We used prospective data from 862 total knee and 716 total hip replacements three years after surgery in order to derive and validate a reduced Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) function scale. The reduced scale was derived using the advice of clinical experts as well as analysis of data. The scale was tested for validity, reliability, and responsiveness.

Items which were retained included: ascending stairs, rising from sitting, walking on the flat, getting in or out of a car, putting on socks, rising from bed, and sitting.

The reduced and full scales had comparable, moderate correlations with other measures of function, confirming convergent validity. Cronbach’s alpha was high (α > 0.85) with the reduced scale confirming reliability. Responsiveness was greater for the reduced scale (full = 1.4, reduced = 1.6).

This reduced version of the WOMAC function scale provides a practical, valid, reliable and responsive alternative to the full function scale for use after total joint replacement. Further work is needed to demonstrate its wider applicability.
lations are expected and are used simply to ensure that spurious results are not confounded. Convergent construct validity examines how strongly a new scale is associated with other measures of the same type.

**Reliability.** This term encompasses the internal consistency of a scale, usually given as Cronbach’s alpha which measures the degree of correlation among items. Values of Cronbach’s alpha of greater than 0.7 indicate adequate reliability for a scale, whereas values above 0.9 may indicate redundancies in the scale. While values greater than 0.9 are necessary for reliable individual scores, they are redundant for group means, the usual focus of research. Reliability also includes interobserver variation and consistency. Since the WOMAC is a self-administered tool, interobserver reproducibility is not relevant in this case, and as the new scale is a subset of the original, consistency is assumed to be inherent and has not been investigated further.

**Responsiveness.** This is the measure which detects true changes in clinical states. There are several way of assessing responsiveness, but in this instance, the method of standardised response means has been used.\(^1\)\(^2\)

**Patients and Methods**

The Kinemax Outcomes Study is a multinational, prospective cohort study of primary total knee replacement (TKR) for patients with osteoarthritis. Patients were recruited between September 1997 and December 1998 and all surgeons used the Kinemax Plus (Stryker Howmedica Osteonics, Mahwah, New Jersey) prosthesis. Data were gathered before and at three and 12 months after operation by physiotherapists, and research personnel. Respondents were requested to indicate which five items from the scale of function they would keep, using the following three criteria: (1) most likely to change after surgery; (2) what patients care about the most; and (3) representative of a broad spectrum of levels of activity (one orthopaedic surgeon gave 6 preferences instead of 5).

Items which were considered to be ambiguous with regard to gender or culture or to be open to misinterpretation were omitted from the reduced scale.

The data were then analysed to confirm that the selected items represented a range of difficulties, were clinically sensitive to detecting change, had few missing values and were applicable to patients with hip and knee symptoms. The responses were examined to identify the difficulty of each question for the patient carrying out the activity before and at three months after operation.

For the analysis of items, responses were coded in the standard fashion for WOMAC with 0 for ‘no difficulty’ and 4 for ‘extreme difficulty’ in carrying out the physical function (intermediate answers being 1 - mild, 2 - moderate, 3 - severe).\(^2\)

The reduced questionnaire was then assessed for validity, responsiveness and reliability using the follow-up data at one year from the TKR dataset. For these analyses summary scores for each patient were calculated for both the full and reduced scales of function by taking the mean of all responses, multiplying by 25, and subtracting from 100. This transformed the raw WOMAC function score to a 0 to 100 scale, worst to best. A score of 0 indicates extreme restriction in all activities, and a score of 100 indicates no restriction for any item. If four or more of the items of function were missing for the full scale, a score was not calculated, as is the standard practice for handling missing responses.

In order to confirm validity, the correlations between the full and reduced WOMAC function scores for the preoperative, three-month and 12-month postoperative data were examined. Subgroups by age, gender and country were also compared in order to determine if there was differential reporting for these factors. At all stages, because of the nature of the scoring system (producing categorical data), we used the non-parametric Spearman’s rank correlation coefficient rather than the parametric Pearson’s correlation coefficient.
We assessed convergent validity by comparing the strength of correlation of the full and reduced scores with several scales measuring similar outcomes. For the TKR patients these included the SF-36 physical component score and physical function score, the Knee Society function score and the Oxford Knee Score. For the THR patients these included the SF-12 physical component score and the Harris hip score.

