VALIDATION OF RHEUMATOID ARTHRITIS IMPROVEMENT CRITERIA THAT INCLUDE SIMPLIFIED JOINT COUNTS

ANKE M. VAN GESTEL, CEES J. HAAGSMA, and PIET L. C. M. VAN RIEL

Objective. To study the validity of response criteria for rheumatoid arthritis (RA) that included 28-joint counts instead of more comprehensive joint counts.

Methods. In a double-blind, placebo-controlled trial of 105 patients treated with methotrexate, sulphasalazine, or both, response was evaluated at week 52. Both European League Against Rheumatism and American College of Rheumatology definitions of response, with comprehensive as well as simplified joint counts, were calculated. We studied the differences between the criteria with and without simplified joint counts, the discriminating capacity between treatment groups, and the association with change in functional capacity and joint damage.

Results. Response criteria that included 28-joint counts classified patients' responses more conservatively. No differences between treatment groups were found with either set of response criteria. The association with change in functional capacity was significant in all cases. All response criteria were significantly associated with radiographic progression of RA.

Conclusion. Improvement criteria that include 28-joint counts are as valid as the original improvement criteria that included more comprehensive joint counts.

For the last several years, researchers in the field of rheumatology have been emphasizing the need for standardized and validated measures to be used in clinical trials. Uniformity of outcome measures will improve the comparability of studies. For studies of patients with rheumatoid arthritis (RA), a core set of disease activity measures has been proposed (1). Two sets of criteria for improvement have recently been developed (2,3). Both the European League Against Rheumatism (EULAR) response criteria and the American College of Rheumatology (ACR) improvement criteria for the evaluation of antirheumatic therapy in RA can be calculated using a simplified joint count instead of an extensive joint count. Simplified, 28-joint counts for tender and swollen joints were included in the core set of disease activity measures, and their use in clinical trials was recommended. Excluding 30–40 joints without loss of useful information (when comparing groups of patients) is highly appreciated by examiners, and probably also by the patients, because of the simplified examination. However, response criteria were developed and validated using extensive joint counts, and changes in components cannot be made without validating the resulting new criteria set (4). We therefore performed the present validation study of the criteria using the simplified joint counts.

PATIENTS AND METHODS

Treatment response was examined and validated in a 52-week, double-blind, placebo-controlled trial in 105 patients with early RA (5). Patients were randomized over 3 treatment groups: methotrexate and placebo (MTX), sulphasalazine and placebo (SSZ), and methotrexate and sulphasalazine (COMBI). Outcome measures were a Dutch version of the Stanford Health Assessment Questionnaire (D-HAQ) (the Functional Index of the HAQ) (6) and radiographs of the hands and feet evaluated according to the modified Sharp method (7). The D-HAQ was measured at months 0 and 12, and radiographs were taken at months 0, 12, and 18. Treatment response was calculated using the preliminary ACR improvement criteria and the EULAR response criteria, both calculated once with an extensive (ACR, EULAR) joint count and once with a simplified (ACR28, EULAR28) joint count.

Using the ACR criteria, a patient is classified as a responder when both the tender joint count and the swollen joint count are ≥20% improved from baseline and 3 of the
In the following 5 variables are also ≥20% improved: the erythrocyte sedimentation rate (ESR; in mm/hour), physician's assessment of disease activity (0 = asymptomatic, 1 = mild, 2 = moderate, 3 = severe), patient's assessment of disease activity (same as for the physician's assessment), patient's assessment of pain (on a 100-mm visual analog scale, where 0 = no pain and 100 = worst possible pain), and functional index score from the D-HAQ (range 0.00–3.00). Some studies report the use of ACR 50% improvement criteria to specify excellent responders (8). We therefore used a 3-category ACR improvement criterion identifying nonresponders (<20% improvement), good responders (≥20%), and excellent responders (≥50%).

The EULAR criteria are based on the Disease Activity Score (DAS), which is calculated as follows:

\[
\text{DAS} = 0.53938 \times \sqrt{\text{RAI}} + 0.06465 \times \text{SW28} + 0.330 \times \ln \text{ESR} + 0.00722 \times \text{GH}
\]

where RAI is the Ritchie articular index (53 joints in 26 units, graded for tenderness on pressure, 0 = no pain, 1 = patient complains of pain, 2 = patient complains of pain and winces, 3 = patient complains, winces, and withdraws; maximum score 78), SW28 is the ungraded count of joints with swelling due to synovitis (maximum score 44), lnESR is the natural logarithm of the ESR (mm/hour; Westergren), and GH is general health, as assessed by the patient using a 100-mm visual analog scale (“How do you feel concerning your arthritis?” 0 = very well and 100 = extremely bad).

