table 2). Looking at separate reader pairs showed heterogeneity among readers with one reader scoring differently from the others (data not shown). Analysis in separate joint groups showed comparable results concerning comparison between methods (appendix 3).

Table 1. Reliability for status scores for the Kellgren-Lawrence grading scale (KL), OARSI atlas (OARSI) and Verbruggen-Veys anatomical phase score (VV) expressed by intraclass correlation coefficient (ICC).

	Reader	KL ICC (95%CI)	OARSI ICC (95%CI)	VV ICC (95%CI)
Intrareader				
Baseline*	1	0.96 (0.92 to 0.98)	0.97 (0.95 to 0.99)	0.97 (0.95 to 0.99)
	2	0.93 (0.87 to 0.96)	0.96 (0.92 to 0.97)	0.97 (0.94 to 0.98)
	3	0.90 (0.81 to 0.95)	0.77 (0.61 to 0.87)	0.88 (0.78 to 0.93)
Interreader				
Baseline*	1-2	0.91 (0.87 to 0.94)	0.95 (0.93 to 0.97)	0.95 (0.88 to 0.97)
	1-3	0.85 (0.76 to 0.90)	0.81 (0.56 to 0.90)	0.84 (0.56 to 0.93)
	2-3	0.84 (0.77 to 0.89)	0.80 (0.46 to 0.91)	0.81 (0.21 to 0.93)
	All	0.87 (0.82 to 0.91)	0.85 (0.71 to 0.91)	0.86 (0.66 to 0.93)

*ICCs for status scores at year 2 and 6 are very similar to those at baseline.

Table 2. Reliability for change scores and sensitivity to change assessed by the smallest detectable change (SDC) and percentage of patients with progression above the SDC for the Kellgren-Lawrence grading scale (KL), OARSI atlas (OARSI) and Verbruggen-Veys anatomical phase score (VV).

	KL		OARSI		VV	
	SDC (%)*	Progression, n (%)	SDC (%)	Progression, n (%)	SDC (%)	Progression, n (%)
Intrareader SDC and progression above this SDC						
2-Year						
Reader 1	2.1 (2.8)	17 (18.9)	1.2 (1.1)	20 (22.2)	1.4 (1.2)	17 (18.9)
Reader 2	2.5 (2.7)	22 (24.7)	3.0 (2.7)	16 (17.8)	3.4 (2.6)	9 (10.0)
Reader 3	7.1 (8.9)	11 (12.4)	10.2 (7.3)	11 (12.2)	7.8 (5.2)	14 (15.6)
6-Year						
Reader 1	3.7 (4.7)	45 (50.6)	3.0 (2.5)	50 (55.6)	3.5 (2.6)	24 (26.7)
Reader 2	4.4 (4.7)	51 (57.3)	4.8 (3.7)	54 (60.0)	6.3 (4.6)	19 (21.1)
Reader 3	8.1 (9.3)	41 (46.1)	11.1 (8.0)	32 (35.6)	9.9 (6.1)	31 (34.4)
Interreader SDC and progression above this SDC						
2-Year						
Reader 1	2.9 (3.2)	17 (18.9)	4.1 (2.9)	6 (6.7)	2.7 (1.8)	12 (13.3)
Reader 2		22 (24.7)		11 (12.2)		12 (13.3)
Reader 3		50 (56.2)		34 (37.8)		47 (52.2)
6-Year						
Reader 1	3.8 (4.1)	45 (50.6)	4.6 (3.3)	30 (33.3)	4.0 (2.5)	24 (26.7)
Reader 2		60 (67.4)		54 (60.0)		29 (32.2)
Reader 3		71 (79.8)		67 (74.4)		59 (65.6)

*SDC expressed as absolute value and as percentage of maximum observed score.

Based on the interreader SDC KL detected most progression (table 2). This was found for all three readers, although the percentages of progression varied between them. The results in the separate joint groups were similar (appendix 4).

The global scoring methods KL and especially VV were fastest to perform and scoring individual features with OARSI took more time (table 3). Each method took more time to perform in patients with higher levels of structural abnormalities.

