Arthritis Care & Research
Vol. 68, No. 8, August 2016, pp 1206–1210
DOI 10.1002/acr.22793

BRIEF REPORT

Further Simplification of the Simple Erosion Narrowing Score With Item Response Theory Methodology

MARTIJN A. H. OUDE VOSHAAR,[1] OLGA SCHENK,[1] PETER M. TEN KLOOSTER,[1] HARALD E. VONKEMAN,[2] HEIN J. BERNELOT MOENS,[3] MAARTEN BOERS,[4] AND MART A. F. J. VAN DE LAAR[2]

Objective. **To further simplify the simple erosion narrowing score (SENS) by removing scored areas that contribute the least to its measurement precision according to analysis based on item response theory (IRT) and to compare the measurement performance of the simplified version to the original.**

Methods. **Baseline and 18-month data of the Combinatietherapie Bij Reumatoide Artritis (COBRA) trial were modeled 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 were studied by comparing their ability to differentiate between the COBRA and sulfasalazine arms. 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 the 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 were $\leq$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.**

Introduction

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, assesses the presence of erosions and joint space narrowing (JSN) in a total of 44 and 42 joints of the feet and hands, respectively (4,5). Previous studies, however, have documented that scoring radiographic films according to the Sharp/van der Heijde 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 the Sharp/van der Heijde method, the simple erosion narrowing score (SENS), was proposed for use in these settings (6). The SENS includes the same joints as the original Sharp/van der Heijde 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

[1]Martijn A. H. Oude Voshaar, PhD, Olga Schenk, MSc, Peter M. ten Klooster, PhD: University of Twente, Enschede, The Netherlands; [2]Harald E. Vonkeman, MD, PhD, Mart A. F. J. van de Laar, MD, PhD: University of Twente and Medisch Spectrum Twente, Enschede, The Netherlands; [3]Hein J. Bernelot Moens, MD, PhD: Ziekenhuisgroep Twente, Enschede, The Netherlands; [4]Maarten Boers, MD, PhD: VU University Medical Center, Amsterdam, The Netherlands.

Address correspondence to Martijn A. H. Oude Voshaar, MD, Department of Psychology, Health & Technology, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands. E-mail: a.h.oudevoshaar@utwente.nl.

Submitted for publication August 4, 2015; accepted in revised form November 17, 2015.

Significance & Innovations

- Our results show that patients with low levels of joint damage are poorly differentiated by the simple erosion and narrowing (SENS) score. As a group, the feet contribute relatively strongly 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.

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 while 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 evaluated the discriminant validity and responsiveness of a refined SENS, where joint groups that contribute little to total measurement precision were omitted.

Patients and methods

Patients. We used baseline and 18-month followup data from the Combinatietherapie Bij Reumatoide Artritis (COBRA) study. COBRA was a multicenter, randomized, double-blind, controlled trial of sulfasalazine (SSZ) monotherapy versus combined step-down prednisolone, methotrexate, and SSZ in early RA. More details regarding the study have been described previously (8).

Instrument: simple erosion and narrowing score (SENS). Erosions are assessed in 16 joints of each hand and wrist and another 6 joints of each foot. JSN is assessed in 15 joints for each hand and wrist and 6 joints for each foot. Each joint is assigned a maximum score of 0–2 (where 0 = no structural damage and 2 = both erosion and narrowing present). The total score ranges 0–86. For the present study, separate analyses were performed for the SENS erosion score (range 0–44) and SENS JSN score (range 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) 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. 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:

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

where P_{1ni} = patient n's probability of a positive rating for structural damage in joint i, and β_i = the position on the θ scale where $P_{1ni} = P_{0ni}$.

