Use of the 28-Joint Count Yields Significantly Higher Concordance Between Different Examiners Than the 66/68-Joint Count

MATHIAS GRUNKE, MATTHIAS N. WITT, MONIKA RONNEBERGER, AMELIE SCHNEZ, RUEDIGER P. LAUBENDER, MATTHIAS ENGELBRECHT, ARTHUR KAVANAUGH, and HENDRIK SCHULZE-KOOPS

ABSTRACT. ***Objective.*** Joint counts are the key outcome measure in rheumatoid arthritis (RA). There is a great variability between different assessors of the same patient; this variability can be reduced by standardized training. The training effect is far less pronounced for the 66/68-joint count compared to the 28-joint count. We evaluated the reason for the higher interrater disagreement in the 66/68 compared to the 28-joint count.

Methods. Participants in joint examination seminars evaluated a patient with RA before and after training in the European League Against Rheumatism technique. Joints were rated positive or negative for tenderness and swelling. The number of positive joints and the variability between examiners before and after the training were compared. Concordance was calculated for every single joint using the Fleiss-Kappa test.

Results. In total, 256 health professionals were instructed in the 66/68-joint count and 84 in the 28-joint count. The disagreement between examiners was higher for swelling than for tenderness. After the training, there was a significant reduction of interrater variability, which was more pronounced in the 28 than in the 66/68-joint count. Comparisons between joint counts revealed that the joints of the feet were more likely to be rated negative, yet interrater disagreement was still high.

Conclusion. Standardization of joint examination significantly reduces variability between assessors. The better performance of the 28-joint count is due to the lower number of joints examined, especially the foot joints, which remain difficult to assess reliably even after training. (First Release June 1 2012; J Rheumatol 2012;39:1334–40; doi:10.3899/jrheum.110677)

Key Indexing Terms:
DISEASE ACTIVITY SCORE-28 OUTCOME RESEARCH RHEUMATOID ARTHRITIS

The tender and swollen joint counts are key elements of the core set of assessments defined by the American College of Rheumatology (ACR) that are recommended for all clinical trials in rheumatoid arthritis (RA)[1] as well as for daily practice[2]. Joint counts are also a critical component of composite disease activity measures such as the Disease Activity Score (DAS)[3] or the ACR response criteria[4]. The joint counts that are most widely used are the 28-joint count, which rates joints of the upper extremities and the knees, and the 66/68-joint count, which includes the joints of the 28-joint count and also the joints of the lower extremities except for the distal interphalangeal (DIP) joints of the feet. Both joint counts consist of the 2 dimensions "tenderness" and "swelling" for every single joint except for the hip in the 66/68-joint count, which is assessed only for tenderness.

Recently, technical developments such as magnetic resonance imaging and ultrasensitive ultrasound[5] have proven to be very sensitive in detecting even small degrees of synovitis, but their use in clinical practice and in the clinical trial setting is limited by constraints of time and costs. Moreover, these methods are not able to measure tenderness, which is perceived by the patients. In consequence, "traditional" joint counts are still the most important components in outcomes research in rheumatology[6,7].

It has been shown that there is high variability in the joint examination results between different assessors, even among experienced rheumatologists[7,8,9,10]. Differences in the evalua-

From the Division of Rheumatology, Medizinische Klinik and Poliklinik IV, University of Munich, Munich; Department of Internal Medicine 3, University of Erlangen-Nuremberg, Erlangen; Institute of Medical Informatics, Biometry, and Epidemiology, University of Munich, Munich, Germany; and Division of Rheumatology, Allergy and Immunology, University of California, San Diego, California, USA.

M. Grunke, MD; M.N. Witt, MD, Division of Rheumatology, Medizinische Klinik and Poliklinik IV, University of Munich; M. Ronneberger, MD, Department of Internal Medicine 3, University of Erlangen-Nuremberg; A. Schnez, MD, Division of Rheumatology, Medizinische Klinik and Poliklinik IV, University of Munich; R.P. Laubender, MA, MPh, Institute of Medical Informatics, Biometry, and Epidemiology, University of Munich; M. Engelbrecht, Dipl-Psych, Department of Internal Medicine 3, University of Erlangen-Nuremberg; A. Kavanaugh, MD, Division of Rheumatology, Allergy and Immunology, University of California, San Diego; H. Schulze-Koops, MD, Division of Rheumatology, Medizinische Klinik and Poliklinik IV, University of Munich.

Dr. Grunke and Dr. Witt contributed equally to this report.

