Universiteit Leiden
The Netherlands

Methodological aspects of outcome assessment in inflammatory rheumatic diseases

Lukas, C.

Citation

Lukas, C. (2012, May 31). *Methodological aspects of outcome assessment in inflammatory rheumatic diseases*. Retrieved from https://hdl.handle.net/1887/19033

Version: Corrected Publisher's Version

License: Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden

Downloaded from: https://hdl.handle.net/1887/19033

Note: To cite this publication please use the final published version (if applicable).

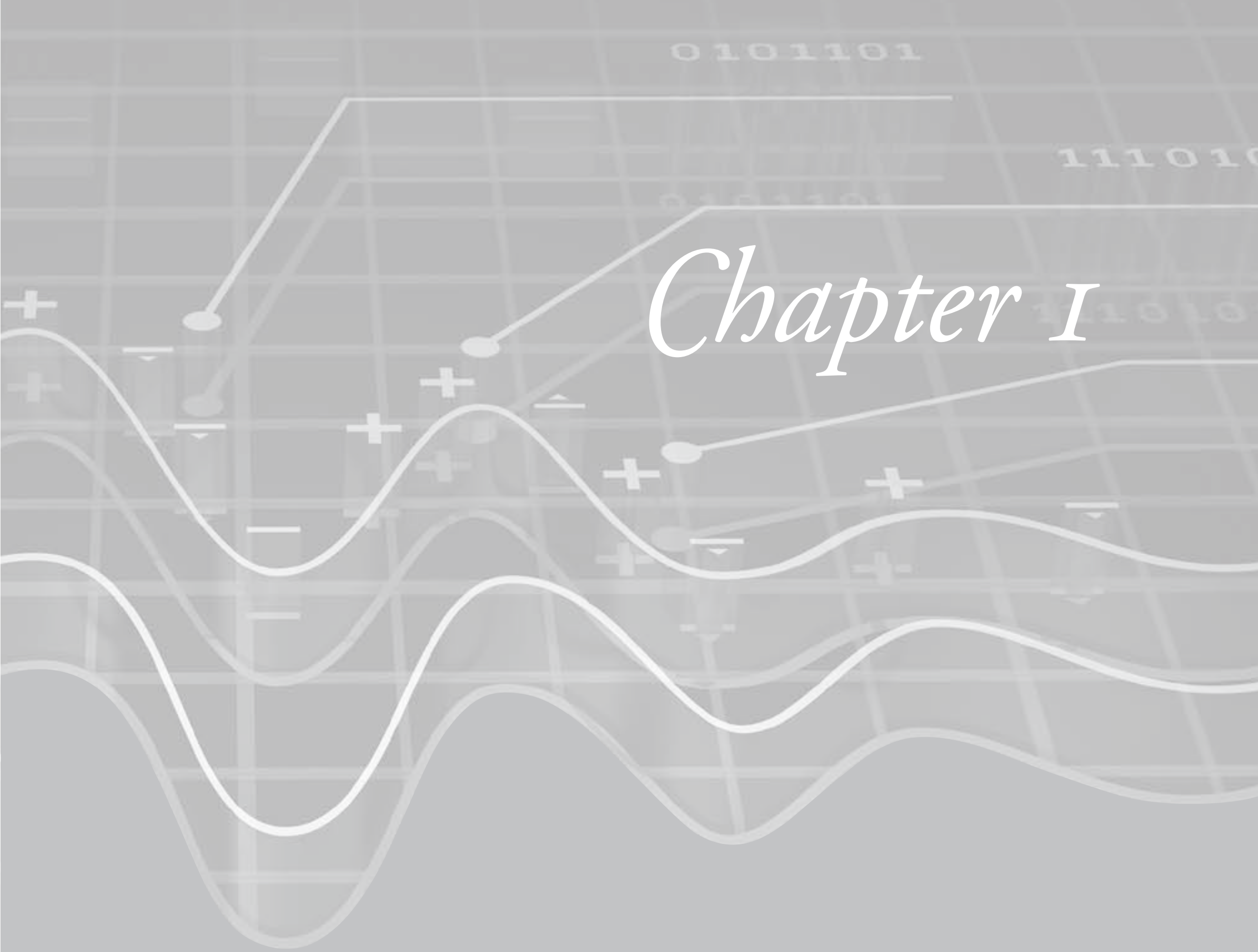
Cover Page

Universiteit Leiden

Leiden University Repository

The handle http://hdl.handle.net/1887/19033 holds various files of this Leiden University dissertation.

