Arthritis Care & Research
DOI 10.1002/acr.22793

Further Simplification of the Simplified Erosion Narrowing Score with Item Response Theory Methodology

Martijn A.H. Oude Voshaar[1]; Olga Schenk[2]; Peter M. Ten Klooster[1] ; Harald E. Vonkeman[1,3] ; Hein J. Bernelot Moens[4] ; Maarten Boers[5] ; Mart A.F.J. van de Laar[1,3]

[1] Arthritis Center Twente and Department of Psychology, Health & Technology, Enschede, University of Twente, The Netherlands

[2] MIRA-Institute for Biomedical Technology and Technical Medicine, University of Twente, Enschede, The Netherlands.

[3] Arthritis Center Twente and Department of Rheumatology and Clinical Immunology, Medisch Spectrum Twente, Enschede, The Netherlands

[4] Department of Rheumatology, Ziekenhuisgroep Twente, The Netherlands.

[5] Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, the Netherlands.

Financial disclosure: None of the authors received financial support or other benefits from commercial sources for the work reported on in the manuscript, nor do any authors have other financial interests which could create a potential conflict of interest or the appearance of a conflict of interest with regard to the work

Corresponding Author: Martijn Oude Voshaar, Department of Psychology, Health & Technology, , University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. Tel.: +31 53 489 4470, E-mail: A.H.Oudevoshaar@utwente.nl

Received: Aug 04, 2015; Revised: Nov 04, 2015; Accepted: Nov 17, 2015

Abstract

Objective: To further simplify the Simple Erosion Narrowing Score (SENS) by removing scored areas that contribute least to its measurement precision according to Item Response Theory (IRT) based analysis and to compare the measurement performance of the simplified version to the original.

Methods: Baseline and 18 months data of the 'Combinatietherapie bij reumatoide artritis' (COBRA) trial were modelled using longitudinal IRT methodology. Measurement precision was evaluated across different levels of structural damage. SENS was further simplified by omitting the least reliably scored areas. Discriminant validity of SENS and its simplification was studied by comparing their ability to differentiate between the COBRA and SSZ arm. Responsiveness was studied by comparing standardized change scores between versions.

Results: SENS data showed good fit to the IRT model. Carpal and feet joints contributed least statistical information to both erosion and joint space narrowing scores. Omitting the joints of the foot reduced measurement precision for the erosion score in cases with below average levels of structural damage (relative efficiency compared with the original version ranged 35%-59%). Omitting the carpal joints had minimal effect on precision (relative efficiency range 77%-88%). Responsiveness of a simplified SENS without carpal joints closely approximated the original version (i.e., all Δ standardized change scores ≤ 0.06). Discriminant validity was also similar between versions for both the erosion score (relative efficiency = 97%) and the SENS total score (relative efficiency = 84%).

Conclusion Our results show that the carpal joints may be omitted from the SENS without notable repercussion for its measurement performance.

Significance and innovation

- Our results show that patients with low levels of joint damage are poorly differentiated by the Simplified Erosion and Narrowing (SENS) score. As a group, the feet contribute relatively much to the measurement precision of SENS scores for patients with low levels of joint damage.
- The Carpal joints contribute relatively little statistical information to the SENS, across the range of possible scores
- We demonstrate that the carpal joints may be omitted from SENS without noticeable repercussion for its responsiveness and discriminant validity. A further simplified SENS may therefore be a clinically more feasible tool, to be used in clinical practice or observational studies.

Rheumatoid arthritis (RA) is a systemic inflammatory disease characterized by progressive inflammation of the connective tissue of the body. Resulting structural damage of joints reflects its severity and progression and can be quantified by scoring radiographic films. Radiographic progression is therefore considered an important, objective outcome domain in RA (1,2). Although all synovial joints can be affected, most scoring systems that have been proposed and refined over time assess a selection of commonly affected or easy to read areas (3). The most popular scoring method in contemporary settings, the Sharp van der Heijde method (SvH), assesses the presence of erosions and joint space narrowing in a total of 44 and 42 joints of the feet and hands, respectively (4,5). Previous studies have documented that scoring radiographic films according to the SvH method may be a time consuming and cumbersome process, which is a disadvantage in long-term observational studies (3).

Since structural damage is rarely assessed in observational studies (3), a simplified version of SvH, the simple erosion narrowing score (SENS), was proposed for use in these settings (6). The SENS includes the same joints as the original SvH score, but simplifies the grading system of the included areas. The feasibility of assessing structural damage could be further facilitated by refinement of the areas to be scored. However, reducing the number of areas to score makes the total score less reliable, potentially undermining its responsiveness or discriminant validity. Therefore, ideally only those areas that contribute least to the reliability of the total score should be removed.