DISCUSSION

This study on the reliability, sensitivity to change and feasibility of three radiographic scoring methods for hand OA shows minor differences between the methods. Reliability was high and sensitivity to change was good over both time periods, with slightly higher values for KL. There were differences in change scores and proportions of progression between readers, despite use of methods to enhance consistency. VV was the quickest method to perform.

To our knowledge, only one previous study has compared the clinimetric properties of radiographic scoring methods in hand OA, showing equal performance for reliability and sensitivity to change over 1 year.[11] Reliability was high in that study. Sensitivity to change expressed by standardised response means (SRMs) was low, whereas we found it to be good based on the SDC. Because different methods were used, meaningful comparison is difficult.

We used the SDC to assess reliability of change scores since it was more suitable than the ICC. The ICC is a measure of relative agreement reflecting signal-to-noise ratio. Therefore it is sensitive to relative subtle interreader discrepancies if the total range of scores is narrow, which was the case in this study.

We found that the global scoring methods VV and KL were faster to perform than OARSI. Recently it was shown that scoring osteophytes, JSN, malalignment and erosions may be sufficient to differentiate subjects with regard to disease severity.[16] This may improve the ease of use of OARSI.

There were differences between readers, despite a training session before starting the study, discussion sessions and use of atlases. The multicentre international study design might have contributed to this finding. The differences did not lead to

Table 3. Performance time for each set of three hand radiographs and the association between performance time and radiographic score for the Kellgren-Lawrence grading scale (KL), OARSI atlas (OARSI) and Verbruggen-Veys anatomical phase score (VV).

	KL	OARSI	VV
Performance time (minutes)			
Mean (SD)	4.3 (2.5)	9.3 (6.0)	2.8 (1.5)
Range	0.9-13.1	1.1-35.0	0.9-9.1
5^{th}-95^{th} Percentile	1.2-9.0	3.4-20.6	1.1-5.7
Association with radiographic score, β-coefficient (95%CI)*	3.9 (1.0 to 6.8)	8.0 (5.3 to 10.7)	21.1 (12.9 to 29.2)

*Number of points in radiographic score associated with one minute increment in performance time.

inconsistency in the comparison of methods. Clinical trials frequently involve multiple international centres, and the use of a central reading centre for radiographs therefore seems appropriate. The question remains: what is the true amount of structural abnormalities in OA? Experts in the field involved in this study scored a range of radiographic OA pathology together and concluded that it is very challenging to define a true score owing to variation in interpretation between readers. The use of quantitative measures, for instance measurement of joint space width, reduces interperson interpretation considerably. Using mean scores from multiple readers will on average be close to the "truth" and increase precision and generalisibility.

This study has a number of potential limitations. First, the level of radiographic abnormalities at baseline was relatively low compared with other samples from patients with hand OA. Although this has no effect on the comparison between methods, they may perform differently in other hand OA phenotypes. Second, we scored in chronological order. This may lead to overestimation of progression, but also to higher sensitivity to change. [17] Since potential overestimation will occur for all scoring methods it has no influence on the conclusions.

In conclusion, based on our findings it is not possible to recommend one of the scoring methods. Rather, based on the different character of the methods, the choice depends on the study objective. Further research on the validity of radiographic scoring methods as well as possibilities for their modification in order to enhance reliability, sensitivity to change and ease of use is warranted.

