RMD Open
Rheumatic & Musculoskeletal Diseases

ORIGINAL ARTICLE

Multireader assessment as an alternative to reference assessment to improve the detection of radiographic progression in a large longitudinal cohort of rheumatoid arthritis (ESPOIR)

Frederique Gandjbakhch,[1,2] Benjamin Granger,[2,3] Romain Freund,[1,2] Violaine Foltz,[1,2] Sandrine Jousse-Joulin,[4] Valerie Devauchelle,[4] Mona Afshar,[5] Jean David Albert,[6] Florian Bailly,[1,2] Elodie Constant,[7] Lisa Biale,[8] Morgane Milin,[4] Marion Couderc,[9] Delphine Denarie,[7] Anne Fradin,[10] Virginie Martaille,[11] Audrey Pierreisnard,[1,2] Nicolas Poursac,[12] Alain Saraux,[4] Bruno Fautrel[1,2]

To cite: Gandjbakhch F, Granger B, Freund R, *et al*. Multireader assessment as an alternative to reference assessment to improve the detection of radiographic progression in a large longitudinal cohort of rheumatoid arthritis (ESPOIR). *RMD Open* 2017;**3**:e000343. doi:10.1136/rmdopen-2016-000343

► Prepublication history and additional material is available. To view please visit the journal (http://dx.doi.org/10.1136/rmdopen-2016-000343).

FG, BG and RF performed a similar amount of work and are considered co-first authors.

Received 27 July 2016
Revised 29 November 2016
Accepted 30 November 2016

For numbered affiliations see end of article.

Correspondence to
Dr Frederique Gandjbakhch; frederique.gandjbakhch@psl.aphp.fr

ABSTRACT

Introduction: Structural damage progression is a major outcome in rheumatoid arthritis (RA). Its evaluation and follow-up in trials should involve radiographic scoring by 1 or 2 readers (reference assessment), which is challenging in large longitudinal cohorts with multiple assessments.

Objectives: To compare the reproducibility of multireader and reference assessment to improve the feasibility of detecting radiographic progression in a large cohort of patients with early arthritis (ESPOIR).

Methods: We used 3 sessions to train 12 rheumatologists in radiographic scoring by the van der Heijde-modified Sharp score (SHS). Multireader scoring was based on 10 trained-reader assessments, each reader scoring a random sample of 1/5 of all available radiographs (for double scoring for each X-ray set) for patients included in the ESPOIR cohort with complete radiographic data at M0 and M60. Reference scoring was performed by 2 experienced readers. Scoring was performed blindly to clinical data, with radiographs in chronological order. We compared multireader and reference assessments by intraclass correlation coefficients (ICCs) for SHS and significant radiographic progression (SRP).

Results: The intrareader and inter-reader reproducibility for trained assessors increased during the training sessions (ICC 0.79 to 0.94 and 0.76 to 0.92), respectively. For the 524 patients included, agreement between multireader and reference assessment of SHS progression between M0 and M60 and SRP assessment were almost perfect, ICC (0.88 (95% CI 0.82 to 0.93)) and (0.99 (95% CI 0.99 to 0.99)), respectively.

Conclusions: Multireader assessment of radiographic structural damage progression is comparable to reference assessment and could be used to improve the feasibility of radiographic scoring in large longitudinal cohort with numerous X-ray evaluations.

Key messages

What is already known about this subject?

- Structural damage progression is a major outcome in rheumatoid arthritis.
- Its evaluation in trials is time-consuming and challenging in large longitudinal cohorts with multiple assessments.

What does this study add?

- After training, multireader assessment of radiographic structural damage progression is comparable to reference assessment.
- Multireader assessment can improve the feasibility of radiographic scoring in large longitudinal cohort with numerous X-ray evaluations.