Item response theory (IRT) is a statistical framework increasingly used to develop and evaluate patient reported- and clinical outcome measures in rheumatology; it is ideally suited to simplify scales whilst preserving the reliability of the original instrument (7). In IRT, so-called information functions describe measurement precision across different levels of the trait being measured. Analysis of these functions allows identification of poorly measured trait levels. In addition, the contribution of individual items, or individual joint scores in the present study, to the measurement precision of the instrument can be quantified

In the current study we first explored the contribution of scored areas included in the SENS to its measurement precision across different levels of structural damage with IRT analysis. Subsequently, we evaluate the discriminant validity and responsiveness of a refined SENS where joint groups that contribute little to total measurement precision were omitted.

Methods

Patients

We used baseline and 18 months follow-up data of the 'Combinatietherapie Bij Reumatoide Artritis' (COBRA) study. COBRA was a multi-center randomized double blind, controlled trial of sulphasalazine (SSZ) monotherapy versus combined step-down prednisolone, methotrexate and SSZ in early rheumatoid arthritis. More details regarding the study have been describe previously (8).

Instrument

Simple erosion and narrowing score (SENS). Erosions are assessed in 16 joints of each hand/wrist and another 6 joints of each foot. Joint space narrowing is assessed in 15 joints for each hand and wrist and six joints for each foot. Each joint is assigned a maximum score of 0-2 with 0 representing no structural damage and 2 representing both erosion and narrowing present. The total score ranges from 0-86. For the present study, separate analyses were performed for the SENS erosion score, ranging from 0-44 and SENS joint narrowing score (JSN), ranging from 0-42 (6).

IRT analysis

The 2 parameter logistic model is an appropriate IRT model to analyze dichotomous data (7). This model gives the probability that structural damage (i.e. erosion or JSN, respectively) is present in a particular joint as a logistic function, called the item characteristic function, of the difference between a patients' overall level of structural damage (θ) and a parameter reflecting the sensitivity of the particular joint to inflammatory damage (β). Each scored area is further characterized by a discrimination parameter, α, which represents its ability to differentiate between different levels of θ. In the case of the logistic model, higher values of α steepen the logistic curve. Discrimination parameters can be interpreted like factor loadings in factor analysis; they reflect the strength of the association of the scored area with the overall structural damage trait. This model is given by:

$$P_{1ni} = \frac{\exp \alpha(\theta_n - \beta_i)}{1+\exp \alpha(\theta_n - \beta_i)} \quad ,$$

where P_{1ni} = patient *n*'s probability of a positive rating for structural damage in joint *i*,

β_i = defined as the position on the θ scale where $P_{1ni} = P_{0ni}$.

Baseline and 18 months data were modelled in a multidimensional generalization of the 2 parameter logistic model, suitable for longitudinal data using the MIRT software package (9,10). In this model both time points are represented by distinct, correlated dimensions and patients are described by 2 time point specific structural damage scores (θ_{T1} & θ_{T2}) but the item parameters are constrained to be equal over time so that each joint is characterized by one ICF that traces the probability that structural damage is present as a function of θ. To evaluate the fit of this model, Lagrange Multiplier statistics and accompanying effect size statistics were obtained, these are described in more detail elsewhere (11). Essentially, the Lagrange Multiplier test evaluates whether item parameters are invariant across the subsample of patients with low, medium and high total scores respectively. If the item parameters vary between subgroups, the observed average item scores within subgroups will also differ from those expected by the model which negatively influences the validity of inferences on the item parameters such as those in this study. The magnitude of this violation can be quantified by the absolute residuals (observed score minus expected score). Therefore effect size statistics that represent the absolute residuals averaged across the three subsamples of patients with low, medium and high levels of overall structural damage were obtained as well. Separate tests were performed to evaluate the ability of a joints' characteristic function to reproduce the observed data at each time point (θ_{T1}, θ_{T2}). In accordance with previous studies, cut-off points for acceptable fit were defined as $p > 0.05$ and effect size < 0.10 (12). The assertion that the item parameters were stable over time was also evaluated within this general framework.

Information functions, which quantify measurement precision of the individual joints across the possible levels of θ, were obtained from the item parameters. Reliability was assessed at the level of individual joints and aggregated for the following groups of joints: PIP, MCP CMC, wrist and feet. Reliability of groups of scored areas was evaluated by summing information functions. Information (I) is inversely related to the standard error of estimation (SEE) for each level of θ (i.e., $SEE\theta = \frac{1}{\sqrt{I(\theta)}}$). SEE < 0.32 is generally considered to reflect sufficient reliability for assessment at the level of individuals (13).

