Responsiveness of the WOMAC osteoarthritis index as compared with the SF-36 in patients with osteoarthritis of the legs undergoing a comprehensive rehabilitation intervention

F Angst, A Aeschlimann, W Steiner, G Stucki

Abstract

Objective—To compare the responsiveness of the condition-specific Western Ontario and McMaster Universities osteoarthritis (OA) index (WOMAC) and the generic Short Form-36 (SF-36) in patients with OA of the legs undergoing a comprehensive inpatient rehabilitation intervention.

Methods—A prospective follow up study of consecutively referred inpatients of a rehabilitation clinic was made. The patients included fulfilled the American College of Rheumatology criteria for knee or hip OA and underwent both passive and, particularly, active physical therapy for three to four weeks. Responsiveness assessment was performed using the standardised response mean (SRM), effect size, and Guyatt’s responsiveness statistic between admission and discharge (end of rehabilitation) and then again between admission and three months later. For pain and function the SRMs were stratified by sex and OA joint. Effects were tested by the t test and SRMs of different scales were compared by the jack knife test.

Results—At the three month follow up, complete data were obtained for 223 patients. In general, the three responsiveness statistics showed a similar order of responsiveness. For both instruments, the pain scales were more responsive than the function scales. The responsiveness of the pain scale of both instruments was comparable (SRM=0.723 for WOMAC and SRM=0.528 for SF-36 at the end of rehabilitation; SRM=0.377 for WOMAC and SRM=0.468 for SF-36 at the three month follow up). In the measurement of function, the WOMAC was significantly more responsive than the SF-36 (SRMs, end of rehabilitation: 0.628 v 0.249; three month follow up: 0.235 v −0.001). Responsiveness tended to be higher in women and in knee OA than in men and hip OA.

Conclusions—Both instruments, the WOMAC and the SF-36, capture improvement in pain in patients undergoing comprehensive inpatient rehabilitation intervention. Functional improvement can be detected better by the WOMAC than by the SF-36. All the other scales of both instruments were more weakly responsive.

Comprehensive assessment of patients with osteoarthritis (OA) of the legs includes both measurement of impairment—for example, symptoms and disability, and health related quality of life. Health related quality of life may be measured by condition-specific and generic health status questionnaires.

Disease-specific instruments are useful for measuring clinically important changes in response to treatments. Because these instruments include measurement of symptoms and abilities most relevant to a particular disease, they are usually better able to detect subtle improvements in health status than generic health status instruments. The most widely used condition-specific instruments for the assessment of hip or knee OA is the Western Ontario and McMaster Universities OA index (WOMAC), which is recommended by the OMERACT (Outcome Measures in Rheumatology Clinical Trials).

General health status instruments measure multiple aspects of health, including, specifically, physical function, social function, and pain, and are suitable for comparison of health status between diseases. The most widely used generic instrument is the Short Form-36 (SF-36). Validity and responsiveness are the most important criteria when deciding which particular instrument to use in a clinical trial or which to recommend to professional societies for inclusion in core sets, as exemplified by the OMERACT guidelines.

Although data are now available on the responsiveness of the WOMAC and the SF-36 in patients undergoing joint arthroplasty and treatment by non-steroidal anti-inflammatory drugs (NSAIDs), there are no similar longitudinal results for patients undergoing a comprehensive rehabilitation intervention.
This study aimed at examining the responsiveness of the WOMAC and the SF-36 in patients with OA of the legs undergoing a comprehensive rehabilitation intervention.

Methods

PATIENTS AND DATA COLLECTION

Patients were recruited from the Zurzach Rheumatology and Rehabilitation Clinic, Switzerland. All patients with hip or knee OA who were consecutively referred for a comprehensive inpatient rehabilitation intervention were invited, by letter, to participate in the study at least four weeks before entry into the clinic. On the day they entered the clinic, a doctor carried out a baseline interview and examination which determined inclusion in or exclusion from the study.