Address correspondence to Dr. M. Grunke, University of Munich, Medizinische Poliklinik und Poliklinik IV, Division of Rheumatology, Ziemssenstr. 1, 80336 Munich, Germany.
E-mail: mathias.gruenke@med.uni-muenchen.de

Accepted for publication April 20, 2012.

tions of affected joints may lead to errors in assessments of disease activity in given patients and can severely confound results of multicenter trials.

It is therefore common practice, and even required by some authorities, to standardize joint examination techniques in clinical trials[11,12,13]. Usually this is done by hands-on training programs during or adjacent to investigator meetings for the respective studies. The examination technique used in this work is based on the recommendations of the EULAR (European League Against Rheumatism) handbook of clinical assessments in rheumatoid arthritis. We have shown that such standardization of training can lead to a significant reduction in variability between joint examiners. However, variability remains higher after training for the 66/68-joint count, which is usually used for determination of the ACR response, than for the 28-joint count used for the Disease Activity Score-28 (DAS28) and EULAR response[9]. In consequence, it has been suggested that the results of standardization training be evaluated on a single-joint level to determine in which joints the training is effective in reducing variability and in which joints it somewhat falls short. We therefore extended the existing data pool and reanalyzed it on a single-joint level.

The hypothesis of this work was that the joints of the 66/68-joint count show a higher variability than those of the 28-joint count, due to not only the higher number of joints assessed, but more importantly to the uncertainties in assessing joints of the lower extremities.

MATERIALS AND METHODS

Joint assessment training was performed by 1 trainer with 15–25 healthcare professionals from different clinical sites and countries. Participants were mostly physicians specializing in rheumatology, along with study nurses and a few medical technicians and physiotherapists. All data included into this evaluation were collected by one of 2 trainers, who were from the same rheumatology center and who used an identical training design, as follows.

Trainees were divided into groups of a maximum size of 6. To ensure independence of assessments for each participant, trainees originating from the same trial investigation site were assigned to different groups. To evaluate the effect of standardization, each of the groups examined 1 patient with RA before and after they were made familiar with the EULAR examination technique. Volunteer patients with varying levels of active disease (i.e., all patients had at least a moderate disease activity, with DAS28 scores ≥ 3.2) were selected for the sessions. Joints were rated positive or negative for tenderness and swelling, without grading. Before the standardization training, participants were invited to perform the examination according to the individual technique they had routinely used in their practices. Results were collected and tabulated.

Subsequently, one of the authors delivered a lecture about the background of joint counts in RA and their importance as the main outcome measures in clinical trials. In addition, a standardized examination technique based on that recommended by EULAR was demonstrated by the trainer for each joint. Depending on the design of the given clinical trial, either the 66/68 or the 28-joint count was applied. The 28-joint count consists of the finger joints excluding the DIP joints, the wrists, elbows, shoulders, and knees. The 66/68-joint count additionally counts the DIP of the fingers, acromioclavicular and sternoclavicular joints, ankles, tarsal joints, and metatarsophalangeal (MTP) and proximal interphalangeal (PIP) joints of the feet. The hips are evaluated only for tenderness, making 68 joints available for tenderness evaluation and 66 joints for swelling. Each group then practiced joint count examination in 1 to 3 additional different patients with RA under the direct supervision of the trainer. Particular joints with differing results for tenderness or swelling within a group were discussed between the groups and the trainer.

Finally, each examiner returned to the first patient and reevaluated the joint count using the standardized examination technique, now without guidance by the trainer. Again, the results were tabulated and compared with the investigations before the seminar with regard to changes in tender and swollen joint counts within the groups.

Changes in overall joint counts were calculated over the whole number of assessments that could be evaluated. Only examinations with a complete dataset of tender and swollen joint counts before and after the training were evaluated.

Statistics. We quantified the concordance of the presence or absence of swelling and of tenderness for each joint and for pre- and post-training by using the kappa statistic. However, because for each joint multiple raters (the assessors per training session) existed per subject (the patient of the session) and the raters for 1 subject were not identical to those for the other subjects, the modified kappa statistic was used for each joint, which combines the variation between subjects and within subjects as outlined by Fleiss[14] for such data (Fleiss: formula 13.44). Further, we calculated the corresponding standard error (formula 13.46)[14] and hence the 95% CI of kappa for each joint. A kappa value ≥ 0.6 was considered a strong agreement. We plotted all kappa statistics of the joints and corresponding 95% CI in a CI plot stratified by swelling/tenderness and pre/post training.

To quantify the effect of the training, we calculated the difference of the kappa statistic of each joint for swelling and for pressure pain. The standard errors for that difference can be obtained by applying bootstrap techniques for each joint. For each bootstrap we generated 999 replicates. As the distribution of the bootstrap replicates was close to normal, we used the normal approximation method for deriving a 95% CI for a difference of kappa statistics for each joint. Again, the 95% CI for each joint stratified by swelling/tenderness were plotted in a CI plot.

