Standardization of Joint Examination Technique Leads to a Significant Decrease in Variability Among Different Examiners

MATHIAS GRUNKE, CHRISTIAN E. ANTONI, ARTHUR KAVANAUGH, VERENA HILDEBRAND, CLAUDIA DECHANT, GEORG SCHETT, BERNHARD MANGER, and MONIKA RONNEBERGER

ABSTRACT. ***Objective.*** To reduce the amount of variability among assessors, we conducted joint examination standardization seminars in conjunction with multicenter clinical trials for patients with rheumatoid arthritis (RA). The examination techniques used were based on the recommendations of the European League Against Rheumatism (EULAR).

Methods. To evaluate the effect of standardization, participants at the seminars examined a given patient with RA before and after they were made familiar with the EULAR examination technique. The number of tender and swollen joints as well as the variance among the examiners before and after the training were compared. Joints were rated positive or negative for tenderness and swelling without grading.

Results. Overall, 553 individuals from a variety of countries in Europe, North America, Asia, and Australia participated. Examiners included different kinds of health professionals, mainly physicians and nurses. We found a substantial variance among examiners before the training in the standardized method. This variance could be significantly reduced by the training. We also found that the number of joints considered active was markedly reduced after the training.

Conclusion. Standardized joint examination training significantly reduces variability among different assessors. (J Rheumatol First Release Feb 15 2010; doi:10.3899/jrheum.090195)

Key Indexing Terms:
OUTCOME ASSESSMENT | RHEUMATOID ARTHRITIS | EDUCATION
DISEASE ACTIVITY | JOINT EXAMINATION

In rheumatoid arthritis (RA), the number of affected joints is the most specific measure to determine actual disease activity. The clinically important aspects of joint inflammation, namely joint tenderness and swelling, are not always congruent and therefore have to be counted separately. The number of affected joints is crucial for both diagnostic and prognostic reasons. Joint counts are also critical components of composite disease activity measures such as the Disease Activity Score (DAS)[1]. Joint counts are the key determinants of response to therapy, for example in the European League Against Rheumatism (EULAR) response criteria[2] or the American College of Rheumatology (ACR) remission criteria[3]. The counts of tender and swollen joints are key elements of the core set of assessments defined by the ACR that are recommended for all clinical trials in RA[4] as well as for daily practice[5].

Different methods exist to count the number of involved joints. They vary in the number of joints assessed, the weighting of certain joints or joint areas, and the grading of tenderness and swelling according to their extent or just as negative or positive[6,7]. Prevoo, *et al* compared 7 of the most widely used methods and did not find substantial differences concerning reliability and validity among them[8]. The ACR 66/68-joint count, the 28-joint count, and the Ritchie Articular Index have the broadest acceptance at present. Smolen, *et al* demonstrated that the 28-joint count, which rates only joints of the upper extremities and the knees, is as sensitive and reliable as the more time-consuming 66/68-joint count, which includes the joints of the lower extremities except for the distal interphalangeal (DIP) joints of the feet[9,10].

Whatever joint count is used, there is a high degree of variability among the examinations done by a single individual and especially among different assessors[11,12]. With the emergence of newer, more effective therapies for RA,

From the Unit for Rheumatology, Ludwig-Maximilians-University, Munich; Department of Internal Medicine 3, University of Erlangen-Nuremberg, Erlangen, Germany; Novartis Pharmaceuticals Corporation, East Hanover, New Jersey; and Division of Rheumatology, Allergy and Immunology, University of California, San Diego, USA.

M. Grunke, MD; C. Dechant, MD, Unit for Rheumatology, Ludwig-Maximilians-University; C.E. Antoni, MD, Novartis Pharmaceuticals Corporation; A. Kavanaugh, MD, Division of Rheumatology, Allergy and Immunology, University of California, San Diego; V. Hildebrand, MD; G. Schett, MD; B. Manger, MD; M. Ronneberger, MD, Department of Internal Medicine 3, University of Erlangen-Nuremberg.