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



Chapter 1

Reliability and Sensitivity to change of the Simple Erosion Narrowing Score compared to the Sharp/van der Heijde method for scoring radiographs in rheumatoid arthritis

Published in Dias EM, Lukas C, Landewe RB, Fatenejad S,
van der Heijde DM.

Ann Rheum Dis 2008;67(3):375-9 Online First (2007 Jul 20)

ABSTRACT

Objective: To compare the performance of a simplified scoring method for structural damage on radiographs of patients with rheumatoid arthritis (RA) (the Simple Erosion Narrowing Score or SENS) with the Sharp-van der Heijde Score (SHS) as reference.

Method: We used the radiographic data from the Trial of Etanercept and Methotrexate with Radiographic Patient Outcomes (TEMPO trial). The SENS was derived from the crude SHS data that were available on a per-joint basis. Inter-observer reliability for status scores and change scores was determined by intra class correlation coefficients (ICC) and by the smallest detectable change (SDC) method. The ability to discriminate between treatment groups was assessed by the Mann-Whitney U test. Stratifying the sensitivity to change and discriminative ability for different levels of disease severity assessed a potential ceiling effect.

Results: Inter-observer reliability was similar for both methods. ICCs were higher for status scores than for change scores. The SDC was 4.98 for SHS and 2.28 for SENS. The sensitivity of the SENS to detect progression above the SDC, with reference SHS, ranged from 45.0 to 88.7 %. The specificity ranged from 81.5 to 97.3 %, and the Kappa coefficient (between-method agreement) ranged from 0.58 to 0.66. Discriminative ability between treatment groups was good and similar for both methods. A ceiling effect could not be detected.

Conclusions: The performance of SENS is as good as that of SHS. This confirms that SENS is a valuable and sufficiently validated method, which is feasible for use in clinical practice.

INTRODUCTION

Rheumatoid arthritis (RA) is an autoimmune mediated disease characterized by chronic inflammation of synovial joints, causing damage in cartilage and periarticular bone and subsequent destruction of these joints. Progression of joint damage as assessed on radiographic films is strongly associated with inflammatory activity and functional disability [1-3]. Radiographic damage and its progression can therefore be used as a measure of the severity of RA at a specific time, the course of RA and the responsiveness of RA to therapy for short or long-term duration [4].

Many scoring methods were designed for the assessment of radiographic damage. The most commonly used methods are the Sharp method, one of its modifications being the Sharp/van der Heijde scoring (SHS) method, and the Larsen method with modifications [4, 5]. All these methods require trained readers to obtain sufficient reliability [4, 6], making them difficult to use in clinical practice. Sharp and SHS methods are comprehensive (provide several types of information about each joint), an additional disadvantage for the use in clinical practice. Effective drugs for the treatment of RA that may stop radiographic progression necessitate the development of feasible methods that assess structural damage in clinical practice.

Previously, we have proposed and tested a simplification of the SHS method, the Simplified Erosion Narrowing Score (SENS) [7]. The SENS method is less time consuming and less comprehensive than the scoring methods described above, and is easier to teach and learn. It assesses erosions and joint space narrowing (JSN) in the same joints as SHS. But instead of grading erosions and JSN for severity, SENS only acknowledges the presence (1 point) or absence (zero points) of erosions and JSN separately.

We have reported previously that SENS was as reliable and sensitive –to change as SHS in a small set of patients with up to 6 years of disease duration. A potential shortcoming of SENS is that it detects only the first erosion or narrowing per joint; it does not take into account an increase in the numbers or the severity of erosions or JSN per joint. We have shown in the context of the Combinatietherapie Bij Reumatoide Artritis [Combination Therapy in Rheumatoid Arthritis] (COBRA) trial that indeed progression in eroded joints makes an important contribution to the progression score [8]. As such, SENS could potentially be prone to a ceiling effect. The objective of the present work was to test the performance of SENS in terms of reliability, sensitivity-to-change, and the ability to discriminate between treatment groups, in comparison with SHS in a large group of patients with a considerable variation in disease duration and severity of the disease. We also investigated whether a ceiling effect plays a role in the performance of SENS.