The DAS has been revised to include a simplified, ungraded 28-joint count for tenderness (TEN28) and for swelling (SW28), called the DAS28 (9), which is calculated as follows:

\[
\text{DAS28} = 0.56 \times \sqrt{\text{TEN28}} + 0.28 \times \sqrt{\text{SW28}} + 0.70 \times \ln \text{ESR} + 0.014 \times \text{GH}
\]

Because the DAS and the DAS28 have different scales, we developed in our longitudinal data set a formula to transform DAS values into DAS28 values. This formula is necessary when different Disease Activity Scores were used between, or even within, studies:

\[
\text{DAS28} = 1.072 \times \text{DAS} + 0.938
\]

Using the DAS, the DAS28, and the DAS28', we defined 3 sets of EULAR response criteria: EULAR, EULAR28, and EULAR28', respectively. A 1.2-point change in the DAS or DAS28 values from baseline was considered a statistically significant change (2 times the measurement error). High activity of disease was defined as a DAS >5.7, or a DAS28 >5.1. Low activity of disease was defined as a DAS ≤2.4, or a DAS28 ≤3.2. Good responders were patients with a significant change and low disease activity. Moderate responders were patients with a significant change and moderate/high disease activity or patients with a change of 1.2 and >0.6 (1 times the measurement error) and low/moderate disease activity. Nonresponders were the remaining patients (2). EULAR28' response, including a transformed DAS, was defined similar to the response criteria including the DAS28.

To validate the 6 sets of response criteria (ACR, ACR28, ACR50%, EULAR, EULAR28, EULAR28'), we studied (a) the differences between the improvement criteria including extensive joint counts and those including 28-joint counts, (b) the potential to discriminate between treatment groups, (c) the association with change in functional capacity, and (d) the association with progression of joint damage. The level of improvement after 52 weeks of therapy was used.

For the data analyses, the time of the last observation of an early drop-out was used as the end point measurement, and patients remained in the original treatment group. The following tests were performed separately for each response criterion. The discriminating capacity was studied using Kruskal-Wallis tests, with treatment group and response classes as variables. Kruskal-Wallis tests were used to study whether there was a difference in the relative change in functional capacity (by D-HAQ) after 12 months (change corrected for baseline value) between the response classes. The association of response with the square root (transformation to normal distribution) of change in total joint damage (radiographic progression) after 12 and 18 months compared with baseline was analyzed with analysis of variance.

**RESULTS**

The characteristics of the patient population and the results of the trial are reported elsewhere (5). The treatment groups were comparable at baseline. The median D-HAQ score at baseline was 0.78 (range 0.00–3.00), and the median Sharp score was 5 (range 0–59). ACR improvement criteria could be calculated in 103 of the 105 patients (1 patient had a D-HAQ score of 0 at baseline and another patient's D-HAQ score at month 12 was missing). EULAR response criteria could be calculated in all except 1 patient, who had no ESR measurement at baseline.

Figure 1 shows the classification of responders according to the 6 sets of criteria. Fourteen patients were classified differently with the EULAR28 criteria than with the EULAR criteria. Generally, the EULAR28 classified patients in a lower response category (10 of 14 patients). The EULAR28' criteria, although transformed, classified patients almost identically (only 2 exceptions) to the classification by the EULAR criteria. Similar to the EULAR28 criteria, the ACR28 criteria classified patients generally lower (i.e., less response to treatment) than did the ACR criteria.

When comparing treatment response in the 3 treatment groups (MTX, SSZ, COMBI), no significant differences were found using any of the improvement criteria. Examining the percentage of nonresponders per treatment group with the EULAR criteria, the SSZ group had more nonresponders (24%) than the MTX (12%) and COMBI (17%) groups. With the ACR
VALIDATION OF RA IMPROVEMENT CRITERIA