REFERENCES

1. van Saase JL, van Romunde LK, Cats A et al. Epidemiology of osteoarthritis: Zoetermeer survey. Comparison of radiological osteoarthritis in a Dutch population with that in 10 other populations. Ann Rheum Dis 1989;48:271-80.
2. Zhang Y, Niu J, Kelly-Hayes M et al. Prevalence of symptomatic hand osteoarthritis and its impact on functional status among the elderly: The Framingham Study. Am J Epidemiol 2002;156:1021-7.
3. Maheu E, Altman RD, Bloch DA et al. Design and conduct of clinical trials in patients with osteoarthritis of the hand: recommendations from a task force of the Osteoarthritis Research Society International. Osteoarthritis Cartilage 2006;14:303-22.
4. Altman RD, Gold GE. Atlas of individual radiographic features in osteoarthritis, revised. Osteoarthritis Cartilage 2007;15 Suppl A:A1-56.
5. Kallman DA, Wigley FM, Scott WW, Jr et al. New radiographic grading scales for osteoarthritis of the hand. Reliability for determining prevalence and progression. Arthritis Rheum 1989;32:1584-91.
6. Kellgren J. The Epidemiology of chronic rheumatism. Atlas of standard radiographs of arthritis. Oxford: Blackwell Scientific 1963:1-13.
7. Kessler S, Dieppe P, Fuchs J et al. Assessing the prevalence of hand osteoarthritis in epidemiological studies. The reliability of a radiological hand scale. Ann Rheum Dis 2000;59:289-92.
8. Lane NE, Nevitt MC, Genant HK et al. Reliability of new indices of radiographic osteoarthritis of the hand and hip and lumbar disc degeneration. J Rheumatol 1993;20:1911-8.
9. Verbruggen G, Veys EM. Numerical scoring systems for the anatomic evolution of osteoarthritis of the finger joints. Arthritis Rheum 1996;39:308-20.
10. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16:494-502.
11. Maheu E, Cadet C, Gueneugues S et al. Reproducibility and sensitivity to change of four scoring methods for the radiological assessment of osteoarthritis of the hand. Ann Rheum Dis 2007;66:464-9.
12. Riyazi N, Meulenbelt I, Kroon HM et al. Evidence for familial aggregation of hand, hip, and spine but not knee osteoarthritis in siblings with multiple joint involvement: the GARP study. Ann Rheum Dis 2005;64:438-43.
13. Altman R, Alarcon G, Appelrouth D et al. The American College of Rheumatology criteria for the classification and reporting of osteoarthritis of the hand. Arthritis Rheum 1990;33:1601-10.
14. Botha-Scheepers S, Riyazi N, Watt I et al. Progression of hand osteoarthritis over 2 years: a clinical and radiological follow-up study. Ann Rheum Dis 2009;68:1260-4.
15. Bruynesteyn K, Boers M, Kostense P et al. Deciding on progression of joint damage in paired films of individual patients: smallest detectable difference or change. Ann Rheum Dis 2005;64:179-82.
16. Haugen IK, Bijsterbosch J, Slatkowsky-Christensen B et al. The construct validity of a modified OARSI system in radiographic hand osteoarthritis using Rasch analysis. Ann Rheum Dis 2010;69:272.
17. Botha-Scheepers S, Watt I, Breedveld FC et al. Reading radiographs in pairs or in chronological order influences radiological progression in osteoarthritis. Rheumatology (Oxford) 2005;44:1452-5.

APPENDIX 1.

Inclusion criteria and sampling

Structural abnormalities were defined as the presence of radiographic hand OA based on a Kellgren-Lawrence score grade ≥2 in at least one interphalangeal (IP) or first carpometacarpal (CMC-1) joint, or the presence of ≥2 Heberden's or Bouchard's nodes on physical examination.

From the group of 102 eligible patients a sample of 90 patients was included to ensure variability in baseline and progression scores based on previous results from the GARP study on progression of hand OA over 2 years. Since progression rates were low, we included all patients with progression over this period (n=33). From the remaining group we included patients to ascertain maximal variability in Kellgren-Lawrence score at baseline; so both patients with low as well as high Kellgren-Lawrence baseline scores are represented.

APPENDIX 2.

Status and change scores for the Kellgren-Lawrence grading scale (KL), OARSI atlas (OARSI) and Verbruggen-Veys anatomical phase score (VV) in 90 hand osteoarthritis patients.

