Limitations of a Quantitative Swollen and Tender Joint Count to Assess and Monitor Patients with Rheumatoid Arthritis

Theodore Pincus, M.D.

Abstract

A quantitative count of swollen and tender joints is a primary measure to assess patients with rheumatoid arthritis (RA), and is weighted of higher value than the other 5 Core Data Set measures in all indices in which it is included. However, a number of limitations of the swollen and tender joint count have been described in the rheumatology literature, including poor reproducibility, with a requirement to be performed by the same observer at each visit; likelihood to improve with placebo treatment as much or more than the other 5 RA Core Data Set measures in clinical trials; improvement over 5 years in clinical care, while joint damage and functional disability may progress; and lower sensitivity to detect inflammatory activity than ultrasound. Most visits to a rheumatologist do not include a formal quantitative joint count. It may be suggested that a careful qualitative joint count, supplemented by quantitative patient self-report questionnaire scores, may be more than adequate to monitor and document changes in patient status in busy clinical settings.

A quantitative count of swollen and tender joints has been a primary measure to assess patients with rheumatoid arthritis (RA) for decades. An abnormal joint examination reflects pathogenetic mechanisms of inflammation and is the most specific measure to establish a diagnosis of RA, particularly with a characteristic distribution. The number of swollen and tender joints is regarded as the most important measure for RA clinical trials to distinguish active from control treatments, and as the best measure of status in usual clinical care.

Nonetheless, a quantitative joint count cannot serve as the only measure—a “gold standard”—to assess and monitor all individual patients with RA. One patient may have many swollen joints but little pain, while another patient may have considerable pain and few swollen joints; however, both may receive identical treatments. Therefore, the joint count has been incorporated into pooled indices to assess individual patients. RA indices may also include laboratory tests, patient self-report questionnaire measures, and global estimates by the physician and patient.

The preeminence of the joint count has been accepted by the rheumatology community, based in large part on its specificity and reflection of pathogenetic mechanisms. The major pooled indices and criteria for improvement of patients with RA, based on the Core Data Set, the disease activity score (DAS), and a clinical disease activity index (CDAI), regard a joint count as more important than the other 5 Core Data Set measures.

Even so, a number of limitations of the joint count have been described in the rheumatology literature (which are seen with any quantitative measure) (Table 1). Recognition of these limitations may be of value in efforts to improve measurement in rheumatic diseases and are summarized here.

1. Joint counts are poorly reproducible. Joint counts appear poorly reproducible in formal studies, although reproducibility can be improved with training. A series of early studies established that grading of joint tenderness in a Ritchie articular index, or swelling or tenderness in a standard ACR index, were characterized by substantial error rates. Therefore, gradations of abnormalities, such as mild, moderate, or severe, are not used,
and joints are scored simply as “normal” or “abnormal” for swelling, tenderness, or both. Although this approach reduces the possible numerical range of the joint count, it is necessary for acceptable reliability.

Lassere and colleagues reported results of examinations of 10 patients by two observers, with evidence of interclass correlation coefficients (ICCs) of 0.52 for 28 and 68 swollen joint counts, 0.64 for a 28 tender joint count, and 0.81 for a 74 tender joint count. These values generally were lower than seen for patient self-report questionnaire and radiographic scores, as ICCs for health assessment questionnaire (HAQ) function, pain, and patient global estimate ranged from 0.75 to 0.91 (Table 2). Their report also included a careful review of the literature, indicating intraclass ICCs of 0.67 to 0.95, and interclass ICCs of 0.48 to 0.85 for tender joint count or Ritchie index (a 14 joint count reported by Hanson and coworkers included an ICC of 0.15) (Table 3). By contrast, ICCs of health assessment questionnaire results ranged from 0.91 to 0.96, including Portuguese and French translations. Lassere and associates commented that “a cursory look at these results is sobering. Clinical measures assumed for years to be robust are shown to be much more unstable, more noisy than one would have expected.”