PATIENTS AND METHODS

Patients:

In this work we used the data from a double-blind, randomized, clinical trial (TEMPO-trial) [9] in which three treatment groups (methotrexate (MTX), etanercept (ETAN) and the combination of both (MTX+ETAN)) were compared for clinical and radiographic efficacy in 680 patients of the 682 included in the clinical trial (see below, *Readers* paragraph). The disease duration ranged from 14 weeks to 26.4 years (mean 6.6 years) in this subset.

Readers:

Radiographs of the hands, wrists and feet taken at baseline and 52-week follow-up were used for this analysis. The digitised radiographs were read in pairs with unknown sequence. Each set was scored twice by two readers from a pool of three trained readers, meaning that every reader scored two thirds of all the radiographs. Data from two readers were used in the analysis, which implies that a random third of the total patient population was available for interreader reliability analyses, and two-thirds of the total patient population for the comparison of SENS and SHS.

Radiographic scoring methods:

Structural damage as seen on radiographs was assessed by the SHS method. The SHS method [10] assesses joint erosions and JSN. Joint erosions are scored in 32 joints in the hands and wrists and 12 joints in the feet. Erosion scores per joint can range from 0 to 5 in the hands and wrists and from 0 to10 in the feet. JSN is scored in 30 joints in the hands and wrists and in 12 joints in the feet. JSN scores per joint range from 0 to 4 in hands, wrists and feet. The maximum total erosion score (the sum of all joint scores for erosion) is 280 and the maximum JSN score (the sum of all joint scores for JSN) is 168. The total score is the summed score of the total erosion and total JSN score, and has a maximum of 448.

The SENS method [7] assesses the same joints. It scores 1 point for each joint when showing at least one eroded location and also 1 point if JSN is present. In fact, the number of eroded joints and the number of narrowed joints are scored. Consequently the maximum total erosion score is 44, the maximum total JSN score is 42 and the maximum total score is 86.

For the purpose of this study, we derived the SENS data from the SHS. A score of 1 or more in SHS was substituted for a 1 in SENS. A 0 in SHS remained a 0 in SENS.

Statistical analysis:

Reliability

The inter-observer reliability was assessed by the intra class correlation coefficient (ICC, absolute agreement, two-way mixed model) per scoring method, for status scores at baseline and 52 weeks separately, and for change scores. Only joints scored by both readers and at both time points (i.e. excluding missing values) were used to calculate the total score per patient.

The Smallest Detectable Change (SDC) was applied as a second method of reliability [11]. The advantage of the SDC is that it is reflected in the units of the measurement scale. The SDC reflects the measurement error due to inter-observer variability. The SDC was calculated according to Bruynesteyn et al, as follows:

$$\mathrm{SDC}\ \frac{1.96\mathrm{xSDdiff}}{\sqrt{k}\ \mathrm{x}\ \sqrt{2}}$$

SDdiff is the standard deviation of the set of differences in change scores obtained by two readers, k is the number of readers whose change-scores are used (here: k=1 since the SDC on the scores of each reader separately was used) and the factor 1.96 represents the 95-percent limits of agreement according to Bland and Altman [12], according to which each change score laying within these limits of agreement is considered to be a consequence of measurement error.

Sensitivity to change

With SHS as a gold standard, we used the SDCs obtained with SHS and SENS as cut off values in order to determine if a change can be explained by measurement error alone, and to compare the sensitivity and specificity of the SENS method. In addition we used kappa statistics to determine the level of agreement with both dichotomized methods. Subsequently, we determined optimal cut off levels for SHS and SENS using Receiver Operating Characteristic (ROC) curves.

Because we were only interested in the performance of SENS in relation to SHS in detecting change we excluded the patients with no change according to the SHS method.

Cumulative probability plots [13] of status and change scores, per method and per reader were drawn to visualize the relationship between SHS and SENS and to detect possible individual outliers.