A. For the joints described in original method together

	KL (0-120)		OARSI (0-198)		VV (0-218.4)	
	Mean (SD)	**Range**	**Mean (SD)**	**Range**	**Mean (SD)**	**Range**
Reader 1						
Baseline	19.2 (12.9)	2 to 75	31.4 (17.5)	6 to 106	27.7 (18.4)	10.4 to 116.6
Year 2	20.4 (13.5)	2 to 75	32.3 (18.5)	6 to 108	28.7 (19.9)	11.6 to 119.1
Year 6	23.8 (15.1)	2 to 79	35.8 (21.4)	6 to 120	32.1 (25.0)	11.6 to 135.5
Change year 2	1.2 (2.1)	-2 to 9	1.0 (1.9)	-2 to 10	1.0 (2.3)	0 to 14.8
Change year 6	4.6 (4.3)	0 to 21	4.5 (5.9)	-4 to 35	4.5 (7.9)	0 to 32.9
Reader 2						
Baseline	20.4 (14.8)	4 to 90	32.6 (18.3)	6 to 113	30.8 (20.1)	5.8 to 129.2
Year 2	22.0 (15.4)	5 to 91	35.0 (20.0)	6 to 113	32.0 (21.4)	7.0 to 130.3
Year 6	26.4 (17.1)	6 to 93	40.0 (22.9)	6 to 129	35.7 (26.0)	8.1 to 136.9
Change year 2	1.6 (2.0)	-2 to 9	2.4 (3.5)	-1 to 21	1.1 (2.4)	0 to 16.8
Change year 6	6.0 (4.8)	-2 to 22	7.4 (6.7)	0 to 38	4.9 (8.2)	-1.2 to 37.3
Reader 3						
Baseline	21.7 (15.9)	2 to 77	24.2 (23.1)	0 to 125	20.4 (25.6)	0 to 138.6
Year 2	25.4 (17.1)	4 to 79	28.7 (25.4)	1 to 139	24.1 (25.9)	0 to 148.8
Year 6	30.8 (19.0)	6 to 87	35.1 (28.3)	1 to 139	29.4 (30.0)	1.2 to 162.6
Change year 2	3.6 (4.1)	-5 to 24	4.4 (6.1)	-5 to 40	3.7 (4.2)	-4.4 to 16.7
Change year 6	9.1 (6.2)	-1 to 28	10.9 (9.6)	-3 to 51	9.0 (8.9)	-2.3 to 39.2

B. For separate joint groups: DIP/PIP joints (KL, OARSI, VV), MCP joints (KL, OARSI), CMC-1 joints (KL, OARSI). VV is not included for MCP and CMC-1 joints since it is most frequently used for assessment of interphalangeal joints.

DIP/PIP joints

	KL (0-64)		OARSI (0-160)		VV (0-124.8)	
	Mean (SD)	**Range**	**Mean (SD)**	**Range**	**Mean (SD)**	**Range**
Reader 1						
Baseline	12.7 (10.4)	0 to 52	24.9 (15.2)	5 to 90	22.4 (15.6)	3.5 to 91.0
Year 2	13.5 (11.0)	0 to 55	25.6 (16.1)	5 to 91	23.3 (17.1)	3.5 to 102.7
Year 6	15.6 (12.5)	0 to 58	27.8 (18.8)	5 to 102	26.4 (22.2)	3.5 to 118.1
Change year 2	0.8 (1.7)	-2 to 9	0.7 (1.7)	-1 to 9	0.8 (2.3)	0 to 14.8
Change year 6	2.9 (3.4)	0 to 15	3.0 (5.2)	-4 to 30	4.0 (7.7)	0 to 30.9
Reader 2						
Baseline	12.7 (11.7)	0 to 57	25.6 (15.7)	3 to 95	23.6 (16.6)	3.5 to 97.5
Year 2	13.8 (12.4)	0 to 59	27.3 (17.0)	3 to 97	24.7 (18.1)	4.6 to 97.6
Year 6	16.8 (13.9)	0 to 62	31.3 (20.0)	3 to 107	27.9 (22.6)	4.6 to 108.9
Change year 2	1.0 (1.6)	-2 to 9	1.7 (3.3)	-6 to 20	1.0 (2.5)	0 to 16.8
Change year 6	4.1 (3.9)	-2 to 19	5.7 (6.2)	0 to 31	4.2 (7.9)	-1.2 to 37.3
Reader 3						
Baseline	14.8 (12.6)	0 to 58	19.1 (19.3)	0 to 102	17.5 (20.4)	0 to 110.4
Year 2	17.2 (13.7)	1 to 60	22.7 (21.6)	0 to 112	20.4 (22.4)	0 to 116.6
Year 6	20.8 (15.1)	2 to 64	27.3 (24.2)	0 to 112	24.6 (26.1)	0 to 121.2
Change year 2	2.4 (3.3)	-5 to 14	3.6 (5.4)	-6 to 30	2.9 (3.7)	-4.4 to 12.6
Change year 6	6.0 (4.8)	-2 to 19	8.2 (8.4)	-3 to 43	7.1 (8.0)	-1.9 to 34.8