INTRODUCTION

Rheumatoid arthritis (RA) is a long-lasting autoimmune disorder marked by synovial membrane inflammation that can cause joint destruction after a few years,[1–3] thereby impairing quality of life and causing disability.[4] Structural damage progression in RA is one major outcome; therefore, the evaluation and follow-up of structural damage progression are internationally recommended.[5]

Plain radiographs of the hands and feet are considered the gold standard to assess

structural damage progression.[5 6] Erosions and joint space narrowing (JSN) are the two typical radiographic lesions found in RA. The most frequently used contemporary scoring system is the Sharp score modified by van der Heijde (SHS),[3 7] one of two reference methods used in most RA clinical trials and longitudinal observational studies. The SHS method evaluates, in each hand, 16 areas for erosions and 15 areas for JSN, and, in each foot, 6 areas for erosions and 6 areas for JSN. The erosion score per hand joint can range from 0 to 5. JSN and joint subluxation or luxation are combined in a single score, from 0 to 4. The maximal score for erosion and JSN are 160 and 120, respectively, for the hands and 120 and 48, respectively, for the feet. The maximal total SHS is 448.

Reproducibility and sensitivity to change are important characteristics in scoring methods. Studies that evaluated the reproducibility and sensitivity to change of the Sharp, Larsen and SHS methods[8–10] found that the SHS method had the best sensitivity to change and very good reproducibility improved by reader training.[11]

To improve the reproducibility and sensitivity to change in trials and observational studies, a methodological consensus has been developed for radiographic scoring and assessment of RA-related joint damage progression. According to this consensus, progression of radiological joint damage is usually based on the simultaneous assessment of a series of X-rays for each patient by one or two readers, who are blinded to clinical data, with known order of radiographs.[12] This consensus is challenging in terms of feasibility in large observational cohorts including a large number of patients and multiple times for assessment because of substantial burden or workload in scoring several hundred hand and foot X-ray sets. For example, in the large longitudinal cohort of early arthritis (ESPOIR), 813 patients were followed during 5 years, for 4065 X-ray sets produced. Using a reference assessment and considering that at least 20 min[13] is needed to interpret one X-ray set, one reader would have to score for 1355 hours (8 hours/day for 170 days). Multireader assessment might be more feasible in detecting radiographic progression in cohorts including a large number of patients and multiple assessment times by dividing the significant workload of radiographic scoring. More readers would facilitate the assessment of structural damage progression but could also imply risk of increased reading error and reduced reproducibility.

The objective of this study was to compare the reproducibility of a multireader and usual reference assessment to possibly improve the feasibility of detecting radiographic progression in a large cohort of patients with early arthritis (ESPOIR).

MATERIALS AND METHODS

Study population

The French Society for Rheumatology initiated a large, national, multicentre, longitudinal, prospective registry known as the ESPOIR cohort of early arthritis.[14] The protocol of the study was approved in July 2002 by the Ethics Committee of Montpellier University (no. 020307). All patients gave their signed informed consent to be included in the study.

All radiographic data, used for reader training and multireader assessment, were from the ESPOIR cohort. Briefly, patients were recruited if they had a clinical diagnosis of definitive or probable RA or undifferentiated arthritis with potential to progress to RA. The inclusion criteria were age 18–70 years, swelling in at least two joints for ≥6 weeks and <6 months, no history of disease-modifying antirheumatic drug therapy, and no history of glucocorticoid therapy. Patients were excluded if they had other clearly defined inflammatory rheumatic or connective tissue disease or early arthritis with no potential to progress to RA. Included patients underwent clinical and biological evaluation every 6 months for 2 years, then once a year for at least 10 years. Radiographs of hands and feet were taken each year from baseline (M0) to 5 years (M60), except M48. All patients of the ESPOIR cohort with complete radiographic data at M0 and M60 were included in the current study.

Reader selection

An information letter was sent to each supervisor of the investigation centres involved in the ESPOIR cohort and to each departmental head of rheumatology of university hospitals to inform them about the project and to propose including a co-worker in the study. The organisation committee selected readers by evaluating motivation letters and curriculum. Twelve hospital rheumatologists were selected to be trained in radiographic scoring and assessing RA-related joint damage progression by the SHS.