Address correspondence to Dr. M. Grunke, University of Munich, Internal Medicine Unit for Rheumatology, Pettenkoferstr. 8a, Munich 80336, Germany. E-mail: mathias.gruenke@med.uni-muenchen.de

Accepted for publication November 1, 2009.

and the increasing number of multicenter trials, standardization of joint examination techniques has become a matter of increasing interest[13,14]. Differences in the evaluation of affected joints may lead to errors in assessments of disease activity in given patients and can severely confound results of multicenter trials.

The most recent published data on a standardization program are from Scott, *et al*[15]. In a cohort of 8 joint assessors, who performed joint counts in the same patient before and after a standardized training, they found an increased sensitivity for detecting affected joints, but still a high degree of variability.

In order to reduce the amount of variability among assessors, we conducted joint assessment standardization seminars in conjunction with multicenter clinical trials for patients with RA. The examination technique used was based on the recommendations of the EULAR *Handbook of Clinical Assessments in Rheumatoid Arthritis*[16].

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 who specialized in rheumatology, along with study nurses and a few medical technicians and physiotherapists. All data in our evaluation were collected by 1 of 3 trainers from the same institution and using an identical training design. Trainees were divided into groups of a maximum 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[16]. Volunteer patients with RA with varying levels of active disease (i.e., nearly 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 (i.e., 0–3). Before the standardization training, participants were invited to perform the examination according to the technique they had customarily 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[16] 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 and proximal interphalangeal joints of the feet. The hips are evaluated only for tenderness, making 68 joints evaluated for tenderness and 66 joints for swelling. Each group then practiced joint-count examining for an additional 1 to 3 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 concerning changes in tender and swollen joint counts within the groups.

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

Variance was calculated within the groups assessing the same patient. For comparability of data, groups of fewer than 3 and more than 6 participants were excluded from statistical evaluation. The values for tenderness and swelling were not equally distributed (Kolmogorov-Smirnov test), because disease activity naturally differed significantly among the participating patients. Therefore, variance was calculated by the nonparametric Wilcoxon signed-rank test for paired samples.

RESULTS

Between August 2002 and November 2006, 553 individuals from a variety of countries in Europe, North 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 that were too small, 106 individuals could not be evaluated. Thus, 447 trainees in 118 groups were included, 251 (71 groups) of them being trained in the 66/68-joint count and the remaining 196 (47 groups) in the 28-joint count.

Among the 251 trainees performing a 66/68-joint count, a mean number of 18 joints was considered positive for tenderness and 10 positive for swelling (standard deviations 15 and 5, respectively). After the standardized training, these numbers decreased to 15 for tenderness and 7 for swelling (SD 15 and 5, respectively; Table 1). This decrease was highly significant ($p < 0.001$).

As the overall joint counts markedly decreased with the training, we calculated the percentage of patients who would have been considered trial-active patients, based on commonly employed inclusion criteria of at least 6 tender and 6 swollen joints before and after the training session. Of note, while 55% would have been rated as having joint counts high enough to be eligible for a study before the training, only 33% of these same patients would have been considered eligible after the training. The variance among assessors examining the same patient (3–6 trainees in 71 groups with 1 patient each) was 21 joints before and 14 after the standardization training for tenderness and 28 before and 6 after the training for swelling (Figure 1).

The 196 trainees who were trained in the 28-joint count rated 11 ± 9 (mean ± SD) joints positive for tenderness before the training. After the training, the number decreased to 10 ± 9 joints. Swelling was detected in 8 ± 5 joints before and 6 ± 4 joints after the training (Table 2). Again, the decrease among the untrained and trained assessments was highly significant ($p = 0.005$ and 0.002, respectively). "Trial-active patients" decreased from 51% before to 34% after the training.

The variances among the assessors examining the same patient (3–6 trainees in 47 groups with 1 patient each) were 7 before and 2 after the training for tenderness and 12 before and 6 after the training for swelling (Figure 2).

Table 1. Results for 251 trainees using the 66/68-joint count method, before and after standardized training.

	Pretraining Values	Posttraining Values
Number of patients (= training groups)	71	
Number of trainees	251	
Number of tender joints, mean (SD)	18 (15)	15 (15)*
Variance (pain)	21	14
Number of swollen joints, mean (SD)	10 (5)	7 (5)*
Variance (swelling)	28	6
Number (%) of evaluations with > 6 tender and swollen joints ("trial-active patients")	139 (55)	82 (33)

* $p < 0.005$.