MCP joints

	KL (0-40)		OARSI (0-70)	
	Mean (SD)	**Range**	**Mean (SD)**	**Range**
Reader 1				
Baseline	1.6 (2.5)	0 to 13	3.1 (4.3)	0 to 29
Year 2	1.7 (2.6)	0 to 13	3.2 (4.4)	0 to 29
Year 6	2.0 (2.8)	0 to 13	3.5 (4.7)	0 to 29
Change year 2	0.1 (0.3)	0 to 2	0.1 (0.3)	-1 to 2
Change year 6	0.4 (0.9)	-1 to 6	0.4 (1.0)	-1 to 6
Reader 2				
Baseline	2.2 (3.5)	0 to 23	4.8 (5.3)	0 to 33
Year 2	2.4 (3.6)	0 to 23	4.9 (5.3)	0 to 33
Year 6	2.8 (4.0)	0 to 23	5.5 (5.8)	0 to 33
Change year 2	0.3 (0.6)	-1 to 3	0.1 (0.6)	-2 to 2
Change year 6	0.6 (1.3)	-2 to 6	0.7 (1.7)	-2 to 7
Reader 3				
Baseline	1.6 (2.5)	0 to 12	1.6 (3.9)	0 to 26
Year 2	1.8 (2.9)	0 to 16	2.0 (4.4)	0 to 26
Year 6	2.4 (3.4)	0 to 18	2.7 (5.1)	0 to 29
Change year 2	0.2 (1.1)	-2 to 5	0.4 (1.4)	-3 to 8
Change year 6	0.8 (1.7)	-2 to 8	1.1 (2.1)	-2 to 10

CMC-1 joints

	KL (0-8)		OARSI (0-20)	
	Mean (SD)	**Range**	**Mean (SD)**	**Range**
Reader 1				
Baseline	2.7 (2.2)	0 to 8	4.0 (2.8)	0 to 14
Year 2	2.8 (2.2)	0 to 8	4.1 (2.9)	0 to 14
Year 6	3.3 (2.4)	0 to 8	4.8 (3.3)	0 to 14
Change year 2	0.1 (0.4)	0 to 2	0.1 (0.5)	-2 to 2
Change year 6	0.6 (0.8)	0 to 2	0.8 (1.3)	-1 to 6
Reader 2				
Baseline	3.3 (1.9)	0 to 8	4.7 (3.1)	0 to 16
Year 2	3.4 (1.9)	0 to 8	5.0 (3.1)	0 to 16
Year 6	3.8 (2.0)	0 to 8	5.6 (3.4)	0 to 17
Change year 2	0.2 (0.6)	-1 to 2	0.3 (0.6)	-1 to 2
Change year 6	0.6 (0.7)	-1 to 3	0.8 (1.3)	-2 to 5
Reader 3				
Baseline	3.4 (2.2)	0 to 8	3.2 (3.4)	0 to 16
Year 2	3.8 (2.3)	0 to 8	3.6 (3.8)	0 to 17
Year 6	4.3 (2.5)	0 to 8	4.6 (4.1)	0 to 16
Change year 2	0.4 (0.9)	-2 to 3	0.5 (1.6)	-3 to 9
Change year 6	0.8 (1.3)	-2 to 4	1.5 (2.1)	-2 to 9

APPENDIX 3.

Reliability for status scores for the Kellgren-Lawrence grading scale (KL), OARSI atlas (OARSI) and Verbruggen-Veys anatomical phase score (VV) expressed by intraclass correlation coefficient (ICC) for separate joint groups; DIP/PIP joints (KL, OARSI, VV), MCP joints (KL, OARSI), CMC-1 joints (KL, OARSI). VV is not included for MCP and CMC-1 joints since it is most frequently used for assessment of interphalangeal joints.