Reader training

Each of the 12 candidates followed a structured training. The training programme included a 2-day session involving theoretical and practical workshops on a standardised scoring methods, software used to score and principal difficulties and 'traps' in scoring. In order to standardise the readings, all readers received the same computer with large screen (iMAC). During the first day, readers were trained in scoring, with immediate correction by the trainers (X-ray sets A and B). These scorings were not used to evaluate reliability because the scoring was not performed individually. At the end of the second day, 30 X-ray sets corresponding to 30 patients with RA with different ages, severity and disease progression at two times, M0 and M12 (sets C and D), were given to candidates. Each candidate had to score sets C and D by the SHS method. After at least 48 hours from the first scoring, candidates scored the same sets once again for assessing intra-rater and inter-rater reliability for each radiographic set by calculating intraclass correlation coefficients (ICCs). The training was complete with sufficient intra-rater and inter-rater reliability (ie, ICC≥0.8). With ICC<0.8, new exercises were organised. Candidates

scored two other radiographic sets (sets E/F and G/H, of 30 and 25 patients, respectively, at two times) separated by training meetings to discuss significant discrepancies and difficulties in scoring.

Structural damage assessment of the cohort by multireader and usual reference scoring

To compare the agreement and reproducibility of multireader and reference assessment, plain radiographs were scored (by the SHS) using the same equipment than during the training for all patients of the ESPOIR cohort with complete radiographic data at M0 and M60, according to two different methods (reference or multireader assessment).

The reference assessment was used as a gold standard and according to recommendations. With blinding from clinical and biological data and with radiographs in chronological order, two experienced trained readers (MM and FB) scored all radiographs from baseline (M0) and M60 by the SHS. The patient score was calculated as the mean of the two scores evaluated by the two experienced readers.

The multireader assessment involving 10 trained readers (AF, MA, MC, LB, JDA, EC, DD, VM, AP, NP) was compared with the reference assessment. For this assessment, all patients included in the study were randomly divided into equal subgroups and their X-ray sets were randomly allocated to the 10 readers. Each X-ray set corresponding to one patient (ie, two radiographs of hands and feet at times M0 and M60) was scored according to the SHS by two different readers of the multireader group with blinding to clinical and biological data and with radiographs in chronological order.

Statistical analysis

Statistical analysis involved use of R Statistical software (V.3.2.0; R Foundation for Statistical Computing, Vienna, Austria). SHSs are presented as median (first quartile (Q1); third quartile (Q3)). ICCs calculated for intrareader and inter-reader reliability involved use of a generalised linear mixed model to measure variances. A bootstrap procedure with 500 replications was used to estimate 95% CIs. To evaluate training performance, the ICCs for intrareader and inter-reader reliability for each training session were calculated, as was an overall ICC taking into account all X-ray sets for patients and all training sessions. Different approaches were proposed to analyse multireader and reference readings. Agreement was evaluated for SHSs (SHS for each time point and ΔSHS corresponding to SHS change between M0 and M60) and for structural damage progression. Two definitions of structural damage progression corresponding to two different thresholds were used: ΔSHS-5 with SHS change between M0 and M60>5, and significant radiographic progression (SRP)[17] with SHS change between M0 and M60 greater than the smallest detectable change (SDC). The SDC is defined as $1.96\times SD_{CHANGE-SCORE}/(\sqrt{2}\times\sqrt{k})$, where k represents the number of readings.[12] Agreement and homogeneity between the multireader and reference assessments were evaluated by ICCs. No agreement was characterised as ICC<0 and slight agreement 0–0.20; fair 0.21–0.40; moderate 0.41–0.60; substantial 0.61–0.80 and almost perfect 0.81–1. Agreements between multireader and reference assessments for the SHS and ΔSHS and for progression (ΔSHS-5 and SRP) were evaluated by analysis of the agreement between multireader and reference assessments taking into account mean scores of the two readings for the multireader assessment compared with mean scores for the two readers of the reference assessment and an assessment of the homogeneity of scoring between readers from the multireader and reference groups. To assess the correlation of SHS between multireader and reference we calculated the Pearson correlation coefficient. Bland and Altman plots were used to visualise the agreement between multireader and reference assessments. Finally, homogeneity between readers within the multireader group was evaluated by calculating ICCs for SHS and ΔSHS and agreement for the score progression (ΔSHS-5 and SRP).

RESULTS

Patient characteristics in the ESPOIR cohort were previously published.[15] In total, 524 patients had radiographic data at M0 and M60 and were included in this study.