In view of intrinsic variation among observers in joint counts scores, several analyses of the potential advantages of training of the rheumatologists or other assessors to standardize joint count scores have been reported. Klinkhoff and colleagues found that percent variation in joint counts among six rheumatologists, who independently examined six patients before and after standardization, was reduced from 16.8% before to 3.2% after standardization. Scott and coworkers found mean coefficients of variation of 82% for swollen joints and 66% for tender joints prior to training, which were reduced to 59% for swollen joints and 65% for tender joints after training. These observers commented that “large inter-observer differences are commonly seen when clinical measurements are critically studied,” and that “standardization appears essential for clinical studies.”

Table 1  Limitations of Joint Counts

1. Joint counts are poorly reproducible.15-27
2. The need for the same observer to perform a joint count at baseline and each subsequent visit excludes monitoring by several health professionals, unlike most standard clinical measures.
3. Joint count measures are at least as likely or more likely to improve with placebo treatment than the other 5 RA Core Data Set measures.29
4. Joint counts have similar or lower relative efficiencies than global and patient measures to document differences between active and control treatments in clinical trials,29,31-40
5. Joint counts may improve over 5 years, while progressive joint damage and functional disability may occur.45
6. Joint counts are not as sensitive to detect inflammatory activity as ultrasound.63-65
7. Most visits to a rheumatologist include a careful joint examination, but do not include a formal joint count.66

Table 2  Reliability and Reproducibility of Joint Examinations (Interobserver Reliability) and Patient Self-Report Clinical Measures (Test-Retest Reliability)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Scale</th>
<th>Mean</th>
<th>ICC</th>
<th>Mean Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Examinations</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tender 28</td>
<td>0 – 28</td>
<td>5.3</td>
<td>0.64</td>
<td>0.9</td>
</tr>
<tr>
<td>Tender 74</td>
<td>0 – 74</td>
<td>10.4</td>
<td>0.81</td>
<td>2.3</td>
</tr>
<tr>
<td>Swollen 28</td>
<td>0 – 28</td>
<td>5.1</td>
<td>0.52</td>
<td>1.0</td>
</tr>
<tr>
<td>Swollen 68</td>
<td>0 – 68</td>
<td>6.7</td>
<td>0.52</td>
<td>0.6</td>
</tr>
<tr>
<td>Patient Self-Report Measures</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HAQ physical function</td>
<td>0 – 3</td>
<td>1.2</td>
<td>0.91</td>
<td>-0.05</td>
</tr>
<tr>
<td>Pain</td>
<td>0 – 100</td>
<td>37</td>
<td>0.75</td>
<td>-4.8</td>
</tr>
<tr>
<td>Patient global</td>
<td>0 – 100</td>
<td>37</td>
<td>0.75</td>
<td>-4.7</td>
</tr>
</tbody>
</table>

of quality of care.

A simple problem arises if the regular treating rheumatologist is not available on a given day. A joint count by another rheumatologist might vary substantially, and, therefore, not be directly comparable to the previous joint count by the regular rheumatologist. In addition, possible collaborative care between rheumatologists and family practitioners or other health professionals to manage patients is limited, as joint counts cannot be compared reliably.

A need for the same observer for an accurate joint count, to compare from one time point to another, indicates a poor method to ascertain changes in clinical status over time. A contrast can be drawn with patient self-report measures, which involve only a single observer, and which can be ascertained by any health professional. Kirwan and coworkers have documented that patients may be monitored with a periodic HAQ, without regular office visits, with similar outcomes to usual care, although this approach may not be optimal for some patients. This type of management may be increased in the future, which is particularly important in view of an apparent, growing worldwide shortage of rheumatologists, but cannot be conducted according to current recommendations emphasizing joint counts. It may be desirable for rheumatology as a specialty to move away from a measure that, although advocated for more than 50 years, has not been implemented clinically in most usual care settings.