DIP/PIP joints

	Reader	KL ICC (95%CI)	OARSI ICC (95%CI)	VV ICC (95%CI)
Intrareader				
Baseline*	1	0.95 (0.91 to 0.97)	0.97 (0.94 to 0.98)	0.98 (0.96 to 0.99)
	2	0.92 (0.86 to 0.96)	0.96 (0.93 to 0.98)	0.97 (0.95 to 0.98)
	3	0.89 (0.80 to 0.94)	0.77 (0.61 to 0.87)	0.90 (0.82 to 0.94)
Interreader				
Baseline*	1-2	0.92 (0.89 to 0.95)	0.96 (0.94 to 0.97)	0.97 (0.95 to 0.98)
	1-3	0.85 (0.76 to 0.90)	0.83 (0.59 to 0.92)	0.85 (0.69 to 0.91)
	2-3	0.85 (0.77 to 0.91)	0.83 (0.52 to 0.92)	0.86 (0.59 to 0.93)
	All	0.87 (0.82 to 0.91)	0.86 (0.74 to 0.92)	0.88 (0.79 to 0.93)

*ICCs for status scores at year 2 and 6 are very similar to those at baseline

MCP joints

	Reader	KL ICC (95%CI)	OARSI ICC (95%CI)
Intrareader			
Baseline*	1	0.91 (0.84 to 0.95)	0.95 (0.90 to 0.97)
	2	0.84 (0.72 to 0.91)	0.88 (0.78 to 0.93)
	3	0.83 (0.69 to 0.90)	0.79 (0.64 to 0.88)
Interreader			
Baseline*	1-2	0.81 (0.72 to 0.87)	0.71 (0.51 to 0.82)
	1-3	0.71 (0.59 to 0.80)	0.70 (0.49 to 0.81)
	2-3	0.57 (0.41 to 0.69)	0.52 (0.11 to 0.74)
	All	0.70 (0.60 to 0.78)	0.63 (0.43 to 0.76)

*ICCs for status scores at 2 two and 6 are very similar to those at baseline.

CMC-1 joints

	Reader	KL ICC (95%CI)	OARSI ICC (95%CI)
Intrareader			
Baseline*	1	0.88 (0.78 to 0.93)	0.89 (0.80 to 0.94)
	2	0.85 (0.75 to 0.92)	0.86 (0.75 to 0.92)
	3	0.80 (0.64 to 0.89)	0.75 (0.58 to 0.86)
Interreader			
Baseline*	1-2	0.78 (0.63 to 0.87)	0.87 (0.75 to 0.92)
	1-3	0.68 (0.49 to 0.80)	0.71 (0.56 to 0.81)
	2-3	0.72 (0.61 to 0.81)	0.68 (0.36 to 0.83)
	All	0.73 (0.62 to 0.81)	0.75 (0.61 to 0.83)

*ICCs for status scores at year 2 and 6 are very similar to those at baseline.

APPENDIX 4.

Reliability for change scores and sensitivity to change assessed by the smallest detectable change (SDC) and percentage of patients with progression above the SDC for the Kellgren-Lawrence grading scale (KL), OARSI atlas (OARSI) and Verbruggen-Veys anatomical phase score (VV) for separate joint groups; DIP/PIP joints (KL, OARSI, VV), MCP joints (KL, OARSI), CMC-1 joints (KL, OARSI). VV is not included for MCP and CMC-1 joints since it is most frequently used for assessment of interphalangeal joints. Panel A. Intrareader SDC and progression above this SDC. Panel B. Interreader SDC and progression above this SDC.

DIP/PIP joints
A.

	KL		OARSI		VV	
	SDC (%)*	Progression, n (%)	SDC (%)	Progression, n (%)	SDC (%)	Progression, n (%)
2-Year						
Reader 1	1.7 (3.2)	17 (18.9)	1.0 (1.1)	25 (27.8)	1.3 (1.4)	13 (14.4)
Reader 2	1.8 (3.0)	25 (27.8)	1.9 (2.0)	32 (35.6)	2.4 (2.5)	14 (15.6)
Reader 3	5.2 (8.7)	15 (16.9)	7.8 (7.0)	14 (15.6)	6.7 (6.1)	15 (16.7)
6-Year						
Reader 1	2.0 (3.5)	40 (44.4)	2.4 (2.3)	30 (33.3)	3.3 (2.8)	22 (24.4)
Reader 2	4.0 (6.4)	32 (35.6)	3.6 (3.3)	49 (54.4)	4.0 (3.6)	24 (26.7)
Reader 3	6.3 (9.9)	31 (34.8)	7.9 (7.1)	35 (38.9)	9.4 (7.7)	26 (28.9)

*SDC expressed as absolute value and as percentage of maximum observed score.