Reader training

After three training sessions, the intra-rater and inter-rater reliability increased considerably (from ICC 0.79 (95% CI 0.68 to 0.85) and 0.76 (0.65 to 0.84) to 0.94 (0.88 to 0.97) and 0.92 (0.81 to 0.96), respectively; table 1) and the objective of training (ICC>0.8) was achieved. The overall reproducibility (including all times for all training sessions: sets C/D, E/F and G/H) was excellent for both intra-rater evaluations (ICC 0.92 (0.87 to 0.93) and inter-rater evaluation (ICC 0.90 (0.84 to 0.93).

Multireader and reference scoring or assessment

The reference group, composed of two trained readers, scored 1048 sets of radiographs (524 patients, M0 and M60). The radiographs for these 524 patients were

Table 1 Evolution of inter-rater and intra-rater reliability during three sessions to train readers in evaluating radiographs

Reliability	First session	Second session	Third session
Inter-rater	0.76 (0.65 to 0.84)	0.91 (0.83 to 0.95)	0.92 (0.81 to 0.96)
Intra-rater	0.79 (0.68 to 0.85)	0.94 (0.88 to 0.97)	0.94 (0.88 to 0.97)

Data are intraclass correlation coefficient (95% CI).

divided and randomly allocated to 2 of the 10 trained readers of the multireader group, who scored sets of radiographs for M0 and M60. In total, 385 patients had full data for M0 and M60, which allowed for evaluating SHS for these two times.

Structural damage progression between M0 and M60

Among the 385 patients with full data at M0 and M60, for the reference group, the median SHS was 1 (Q1;Q3 0;3) at M0 and 3 (Q1;Q3 0.5;10.5) at M60 and median SHS change between M0 and M60 1.5 (Q1;Q3 0;7). Structural damage progression (ΔSHS-5) was observed for three patients in the multireader group and one patient in the reference group. In our cohort, the SDC between M0 and M60 was 11. Each method showed one patient with SRP. No patient with structural progression was identified by both methods, using ΔSHS-5 or SRP (table 2).

Agreement between multireader and reference assessments

For the SHS, we found good correlation between multireader and reference assessments (r=0.87, p<0.001; figure 1). The overall agreement between multireader and reference assessments was good (ICC 0.69 (95% CI 0.62 to 0.75)). Results were similar whichever the joint (ICC for hands and feet, 0.69 (0.62 to 0.75) and 0.65 (0.52 to 0.75), respectively) or the lesion assessed (ICC for JSN and erosion, 0.71 (0.65 to 0.77) and 0.65 (0.57 to 0.73), respectively; table 3).

The Bland and Altman plot showed the absence of systematic bias between the two scoring methods (mean difference=−0.0062, p=0.977; figure 2). We found a proportional negative bias showing that the agreement differed by the level of score (ie, agreement was less for patients with high SHS (slope=−0.2036, p<0.001)).

More interestingly, the agreement between multireader and reference assessments for SHS change between M0 and M60 were excellent (ΔSHS-5 ICC 0.99 (95% CI 0.95 to 0.99); SRP ICC 0.99 (0.99 to 0.99); table 2). These results were consistent whichever the location (hands or feet) or the elementary lesion assessed.

Homogeneity of scores between readers (multireader and reference groups)

Similar results were found when evaluating the homogeneity of the scores between all readers from the multireader and reference group: SHS for total score, erosion score and JSN score (ICC 0.67 (95% CI 0.63 to 0.72),

Table 2 Number of patients with structural damage progression (ΔSHS-5 and SRP) in the multireader group and in the reference group

		Multireader group	
Structural damage progression		**ΔSHS<5**	**ΔSHS≥5**
Reference group	ΔSHS<5	382	3
	ΔSHS≥5	1	0
		No SRP (ΔSHS<SDC)	**SRP (ΔSHS≥SDC)**
Reference group	No SRP (ΔSHS<SDC)	384	1
	SRP (ΔSHS≥SDC)	1	0

SHS, van der Heijde-modified Sharp score; ΔSHS, SHS change between month 0 (M0) and M60; ΔSHS-5, SHS change between M0 and M60>5; SRP, significant radiographic progression.