Responsiveness was assessed using standardised response means calculated as the mean change in score from preoperative to 12 months divided by the standard deviation of the change in score. This variable assesses the extent of improvement. Therefore, patients who indicated that their quality of life (taken from question 2 in the SF-36) was the same or worse since their operation were excluded from this section of analysis. Given that TKR has dramatic effects on pain and function, values of greater than 1 were indicative of adequate responsiveness. In order to evaluate whether the change is relevant to the patient, responsiveness was also assessed by determining whether changes in the full and reduced scores correlated with other indicators of change in the patients' clinical status including the quality of life, general health and satisfaction with outcome. Higher correlations indicated greater responsiveness.

Results

Data sources

Clinical opinion of orthopaedic personnel. The respondents comprised 21 surgeons, six research personnel, five nurses and four physiotherapists, of which 24 were from the UK, and 12 from the USA.

TKR data. This dataset contained 862 primary TKR patients. A total of 806 (94%) of these had valid WOMAC scores at their three-month review, and 762 (88%) at 12 months. Their mean age was 70 years (38 to 90), and 59% were women. Most (50%) were from the UK with 31% from the USA and 19% from Australia.

THR data. There were 716 patients with a diagnosis of osteoarthritis included in the analysis. Their mean age was 73.6 years (65 to 93), with a standard deviation of 5.5 years, and there were 665 valid WOMAC function scores for this dataset.

Derivation of scale

Clinical opinion of orthopaedic personnel. The results of this survey are given in Table I. Those items which were eventually selected for retention in the final reduced model, are indicated in bold type. Four of the top seven items were kept in the reduced scale. Descending stairs was not included as one stair item was already in the scale, and the item for ascending stairs proved to be more responsive to change over time. In order to avoid items which would not apply to one gender or cultural group, domestic duties, shopping and bathing were not included. The patient-based data were then examined to justify the inclusion of three further items, which created a broad range of difficulty. These activities were also considered to be basic for daily living. Data driven analysis using TKR data. The analysis of the TKR cohort gave the mean item values indicated in Table I. For the preoperative data the mean score for the easiest question is 1.5, 1 being mild restriction, and 2 being moderate restriction of activity. The most difficult item has a mean score of 2.7, where 3 is severe restriction of activity.

Table I. Results of the survey of the clinical advisors and the data-driven results including the mean response, the difficulty rankings and the number of missing responses per item (preoperative and three-month postoperative TKR data).

<table>
<thead>
<tr>
<th>Item</th>
<th>Clinical</th>
<th>Preoperative</th>
<th>Three-month</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Rank (most to least)</td>
<td>Rank (hardest to easiest)</td>
</tr>
<tr>
<td>1. Descending stairs</td>
<td>16</td>
<td>4</td>
<td>2.72 1 9 1 52</td>
</tr>
<tr>
<td>2. Ascending stairs*</td>
<td>20</td>
<td>3</td>
<td>2.66 3 11</td>
</tr>
<tr>
<td>3. Rising from sitting*</td>
<td>27</td>
<td>1</td>
<td>2.40 7 2</td>
</tr>
<tr>
<td>4. Standing</td>
<td>7</td>
<td>10</td>
<td>2.19 10 8 0.99 12 10</td>
</tr>
<tr>
<td>5. Bending to floor</td>
<td>5</td>
<td>12</td>
<td>2.39 8 12</td>
</tr>
<tr>
<td>6. Walking on flat*</td>
<td>27</td>
<td>1</td>
<td>2.25 9 15</td>
</tr>
<tr>
<td>7. Getting in/out of car*</td>
<td>13</td>
<td>6</td>
<td>2.47 6 9</td>
</tr>
<tr>
<td>8. Going shopping</td>
<td>8</td>
<td>9</td>
<td>2.60 4 22</td>
</tr>
<tr>
<td>9. Putting on socks*</td>
<td>12</td>
<td>7</td>
<td>1.98 12 9</td>
</tr>
<tr>
<td>10. Rising from bed*</td>
<td>5</td>
<td>12</td>
<td>2.01 11 8</td>
</tr>
<tr>
<td>11. Taking off socks</td>
<td>2</td>
<td>15</td>
<td>1.88 13 9</td>
</tr>
<tr>
<td>12. Lying in bed</td>
<td>4</td>
<td>14</td>
<td>1.62 15 9</td>
</tr>
<tr>
<td>13. Getting in/out of bath</td>
<td>11</td>
<td>8</td>
<td>2.52 5 80</td>
</tr>
<tr>
<td>14. Sitting*</td>
<td>2</td>
<td>15</td>
<td>1.54 17 9</td>
</tr>
<tr>
<td>15. Getting on/off toilet</td>
<td>6</td>
<td>11</td>
<td>1.59 16 7</td>
</tr>
<tr>
<td>16. Heavy domestic duties</td>
<td>2</td>
<td>15</td>
<td>2.68 2 42</td>
</tr>
<tr>
<td>17. Light domestic duties</td>
<td>14</td>
<td>5</td>
<td>1.71 14 35</td>
</tr>
</tbody>
</table>