All statistical analyses were performed using R, version 2.13.2[15].

RESULTS

Between August 2002 and August 2008, 600 individuals from a variety of countries in Europe, North and South America, Asia, and Australia were trained according to the standardized training method described. Most of the training sessions were an integral part of investigator meetings for clinical trials of novel RA therapies organized by different sponsors. Because of incomplete data or inclusion in groups of < 3 participants, 260 individuals could not be evaluated. Thus, 340 trainees in 82 groups were included, 256 (62 groups) of them being trained in the examination of the 66/68-joint count and the remaining 84 (20 groups) in the 28-joint count.

Descriptive statistics for the 66/68-joint count. The mean number of tender joints was 17.27 (± 12.88) before and 15.25 (± 12.38) after the training. The 95% CI was between 15.68 and 18.86 before and between 13.66 and 16.82 after the training. The respective numbers for swollen joints were 11.5 (± 7.23) before and 8.99 (± 6.74) after the training, with 95% CI declining from 10.61–12.39 to 8.13–9.85.

Data for the 28-joint count. The mean number of tender joints was 9.69 (± 7.32) before and 8.46 (± 7.11) after the training. The 95% CI was 8.1–11.28 before and 6.9–10.03 after the training. The respective numbers for swollen joints were 10.21 (± 5.63) before and 7.85 (± 3.15) after the training, with 95% CI declining from 8.99–11.44 to 7.16–8.55.

Figure 1 shows the frequency of positive ratings for swelling for every single joint of the 66/68-joint count after the training and gives an impression of the distribution pattern of joint involvement in the patients examined.

Kappa statistics. Agreements between assessors of single joints are shown in Figures 2A and 3A, displayed as kappa values for tenderness and swelling before the training. Comparing tenderness and swelling, there was a higher agreement for tenderness, whatever joint count was used. Looking at the dimension of swelling, there was a significantly higher agreement for the joints of the 28-joint count (circles) compared to those assessed only in the 66/68-joint count. While agreement improved with training, the differences between the 2 joint counts remained obvious.

Figures 2B and 3B show the training effect on the evaluation of every single joint of the 66/68-joint count as the difference between interrater agreement before and after the training. There was a tendency to improvement in almost all examined joints for tenderness as well as for swelling. Statistical significance (indicated by CI that separate from zero) was reached in the elbows, wrists, metacarpophalangeal (MCP1), MCP2, PIP2, PIP3, PIP4, tarsus, MTP1, and MTP4 joints for the dimension of tenderness. The agreement on swelling improved significantly in the sternoclavicular joints, wrists, MCP2, MCP4, and PIP3 joints.

DISCUSSION

Standardization of the joint examination technique significantly reduces variability between different examiners. In a previous study we showed that even after standardization training, the overall variability is significantly higher when the 66/68-joint count is used rather than the 28-joint count[9]. To investigate this discrepancy, we evaluated the results of our training seminars on individual joint levels. This offers the possibility to investigate which joints were counted positive and which were associated with the highest agreements and disagreements between different assessors.