Discrimination between treatment groups

Discriminative ability for differentiating between treatment groups was assessed by calculating the Mann-Whitney U test comparing the change in total scores in the groups treated with MTX and with MTX+ETAN.

Ceiling effect

To assess a possible ceiling effect we determined the sensitivity –to change and discriminative ability by different levels of disease severity. This was achieved by stratifying the patient population in quartiles based on the increasing baseline radiographic SHS scores.

RESULTS

Table 1 describes the observed scores per reader, per scoring method for status and 52-week change scores. The patient population scored by reader 1 only partially overlapped the patient population scored by reader 2. Consequently, only comparison of SHS and SENS scores per reader but not comparison of absolute values between readers, is informative.

Figure 1 presents the scores of SHS and SENS expressed as percentage of the maximum possible score of the respective methods. The SHS scores are plotted from the lowest to the highest value against their cumulative probability. Each value of SHS corresponds to the SENS value of the same patient. The plot of the status score shows that for each case SENS is higher than SHS if expressed as the percentage of the maximum possible total score. The plot of the change scores visualizes the positive and negative changes. It also shows that a change in SENS is almost never negative when SHS change is positive, and vice versa.

TABLE 1. Observed values in total patient population, per scoring method for Total scores

Reader		Radiographic abnormality	SHS method		SENS method	
			Mean (SD)	Range	Mean (SD)	Range
1	**Status scores Baseline** N = 451	Erosion	14.8 (18.8)	0 – 182	7.3 (6.2)	0 – 33
		JSN	22.9 (28.4)	0 – 138	8.6 (9.9)	0 – 42
		Total score	37.7 (43.7)	0 – 320	15.9 (14.9)	0 – 75
	Status scores 52 weeks N = 358	Erosion	15.0 (19.7)	0 – 182	7.5 (6.1)	0 – 33
		JSN	24.4 (29.3)	0 – 138	9.0 (10.1)	0 – 42
		Total score	39.4 (45.3)	0 – 320	16.5 (15.0)	0 – 75
	52-week change scores N = 356	Erosion	-0.34 (3.8)	-14 – 33	-0.03 (1.8)	-10 – 9
		JSN	0.03 (3.0)	-30 – 28	-0.04 (1.2)	-11 – 13
		Total score	-0.32 (5.2)	-28 – 41	-0.08 (2.4)	-13 – 16
2	**Status scores Baseline** N = 458	Erosion	19.0 (31.6)	0 – 232	5.7 (7.2)	0 – 41
		JSN	20.2 (25.8)	0 – 148	9.2 (10.0)	0 – 40
		Total score	39.3 (54.6)	0 – 380	14.8 (16.3)	0 – 81
	Status scores 52 weeks N = 348	Erosion	19.5 (32.1)	0 – 232	5.8 (7.4)	0 – 41
		JSN	20.8 (26.5)	0 – 148	9.5 (10.2)	0 – 40
		Total score	40.3 (55.8)	0 – 380	15.4 (16.5)	0 – 81
	52-week change scores N = 347	Erosion	-0.15 (4.1)	-31 – 52	0.05 (1.1)	-7 – 12
		JSN	0.14 (3.4)	-20 – 30	-0.05 (1.2)	-10 – 10
		Total score	-0.01 (6.6)	-43 – 82	0.00 (1.9)	-10 – 17

SHS, Sharp/van der Heijde Score; SENS, Simplified Erosion Narrowing Score;
SD, standard deviation; JSN, joint space narrowing

Reliability:

The between-reader ICCs at baseline and 52 weeks were 0.81 and 0.77 for SHS and 0.91 and 0.89 for SENS, respectively. The ICCs for change were 0.30 for SHS and 0.22 for SENS. Probability plots with the scores of reader 1 and reader 2 plotted together showed no systematic differences (probability plot with the scores of reader 2 are not shown). The SDCs were calculated by using the set of overlapping data of reader 1 and 2 (n=181). The SDC for SHS was 4.98 units (which is 1.12% of the maximum score) and for SENS 2.28 units (which is 3.49% of the maximum score).