One way analysis of variance (ANOVA) evaluated the ability of the SENS and a further simplified version that omits poorly performing joint areas, to identify differences in structural progression between groups over the first 18 months in the COBRA trial. Responsiveness was evaluated by comparing standardized change scores (i.e. $\frac{\bar{x}_{bl}-\bar{x}_{t2}}{SD_{pooled}}$) between original and simplified SENS scores.

Results

At the 18 months evaluation patients in the COBRA group had a significantly lower median radiologic damage (Sharp) score compared with those in the SSZ monotherapy group. Joint damage outcomes of the study have been described in more detail elsewhere (9). The results of the analysis of model fit are presented in supplementary tables 1 and 2. Although a number of individual Lagrange Multiplier tests indicated statistically significant lack of fit for SENS erosion, particularly at T2, none of the joint scores showed statistically significant misfit at both time points. Moreover, the magnitude of misfit was modest according to the effect size statistics, with no effect size > 0.10. Finally none of the joint scores were flagged for longitudinal differential item functioning , indicating that item parameters were indeed stable over time . Similarly, for SENS JSN, 3/42 joints had statistically significant Lagrange Multiplier tests across time points, but the magnitude of misfit was again modest and no items were flagged for longitudinal bias. From these results we concluded that model fit was acceptable for both the erosion and JSN scores.

Figure 1 presents information functions for the PIP, CMC, MCP, carpal and feet joints included in the SENS erosion score, as well as the distribution of patients across different levels of structural damage. More detailed information about SEE for erosion and JSN scores is provided in S3. For all scored joint areas, SEE was lowest above the mean of the θ-scale, while the structural damage scores were distributed around the mean of the θ-scale. This reflects the low prevalence of damage in individual joints in this sample and indicates that individual joint scores discriminated better between patients with more severe structural damage than observed in the current sample. Furthermore, for SENS erosion, the MCP, CMC and PIP joints all contributed comparatively strongly to the reliability of the total SENS erosion score, while the carpal joints and joints of the feet, with the exception of the right M4, performed relatively poorly. (SEE for individual joints available on request

from the corresponding author). However, compared with the carpal joints, the feet added more to measuring the lower levels of joint erosion. For the JSN score, the feet as a group, contributed most to the reliability of the total score, with particularly the right MTP2 contributing much to the overall reliability of the tool. Again, the carpal joint contributed minimally compared with the individual PIP, CMC and MCP joints. In table 1 it can be seen that only JSN scores > 1 SD above the mean and erosion scores > 2 SDs above the mean yielded a SEE < 0.30

Combined over the analyses, the carpal joints and joints of the feet appeared to perform worse than the joints of the PIP, CMC and MCP. Figure 1 presents the relative efficiency (RE) of θ estimates of a SENS score without the joints of the feet and a SENS score without the carpal joints, both compared with the SENS total score. Since the efficiency of an estimate depends on the number of items and their quality, the total score logically performs best. It can be seen in figure 1 that θ estimates without the carpals yielded RE > 0.77 for SENS erosion and RE >0.69 for JSN. In contrast, a SENS erosion score without the feet resulted in substantial loss of efficiency for patients with low levels of structural damage. For the SENS score omitting only the carpals SEE were generally similar to the original SENS, except for JSN scores 3 SD's below the mean (table 1). Based on these results we proceeded with this simplification.

Discriminant validity and responsiveness

The efficiency of the simplified total and erosion scores to discriminate between COBRA and SSZ was 84 resp. 96% of the original SENS (Table 2). Furthermore, the standardized change scores were only marginally lower than the original scores. The results for JSN are not shown since in the original trial it did not significantly discriminate between COBRA and SSZ at this time point (F=0.282, p =0.59).

Discussion

In the current study, analysis with IRT methodology suggests that scoring of RA structural damage can be further simplified. The results indicate that a shortened version of the SENS that omits the carpal joints closely approximates the measurement performance (discriminant validity) of the original SENS. This will improve feasibility of structural

damage assessment A refined SENS with 16 less areas to score may be a more feasible tool in observational studies.

Although the results of this study revealed that the carpal joints as a group contributed least to the measurement precision of SENS, these joints might be more important in the original SvH score, since several previous studies have found that erosion and joint space narrowing in the carpal joints show relatively rapid rate of progression compared to other joints (14). The feet are frequently involved in RA and believed to contribute valuable information regarding the progression of RA. In this study the individual foot joints provided little statistical information, but as a group they contributed strongly to the precision of joint scores in patients with low levels of structural damage. This finding provides further support for the relevance of scoring the feet for structural damage (5).