B.

	KL		OARSI		VV	
	SDC (%)*	Progression, n (%)	SDC (%)	Progression, n (%)	SDC (%)	Progression, n (%)
2-Year						
Reader 1		9 (10.0)		6 (6.7)		12 (13.3)
Reader 2	2.3 (3.9)	12 (13.3)	3.6 (3.2)	14 (15.6)	2.4 (2.0)	14 (15.6)
Reader 3		37 (41.6)		32 (35.6)		39 (43.3)
6-Year						
Reader 1		40 (44.4)		20 (22.2)		22 (24.4)
Reader 2	2.7 (4.3)	55 (61.1)	3.9 (3.5)	49 (54.4)	3.5 (2.9)	24 (26.7)
Reader 3		67 (75.3)		59 (65.6)		48 (53.3)

*SDC expressed as absolute value and as percentage of maximum observed score.

MCP joints
A.

	KL		OARSI	
	SDC (%)*	**Progression, n (%)**	**SDC (%)**	**Progression, n (%)**
2-Year				
Reader 1	0.2 (1.7)	5 (5.6)	0.6 (2.0)	6 (6.7)
Reader 2	1.3 (5.8)	5 (5.6)	0.9 (2.8)	3 (3.3)
Reader 3	2.7 (16.9)	4 (4.5)	2.8 (11.0)	6 (6.7)
6-Year				
Reader 1	0.8 (6.1)	20 (22.2)	1.0 (3.6)	19 (21.1)
Reader 2	1.8 (7.7)	22 (12.2)	1.5 (4.6)	16 (17.8)
Reader 3	2.3 (12.8)	10 (11.2)	4.0 (13.9)	7 (7.8)

*SDC expressed as absolute value and as percentage of maximum observed score.

B.

	KL		OARSI	
	SDC (%)*	**Progression, n (%)**	**SDC (%)**	**Progression, n (%)**
2-Year				
Reader 1		5 (5.6)		6 (6.7)
Reader 2	0.9 (2.1)	19 (21.1)	0.9 (2.8)	16 (17.8)
Reader 3		21 (23.6)		19 (21.1)
6-Year				
Reader 1		8 (8.9)		8 (8.9)
Reader 2	1.3 (3.1)	11 (12.2)	1.4 (4.1)	16 (17.8)
Reader 3		18 (20.2)		22 (24.4)

*SDC expressed as absolute value and as percentage of maximum observed score.

CMC-1 joints
A.

	KL		OARSI	
	SDC (%)*	Progression, n (%)	SDC (%)	Progression, n (%)
2-Year				
Reader 1	0.8 (9.4)	6 (6.7)	0.6 (4.1)	11 (12.2)
Reader 2	0.8 (9.5)	18 (20.0)	1.2 (7.5)	4 (4.4)
Reader 3	1.5 (19.2)	12 (13.5)	3.1 (18.0)	4 (4.4)
6-Year				
Reader 1	0.9 (11.7)	35 (38.9)	1.5 (10.5)	21 (23.3)
Reader 2	1.0 (12.9)	41 (45.6)	1.7 (10.1)	24 (26.7)
Reader 3	2.2 (28.0)	8 (9.0)	4.0 (25.1)	16 (17.8)

*SDC expressed as absolute value and as percentage of maximum observed score.

B.

	KL		OARSI	
	SDC (%)*	Progression, n (%)	SDC (%)	Progression, n (%)
2-Year				
Reader 1		6 (6.7)		2 (2.2)
Reader 2	0.7 (8.7)	18 (20.0)	1.8 (6.9)	4 (4.4)
Reader 3		27 (30.3)		14 (15.6)
6-Year				
Reader 1		35 (38.9)		21 (23.3)
Reader 2	0.9 (11.4)	41 (45.6)	1.5 (8.8)	24 (26.7)
Reader 3		45 (50.6)		38 (42.2)

*SDC expressed as absolute value and as percentage of maximum observed score.