<table>
<thead>
<tr>
<th></th>
<th>EULAR28 G</th>
<th>EULAR28' G</th>
<th>ACR G 50%</th>
<th>ACR G 20%</th>
<th>ACR28 G</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>N</td>
<td>G</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td>EULAR G</td>
<td>36</td>
<td>5</td>
<td>0</td>
<td>41</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>39</td>
<td>5</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>3</td>
<td>15</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>EULAR28</td>
<td>36</td>
<td>1</td>
<td>0</td>
<td>26</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>40</td>
<td>2</td>
<td>8</td>
<td>27</td>
</tr>
<tr>
<td>N</td>
<td>0</td>
<td>6</td>
<td>14</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>ACR</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>G 50%</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>G 20%</td>
<td>36</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Sets of European League Against Rheumatism (EULAR) and American College of Rheumatology (ACR) criteria for improvement after 52 weeks of therapy. EULAR and ACR criteria included comprehensive joint counts for swelling and tenderness. EULAR28 and ACR28 criteria included the simplified 28-joint counts for swelling and tenderness. EULAR28' criteria included the DAS28' transformed data to adjust for differences in the Disease Activity Score (DAS) scales used in the EULAR and the EULAR28 criteria sets (see Patients and Methods for details). G = good response; M = moderate response; N = no response. Values are the number of patients.

criteria, the SSZ and MTX groups were comparable (both 26%), while the COMBI group had a lower percentage of nonresponders (20%).

The median D-HAQ score at the end of the study was 0.22 (range 0.00–2.44). The association of improvement with change in functional capacity after 52 weeks of therapy was highly significant for all criteria sets ($P = 0.0001$; for the ACR28 $P = 0.0002$) (Figure 2).

The median Sharp score at 12 months was 23 (range 2–120). The 3 EULAR criteria were all significantly associated with radiographic progression, as were nearly all of the ACR criteria. Among the EULAR criteria, good responders showed less progression at both 12 and 18 months ($P = 0.0001$). At months 12 and 18, progression was also less for good responders according to the ACR criteria ($P = 0.0322$ and $P = 0.0110$, respectively), the ACR28 criteria ($P = 0.0075$ and $P = 0.0005$), and the ACR 50% criteria ($P = 0.0624$ and $P = 0.0138$) (Figure 3).

DISCUSSION

In this study, we were primarily interested in knowing whether improvement criteria with extensive joint counts perform as well in evaluating clinical trials as improvement criteria with simplified joint counts. Two secondary research questions were (a) whether a mathematical transformation of the DAS into the DAS28 would influence the response results, and (b) whether an ACR criteria set with 3 categories (<20%, ≥20%, and ≥50% improvement) would be as valid as the original 2-category criteria set.

When comparing the improvement criteria with comprehensive and simplified joint counts, we found that classification differences were generally explained by a lower response classification with the 28-criteria set. Any change in tenderness or swelling of the joints of the feet will be missed when the 28-joint counts are used, since these joints are not included. This exclusion will make the joint counts less sensitive to change, which might result in a lower response classification.

Do these differences influence the validity of the improvement criteria? The EULAR28 response definition showed results equal to those of the EULAR criteria in comparing treatment groups and in the association with functional capacity and radiographic progression. The chance that there was no difference in radiographic progression between good responders and nonresponders with the ACR28 criteria ($P = 0.0075$) was somewhat smaller than with the ACR criteria ($P = 0.0322$). With two-way analysis of variance (including both the ACR and the ACR28 criteria in the model with radiographic progression), the values were $P = 0.0064$ and $P = 0.0268$, respectively.

The discriminating potential of the response cri-
The relative change in functional capacity was calculated as the percentage of change in the Dutch Health Assessment Questionnaire (HAQ) score at month 12 compared with baseline. Bars show 10–90% of the group; squares show the median (50%); lines to the left show 25% of the group; lines to the right show 75% of the group. ACR50% = ACR improvement criteria defining 3 classes: ≥50% improvement, ≥20% improvement, and <20% improvement; E = excellent response. See Figure 1 for other definitions.

The percentage of nonresponders according to both the EULAR and the ACR criteria seemed to differ with respect to the MTX group (12% and 26%, respectively), while the percentages for the other 2 treatment groups were comparable. Of the 9 ACR nonresponders, 5 were classified as moderate responders using the EULAR criteria. Apparently, the MTX nonresponder group (ACR) contained more partial responders than did the SSZ or COMBI nonresponder groups. Is this partial response clinically relevant, or is it only a result of a perhaps too-sensitive EULAR response classification? The actual sensitivity and specificity of the response...
VALIDATION OF RA IMPROVEMENT CRITERIA

With both the ACR and EULAR criteria, there is always the question about "false-negative" and "false-positive" results. The ACR criteria were developed to distinguish between active treatment and placebo. The placebo group is biased by both regression to the mean and the use of prednisone, resulting in decreasing disease activity. Therefore, it is very likely that the "no response" group will comprise a number of "false-negative" patients. This might impair the sensitivity of the ACR criteria for discriminating between active treatments. On the other hand, the EULAR criteria might be too sensitive by including "false-positive" patients in the moderate response group. This question still needs to be addressed in studies with a longer follow-up to see whether the difference in outcome between EULAR moderate responder and nonresponder groups will become more pronounced.