According to the American College of Rheumatology (ACR) guidelines, inclusion criteria were as follows: (a) knee pain for more than 25 of the past 30 days, morning stiffness of less than 30 minutes, and crepitation in the knee or (b) pain for more than 25 of the past 30 days and osteophytes on x-ray examination of the knees indicating knee OA.2 Patients with hip OA were included who had had pain for more than 25 of the past 30 days and at least two of the following three criteria: erythrocyte sedimentation rate <20 mm/1st h, osteophytes on x-ray examination, or obliteration of joint space.23 Exclusion criteria were as follows: history of drug abuse, non-compliance, difficulty completing questionnaires, having a severe illness, or arthroplasty of the joint in question.

Patients were sent or given a set of questionnaires, including the WOMAC and the SF-36, four weeks before entry into the clinic, on the day of entry into the clinic (baseline examination), at the day of discharge from the clinic (end of rehabilitation), and three months after the baseline examination.

The comprehensive rehabilitation intervention of usually three to four weeks’ duration consisted of passive and, especially, active physical therapy and a reduction in the use of NSAIDs as far as possible. Active cinesthesitherapy was performed individually and in groups to strengthen and stretch the musculature and the passive structures, and to recreate the regular joint mobility. Passive treatments included electrotherapy, hydrotherapy, thermotherapy, such as cold or warm compresses, massage, and so forth. Instructions for relaxation techniques and consultations for preventive measures were further elements of the rehabilitation programme. Finally, each patient was given instructions for an individual home rehabilitation programme to be continued after discharge.

MEASURES

The WOMAC is a multidimensional measure of pain, stiffness, and physical functional disability.2–4 This index has gained growing acceptance in OA assessment since its introduction in 1986. The pain dimension or scale includes five items asked about pain at activity or rest (P1–P5). The stiffness dimension includes two questions (St1, St2). The function dimension asks about the degree of difficulty in 17 activities (F1–F17). All 24 WOMAC items are rated on a numerical rating scale (in cm) ranging from 0 (“no symptoms/no limitation”) to 10 (“maximal symptoms/maximal limitation”), which is the preferred format in our population.24 Similar to the Visual Analogue Scale (VAS), this rating provides interval-type data. To score each scale we calculated the mean of the corresponding unweighted item scores. The results thus equal standardised WOMAC scores because standardisation of WOMAC scores is achieved by division of the scale sum score by the number of items.2 The global WOMAC score was calculated as the unweighted mean of all 24 items in order to compare it with the SF-36.12

The SF-36 includes eight multi-item scales containing two to 10 items each plus a single item to assess health transition.16 25 The scales cover the dimensions of physical functioning, role physical, bodily pain, general health, vitality, social functioning, role emotional, and mental health, ranging from 0 (“maximal symptoms/maximal limitations/poor health”) to 100 (“no symptoms/no limitations/excellent health”). The SF-36 is the most widely used general health status instrument and has been translated into many languages. The instrument is suitable for subjects aged 14 years and older and takes approximately 10 minutes to complete. Studies show excellent psychometric properties and there seems to be good responsiveness to change in patients with rheumatic conditions, compared with some longer instruments.26 The SF-36 allows scoring of the eight above mentioned scales and the construction of two summary scales, the physical component summary (PCS) and the mental component summary (MCS) scales.16

ANALYSIS

Patients included in the analysis filled out the questionnaires in accordance with the missing rules of the user’s guide, which specifies completion of at least four of the five pain items, one of the two stiffness items, and 14 of the 17 function items in WOMAC.2 Further, completion of the SF-36 was required so that all eight scores and the PCS and MCS could be calculated.16 To be calculable, each scale must have at least 50% of the corresponding items answered.

A variety of responsiveness statistics (RS) is available. However, it is not yet known which of these statistics is better for assessing responsiveness.10 27 Most commonly, the responsiveness of health status instruments has been compared using the standardised response mean (SRM),28 the effect size (ES),29 and, more rarely, Guyatt’s coefficient.30 The SRM is equal to the mean change in score divided by the standard deviation of the change in scores. The ES equals the mean change in score divided by the standard deviation of the baseline scores. The ES thus relates the change to the initial variation in scores. Finally, Guyatt’s coefficient is equal to the mean of change in scores divided by the between-subject variability or the...
Table 1 Characteristics of patients followed up with complete WOMAC (n=223) at three months after entry into the clinic.