3. Joint count measures are at least as likely or more likely to improve in clinical trials with placebo treatment than the other 5 RA Core Data Set measures. The natural history of RA is that of a progressive disease. However, in clinical trials, patients who receive placebo or control treatment almost invariably show improvement according to swollen and tender joint counts as great as or greater than according to other Core Data Set measures.

One of the last RA clinical trials conducted with a placebo [without “background” methotrexate or other disease-modifying antirheumatic drug (DMARD) treatment] compared leflunomide or methotrexate to placebo; the change from baseline to endpoint was analyzed for each of the 7 Core Data Set measures. An improvement in patient status of 21.4% was seen for swollen joint count, 20.3% for tender joint count, 11.7% for physician global estimate of status, and 11.6% for patient global estimate of status, while a worsening of status of 20.4% was seen for pain, 9.3% for function, and 21.5% for erythrocyte sedimentation rate (ESR) (Fig. 1).

These data are typical of analyses of other clinical trials, which indicate that swollen and tender joint counts are likely to improve with placebo or control treatment as much as or more than any other measure. These observations suggest a possible unconscious (non-deliberate) bias on the part of an assessor to identify more abnormal joints to meet inclusion criteria at screening, and fewer abnormal joints after 4 to 12 months of observation in a clinical trial. The data are consistent with poorer reliability of joint count measures, compared to patient self-report and laboratory measures.

4. Joint counts have similar or lower relative efficiencies than global and patient measures to document differences between active and control treatments in clinical trials. One consequence of apparent improvement of patient status associated with control treatment is that
joint counts have similar or lower relative efficiencies to distinguish between patients who receive placebo versus control treatment, compared to other Core Data Set measures. In an analyses of four adalimumab clinical trials (with arithmetic and percentage changes, resulting in 8 comparisons), the measures with the three highest relative efficiencies among the 7 RA Core Data Set measures included swollen joint count in 2 of 8 comparisons, and tender joint count in 0 of 8 comparisons, physician-assessor global estimate in 8 of 8 comparisons, patient global estimate in 4 of 8, C-reactive protein (CRP) in 4 of 8, physical function in 3 of 8, and pain in 3 of 8 comparisons (Fig. 2). Therefore, relative efficiencies for joint count measures were in similar ranges or lesser than for global, patient self-report, and laboratory measures to distinguish active from control treatments.

The magnitude and rank of relative efficiencies of the 7 Core Data Set measures vary in different clinical trials. The details concerning this variation are considerably less important than the evidence that joint count measures are not associated with greater relative efficiencies to distinguish active from control treatments, compared to the other 5 Core Data Set measures. These findings reinforce the rationale for a pooled index of multiple measures to assess patients with RA. The data further suggest that weighting of joint count measures at higher levels than other Core Data Set measures in ACR criteria, DAS, and CDAI is not supported by statistical evidence, although this weighting may add to specificity, comprehensiveness, and biological plausibility of clinical trial results.

5. Joint counts may improve over 5 years in usual clinical care while progressive joint damage and functional disability may occur. The numbers of affected joints may be improved over 5 to 15 years, while joint damage and functional declines are seen, leading to work disability and premature death. In one study, the mean number of swollen joints on a 28 joint count was 10.8 at

![Figure 1](image1.png) **Figure 1** Mean of the individual percentage changes from baseline in patient- and physician-reported measures for 92 placebo patients at 6 months (trial MN301). The arrow to the left denotes improvement, indicating that deterioration in pain, physical function, and laboratory assessments, was reported with placebo treatment. CRP data exclude one outlier subject in the placebo group (percentage change more than 2 standard deviations from the mean). SJC: swollen joint count. TJC: tender joint count. MDGL: physician global estimate of status. PTGL: patient global estimate of status. VAS: visual analog scale. HAQ: health assessment questionnaire. HAQ DI: HAQ disability index. ESR: erythrocyte sedimentation rate. CRP: C-reactive protein. (Reproduced from: Strand V, Cohen S, Crawford B, et al. Rheumatology. 2004;43:640-7. © Oxford University Press Journal. The British Society of Rheumatology. With permission.)