Sensitivity to change:

ROC-based optimal cut off levels were 2 units for SHS and 1 unit for SENS for the scores of reader 1, and 3 units for SHS and 1 unit for SENS for reader 2. Similarly, optimal cut-off levels for negative change scores were determined: They were –5 units for SHS and –1 unit for SENS for reader 1, and -2 units for SHS and –1 unit for SENS for reader 2.

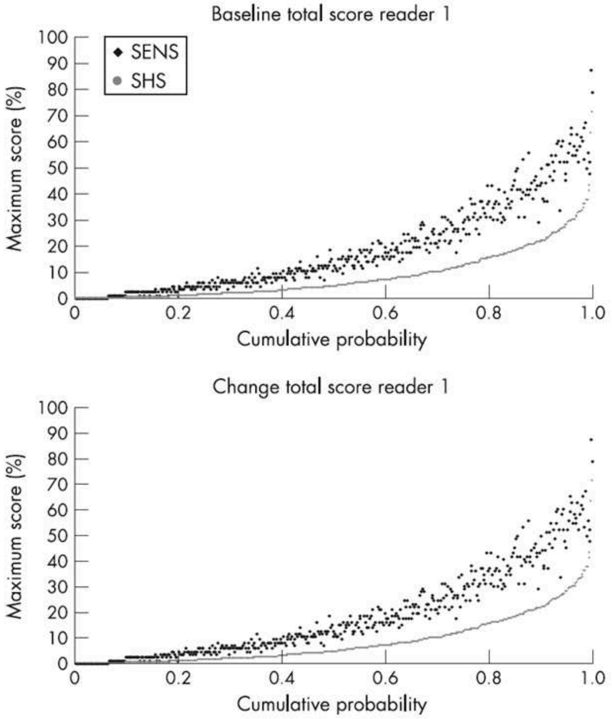


FIGURE 1. Probability plots of Baseline score and Change by SHS and SENS (SHS, Sharp van der Heijde Score; SENS, Simplified Erosion Narrowing Score)

Table 2 compares the performance of both methods with respect to showing change or no change, as based on the different cut off levels. The data show that for both readers, either based on the SDC or on the optimal ROC-based cut-off levels, the performance of the SHS and the SENS is approximately similar with respect to the proportion of patients with change.

TABLE 2. Number and percentage of patients with real or no change of Total score based on SDC and ROC based cut offs.

Reader (N pat)	Cut off level		SHS		SENS	
	SHS	SENS	Real change N(%)	No Change N(%)	Real change N(%)	No change N(%)
1 (n=251)	SDC = 4.98	SDC = 2.28	26 (10.4%)	225 (89.6%)	23 (9.2%)	228 (90.8%)
	2*	1*	62 (24.7%)	189 (75.3%)	90 (35.9%)	161 (64.1%)
2 n=(189)	SDC = 4.98	SDC = 2.28	20 (10.6%)	169 (89.4%)	10 (5.3%)	179 (94.7%)
	3*	1*	34 (18.0%)	155 (82.0%)	50 (26.5%)	139 (73.5%)

SHS, Sharp-van der Heijde Score; SENS, Simplified Erosion Narrowing Score; SDC, smallest detectable change; N, number of patients. * ROC based cut offs.

Table 3 presents the sensitivity, specificity and agreement of the SENS method compared with the SHS method (gold standard).

Table 3. Sensitivity to change in Total score for SENS compared to gold standard SHS based on SDC and ROC based cut offs.

Reader (N)	Cut-off level		Sensitivity %	Specificity %	Kappa
	SHS	SENS			
1 (n=251)	SDC = 4.98	SDC = 2.28	65.4	97.3	0.66
	2*	1*	88.7	81.5	0.61
2 (n=189)	SDC = 4.98	SDC = 2.28	45.0	99.4	0.57
	3*	1*	82.4	85.8	0.58

SHS, Sharp/van der Heijde Score; SENS, Simplified Erosion Narrowing Score;
SDC, smallest detectable change; N, number of patients. * ROC based reader specific cut offs.

Discrimination:

Both the SENS and the SHS total scores discriminated between the MTX-group and the MTX+ETAN group (P<0.001 for both methods performed by both readers). After stratification into quartiles based on SHS score at baseline, discrimination between MTX and combination therapy remained similar in all quartiles for both methods.