Figure 1 shows the joints rated positive for tenderness and

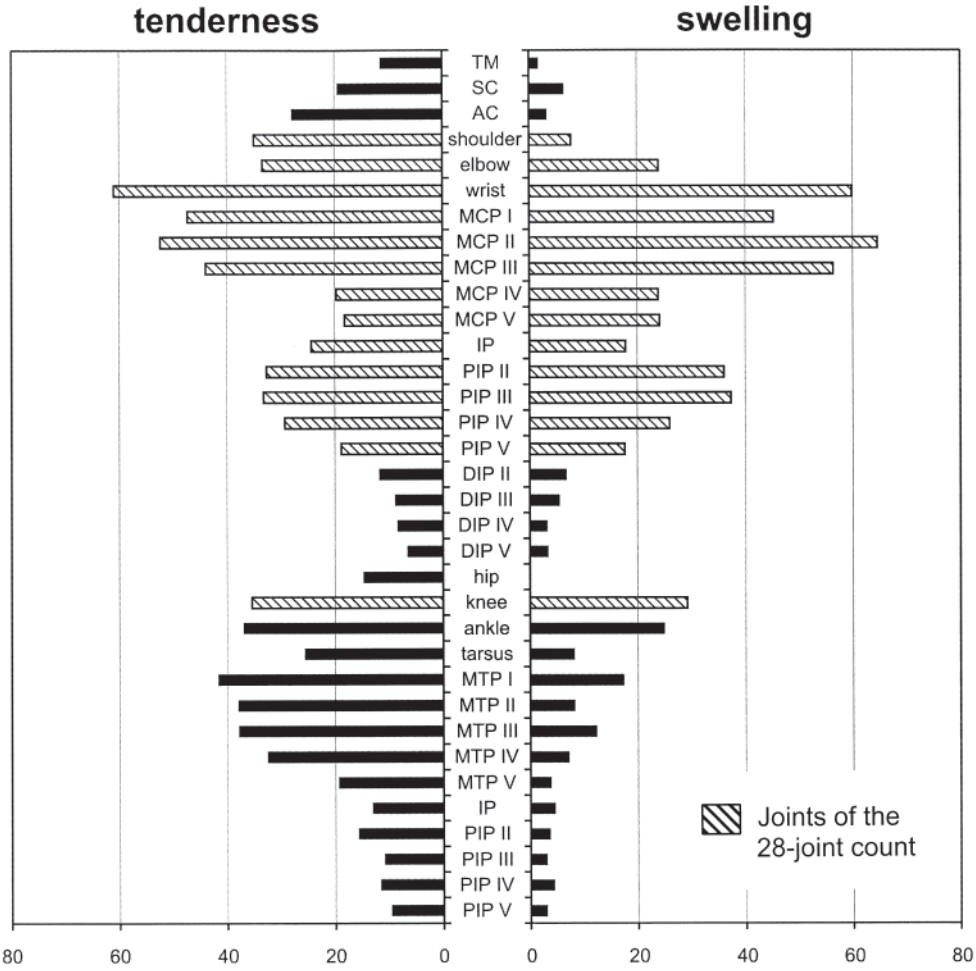

Figure 1. Frequencies of joints rated positive for tenderness and swelling in patients with moderately active RA. TM: temporomandibular; SC: sternoclavicular; AC: acromioclavicular; MCP: metacarpophalangeal; IP: interphalangeal; PIP: proximal interphalangeal; DIP: distal interphalangeal; MTP: metatarsophalangeal.