The SEE quantify the contribution of individual joints and groups of joints to the precision of the instrument, across different levels of structural damage. Higher measurement precision means that the instrument can reliably detect smaller changes in structural damage. Our findings illustrate that both the individual joints and the various total scores have poor precision for most of the levels of structural damage observed in the current sample at baseline as well as at 18 months. In fact information was optimal at 2SDs above the mean level of structural damage of patients in the COBRA study after 18 months. None of the individual scored areas contributed much to precision of below average levels of structural damage. These findings suggest that both SENS and the original SvH work best to evaluate change in patients that already have relatively severe structural damage. This underlines that even better tools are needed to evaluate structural damage in early or well treated disease.

The performance of the further simplified SENS is encouraging. Nevertheless, due to its lower reliability we recommend against its use in observational studies with small sample sizes or in high-stakes studies.

In summary, in this study we show that carpus scoring can be omitted in the commonly used structural damage scoring system SENS without notable repercussions in measurement performance.

References

1. Heijde DM van der. Plain X-rays in rheumatoid arthritis: overview of scoring methods, their reliability and applicability. *Baillieres Clin Rheumatol* 1996;10:435–53.

2. Sharp JT. Radiologic assessment as an outcome measure in rheumatoid arthritis. *Arthritis Rheum* 1989;32:221–9.

3. Boini S, Guillemin F. Radiographic scoring methods as outcome measures in rheumatoid arthritis: properties and advantages. *Ann Rheum Dis* 2001;60:817–27.

4. Heijde D van der. How to read radiographs according to the Sharp/van der Heijde method. *J Rheumatol* 2000;27:261–3.

5. Heijde DM van der, Riel PL van, Nuver-Zwart IH, Gribnau FW, vad de Putte LB. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. *Lancet (London, England)* 1989;1:1036–8.

6. Heijde D van der, Dankert T, Nieman F, Rau R, Boers M. Reliability and sensitivity to change of a simplification of the Sharp/van der Heijde radiological assessment in rheumatoid arthritis. *Rheumatology (Oxford)* 1999;38:941–7.

7. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory. Measurement methods for the social sciences series, Vol. 2*. CA, US: Sage Publications, Inc; 1991.

8. Boers M, Verhoeven AC, Markusse HM, Laar MA van de, Westhovens R, Denderen JC van, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. *Lancet (London, England)* 1997;350:309–18.

9. Glas C. Preliminary manual of the software program Multidimensional Item Response Theory (MIRT). *Univ Twente, Enschede, Netherlands* 2010.

10. Marvelde JM te. Application of Multidimensional Item Response Theory Models to Longitudinal Data. *Educ Psychol Meas* 2006;66:5–34.

11. Glas CAW. Modification indices for the 2-PL and the nominal response model. *Psychometrika* 1999;64:273–294.

12. Groen MM van, Klooster PM ten, Taal E, Laar MAFJ van de, Glas CAW. Application of the health assessment questionnaire disability index to various rheumatic diseases. *Qual Life Res* 2010;19:1255–63.

13. Thissen D. Reliability and measurement precision. In: Wainer H, ed. *Computerized adaptive testing: A primer*. 2nd ed. NJ, US: Lawrence Erlbaum Associates Publishers; 2000:159–184.

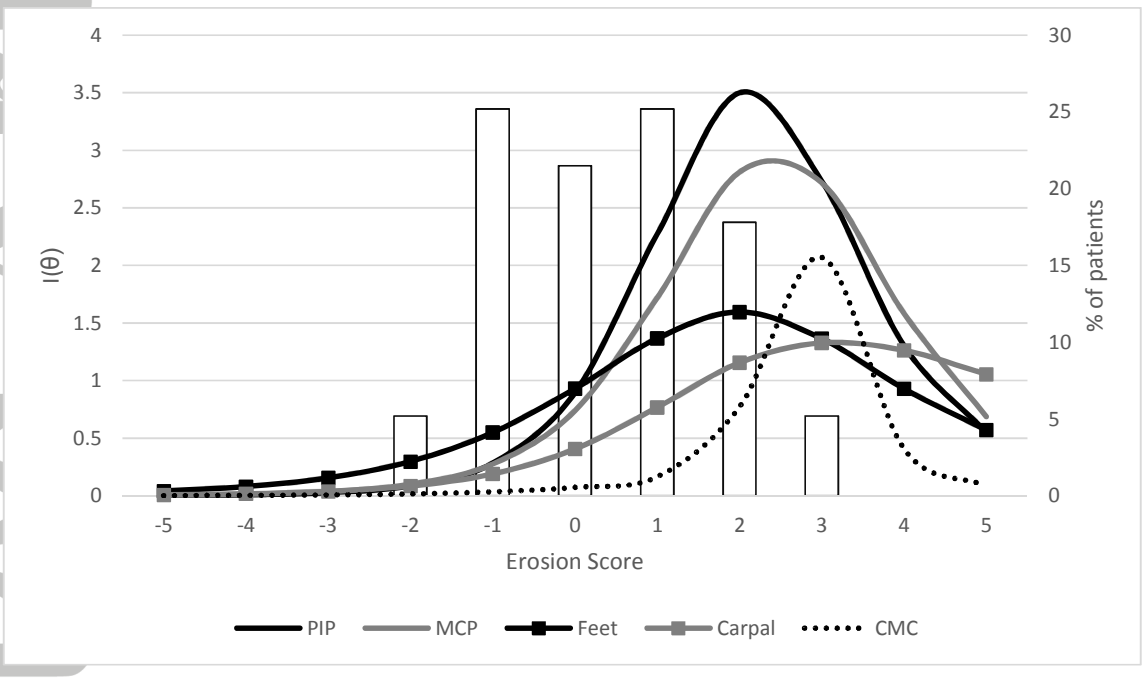
14. Leak RS, Rayan GM, Arthur RE. Longitudinal radiographic analysis of rheumatoid arthritis in the hand and wrist. *J Hand Surg Am* 2003;28:427–34.