Ceiling effect

Table 4 presents the sensitivity, specificity and kappa-statistics of SENS calculated with the cut off levels based on ROC analysis. The sensitivity of

Table 4. Sensitivity to change in Total score based on the Cut-off levels from ROC, categorized by baseline score.

Reader	Reader-specific cut-off level		Quartile: (range of baseline scores in SHS units) (N patients)	Sensitivity (%)	Specificity (%)	Kappa
	SHS	SENS				
1	2	1	1: (0 - 10) (61)	100.0	65.0	0.56
			2: (11 - 29) (62)	100.0	77.1	0.60
			3: (30 - 67) (64)	83.3	84.6	0.57
			4: (68 - 284) (64)	66.7	95.9	0.67
2	3	1	1: (0 - 14) (47)	33.3	72.7	0.02
			2: (15 - 32) (47)	75.0	94.9	0.70
			3: (33 - 79) (47)	90.9	91.7	0.78
			4: (80 - 284) (48)	91.7	86.1	0.70

N, number of patients; SHS, Sharp van der Heijde Score;
SENS, Simplified Erosion Narrowing Score

SENS by reader 1 decreased with an increasing level of the baseline score, but the sensitivity of SENS by reader 2 showed the opposite trend. We saw the same difference in the sensitivity to change analysis between reader 1 and 2 for different levels based on increasing disease duration or increasing annual progression rate (data not shown).

DISCUSSION

This work was intended to evaluate the usefulness of the SENS method in accurately measuring radiographic damage and progression in patients with RA. In order to do so, we measured important "psychometric characteristics" such as inter-reader reliability, sensitivity to change and discriminatory ability, and compared with the actual standard of measuring radiographic damage and progression in clinical trials, the Sharp-van der Heijde method. This comparison aimed at confirming a comparable performance of the more feasible SENS in the evaluation of structural damage and progression. Indeed, SENS does not require specific training (recognition of usual abnormalities caused by RA is sufficient) and is far less time-consuming than other methods applied in clinical trials, which additionally require semi-quantitative evaluation of the extent of lesions. A routine application by every practitioner could consequently be conceivable in clinical practice, provided that a gain in feasibility is not at the cost of a loss in validity.

Firstly, the reliability of SENS, as tested between the two independent readers, was very good for status scores, and even better than the more comprehensive SHS method. The reliability of measuring change scores, however, was unexpectedly low for both methods. An artefact may be that patients studied in the present analysis took part in a clinical trial with very effective drugs and consequently a very low rate of progression, while the first report about the SENS method by van der Heijde et al. [7], as well as a later comparison of five scoring methods – including SHS and SENS – by Guillemin et al [14] included patients with substantial radiographic progression. Indeed, patients selected for the initial validation of the SENS had early but rapidly progressing RA, and were assessed after 5 or 10 years of follow-up. Guillemin et al selected radiographs from a survey conducted in the nineties, when drugs of major structural efficacy (such as the TNF-blockers) were not yet available. The ICC, which is a measure of relative agreement, is sensitive to relatively subtle inter-reader discrepancies if the total range of observed change scores is narrow, and becomes less sensitive to such discrepancies if the total range of change scores is wide [15]. So the low ICCs are probably reflecting the small range of change scores with many patients showing no change rather than real poor inter-reader reliability. This view is further supported by the fact that a completely independent reread of the data during the read of the second year TEMPO trial fully confirmed the results of the data of the first read, used in the present analysis [16].

Secondly, concerning sensitivity to change of SENS when compared with SHS, the results were also comparable across both methods: Taking a cut-off level that accounts for measurement error (SDC), a comparable proportion of patients were scored as "progressing" across both methods and readers, with an acceptable rate of concordance between SENS and SHS. This observation was not dependent on the value of the cut-off level for "true progression": Regardless of whether the cut-off level was based on inter-reader reliability (SDC method) or on statistical arguments (ROC-analysis), the performance of both methods was similar.

Thirdly, the discriminatory ability for discerning structural change in patients treated with methotrexate vs. methotrexate in combination with etanercept was high and comparable for both methods, even though the primary intention of SENS is its use in clinical practice rather than in comparative clinical trials.