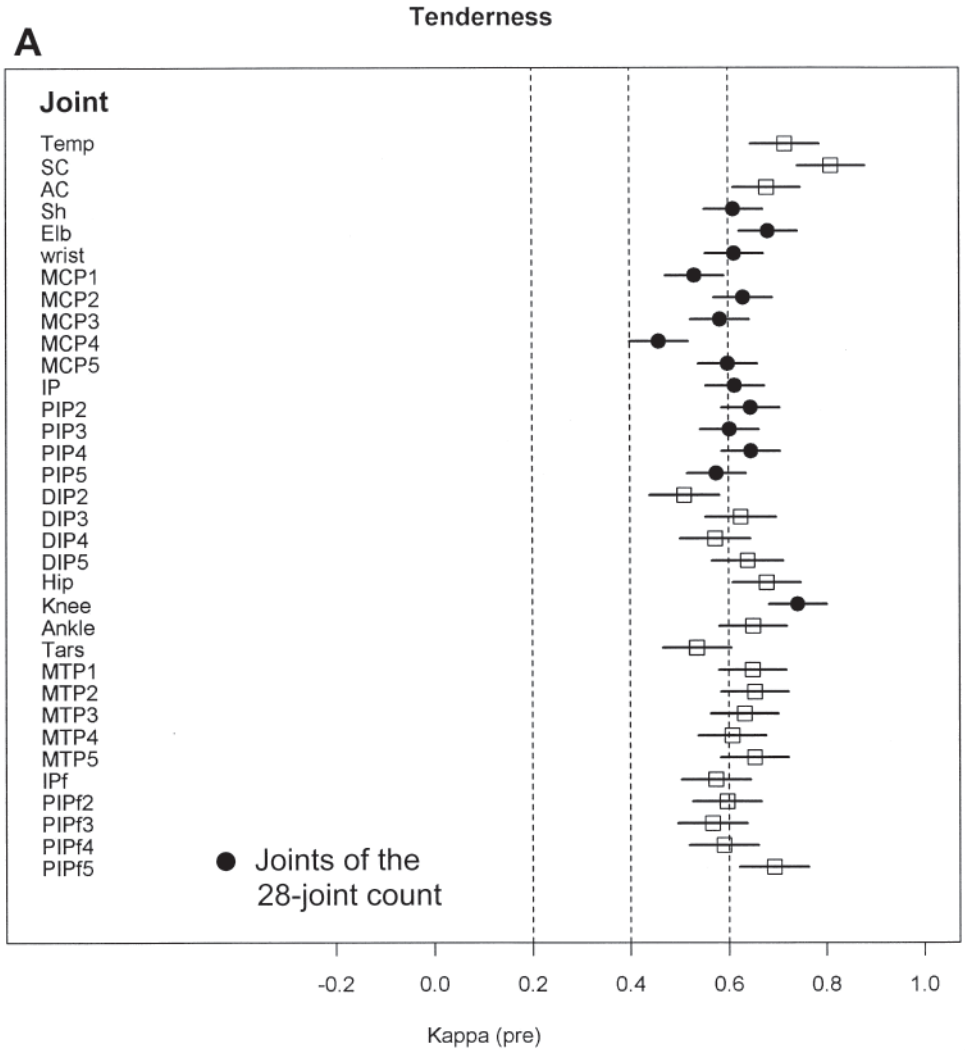

Figure 2A. Interrater agreement on joint tenderness before the standardized training. Positive values indicate improvement by training. Symbols represent mean values with 95% CI. Kappa values < 0.2 indicate poor agreement, 0.2–0.4 slight agreement, 0.4–0.6 moderate agreement, 0.6–0.8 substantial agreement, > 0.8 perfect agreement. For definitions see Figure 1.

swelling after our training intervention. To our knowledge, our work represents the highest number of experts rating the distribution of joint involvement in patients with established RA. Our findings confirm that wrists and the MCP joints I–III are the joints most affected in RA, followed by PIP joints of the hands, elbows, knees, and ankles. In the latter joints, tenderness and swelling were rated in close concordance, whereas ratings contrasted significantly in the MTP joints, the tarsus, and the shoulder. In these joints, tenderness was detected in a significant number of patients, while swelling appeared to occur very rarely.

There are different explanations for this discrepancy. One is the sensitivity of examination in these areas. In the RA population, which usually is older, edema and subcutaneous fat can severely confound physical examination, especially in the lower extremities. Concerning the shoulder, it is generally accepted that swelling is very hard to detect by physical examination and is optimally determined by the use of ultrasound[16,17]. Another explanation for the discrepancy between tenderness and swelling is that tenderness may be due to damage rather than actual synovitis. In RA it is known that destruction in the MTP joints and ankles can occur early in the course of the disease and lead to pain independently of active arthritis. The weightbearing of the lower extremities surely aggravates this situation. Damage, however, is not the dimension we want to measure with our joint counts, which are meant to detect true synovitis.

One effect of our standardization training was that the frequencies of positively rated joints decreased. An explanation for this trend, which was significant in a substantial number of joints, may be that during the training it was stressed that joints should be rated positive only when assessors were quite