![Figure 2](image2.png) **Figure 2** Relative efficiencies of the 7 Core Data Set measures to distinguish adalimumab from control treatment in four clinical trials according to arithmetic and percentage changes. (Reproduced from: Pincus T, Amara I, Segurado OG, et al. J Rheumatol. 2008;35:204. © The Journal of Rheumatology Publishing Company Limited. With permission.)
The mean number of tender joints was 10.6 at baseline and 9.6 at 5-year review, with an effect size of 0.14. By contrast, the mean number of joints with limited motion increased from 6.7 to 10.6, with an effect size of -0.55, and mean number of deformed joints increased from 5.1 to 8.1, with an effect size of -0.50, indicating progression. Mean radiographic scores over the 5-year period indicate progression over the 5-year period, consistent with joint deformity scores and indicating progression of damage (Fig. 3).

These data indicate that improvement in a count of tender or swollen joints, or both, even at a 20% level, may nonetheless be associated with further joint damage over time, detected as limited motion or deformity, or both, on physical examination or on a radiograph.56,57 Reduction of swollen and tender joint counts by 20% does not assure the absence of disease progression. These observations have contributed to a rationale for current approaches to aim for “tight control” of patient status, with a goal of remission.58-62 Although these observations may not necessarily be regarded as a limitation of a swollen and tender joint count, earlier interpretations that a statistically significant 20% improvement versus placebo indicated an important treatment effect are not supported by long-term observations. It would appear desirable that joint counts for any study longer than 1 year should include baseline limited motion and deformity data for evaluation of possible future joint damage.

6. Joint counts are not as sensitive to detect inflammatory activity as ultrasound. A number of studies indicate that formal swollen and tender joint counts are not as sensitive as ultrasonography to detect synovitis. Kane and colleagues analyzed 44 knees at 130 sites for suprapatellar bursitis, knee effusion, or Baker’s cyst.63 Overall, an abnormality was detected at 54 of the 130 sites (42%) by ultrasound, compared to 36 of 130 (28%) detected by clinical examination, including 16% for suprapatellar bursitis, 61% versus 36% for knee effusion, and 23% versus 4% for Baker’s cyst. Karim and colleagues found that ultrasound had a higher sensitivity (98% vs. 85%), specificity (88% vs. 25%), and accuracy (97% vs. 77%) to detect synovitis in the knee, compared to a clinical examination, documented by arthroscopy.64 Wakefield and coworkers reported on 644 painful joints in 80 patients with early untreated oligoarthritis involving fewer than five joints.65 Synovitis was documented by ultrasound in 147 (79%) of 185 joints with clinical synovitis, all of which were painful. However, ultrasound also detected synovitis in 150 (33%) of 439 joints that were painful but not swollen, and 107 (13%) of 826 joints that were clinically normal (Fig. 4). The investigators suggested that patients with apparent oligoarthritis may have more joints with apparent synovitis than is detectable by clinical examination. These studies indicate that tender and swollen joint counts may be limited for optimal detection of synovitis or of responses to
Across all routine visits of patients with RA under your care (not including clinical trials), what % of these visits includes a formal tender and swollen joint count?