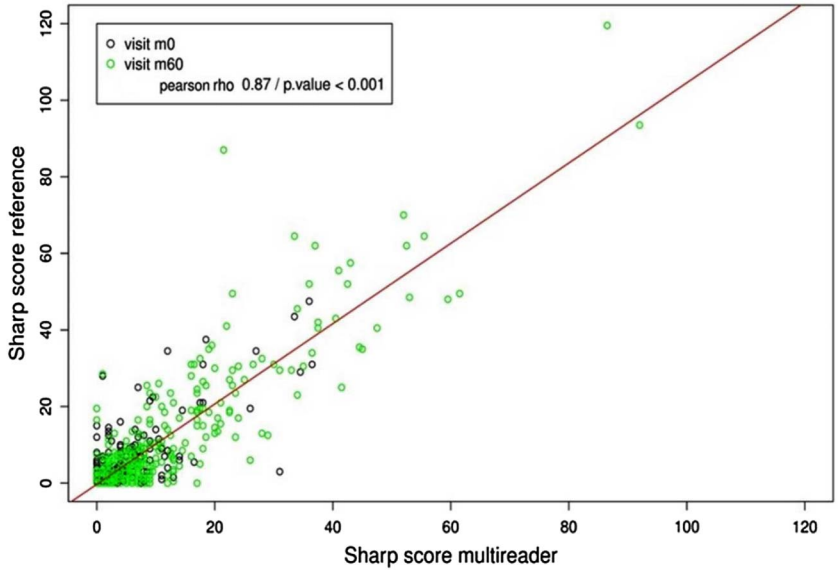

Figure 1 Pearson correlation of van der Heijde-modified Sharp score (SHS) between multireader and reference assessments (all patients, all times).

0.63 (0.59 to 0.67) and 0.69 (0.62 to 0.75), respectively), and for structural damage progression, ΔSHS, ΔSHS-5 and SRP (ICC 0.86 (0.79 to 0.89), 0.89 (0.84 to 0.98) and 0.99 (0.98 to 0.99), respectively; table 4).

Homogeneity between readers within the multireader group

The agreement was substantial for SHS (ICC 0.67 (95% CI 0.62 to 0.73)). Agreement was high for structural damage progression between readers within the multireader group (ICC 0.87 (0.79 to 0.92), 0.95 (0.84 to 0.99) and 0.96 (0.84 to 0.99) for ΔSHS, ΔSHS-5 and SRP, respectively).

DISCUSSION

Here, we aimed to improve the feasibility of use of X-ray assessment to detect structural damage progression in a large longitudinal RA cohort by comparing multireader assessment to the usual reference assessment. Multireader evaluation showed good reproducibility as compared with the reference method. The overall agreement between multireader and reference assessment was

Table 3 Agreement in SHSs between multireader and reference assessment of radiographs

Score	SHS	ΔSHS	ΔSHS-5	SRP
Total	0.69 (0.62 to 0.75)	0.88 (0.82 to 0.93)	0.99 (0.95 to 0.99)	0.99 (0.99 to 0.99)
Hands	0.69 (0.62 to 0.75)	0.91 (0.85 to 0.95)	0.99 (0.99 to 0.99)	0.99 (0.99 to 0.99)
Feet	0.65 (0.52 to 0.75)	0.84 (0.74 to 0.90)	0.99 (0.99 to 0.99)	0.99 (0.99 to 0.99)
Erosion	0.71 (0.65 to 0.77)	0.87 (0.82 to 0.90)	0.99 (0.99 to 0.99)	0.99 (0.99 to 0.99)
JSN	0.65 (0.57 to 0.73)	0.89 (0.84 to 0.93)	0.99 (0.99 to 0.99)	0.99 (0.99 to 0.99)

Data are ICC (95% CI).
ICC, intraclass correlation coefficient; JSN, joint space narrowing; SHS, van der Heijde-modified Sharp score; ΔSHS, SHS change between month 0 (M0) and M60; ΔSHS-5, SHS change between M0 and M60>5; SRP, significant radiographic progression.