Table 1. Discriminant validity of simplified SENS compared with original version

	Baseline Mean (s.d.)	18 months Mean (s.d.)	*ES*	*F**	*p*	RE
SENS Total				4.75	0.03	1.00
Combination therapy	6.8 (7.2)	14.3 (11.4)	0.79			
Monotherapy	5.8 (7.1)	10.9 (10.7)	0.56			
Simplified Total				3.98	0.04	0.84
Combination therapy	5.6 (6.1)	11.6 (9.4)	0.76			
Monotherapy	5.0 (6.3)	8.9 (8.0)	0.54			
SENS Erosion				9.43	<0.01	1.00
Combination therapy	4.8 (5.5)	10.3 (7.9)	0.81			
Monotherapy	4.1 (5.0)	7.2 (7.1)	0.50			
Simplified Total				9.12	<0.01	0.97
Combination therapy	4.2 (4.6)	8.8 (6.9)	0.79			
Monotherapy	3.7 (4.6)	6.1 (6.1)	0.44			

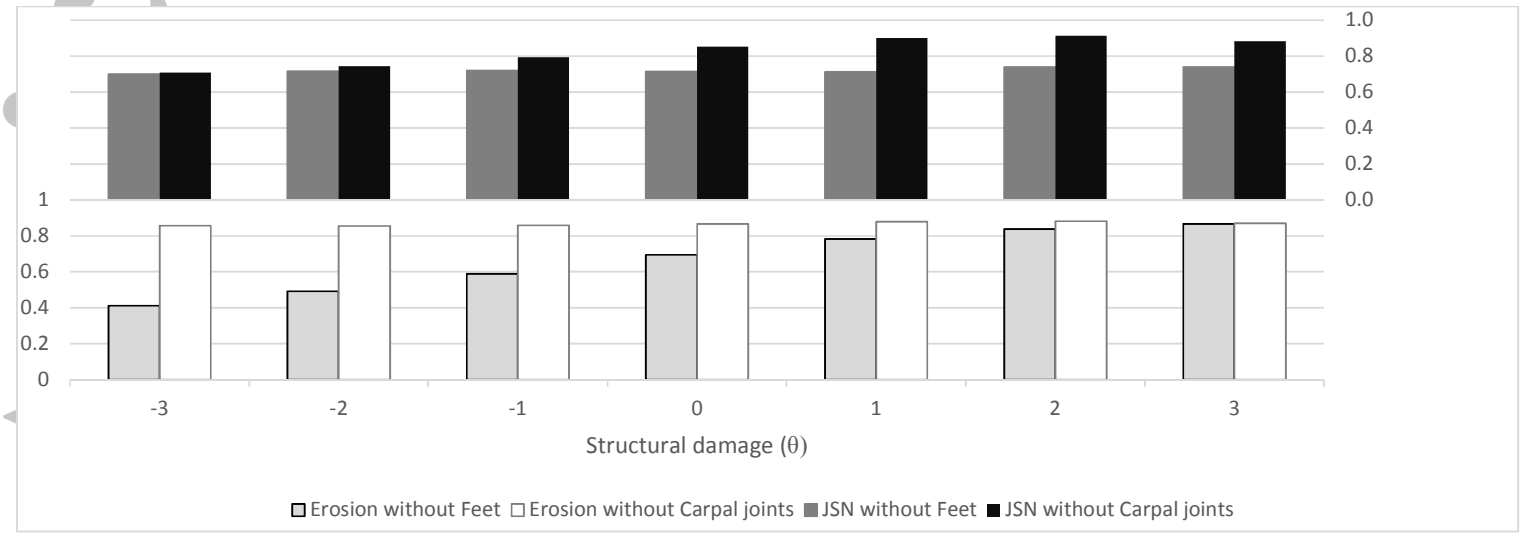
ES = Effect size (M1-M2/ pooled σ); F = comparison of mean change scores between conditions; RE = relative efficiency coefficient (F/F(total); SSZ = sulphasalazine group; COBRA = combination therapy group