*retained items
Similarly, for the three-month postoperative data, mean item scores are presented. There is a shift down the scale, as expected after surgery, from moderate to no difficulty with activity, with the easiest item obtaining a mean score of 0.7, and the most difficult of about 1.8.

In addition, the numbers of missing values for each item were taken into account when considering whether or not it should be retained. These missing responses are also shown in Table I. The number of missing responses for item 8 (going shopping), 13 (getting in or out of the bath), 16 (heavy housework) and 17 (light housework) are particularly high.

The items in the reduced scale were: ascending stairs (Q2), rising from sitting (Q3), walking on flat (Q6), getting in or out of a car (Q7), putting on socks (Q9), rising from bed (Q10), sitting (Q14).

**Validation of scale using TKR and THR data.** The reduced scale was assessed with a variety of methods to examine validity, reliability and responsiveness.

**Criterion validity.** Spearman’s correlation coefficient between the two scales was 0.96 for the knee dataset and 0.97 for the hip dataset. At all time intervals for both TKR and THR patients, comparisons of the full and reduced WOMAC function scores gave remarkably similar mean values and standard deviations (Table II). The very strong correlation between the two scales and high agreement in scores support the hypothesis that the reduced scale captures functional status as well as the original version.

Mean scores were calculated, stratifying the data by gender, age (by quartiles) and country to indicate if there was differential reporting for these factors. The findings indicate similar mean scores for the reduced and full scales, irrespective of assessment time and subdivision. In fact, the mean scores do not differ by more than two points. The results obtained when stratifying by gender are presented graphically in Figure 1. Similar findings emerge when stratifying by country and age quartiles.

**Convergent construct validity.** This was assessed by determining whether the reduced scale has similar strength of correlation with other scales. For this section of analysis, all the preoperative data for TKR were used, as well as those for THR in order to compare the WOMAC scale with the Harris hip scores and the SF-12 physical component scores. For the data on the knee, both full and reduced scales were correlated with the SF-36 physical component score, the SF-36 physical function score, the Knee Society function score and the Oxford knee score. These are shown in Table III. All Spearman rank correlation coefficients are significant at the 1% level, and for the reduced scale, are a mean of only 0.035 less than those for the full scale (95% confidence interval (CI) for the differences 0.025 to 0.045), supporting the hypothesis that the reduced scale is valid. Furthermore,
Table IV. Spearman’s correlation coefficients for the associations between perceived change in health status and quality of life and satisfaction with outcome and change in WOMAC function (calculated as the change in scores one year after surgery for 762 patients who underwent TKR). All correlations are significant at the 1% level. The difference in correlations between full and reduced scores and each outcome are not statistically significant at the 5% level.

<table>
<thead>
<tr>
<th></th>
<th>Change in full WOMAC</th>
<th>Change in reduced WOMAC</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health compared with one year ago</td>
<td>0.37</td>
<td>0.36</td>
<td>0.91</td>
</tr>
<tr>
<td>Change in quality of life since TKR</td>
<td>0.38</td>
<td>0.42</td>
<td>1</td>
</tr>
<tr>
<td>Satisfaction score</td>
<td>0.30</td>
<td>0.35</td>
<td>1</td>
</tr>
</tbody>
</table>

Discussion

Both psychometric and clinical approaches were used to develop a shortened version of the WOMAC function scale. The reduced scale compares favourably with the full scale overall.