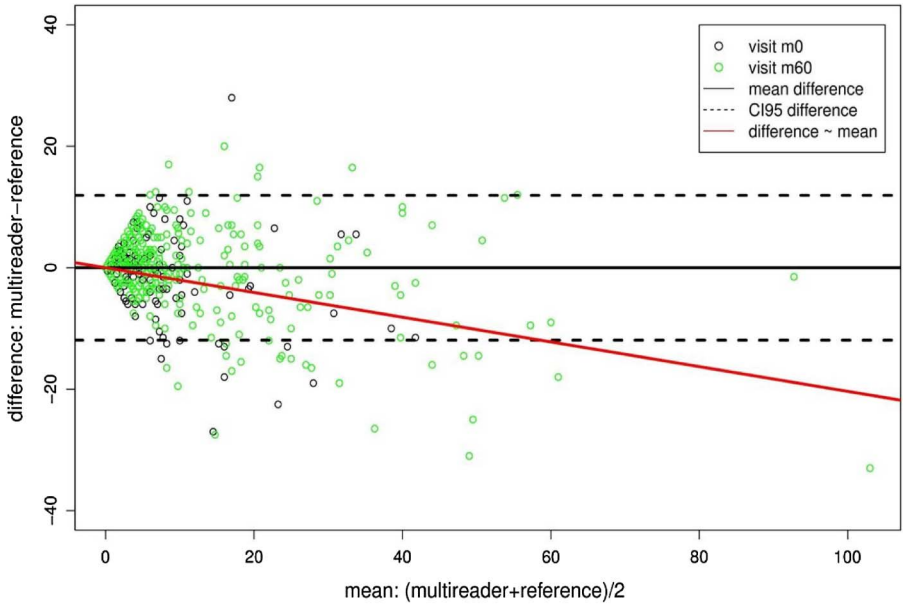

Figure 2 Bland and Altman plot of multireader and reference assessments.

Table 4 Homogeneity of SHSs between readers (multireader and reference assessments)

Score	SHS	ΔSHS	ΔSHS-5	SRP
Total	0.67 (0.63 to 0.72)	0.86 (0.79 to 0.89)	0.89 (0.84 to 0.98)	0.99 (0.98 to 0.99)
Hands	0.65 (0.59 to 0.69)	0.85 (0.77 to 0.90)	0.98 (0.98 to 0.99)	0.98 (0.95 to 0.99)
Feet	0.68 (0.59 to 0.75)	0.81 (0.72 to 0.88)	0.98 (0.97 to 0.99)	0.99 (0.95 to 0.99)
Erosion	0.63 (0.59 to 0.67)	0.76 (0.68 to 0.82)	0.96 (0.87 to 0.98)	0.98 (0.96 to 0.99)
JSN	0.69 (0.62 to 0.75)	0.85 (0.79 to 0.89)	0.98 (0.98 to 0.99)	0.96 (0.87 to 0.99)

Data are ICC (95% CI).
ICC, intraclass correlation coefficient; JSN, joint space narrowing; SHS, van der Heijde-modified Sharp score; ΔSHS, SHS change between month 0 (M0) and M60; ΔSHS-5, SHS change between M0 and M60>5; SRP, significant radiographic progression.

good. More interestingly, the agreement between these two methods was excellent for change in SHS between M0 and M60. These results suggest that structural damage progression can be evaluated with similar results whatever the reader method used. Multireader assessment presents the advantage of the greatest feasibility for a large cohort (because each reader has to score a reduced number of sets) and allows for detecting structural damage progression with similar results as with the usual reference method.

Our study allowed us to evaluate the training duration needed to obtain good reliability. After three training sessions, readers reached satisfactory reliability. A significant increase in reproducibility resulted in excellent ICCs for intrareader reliability (0.79 to 0.94) and inter-rater reliability (0.76 to 0.92) in our training group. Moreover, our results highlighted the rapidity of the training (only 2 days) to achieve almost perfect agreement for intra-rater and inter-rater reliability.[11]

Several studies evaluated the reproducibility of inter-rater reliability in radiographic evaluation in RA. This reproducibility depends on reader experience, number of readers, joint training of the readers, use of progression score or absolute score, and time of reproducibility evaluation during the follow-up of the patient.[9] The results of different studies evaluating inter-rater reliability in RA scoring are shown in the online supplementary table S1. These results highlight that reproducibility is never poor (<0.6) but can range from correct[16 17] and good[18 19] to excellent.[2 7 9 20–26] Of note, the reference statistic used to evaluate the reproducibility is the ICC. Only a few studies evaluated inter-rater reliability in radiographic evaluation in RA with >2 readers.[16 20 22] In those studies, the reproducibility was heterogeneous (from 0.58 to 0.97). In our study, the reproducibility of inter-rater reliability in radiographic evaluation was comparable to that from the Sharp *et al*[20] study.