Figure 1: Distribution of patients and standard information functions of scored areas included in SENS Erosion scores across different levels of structural damage



I(θ) : amount of statistical information provided per scored area; ** The amount of statistical information provided by different scored areas is expressed on a scale with mean = 0 (SD=1), range = -5SD to +5SD. higher values of I(θ) indicate higher measurement precision. Bars represent % of patients at each of the depicted score levels. Higher scores reflect more severe structural damage.

Figure 2: Relative Efficiency of Refined SENS scores without carpal joints and without feet joints compared with original SENS JSN (upper panel) and erosion scores (lower panel)



Relative efficiency = $I_{refinedSENS}$ (θ)/ I_{Sens} (θ). Relative efficiency is presented for different levels of structural damage expressed on a scale with mean = 0 (SD=1), range = -3SD to +3SD. Higher scores reflect more severe structural damage.

S1 1Item fit and longitudinal differential item functioning for SENS Erosions score

Joint location	T1			T2			Longitudinal DIF		
	LM	*p*	*ES*	*LM*	*p*	*ES*	*LM*	*p*	*ES*
PIP1*	0.97	0.62	0.02	26.10	0.00	0.04	0.24	0.63	0.01
PIP2*	0.23	0.89	0.03	0.77	0.68	0.03	0.50	0.48	0.02
PIP3*	3.52	0.17	0.08	3.26	0.20	0.05	0.08	0.78	0.01
PIP4*	0.18	0.91	0.02	1.40	0.50	0.04	0.97	0.32	0.03
PIP5*	13.86	0.00	0.02	29.35	0.00	0.01	0.00	0.95	0.01
PIP1	0.07	0.97	0.02	0.37	0.83	0.03	0.04	0.85	0.01
PIP2	0.70	0.70	0.03	0.00	1.00	0.00	0.13	0.72	0.01
PIP3	0.47	0.79	0.02	0.85	0.66	0.02	0.04	0.84	0.01
PIP4	0.97	0.62	0.04	0.23	0.89	0.01	0.02	0.89	0.01
PIP5	0.45	0.80	0.03	0.08	0.96	0.01	0.02	0.88	0.02
MCP1*	1.50	0.47	0.08	5.81	0.05	0.05	0.00	0.98	0.01
MCP2*	0.75	0.69	0.07	1.47	0.48	0.03	0.07	0.79	0.01
MCP3*	0.40	0.82	0.02	1.01	0.60	0.02	0.00	0.96	0.00
MCP4*	0.10	0.95	0.03	0.09	0.96	0.02	0.26	0.61	0.01
MCP5*	0.75	0.69	0.03	0.22	0.90	0.01	0.01	0.91	0.01
MCP1	0.83	0.66	0.05	0.21	0.90	0.02	1.01	0.31	0.03
MCP2	1.44	0.49	0.06	0.85	0.65	0.01	0.03	0.87	0.01
MCP3	0.05	0.97	0.04	0.65	0.72	0.02	0.38	0.54	0.02
MCP4	0.13	0.94	0.03	1.18	0.55	0.02	0.05	0.82	0.01
MCP5	1.96	0.37	0.05	34.52	0.00	0.02	0.18	0.67	0.02
CMC1*	10.88	0.00	0.06	15.42	0.00	0.01	1.00	0.32	0.02
CMC1	4.07	0.13	0.08	3.14	0.21	0.03	0.82	0.37	0.03
Trapezius*	1.85	0.40	0.07	0.03	0.98	0.01	0.07	0.80	0.02
Trapezius	3.06	0.22	0.07	3.53	0.17	0.04	0.10	0.75	0.01
Scaphoid*	6.09	0.05	0.11	10.64	0.00	0.06	0.84	0.36	0.04
Scaphoid	18.11	0.00	0.03	40.09	0.00	0.01	0.53	0.46	0.01
Lunate*	2.25	0.32	0.05	44.20	0.00	0.02	0.42	0.52	0.02
Lunate	0.47	0.79	0.02	0.68	0.71	0.01	0.46	0.50	0.02
Radius*	24.22	0.00	0.04	84.35	0.00	0.02	0.29	0.59	0.01
Radius	1.73	0.42	0.04	0.10	0.95	0.01	0.02	0.88	0.00
Ulnar*	2.16	0.34	0.05	39.69	0.00	0.06	0.17	0.68	0.01
Ulnar	3.13	0.21	0.06	0.57	0.75	0.04	1.57	0.21	0.05
IP*	0.57	0.75	0.03	0.21	0.90	0.01	0.02	0.90	0.02
IP	0.65	0.72	0.05	2.96	0.23	0.04	0.01	0.94	0.02
MTP1*	0.14	0.93	0.02	1.62	0.44	0.04	0.17	0.68	0.01
MTP2*	0.80	0.67	0.04	1.79	0.41	0.05	0.39	0.53	0.05
MTP3*	2.02	0.36	0.07	2.43	0.30	0.04	1.33	0.25	0.03
MTP4*	0.55	0.76	0.04	0.02	0.99	0.00	0.04	0.84	0.02
MTP5*	3.26	0.20	0.07	0.96	0.62	0.03	0.16	0.69	0.03
MTP1	0.91	0.63	0.07	0.38	0.83	0.02	0.11	0.74	0.02
MTP2	1.70	0.43	0.06	2.59	0.27	0.03	0.97	0.32	0.05
MTP3	4.81	0.09	0.09	2.21	0.33	0.05	1.97	0.16	0.08
MTP4	1.11	0.57	0.05	0.13	0.94	0.01	0.05	0.83	0.01
MTP5	34.06	0.00	0.01	19.74	0.00	0.02	0.00	0.99	0.01