Another important issue is the absence of a clear ceiling effect. Indeed, because the number of potentially affected joints that are scored is limited in RA, one could theoretically achieve the maximal value relatively early in the course of the disease, and the consequence of this would be a decreased sensitivity to change in patients with longer disease duration and a lot of structural damage already present. However, although no definite conclusion can be made from our data, the opposite trends that were observed for readers 1 and 2 when sensitivity to change was compared in different groups of patients based on the SHS score at baseline leads us to the conclusion that a systematic tendency is unlikely.

Potential limitations of this work are related to the origin of the database we used: Because the radiographs were obtained from patients included in a clinical trial comparing highly efficient drugs, during a short time period and scored by trained readers, several potential weaknesses may be raised. The database includes a high proportion of patients with no progression at all or low progression rates. This hinders the use of statistical analyses that are based on a Gaussian distribution, such as the ICCs, especially when change scores are evaluated. On the other hand, both methods were hindered by these unfavourable conditions, and it is the comparison between the methods in which we are really interested.

Another limitation of the results is that we derived the SENS-scores directly from the SHS scores. One could argue that this may increase the agreement between the two methods in an artificial manner. However, if we would have used an independent read for the SENS we would have ended up with data comparing SHS with SENS but it would not have been possible to disentangle variation caused by intrareader variation from differences between the methods. Information that is still needed is the repeatability of results if SENS is applied in practice by clinicians.

In summary, we were able to demonstrate in a large database that SENS was at least as reliable, sensitive to change and discriminatory as SHS, even in this context of a short-term clinical trial comparing very efficient drugs, and when the time sequence of the images scored were not known. The improved feasibility of the SENS method in comparison with the SHS method, together with its comparable psychometric properties, allows us to recommend using it to monitor radiographic progression in clinical practice.

REFERENCES

1. van der Heijde D. Radiographic progression in rheumatoid arthritis: does it reflect outcome? Does it reflect treatment? Ann Rheum Dis 2001;60 Suppl 3:iii47-50.
2. Landewe R, van der Heijde D. Radiographic progression in rheumatoid arthritis. Clin Exp Rheumatol 2005;23(5 Suppl 39):S63-8.
3. Odegard S, Landewe R, van der Heijde D et al. Association of early radiographic damage with impaired physical function in rheumatoid arthritis: a ten-year, longitudinal observational study in 238 patients. Arthritis Rheum 2006;54(1):68-75.
4. van der Heijde DM. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. Baillieres Clin Rheumatol 1996;10(3):435-53.
5. Boini S, Guillemin F. Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. Ann Rheum Dis 2001;60(9):817-27.
6. Fries JF, Bloch DA, Sharp JT et al. Assessment of radiologic progression in rheumatoid arthritis. A randomized, controlled trial. Arthritis Rheum 1986;29(1):1-9.
7. van der Heijde D, Dankert T, Nieman F et al. Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. Rheumatology (Oxford) 1999;38(10):941-7.
8. Bruynesteyn K, Van Der Heijde D, Boers M et al. Contribution of progression of erosive damage in previously eroded joints in early rheumatoid arthritis trials: COBRA trial as an example. Arthritis Rheum 2002;47(5):532-6.
9. 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(9410):675-81.
10. van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27(1):261-3.
11. 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(2):179-82.
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307-10.
13. van der Heijde D, Landewe R. Imaging: do erosions heal? Ann Rheum Dis 2003;62 Suppl 2:ii10-2.
14. 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(5):778-86.
15. Lukas C, Braun J, van der Heijde D et al. Scoring Inflammatory Activity of the Spine by Magnetic Resonance Imaging in Ankylosing Spondylitis. A Multi-Reader Experiment. J Rheumatol 2007;34(4):862-70.
16. van der Heijde D, Klareskog L, Rodriguez-Valverde V et al. Comparison of etanercept and methotrexate, alone and combined, in the treatment of rheumatoid arthritis: two-year clinical and radiographic results from the TEMPO study, a double-blind, randomized trial. Arthritis Rheum 2006;54(4):1063-74.