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Never</td>
<td>13%</td>
</tr>
<tr>
<td>1–24% of visits</td>
<td>32%</td>
</tr>
<tr>
<td>25–49% of visits</td>
<td>11%</td>
</tr>
<tr>
<td>50–74% of visits</td>
<td>14%</td>
</tr>
<tr>
<td>75–99% of visits</td>
<td>16%</td>
</tr>
<tr>
<td>Always</td>
<td>14%</td>
</tr>
</tbody>
</table>

**Figure 5** Self report responses for performance of formal tender and swollen joint counts by 550 rheumatologists from 17 European countries. Note: 45% of rheumatologists indicate that fewer than one-fourth of visits include a formal joint count, while only 30% include a formal joint count in 75% or more of visits. (Reproduced from: Pincus T, Segurado OG. Ann Rheum Dis. 2006;65:820-2. © BMJ Publishing Group. With permission.)

therapy, possibly explained in part on the basis of poor reliability, as discussed above.

7. **Most visits to a rheumatologist include a careful joint examination, but do not include a formal joint count.** Rheumatologists have been taught extensively that a formal quantitative joint count should be included in each visit of an RA patient. However, a careful joint examination, but not a formal quantitative joint count, is performed at most visits of patients with RA. Even most European rheumatologists who were investigators in clinical trials do not perform a formal joint count at most patient visits (Fig. 5).

Formal joint counts, in contrast to a brief qualitative estimate of joint inflammation, are rather tedious to perform, requiring about 1.5 minutes, during which time normal conversation between the patient and rheumatologist is interrupted. A joint count may thus occupy up to 10% of a 15 to 20-minute visit, during which it might be preferable for the rheumatologist and patient to discuss drug therapy, psychosocial matters, work issues, and other concerns, rather than to obtain a poorly reproducible, detailed formal joint count. By contrast, a RAPID3 (routine assessment of patient index data) score, which includes only patient self-report measures of physical function, pain, and global status, requires about 10 seconds. RAPID3 is at least as informative to distinguish active from placebo treatments in clinical trials, and as informative as the DAS in clinical care. It may be that a careful qualitative joint count to confirm patient self-report data, including a possible rheumatoid arthritis disease activity index (RADAI) self-report joint count, is adequate in busy usual clinical care.

**Summary and Conclusions**

This brief overview has summarized some limitations of a formal swollen and tender joint count as a measure of inflammation and clinical status in patients with RA. It should be emphasized that all measures include limitations, and that all measures in RA are surrogates for direct measurement of pathogenetic cytokines associated with inflammatory activity. Nonetheless, the data suggest that the preeminence of the joint count in RA over patient questionnaire and global measures is not justified on the basis of statistical evidence, although, as noted above, may be appropriate, on the basis of specificity, comprehensiveness, and biological plausibility.

It is recommended that a formal joint count be included in clinical trials, to provide specificity of treatment effects that is not available from patient self-report measures, laboratory tests, or global measures. However, in busy clinical settings, a careful joint examination, complemented by quantitative measures from a patient self-report questionnaire, may provide sufficient and perhaps even more prognostic quantitative data. Of course, inclusion of patient questionnaires for all patients as a regular practice in the infrastructure of clinical care may be supplemented by formal quantitative joint counts, if desired.

A reassessment of the place of a joint count in rheumatology care might be considered. This reassessment might recognize limitations of the joint count, including poorer reproducibility than patient measures, and the requirement that it be performed by the same observer for reliable comparison. At this time, both a formal joint count and patient questionnaire are included at all patient visits for clinical trials, but neither is included at most visits in usual care. Rather than continued insistence on a formal quantitative joint count, which is neglected by rheumatologists unless required, perhaps advocacy of patient questionnaires will advance measurement to improve care for patients with RA and other rheumatic diseases.

**Disclosure Statement**

The development of this article was supported in part by the Arthritis Foundation, the Jack C. Massey Foundation, and Health Report Services, Inc.

**References**

5. Pincus T. The DAS is the most specific measure, but a patient questionnaire is the most informative measure to assess rheumatoid arthritis. J Rheumatol. 2006;33:834-7.


66. Pincus T, Segurado OG. Most visits of most patients with rheumatoid arthritis to most rheumatologists do not include a formal quantitative joint count. Ann Rheum Dis. 2006;65:820-2.