Baseline and 18-month data were modeled in a multidimensional generalization of the 2-parameter logistic model, suitable for longitudinal data using MIRT software (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} and θ_{T2}), but the item parameters are constrained to be equal over time so that each joint is characterized by 1 International Classification of Functioning, Disability, and Health category 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, described in 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. 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 3 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 joint's characteristic function to reproduce the observed data at each time point (θ_{T1}, θ_{T2}). In accordance with previous studies, cutoff points for acceptable fit were defined as P greater than 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

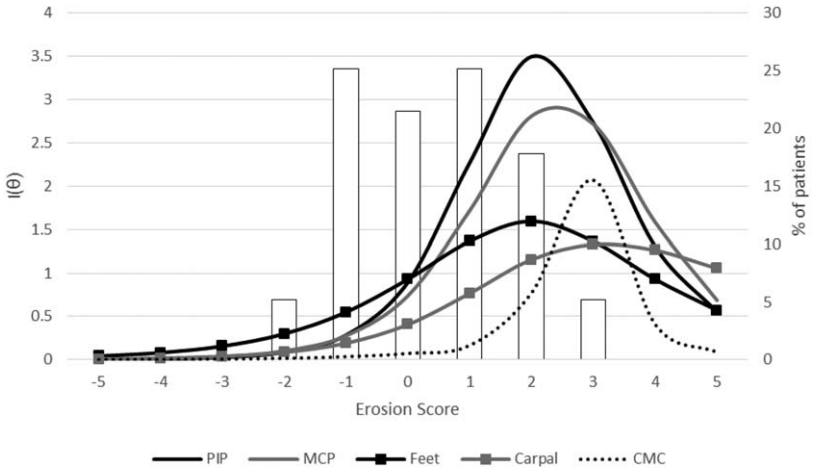

Figure 1. Distribution of patients and standard information functions of scored areas included in simple erosion narrowing (SENS) scores across different levels of structural damage. I(θ) = amount of statistical information provided per scored area. The amount of statistical information provided by different scored areas is expressed on a scale with mean ± SD 0 ± 1 (range − 5 to 5). Higher values of I(θ) indicate higher measurement precision. Bars represent the percentage of patients at each of the depicted score levels. Higher scores reflect more severe structural damage. PIP = proximal interphalangeal; MCP = metacarpophalangeal; CMC = carpometacarpal.

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: proximal interphalangeal (PIP), metacarpophalangeal (MCP), carpometacarpal (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 the estimate (SEE) for each level of θ (i.e., $\theta = \frac{1}{\sqrt{I(\theta)}}$) . An SEE <0.32 is generally considered to reflect sufficient reliability for assessment at the level of individuals (13).

One-way analysis of variance 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., $\frac{\bar{x}_{T1}-\bar{x}_{t2}}{SD_{pooled}}$) between original and simplified SENS scores.

Results

At the 18-month evaluation, patients in the COBRA group had a significantly lower mean radiologic damage (Sharp/van der Heijde 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 shown in Supplementary Tables 1 and 2 (available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22793/abstract). Although a number of individual Lagrange Multiplier tests indicated statistically significant lack of fit for SENS erosion, particularly at time point 2, 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 of 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 shows 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 the SEE for erosion and JSN scores is provided in Supplementary Table 3 (available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22793/abstract). For all scored joint areas, the 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 result 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 fourth metatarsal, performed relatively poorly (the SEE for individual joints is 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 the most to the reliability of the total score, with particularly the right second metatarsophalangeal joint contributing strongly to the overall reliability of the tool. Again, the carpal joint contributed minimally compared with the individual PIP, CMC, and MCP joints. It can be

Table 1. Discriminant validity of simplified simple erosion narrowing score (SENS) compared with original version*

SENS version	Baseline	18 months	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			

* Values for baseline and 18 months are mean ± SD. ES = effect size (first metacarpal–second metacarpal/pooled σ); F = comparison of mean change scores between conditions; RE = relative efficiency coefficient (F/F [total]).

seen in the Supplementary material (available on the *Arthritis Care & Research* web site at http:onlinelibrary.wiley.com/doi/10.1002/acr.22793/abstract) that only JSN scores >1 SD above the mean and erosion scores >2 SDs above the mean yielded an 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 2 shows 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. Figure 2 shows 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 SEEs were generally similar to the original SENS, except for JSN scores 3 SDs below the mean (Table 1). Based on these results, we proceeded with this simplification.