<table>
<thead>
<tr>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA joint</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hip</td>
<td>32</td>
<td>34</td>
<td>66</td>
</tr>
<tr>
<td>Knee</td>
<td>32</td>
<td>25</td>
<td>57</td>
</tr>
<tr>
<td>Total</td>
<td>64</td>
<td>69</td>
<td>133</td>
</tr>
<tr>
<td>Use of drugs:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSAID</td>
<td>85</td>
<td>38</td>
<td>123</td>
</tr>
<tr>
<td>Analgesic</td>
<td>17</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td>Chondroitin sulphate</td>
<td>7</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>116</td>
<td>52</td>
<td>168</td>
</tr>
<tr>
<td>Age (years)</td>
<td>65.6</td>
<td>65.1</td>
<td>65.2</td>
</tr>
<tr>
<td>Education (years)</td>
<td>11</td>
<td>—</td>
<td>6-18</td>
</tr>
</tbody>
</table>

Results

Patients
A total of 433 patients were referred consecutively to the rehabilitation clinic between February 1997 and July 2000 with the diagnosis of OA and were asked to participate in the study at least one month before their entry into the clinic. Eighty one (19%) of them did not fulfill the ACR inclusion criteria. A further 12 (3%) patients could not be included because of severe illness or death before the entry into the clinic. Another six (1%) patients had to be excluded owing to arthroplasty of the OA joint before the clinic stay. Of the remaining 334 patients, 76 (23%) refused to participate until entry into the clinic, or were unable to participate for other reasons (age, language, loss of address, etc) or because they had not completed the questionnaires according to the missing rules of the WOMAC and the SF-36. At entry into the clinic, 258 patients were examined (baseline examination).

At three months after the baseline examination, 223 patients were re-examined with complete WOMAC sets and 211 with both WOMAC and SF-36 questionnaire sets. Between the baseline examination and the three month follow up, three patients had died or become severely ill (for reasons unrelated to the OA), seven patients received arthroplasty (after the clinic stay), six patients were excluded by the predetermined exclusion criteria (as listed before), and 19 (WOMAC) and 31 (SF-36) patients, respectively, returned incomplete forms according to the missing rules or refused to participate further.

Table 1 lists the characteristics of the 223 patients followed up with complete WOMAC sets. The mean age of study subjects was 65.1 years and 159/223 (71%) patients were female. Knee OA was present in 130/223 (58%) patients. Most of the patients stayed three to four weeks in the rehabilitation clinic. One hundred and sixteen (52%) patients used NSAIDs or analgesics, or both, or chondroitin sulphate at baseline examination. Most of them were able to reduce or omit these substances until the end of the rehabilitation stay.

The patients who could not be included in the study or could not be followed up at three months (a total of 210 for the WOMAC) were a median of 4.9 years older than the study patients, but there was no significant difference between the groups in their sex or distribution of joints affected.

The distribution of the baseline and follow up data showed a few patients (0-5% of all cases) with submaximal or maximal score values (no symptoms, perfect health), thus there was a mild ceiling effect for the WOMAC and a mild floor effect for the SF-36.

Responsiveness statistics (globally)
Baseline to end of rehabilitation (3–4 weeks after baseline) (table 2)
The most responsive scale was the pain scale, showing comparable values in the WOMAC (SRM=0.723) and in the SF-36 (SRM=0.528). Also the WOMAC function scale attained a high responsiveness (SRM=0.628), whereas the SF-36 physical functioning (SRM=0.249), role physical (SRM=0.196), and PCS (SRM=0.357) resulted in much lower responsiveness statistics. The SF-36 vitality scale (SRM=0.365) and
mental health (SRM=0.348) also had a relatively high responsiveness, whereas the other psychometric dimensions, such as role emotional (SRM=−0.041) and mental component summary (SRM=0.138), attained substantially lower values. Except in the psychometric dimensions of role emotional and MCS there was significant improvement of health status at the end of rehabilitation.