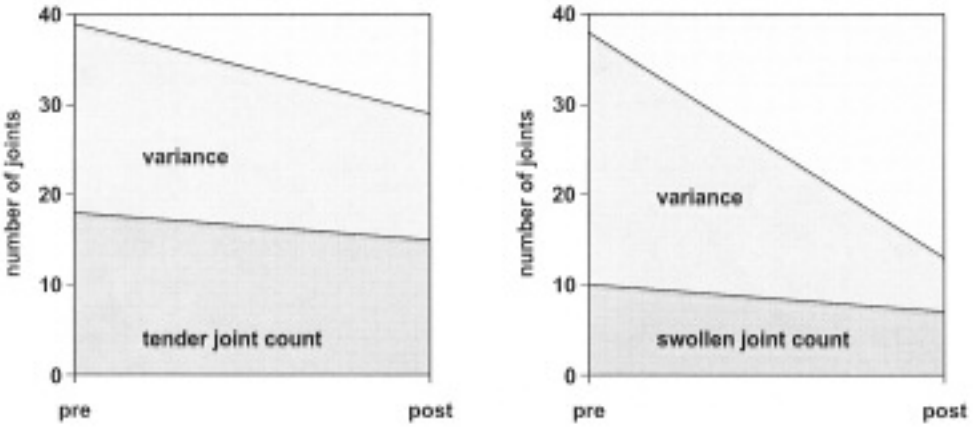

Figure 1. The 66/68-joint count: mean values and variance within groups of tender and swollen joints before and after a standardized training program, assessed by 251 assessors in 71 groups.

Table 2. Results for 196 trainees using the 28-joint count method, before and after standardized training.

	Pretraining Values	Posttraining Values
Number of patients (= training groups)	47	
Number of trainees	196	
Number of tender joints, mean (SD)	11 (9)	10 (9)*
Variance (pain)	7	2
Number of swollen joints, mean (SD)	8 (5)	6 (4)*
Variance (swelling)	12	6
Number (%) of evaluations with > 6 tender and swollen joints ("trial-active patients")	99 (51)	67 (34)

* $p < 0.005$.

DISCUSSION

In this large cohort of health professionals performing joint count assessments, we confirmed the high variability among different assessors when examining the same patients with active RA. This confirms what has been described[17]. With the standardized training method we used, the mean number of positively rated joints decreased significantly. This is in contrast to a recent publication of a standardized training, which showed an increase in the numbers of tender and swollen joints[15]. An explanation for this discrepancy may be that the training sessions described in our study were mostly part of investigator meetings for clinical trials. It is supposed that one reason for high placebo effects in clinical trials is the inclusion of patients who are not as active as required by the protocol. It was therefore stressed during the training sessions that joints should only be rated positive when assessors were sure about tenderness or swelling.

We believe that this conservative approach is valuable not only for the purpose of a clinical trial but for daily prac-

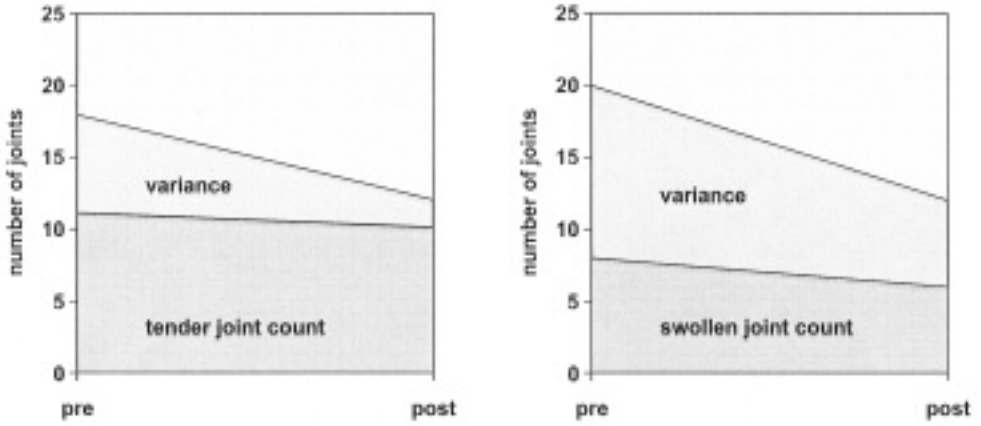

Figure 2. The 28-joint count: mean values and variance within groups of tender and swollen joints before and after a standardized training program, assessed by 196 assessors in 47 groups.

tice as well, as overestimation of affected joints may lead to inappropriate treatment decisions.