*=Left side; LM = Lagrange multiplier statistic; ES = Effect size. T1 Results of item fit analysis for Baseline data; T2 results of item fit analysis for follow-up data; DIF = differential item functioning over time;Cut-off points for acceptable fit were defined as $p > 0.05$ and $ES < 0.10$.pip = proximal interphalangeal joint; mcp =metacarpophalangeal joint; CMC = carpometacarpal; IP = interphalangeal joint of the toe; MTP = metatarsophalangeal joint of the toe;

S2 1tem fit and longitudinal differential item functioning for SENS joint narrowing score

Joint location	T1			T2			Longitudinal DIF		
	lm	p	es	lm	p	es	lm	p	es
PIP2*	25.96	0.00	0.01	0.47	0.49	0.00	0.47	0.49	0.00
PIP3*	46.63	0.00	0.01	0.11	0.74	0.01	0.10	0.76	0.00
PIP4*	0.10	0.76	0.00	0.22	0.64	0.01	0.02	0.88	0.00
PIP5*	0.10	0.76	0.01	0.01	0.93	0.01	0.00	0.95	0.00
PIP2	65.09	0.00	0.00	0.27	0.60	0.01	0.03	0.86	0.00
PIP3	29.44	0.00	0.00	0.59	0.44	0.01	0.01	0.93	0.00
PIP4	38.55	0.00	0.01	0.02	0.89	0.01	0.21	0.65	0.01
PIP5	0.63	0.43	0.01	0.59	0.44	0.01	0.19	0.66	0.01
MCP1*	0.28	0.60	0.02	0.44	0.51	0.01	0.18	0.67	0.00
MCP2*	41.80	0.00	0.01	0.02	0.89	0.02	0.69	0.41	0.02
MCP3*	0.18	0.67	0.01	0.14	0.71	0.00	0.01	0.91	0.01
MCP4*	0.14	0.71	0.00	0.07	0.79	0.01	0.11	0.74	0.00
MCP5*	68.86	0.00	0.01	0.65	0.42	0.03	0.43	0.51	0.01
MCP1	3.49	0.06	0.02	0.00	1.00	0.01	0.01	0.91	0.01
MCP2	0.01	0.92	0.00	0.91	0.34	0.03	0.07	0.80	n/a
MCP3	0.24	0.62	0.00	0.02	0.87	0.00	0.02	0.90	0.01
MCP4	58.64	0.00	0.01	0.00	0.98	0.01	0.18	0.67	0.01
MCP5	62.01	0.00	0.01	0.00	0.98	0.02	1.14	0.29	0.02
CMC3*	46.49	0.00	0.03	0.76	0.38	0.03	3.61	0.06	0.02
CMC4*	0.08	0.78	0.02	1.04	0.31	0.01	0.91	0.34	0.01
CMC5*	26.67	0.00	0.00	0.55	0.46	0.01	0.00	0.96	0.00
CMC3*	30.05	0.00	0.02	0.68	0.41	0.02	13.12	0.00	0.01
CMC4*	0.27	0.60	0.01	0.08	0.78	0.01	0.06	0.80	0.01
CMC5*	23.57	0.00	0.01	26.77	0.00	0.01	0.46	0.50	0.00
MN*	0.10	0.75	0.00	0.56	0.45	0.03	0.00	0.95	0.01
MN	0.05	0.83	0.03	0.06	0.80	0.03	0.59	0.44	0.03
CNL*	0.13	0.72	0.01	0.74	0.39	0.02	0.11	0.74	0.01
CNL	34.79	0.00	0.02	1.30	0.25	0.02	3.99	0.05	0.01
RC*	0.09	0.76	0.01	0.33	0.56	0.01	0.01	0.94	0.00
RC	0.85	0.36	0.03	0.02	0.88	0.00	4.89	0.03	0.02
IP*	1.90	0.17	0.01	0.58	0.44	0.01	1.37	0.24	0.01
IP	2.34	0.13	0.02	0.02	0.90	0.01	0.08	0.78	0.00
MTP1*	0.11	0.74	0.04	0.38	0.54	0.02	0.72	0.40	0.02
MTP2*	1.22	0.27	0.00	0.29	0.59	0.01	0.08	0.78	0.01
MTP3*	0.21	0.65	0.01	0.77	0.38	0.03	0.01	0.92	0.00
MTP4*	24.92	0.00	0.02	27.53	0.00	0.01	0.36	0.55	0.01
MTP5*	0.64	0.42	0.01	0.46	0.50	0.01	0.06	0.80	0.01
MTP1	1.17	0.28	0.02	1.29	0.26	0.01	0.17	0.68	0.02
MTP2	18.21	0.00	0.01	15.52	0.00	0.01	0.29	0.59	0.00
MTP3	0.00	0.97	0.00	0.18	0.67	0.01	0.10	0.75	0.01
MTP4	0.06	0.80	0.01	0.59	0.44	0.01	0.11	0.74	0.01
MTP5	0.11	0.74	0.02	1.10	0.29	0.03	0.38	0.54	0.01