Convergent validity was demonstrated by a moderately strong correlation between various scales of physical function for both the TKR and THR patients. These included the SF-36 physical component score, the SF-36 physical function score, the Knee Society function score and the Oxford knee score for knees, and the Harris hip score and SF-12 physical component score for the hips.

A Cronbach’s alpha reliability coefficient of greater than 0.7 is generally required for group comparisons. The values obtained in this study, although marginally less than those for the full scale, remain more than adequate while also eradicating redundancy. Indeed, these results further support the work of Ryser et al., which indicated that some redundancy occurred in the scale and there was scope for reduction by omission of redundant items.

Standardised response means obtained in our study indicate that the two scales have similar responsiveness, reinforcing the concept that the reduced scale adequately represents the full scale. Indeed, those for the reduced scale are slightly higher than those for the full scale, which may indicate that it is slightly more responsive. Further studies are required to support this.

The reduced scale also correlates significantly with various other measures of perceived functional change and satisfaction with outcome. This further indicates that the scale is responsive to meaningful changes, as observed by the patient.

Since the reduced scale is a subset of the full scale, it will be relatively simple to compare results across studies using either form, especially as the WOMAC is the recommended disease-specific outcome measure. This will increase its acceptability and usefulness within the orthopaedic community. There are other shortened measures of outcome, for example, the Oxford hip and knee scores and the Bristol knee score, but all have their disadvantages. The Oxford
scores, although brief and simple, may not be sufficiently specific for use after total joint replacement and there are concerns regarding missing values. The Bristol knee score has not been validated and is used almost exclusively by the Bristol Knee Group.

Further work should be done in order to validate further this reduced scale, in particular for patients who are treated non-operatively as well as those undergoing THR and revision total joint arthroplasty. Reproducibility needs to be established for the reduced scale. Also, further studies investigating compliance and missing values are needed.

This study was supported by the Stryker Howmedica Osteonics and NIH Grants P01 AR 47782 and K24 AR 02123.

We thank the Kinemax Outcomes Group for their tireless work, in particular, the research assistants involved in data collection. We also wish to acknowledge the work of Dr Elizabeth Wright who provided computer programming support for these analyses. The Kinemax Outcomes Group consists of the following members: William Gillespie, Colin Howie, Ian An-nan, Alastair Gibson, Judith Lane and Peter Abernathy of the Princess Margaret Rose Hospital, Edinburgh, UK; Ian Pinder, David Weir, Nigel Brewster and Karen Bettinson of the Freeman Hospital, Newcastle-upon-Tyne, UK; Maurice Needoff and Roz Jackson of the King’s Mill Clinic, Mansfield, UK; Tim Wilton and Peter Howard of the Derbyshire Royal Infirmary, Derby, UK; Ian Forster, Paul Szyprt, Chris Moran, David Whittaker, Mike Bullock and Zena Hinchcliffe of the Queen’s Medical Centre, Nottingham, UK; Ian Learmonth, John Newman, Chris Ackroyd, George Langkamer, Robert Spencer, Mark Shannon, Evert Smith, John Dixon and Sarah Whitehouse of the Avon Orthopaedic Centre, Bristol, UK; Clement Sledge, Frederick Ewald, Robert Poss, John Wright, Scott Martin, John Kwon, Yvette Valderrama of the Brigham and Women’s Hospital, Boston, USA; Steven Harwin and Michael Lichardt of the Beth Israel Medical Centre, New York, USA; Mark Mehlhoff, Linda Weiler and Tom Cahalan of the Iowa Medical Clinic, Cedar Rapids, USA; Richard Cronk and Allyson Sandago of the Neumuscular and Joint Centre, Corvallis, Canada; Stephen Rackmann and Emma McLaughlin of the Knee Centre, Gold Coast, Australia; and Peter Lewis, Robert Bautze, Tom Stevenson, Gordon Morrison and Jane Clasohm of the Queen Elizabeth Hospital, Adelaide, Australia.

Although none of the authors has received or will receive benefits for personal or professional use from a commercial party related directly or indirectly to the subject of this article, benefits have been or will be received but will be directed solely to a research fund, foundation, educational institution, or other nonprofit organisation with which one or more of the authors are associated.

References