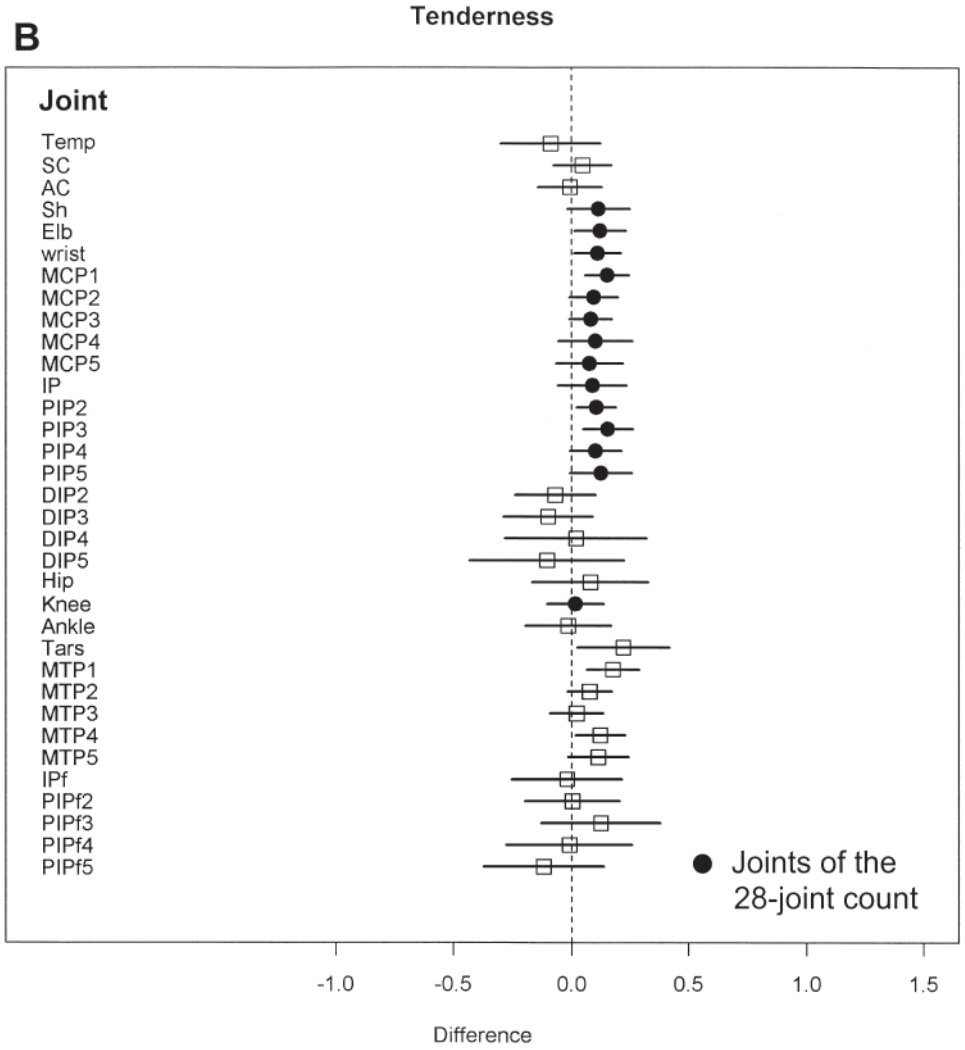

Figure 2B. Training effects on interrater agreement on joint tenderness. Positive values indicate improvement by training. For definitions see Figure 1.

sure about tenderness or swelling. We believe that this conservative approach is valuable not only for the purposes of a clinical trial but for daily practice as well, because overestimation of affected joints may lead to inappropriate treatment decisions. The training effect was similarly pronounced across all joints, in particular with no significant differences between the joints of the 28-joint count and those that are assessed only with the 66/68-joint count.

The level of agreement on tenderness was better than that for swelling. This difference is well known and most probably occurs because determination of tenderness relies upon the patient's information, while swelling must be felt solely by the assessor. Concerning tenderness, there was no substantial difference between the 28- and the 66/68-joint counts and there was a clear improvement after the standardization training. This was markedly different for the dimension of swelling. Before the training session, almost all joints that are assessed only with the 66/68-joint count reached kappa values below 0.2, i.e., insufficient interrater agreement. After the training, there was an increase to values between 0.2 and 0.4 in some of these joints. The kappa values for joints of the 28-joint count were almost all between 0.2 and 0.4 before the training and the majority improved to more than 0.4 after the training.

Altogether, there was insufficient interrater agreement and a low incidence of clinically detected synovitis swelling in the DIP joints of the hands and the small joints of the lower extremities. The same was true for the shoulders. There is no doubt, of course, that synovitis of the shoulders and the joints of the feet are important manifestations of many inflammatory arthritides including RA. In light of our data, it is questionable, however, whether these joints are valuable components of outcome measurements in rheumatology research and practice.