We studied the association of improvement with change in D-HAQ score. A relative change was chosen for 2 reasons. First, despite high initial disease activity in this patient population, baseline D-HAQ scores showed considerable variation. Second, patients with high D-HAQ scores at baseline have more potential to achieve higher absolute changes than patients with low initial values. An absolute change of only 0.50 means maximal improvement in a patient with a D-HAQ score of 0.50 at baseline, while it means only slight improvement in a patient with a baseline D-HAQ score of 2.00. The relative change in functional capacity was used as a different measure of change in disease activity/treatment response. Since the D-HAQ is a component of the ACR improvement criteria, this might influence the results. Therefore, we calculated improvement while excluding the D-HAQ score from the ACR criteria. Four patients were classified differently: with the inclusion of the D-HAQ score they were nonresponders; without the D-HAQ score they were good responders. The association of these ACR criteria (without the D-HAQ component) with relative change in D-HAQ scores from baseline was still significant at \( P = 0.0018 \).

Originally a Disease Activity Score was developed that included the Ritchie articular index, a graded index examining 53 joints, and a swollen joint count based on 44 joints. Because of the advantages of a (ungraded) 28-joint count for tenderness and swelling, a DAS using these simplified joint counts was developed \( (9) \). This latter DAS (DAS28) gives higher values than the original DAS. For comparability of studies, a transformation formula for the DAS into the DAS'28' was developed using linear regression. Since the correlation coefficient of this regression model was \( <1.0 (r = 0.93) \), we studied the influence of this transformation on response results. The EULAR'28' (with the DAS transformed into the DAS'28') classified patients almost identically to the EULAR criteria (with the original DAS). Both the EULAR and EULAR'28' differed from the EULAR28 in 14 patients. The results of the EULAR'28' with regard to discriminating between treatment groups, association with change in D-HAQ scores, and radiographic progression were comparable with those of the EULAR and EULAR28 response criteria. The transformation therefore does not seem to interfere with the validity of the response criteria.

The ACR preliminary criteria for improvement classify response dichotomously. However, there seems to be a need for further specification of response, since several studies have used ACR 50% improvement criteria to define excellent responders \( (8) \). Because of this need, we included the 3-class ACR improvement criteria \( (ACR50\%) \) in our study. EULAR and ACR criteria differ in many respects: the way they were developed and validated, the disease activity variables that are included, the definition of improvement (percentage of change versus absolute change and level of disease activity), and the number of response classes. By excluding 1 of these differences, a comparison between the EULAR and the ACR improvement criteria might be more sensible. The ACR50\% criteria performed as well as the ACR criteria concerning its discriminating capacity and association with change in D-HAQ scores and radiographic damage. Although both the EULAR and the ACR improvement criteria were significantly associated with radiographic progression, the different distribution of joint damage between the response classes \( (\text{Figure 3}) \) seemed to be more pronounced for the EULAR response criteria than for all sets of ACR improvement criteria. An explanation for this observation might be the inclusion of the level of disease activity attained in the EULAR response definition.

Finally, we would like to mention that when using 1 of these improvement criteria sets (EULAR or ACR), trial inclusion criteria are required. With the EULAR criteria, a DAS >2.2 (DAS28 >3.2) at the beginning of the study will be necessary, since a DAS <1.0 (DAS28 <2.0) indicates the absence of disease activity, and a change of 1.2 should be possible. With the ACR criteria, all included variables should be >0 at baseline, since dividing by zero (to calculate relative change) is not possible. In our experience, the D-HAQ score at
Baseline is the main problem with calculating ACR improvement.

Response criteria that include simplified joint counts generally classify patients' responses lower than do response criteria that include comprehensive joint counts. No difference between the treatment groups was found with either set of response criteria. The association with change in function was significant in all cases. The response criteria were significantly associated with radiographic progression. The validity of the EULAR response criteria is not influenced by including a transformed DAS. ACR criteria with 3 classes perform as well as the original 2-group ACR criteria. For future studies we recommend the use of the EULAR or the ACR criteria that include the simplified 28-joint counts.

REFERENCES