Baseline to three month follow up (table 3)
All responsiveness statistics resulted in lower values than those at the end of rehabilitation. Whereas the WOMAC described significant improvement of health status in all of its scales, in the SF-36 this was only the case for role physical, bodily pain, vitality, and PCS. As was seen at the end of rehabilitation, the pain scale was the most responsive dimension in both instruments, showing SRM=0.377 for the WOMAC and SRM=0.468 for the SF-36. The WOMAC function scale (SRM=0.235), the SF-36 role physical (SRM=0.188), and the PCS (SRM=0.228) attained comparable low to moderate values, as did the psychometric vitality (SRM=0.157). All other dimensions in both instruments resulted in minimal responsiveness statistics (SRMs ~0.029 to 0.083).

RESPONSIVENESS STATISTICS STRATIFIED BY SEX AND OA JOINT: COMPARISON OF WOMAC AND SF-36 (TABLE 4)
At the end of rehabilitation (3–4 weeks after entry/baseline) the WOMAC captured the effects of pain significantly more responsively than the SF-36 only in knee OA of women (SRM 0.867 vs 0.529, p=0.002), whereas in the other strata the SRMs of both instruments were comparable (around 0.4–0.8). In all patient groups the WOMAC’s function attained SRMs at least twice those of the SF-36’s physical functioning, leading to high significance. On average, the effects and SRMs were higher in the women than in the men in both pain and function. The WOMAC resulted in higher SRMs of function in hip OA (around
At the three month follow up, the SF-36 was slightly, but not significantly, more responsive in measuring pain, particularly in women, with the exception of men's knee OA (SRM 0.452 vs 0.137, p=0.040). The contrary was found in the function effects, where the WOMAC was significantly more responsive, except in the men's knee OA. Whereas the WOMAC reflected positive effects of the rehabilitation in hip OA (SRM 0.129–0.304), the SF-36 showed a worsening of patient's health (SRMs around –0.22), leading to highly significant differences. Also, at the three month follow up, the SRMs were higher in the women's groups than in the men's in pain and function (with one exception). Most effects were more positive in knee OA than in the corresponding groups of hip OA in both instruments.

Discussion
Both the condition-specific WOMAC and the generic SF-36 capture the improvement in pain in patients undergoing comprehensive inpatient rehabilitation intervention sufficiently well. The pain scales gave the highest responsiveness both at the end of rehabilitation and at the three month follow up in both instruments (SRMs 0.377–0.723).

The responsiveness of the function scales was limited in both instruments, but the WOMAC showed the effects significantly more sensitively than the SF-36 in both follow ups. At the end of rehabilitation, the WOMAC was significantly more responsive, especially in the measurement of function (overall SRM WOMAC 0.628, SF-36 0.249, p=0.000). At the three month follow up, the SF-36 was slightly, but not significantly, more responsive in pain (overall SRM of SF-36 was 0.468, WOMAC 0.379, p=0.154), but the SRMs of physical functioning of the SF-36 were partly negative (overall SRM=−0.001), whereas the function scale of the WOMAC still showed a moderate improvement (SRM=0.235, p=0.000). In both follow ups, the responsiveness in the WOMAC was higher than in the SF-36 in most of the scales. Therefore, the disease-specific WOMAC is better for measuring functional limitations than the generic SF-36, which is consistent with other findings.19 21

The stratification of the effects and the responsiveness resulted in some differences. It seems that the positive effects tend to be higher in women than in men, and higher in knee OA than in hip OA, especially in the function scales (for example, SRMs of WOMAC at the three month follow up: F: 0.318 v M: 0.070 and knee OA: 0.270 v hip OA: 0.182). However, these findings may be debatable in view of the small numbers of some strata (around 30).

A negative SRM reflects worsening of health status, a positive SRM improvement.