The major goal of the training sessions was to decrease variability among different assessors. This goal was reached, although there was never 100% agreement. The most consistent results were achieved with the 28-joint count, with a variance decreasing from 7 before the training to 2 for tenderness and from 12 to 6 for swelling. The results for the 66/68-joint count were comparably positive for the dimension of joint swelling, with a variance of 6 after the training in contrast to 28 in untrained assessments. The variance in tender joint counts was still somewhat high (15) after the training, although not to the extent it was before standardization. It would be interesting to see whether this higher variability in the 66/68-joint count is due just to the higher number of joints counted or to a higher disagreement in the joints of the lower extremity. As this data set reflects just the total numbers, it cannot clarify this question. Therefore, further investigation should address this issue.

Even when using the same technique, determination of whether a joint is tender or swollen is something that is likely to vary slightly among individuals. We therefore decided to compare just the disagreement or agreement within the groups of examiners instead of defining the personal experience of the trainer as the gold standard. One possibility for an objective evaluation would be the use of high-resolution ultrasound. This method can verify only the dimension of swelling. Of note, swelling has been shown to be a source of greater variability than tenderness.

Our data show that the perceptions of joint tenderness and swelling are still very different among examiners. Our report is the first to show that consistency can be substantially improved by standardization training. We therefore believe that the training of joint examination technique should be an essential component of the preparation for any clinical trial involving patients with RA or other inflammatory joint diseases.

REFERENCES

1. 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.
2. 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.
3. Pinals RS, Masi AT, Larsen RA. Preliminary criteria for clinical remission in rheumatoid arthritis. Arthritis Rheum 1981;24:1308-15.
4. 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.
5. Scott DL, Antoni C, Choy EH, Van Riel PC. Joint counts in routine practice. Rheumatology 2003;42:919-23.
6. Sokka T, Pincus T. Quantitative joint assessment in rheumatoid arthritis. Clin Exp Rheumatol 2005;23:S58-62.
7. Hart LE, Tugwell P, Buchanan WW, Norman GR, Grace EM, Southwell D. Grading of tenderness as a source of interrater error in the Ritchie Articular Index. J Rheumatol 1985;12:716-7.
8. Prevoo ML, van Riel PL, van 't Hof MA, van Rijswijk MH, van Leeuwen MA, Kuper HH, et al. Validity and reliability of joint indices. A longitudinal study in patients with recent onset rheumatoid arthritis. Br J Rheumatol 1993;32:589-94.
9. Smolen JS, Breedveld FC, Eberl G, Jones I, Leeming M, Wylie GL, et al. Validity and reliability of the twenty-eight-joint count for the assessment of rheumatoid arthritis activity. Arthritis Rheum 1995;38:38-43.
10. Fransen J, Antoni C, Mease PJ, Uter W, Kavanaugh A, Kalden JR, et al. Performance of response criteria for assessing peripheral arthritis in patients with psoriatic arthritis: analysis of data from randomised controlled trials of two tumour necrosis factor inhibitors. Ann Rheum Dis 2006;65:1373-8.

11. Scott DL, Houssien DA. Joint assessment in rheumatoid arthritis. Br J Rheumatol 1996;35 Suppl 2:S14-8.
12. Lewis PA, O'Sullivan MM, Rumfeld WR, Coles EC, Jessop JD. Significant changes in Ritchie scores. Br J Rheumatol 1988;27:32-6.
13. 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.
14. 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.
15. 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.
16. van Riel PLCM, Scott DL. EULAR handbook of clinical assessment in rheumatoid arthritis. Alphen Aan Den Rijn, The Netherlands: Van Zuiden Communications; 2000.
17. Manger B, Antoni C, Hantschel M, Kalden JR, Kavanaugh A. Standardisation if disease activity assessments in randomized clinical trials. Ann Rheum Dis 2003;62 Suppl I:S10-11.