To the best of our knowledge, this is the first study evaluating the feasibility of multireader assessment as an alternative to the time-consuming reference assessment in a large cohort of patients with RA. A study limitation is the detection of patients with structural progression with both methods due to the less number of patients with structural progression. Thus, our study should be replicated and validated in another population containing a higher number of patients with structural damage progression. Nevertheless, the overall agreement on the change in SHS was almost perfect.

In conclusion, our study highlighted the efficacy and rapidity of training a group of readers for radiographic scoring using the SHS in a large cohort of patients with RA. This method could be proposed as an alternative to monoreader evaluation to improve the feasibility of radiographic scoring in cohorts including a large number of patients and multiple time points. Further validation of multireader assessment of radiographic structural damage progression in RA is needed.

Author affiliations
[1]Department of Rheumatology, APHP, CHU Pitie-Salpetriere, Paris, France
[2]Paris 6 University, GRC-UPMC 08, Pierre Louis Institute of Epidemiology and Public Health, Paris, France
[3]Department of Statistics, CHU Pitie Salpetriere, APHP, Paris, France
[4]Department of Rheumatology, CHU la Cavale Blanche, Brest, France
[5]Department of Rheumatology, Hopital Jean Verdier, Bondy, France
[6]Department of Rheumatology, CHU Rennes, Rennes, France
[7]Department of Rheumatology, CHU Saint Etienne, Saint Etienne, France
[8]Department of Rheumatology, Hopital Begin, Saint Mande, France
[9]Department of Rheumatology, CHU Clermont Ferrand, Clermont Ferrand, France
[10]Department of Rheumatology, CHU Poitiers, Poitiers, France
[11]Department of Rheumatology, Hopital d'Orleans, Orleans, France
[12]Department of Rheumatology, CHU Bordeaux, Bordeaux, France

Twitter Follow Alain Saraux @alain.saraux

Acknowledgements The authors thank the steering committee: A Cantagrel, Toulouse; B Combe, Montpellier; M Dougados, Paris-Cochin; BF, Paris-Pitié; F Guillemin, Nancy; X Le Loet, Rouen; I Logeart, Paris; AS, Brest, J Sibilia, Strasbourg, P Ravaud, Paris-Bichat. Sixteen regional investigation centres: F Berenbaum, Paris- Saint Antoine; MC Boissier, Paris-Bobigny; A Cantagrel, Toulouse; B Combe, Montpellier; M Dougados, Paris-Cochin; P Fardelonne, P Boumier, Amiens; BF, P Bourgeois, Paris-La Pitié; RM Flipo, Lille; Ph Goupille, Tours; F Liote, Paris-Lariboisière; X Le Loet, O Vittecoq, Rouen; X Mariette, Paris Bicetre; O Meyer, Paris Bichat; AS, Brest; Th Schaeverbeke, Bordeaux; J Sibilia, Strasbourg. Coordination centre: JP Daures, Montpellier; N Rincheval, Montpellier; B Combe, Montpellier. X-ray centre: AS, Brest; VD—PENSEC, Brest; C Lukas, Montpellier. Biobank: J Benessiano, Paris-Bichat.

Funding This study was supported by Roche Pharma. (France).

Competing interests None declared.

Patient consent Obtained.

Ethics approval The study was authorised by the Ethics Committee of Paris Ile de France (CPP Ile de France II, no. 2008-07-04).

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Fex E, Jonsson K, Johnson U, *et al*. Development of radiographic damage during the first 5–6 yr of rheumatoid arthritis. A prospective follow-up study of a Swedish cohort. *Rheumatology* 1996;35:1106–15.
2. Sharp JT, Wolfe F, Mitchell DM, *et al*. The progression of erosion and joint space narrowing scores in rheumatoid arthritis during the first twenty-five years of disease. *Arthritis Rheum* 1991;34:660–8.
3. van der Heijde DM, van Leeuwen MA, van Riel PL, *et al*. Biannual radiographic assessments of hands and feet in a three-year prospective followup of patients with early rheumatoid arthritis. *Arthritis Rheum* 1992;35:26–34.
4. Fries JF, Spitz P, Kraines RG, *et al*. Measurement of patient outcome in arthritis. *Arthritis Rheum* 1980;23:137–45.
5. Colebatch AN, Edwards CJ, Østergaard M, *et al*. EULAR recommendations for the use of imaging of the joints in the clinical management of rheumatoid arthritis. *Ann Rheum Dis* 2013;72:804–14.
6. van der Heijde DM. Radiographic imaging: the "gold standard" for assessment of disease progression in rheumatoid arthritis. *Rheumatol Oxf Engl* 2000;39(Suppl 1):9–16.