*Test of difference: comparison of WOMAC and SF-36 per stratum by the modified jack knife test. Null hypothesis: The two SRMs are not different. p Value = type I error = error probability of rejection of the null hypothesis.
Among the different responsiveness measurement concepts, the Guyatt’s statistics attained the highest figures, followed by the SRM, and then the ES. Both questionnaires, the WOMAC and the SF-36, showed the best responsiveness in the pain scales, followed by the function measurement, and then all other dimensions. This differs slightly from a previous arthroplasty study of the WOMAC and the SF-36, in which SF-36 physical functioning was more responsive than SF-36 bodily pain. The figures of responsiveness statistics were about half as high as those after arthroplasty, indicating a lower impact of conservative treatment than of causal surgical treatment (WOMAC, six month follow up: SRM 0.8–3.1). 19

The responsiveness statistics are dependent on the size of the intervention’s effect and its variance, as is obvious by their defining formulas. If it is assumed that patients who reach rehabilitation intervention are less disabled by their OA on average, it has to be expected that the size of the effect will be smaller than that of patients who have had surgery, because of the possible ceiling effect and the probability of a “regression-to-the-mean-effect”. For the same reasons, selection bias introduced by exclusion of severely ill patients may reduce the size of the effect and, therefore, the responsiveness. Thus a small amount of reduction of the responsiveness can be expected in our sample for these reasons.

Another limitation of our results is caused by the fact that the effect of the rehabilitation cannot be separated from the effect of the use of NSAIDs, or other inflammation or pain modulating substances. Treatment with moderate doses of NSAIDs together with low doses of analgesics led to ES values of 0.5–0.8. 11 In our sample, almost every second patient took NSAIDs or analgesics at baseline. In contrast, part of the reduction of the effect and the responsiveness can be attributed to the reduction of the use of anti-inflammatory substances during the clinic stay. Because in practice most patients with OA receive a combination of different treatments, the effects will be confounded in studies of comprehensive treatment strategies.

Responsiveness has important implications for clinical research and practice. When effect sizes below 0.4 are expected, detection of effects requires large sample sizes. For example, the necessary sample size (n) for detecting a minimal clinically important difference, assuming a two sided type I error of 0.05 and a power of 0.8, would be 64 for ES=0.4, 178 for ES=0.3, or 395 for ES=0.2 for each treatment arm. 10 15 35–36 This is true if one assumes that the two samples at baseline and at follow up are independent. In the case of paired follow up data, the required number is reduced by a factor of two using ES2: 32 for ES=0.4, 89 for ES=0.3, or 198 for ES=0.2. An ES of 0.28 or higher can be designated as sufficient, resulting in required sample sizes below 100 when testing with type I error of 5% and power 80%.

However, both the WOMAC and the SF-36 had limited responsiveness in measuring effects of comprehensive rehabilitation interventions. It is unclear whether this is because of the responsiveness of the instruments, or the size of effect of the rehabilitation intervention, or both. This has to be examined in a future analysis of minimal clinically important differences. We may improve the responsiveness by adaptation of the instruments. In patients who have been treated with NSAIDs, the WOMAC’s signal questions were as responsive as the full version of the WOMAC, meaning that a short instrument was usable and resulted in more compliance. 17 Supplementary specific questions to the WOMAC and the SF-36 and patient-specific indices, such as MAC-TAR, 19 may improve the responsiveness. The detection of problem oriented items may attain more responsive statistics, as shown in patients with knee OA, assessed by condition-specific scales of the generic SF-36. 18 Another way to achieve better responsiveness might be to assess new weighted or unweighted linear combinations of items in the WOMAC or the SF-36 determined by factor analysis.

In conclusion, both instruments, the WOMAC and the SF-36, capture improvement in pain in patients undergoing comprehensive inpatient rehabilitation intervention. Functional improvement can be detected better by the WOMAC than by the SF-36. All the other scales of both instruments were more weakly responsive. It needs to be determined whether this is due to limited instrument responsiveness, an inability to detect change, or a limited effectiveness of the rehabilitation intervention.

This study was supported by the Spa-Zurzach Rehabilitation Foundation.

We thank Stephan Mariacher, and Susanne Lehmann for the planning, management, and implementation of the database, and Robin Kyburg and Diane Fassett for help in preparing the manuscript.