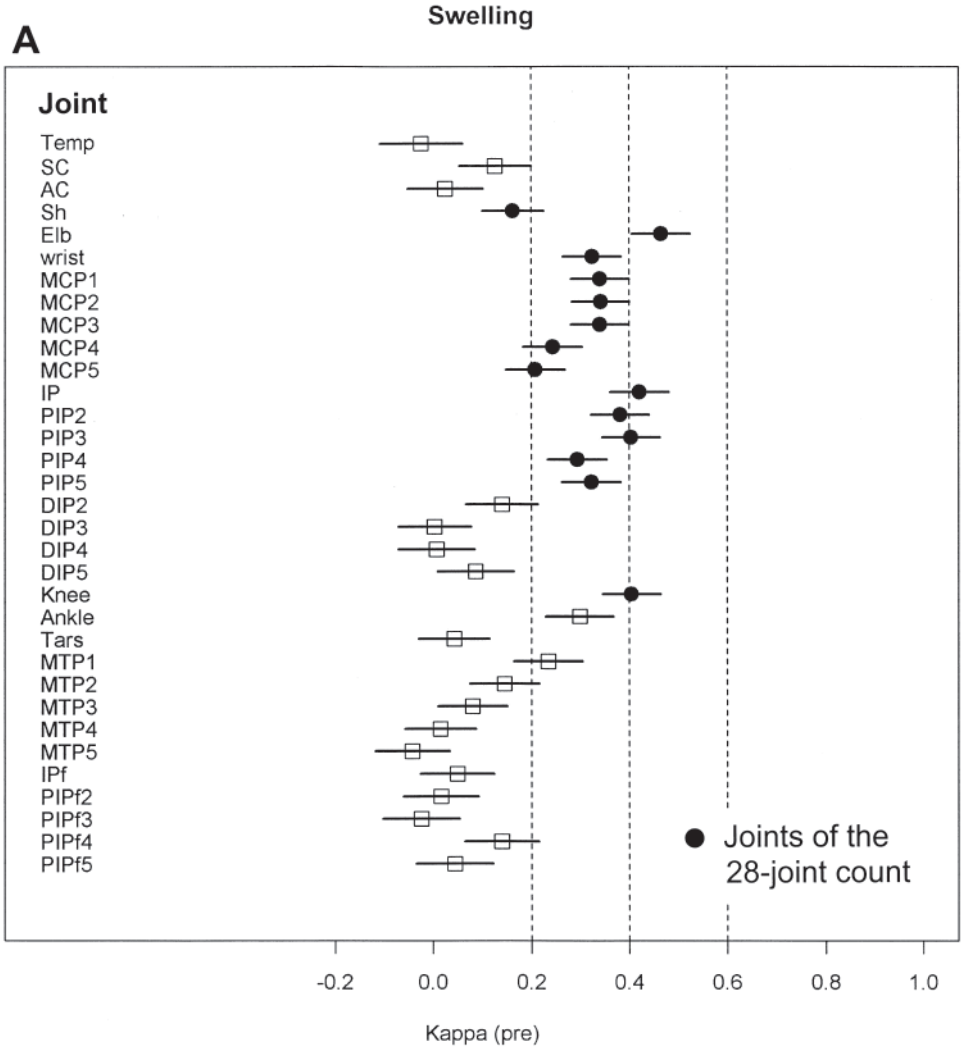

Figure 3A. Interrater agreement on joint swelling before the standardized training. Symbols represent mean values with 95% CI. Kappa values < 0.2 indicate poor agreement, 0.2–0.4 slight agreement, 0.4–0.6 moderate agreement, 0.6–0.8 substantial agreement, > 0.8 perfect agreement. For definitions see Figure 1. Sh: shoulder; Elb: elbow; tars: tarsus; PIPf: proximal interphalangeal joint of the foot.

REFERENCES

1. Felson DT, Anderson JJ, Boers M, Bombardier C, Chernoff M, Fried B, et al. The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials. The Committee on Outcome Measures in Rheumatoid Arthritis Clinical Trials. Arthritis Rheum 1993;36:729-40.
2. Scott DL, Antoni C, Choy EH, Van Riel PC. Joint counts in routine practice. Rheumatology 2003;42:919-23.
3. van der Heijde DM, van 't Hof M, van Riel PL, van de Putte LB. Validity of single variables and indices to measure disease activity in rheumatoid arthritis. J Rheumatol 1993;20:538-41.
4. van Gestel AM, Prevoo ML, van 't Hof MA, van Rijswijk MH, van de Putte LB, van Riel PL. Development and validation of the European League Against Rheumatism response criteria for rheumatoid arthritis. Comparison with the preliminary American College of Rheumatology and the World Health Organization/International League Against Rheumatism Criteria. Arthritis Rheum 1996;39:34-40.
5. Brown AK. Using ultrasonography to facilitate best practice in diagnosis and management of RA. Nat Rev Rheumatol 2009;5:698-706.
6. Sokka T, Pincus T. Quantitative joint assessment in rheumatoid arthritis. Clin Exp Rheumatol 2005;23 Suppl 39:S58-62.
7. Scott DL, Houssien DA. Joint assessment in rheumatoid arthritis. Br J Rheumatol 1996;35 Suppl 2:14-8.
8. Lewis PA, O'Sullivan MM, Rumfeld WR, Coles EC, Jessop JD. Significant changes in Ritchie scores. Br J Rheumatol 1988;27:32-6.
9. Grunke M, Antoni CE, Kavanaugh A, Heldebrand V, Dechant C, Schett G, et al. Standardization of joint examination technique leads to a significant decrease in variability among different examiners. J Rheumatol 2010;37:860-4.
10. Sokka T, Pincus T. Joint counts to assess rheumatoid arthritis for clinical research and usual clinical care: Advantages and limitations. Rheum Dis Clin North Am 2009;35:713-22, v-vi.
11. Bellamy N, Anastassiades TP, Buchanan WW, Davis P, Lee P, McCain GA, et al. Rheumatoid arthritis antirheumatic drug trials. I. Effects of standardization procedures on observer dependent outcome measures. J Rheumatol 1991;18:1893-900.