7. van der Heijde DM, van Riel PL, Nuver-Zwart IH, *et al*. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. *Lancet* 1989;1:1036–8.
8. Plant MJ, Saklatvala J, Borg AA, *et al*. Measurement and prediction of radiological progression in early rheumatoid arthritis. *J Rheumatol* 1994;21:1808–13.
9. Cuchacovich M, Couret M, Peray P, *et al*. Precision of the Larsen and the Sharp methods of assessing radiologic change in patients with rheumatoid arthritis. *Arthritis Rheum* 1992;35:736–9.
10. van der Heijde DMFM. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. *Baillières Clin Rheumatol* 1996;10:435–53.
11. Guillemin F, Billot L, Boini S, *et al*. Reproducibility and sensitivity to change of 5 methods for scoring hand radiographic damage in patients with rheumatoid arthritis. *J Rheumatol* 2005;32:778–86.
12. 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.
13. Boini S, Guillemin F. Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. *Ann Rheum Dis* 2001;60:817–27.
14. Combe B, Benessiano J, Berenbaum F, *et al*. The ESPOIR cohort: a ten-year follow-up of early arthritis in France: methodology and baseline characteristics of the 813 included patients. *Joint Bone Spine* 2007;74:440–5.
15. Tobón G, Saraux A, Lukas C, *et al*. First-year radiographic progression as a predictor of further progression in early arthritis: results of a Large National French Cohort. *Arthritis Care Res* 2013;65:1907–15.
16. Fries JF, Bloch DA, Sharp JT, *et al*. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. *Arthritis Rheum* 1986;29:1–9.
17. Taylor PC, Steuer A, Gruber J, *et al*. Comparison of ultrasonographic assessment of synovitis and joint vascularity with radiographic evaluation in a randomized, placebo-controlled study of infliximab therapy in early rheumatoid arthritis. *Arthritis Rheum* 2004;50:1107–16.
18. Bathon JM, Martin RW, Fleischmann RM, *et al*. A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis. *N Engl J Med* 2000;343:1586–93.
19. Proudman SM, Conaghan PG, Richardson C, *et al*. Treatment of poor-prognosis early rheumatoid arthritis. A randomized study of treatment with methotrexate, cyclosporin A, and intraarticular corticosteroids compared with sulfasalazine alone. *Arthritis Rheum* 2000;43:1809–19.
20. Sharp JT, Bluhm GB, Brook A, *et al*. Reproducibility of multiple-observer scoring of radiologic abnormalities in the hands and wrists of patients with rheumatoid arthritis. *Arthritis Rheum* 1985;28:16–24.
21. van der Heijde DM, van Riel PL, van Leeuwen MA, *et al*. Prognostic factors for radiographic damage and physical disability in early rheumatoid arthritis. A prospective follow-up study of 147 patients. *Br J Rheumatol* 1992;31:519–25.
22. Swinkels HL, Laan RF, van 't Hof MA, *et al*. Modified sharp method: factors influencing reproducibility and variability. *Semin Arthritis Rheum* 2001;31:176–90.
23. 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;350:309–18.
24. de Jong Z, Munneke M, Zwinderman AH, *et al*. Is a long-term high-intensity exercise program effective and safe in patients with rheumatoid arthritis? Results of a randomized controlled trial. *Arthritis Rheum* 2003;48:2415–24.
25. Goekoop-Ruiterman YP, de Vries-Bouwstra JK, Allaart CF, *et al*. Clinical and radiographic outcomes of four different treatment strategies in patients with early rheumatoid arthritis (the BeSt study): a randomized, controlled trial. *Arthritis Rheum.* 2005;52:3381–90.
26. Klareskog L, van der Heijde D, de Jager JP, *et al*. Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial. *Lancet* 2004;363:675–81.