The efficiency of the simplified total and erosion scores to discriminate between COBRA and SSZ was respectively 84% and 97% of the original SENS (see Supplementary Table 3 available on the *Arthritis Care & Research* web site at http://onlinelibrary.wiley.com/doi/10.1002/acr.22793/abstract). Furthermore, the standardized change scores were only marginally lower than the original scores. The results for JSN are not shown, because in the original trial they 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 shortened version will improve feasibility of structural damage assessment, as a refined

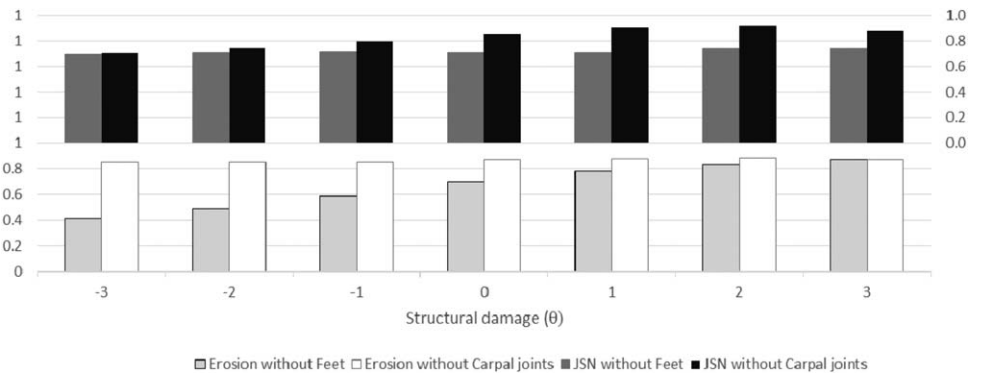

Figure 2. Relative efficiency of refined simple erosion narrowing (SENS) scores without carpal joints and without feet joints compared with original SENS joint space narrowing (JSN) (**upper row**) and erosion scores (**lower row**). Relative efficiency = $I_{refinedSENS}$ (θ)/I_{Sens} (θ). Relative efficiency is shown for different levels of structural damage, expressed on a scale with mean ± SD 0 ± 1 (range −3 to 3). Higher scores reflect more severe structural damage.

SENS with 16 fewer 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 Sharp/van der Heijde score, since several previous studies have found that erosion and JSN in the carpal joints show a relatively rapid rate of progression compared to other joints (14). The feet are frequently involved in RA and may 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 2 SDs 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 in assessing below-average levels of structural damage. These findings suggest that both SENS and the original Sharp/van der Heijde method work best to evaluate change in patients who already have relatively severe structural damage. This suggestion underlines the fact 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.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Oude Voshaar had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Oude Voshaar, Schenk, ten Klooster, Vonkeman, Bernelot Moens, van de Laar.

Acquisition of data. Boers.

Analysis and interpretation of data. Oude Voshaar.

REFERENCES

1. Van der Heijde DM. 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. Van der Heijde D. How to read radiographs according to the Sharp/van der Heijde method. J Rheumatol 2000;27:261–3.
5. Van der Heijde DM, van Riel PL, Nuver-Zwart IH, Gribnau FW, van de Putte LB. Effects of hydroxychloroquine and sulphasalazine on progression of joint damage in rheumatoid arthritis. Lancet 1989;1:1036–8.
6. Van der Heijde D, 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. Los Angeles: Sage Publications; 1991.
8. Boers M, Verhoeven AC, Markusse HM, van de Laar MA, Westhovens R, van Denderen JC, et al. Randomised comparison of combined step-down prednisolone, methotrexate and sulphasalazine with sulphasalazine alone in early rheumatoid arthritis. Lancet 1997;350:309–18.
9. Glas C. Preliminary manual of the software program Multidimensional Item Response Theory (MIRT). Enschede (The Netherlands): University of Twente; 2010.
10. Marvelde JM. Application of multidimensional item response theory models to longitudinal data. Educ Psychol Meas 2006;66:5–34.
11. Glas CA. Modification indices for the 2-PL and the nominal response model. Psychometrika 1999;64:273–94.
12. Van Groen MM, ten Klooster PM, Taal E, van de Laar MA, Glas CA. 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. Mahwah (NJ): Lawrence Erlbaum Associates; 2000. pp. 159–84.
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.