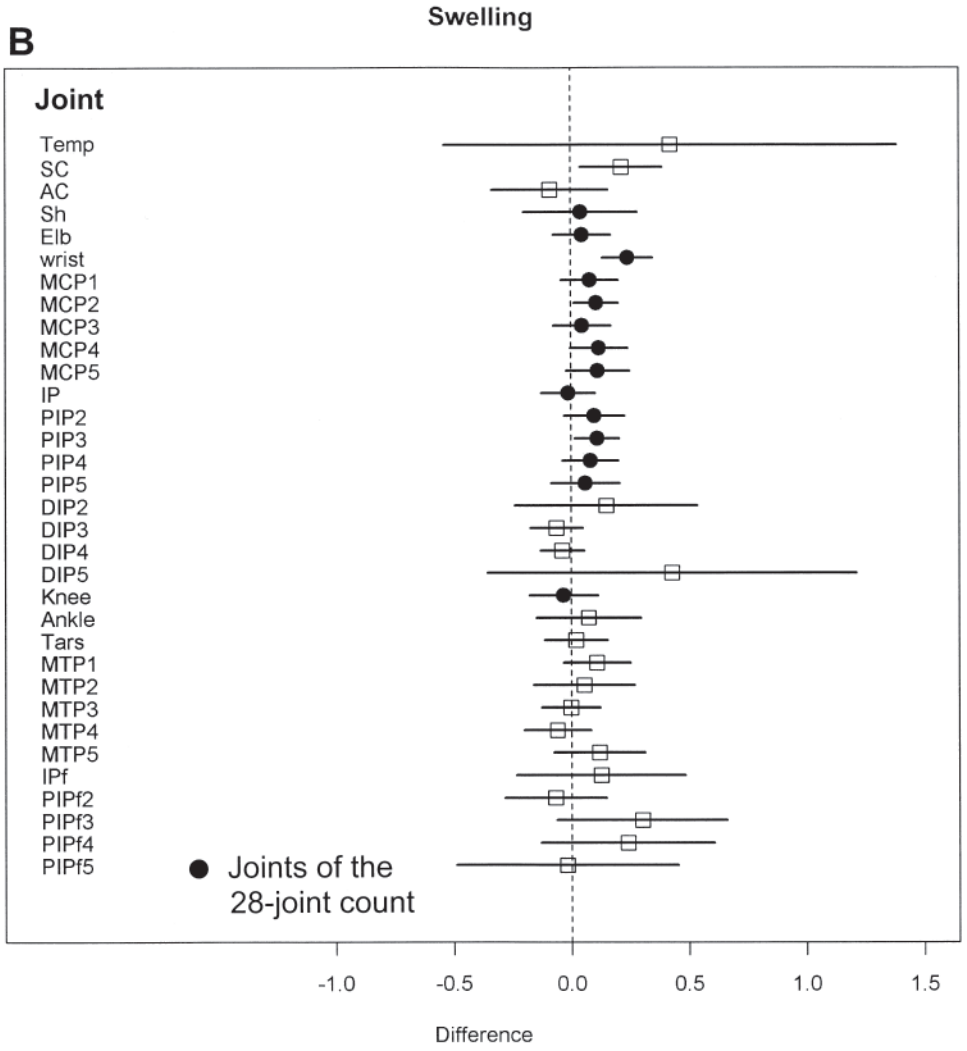

Figure 3B. Training effects on interrater agreement on joint swelling. Positive values indicate improvement by training. For definitions see Figure 1. Sh: shoulder; Elb: elbow; tars: tarsus; IPf: interphalangeal joint of the foot.

12. Klinkhoff AV, Bellamy N, Bombardier C, Carette S, Chalmers A, Esdaile JM, et al. An experiment in reducing interobserver variability of the examination for joint tenderness. J Rheumatol 1988;15:492-4.
13. Scott DL, Choy EH, Greeves A, Isenberg D, Kassinor D, Rankin E, et al. Standardising joint assessment in rheumatoid arthritis. Clin Rheumatol 1996;15:579-82.
14. Fleiss J. Statistical methods for rates and proportions. 2nd ed. New York: Wiley-Interscience Chichester; 1982.
15. Team RDC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2009. [Internet. Accessed April 24, 2012.] Available from: http://www.R-project.org
16. Kim HA, Kim SH, Seo YI. Ultrasonographic findings of the shoulder in patients with rheumatoid arthritis and comparison with physical examination. J Korean Med Sci 2007;22:660-6.
17. Luukkainen R, Sanila MT, Luukkainen P. Poor relationship between joint swelling detected on physical examination and effusion diagnosed by ultrasonography in glenohumeral joints in patients with rheumatoid arthritis. Clin Rheumatol 2007;26:865-7.