*=Left side; LM = Lagrange multiplier statistic; ES = Effect size. T1 Results of item fit analysis for Baseline data; T2 results of item fit analysis for follow-up data; DIF = differential item functioning over time;Cut-off points for acceptable fit were defined as $p > 0.05$ and $ES < 0.10$.pip = proximal interphalangeal joint; mcp =metacarpophalangeal joint; CMC = carpometacarpal; MN = multangular navicular joints ; CNL = capitate-navicular-lunate joint IP = interphalangeal joint of the toe; MTP = metatarsophalangeal joint of the toe;

S3: Distribution of patients and standard error of estimation (SEE) of scored areas included in SENS joint space narrowing and SENS Erosion scores across different levels of structural damage

	Structrual damage score (0)						
	-3	-2	-1	0	1	2	3
Erosions							
Feet	2.5	1.8	1.3	1.0	0.9	0.8	0.9
PIP	6.1	3.4	1.9	1.1	0.7	0.5	0.6
MCP	5.2	3.2	1.9	1.2	0.8	0.6	0.6
Carpal	5.1	3.4	2.3	1.6	1.1	0.9	0.9
CMC	11.9	7.9	5.4	3.7	2.5	1.1	0.7
Erosion score total	1.9	1.3	0.9	0.6	0.4	0.3	0.3
Erosion socre without carpal joints	2.1	1.4	0.9	0.6	0.4	0.3	0.3
% of patients (baseline)	12%	19%	28%	12%	23%	2%	0%
% of patients (18 months)	0%	5%	25%	22%	25%	18%	5%
Joint space Narrowing							
Feet	6.4	3.7	2.0	0.9	0.5	0.4	0.6
PIP	14.9	7.6	3.7	1.6	0.7	0.4	0.6
MCP	14.6	6.0	2.4	1.0	0.4	0.3	0.6
CMC	6.5	3.8	2.1	1.1	0.7	0.6	1.0
Carpal	6.5	3.9	2.3	1.3	0.8	0.7	0.9
Narrowing score total	3.5	2.0	1.0	0.5	0.3	0.2	0.3
Narrowing score without carpal joints	4.2	2.3	1.2	0.5	0.3	0.2	0,3
% of patients (baseline)	0%	29%	42%	10%	18%	1%	0%
% of patients (18 months)	0%	0%	0%	42%	55%	3%	0%

** The standard error of estimation (SEE) is presented for different levels of structural damage expressed on a scale with mean = 0 (SD=1), range = -3SD to +3SD. Lower values of SEE indicate higher measurement precision. Values in parentheses at the bottom of each of both panels represent the percentage of patients at the level of structural damage represented by the column. Higher scores reflect more severe structural damage.