The use of the Oxford hip and knee scores

D. W. Murray, R. Fitzpatrick, K. Rogers, H. Pandit, D. J. Beard, A. J. Carr, J. Dawson

From the University of Oxford, Oxford, England

The Oxford hip and knee scores have been extensively used since they were first described in 1996 and 1998. During this time, they have been modified and used for many different purposes. This paper describes how they should be used and seeks to clarify areas of confusion.

It is now almost ten years since the Oxford hip and knee scores were introduced. The use of patient-reported outcomes in orthopaedics at that time was extremely limited. The scores were developed to assess the outcome of hip and knee replacements in randomised trials and were designed to be completed by the patient in order to minimise potential bias unwittingly introduced by surgeons when assessing the results themselves. An additional advantage in using patient-reported outcome measures was that they could be completed at a remote location by post, thereby avoiding inconvenience and cost.

The Oxford hip and knee scores were devised as joint specific instruments aimed to minimise the influence of comorbidity. They underwent rigorous assessment of reliability, validity and responsiveness in prospective studies. Their use has steadily increased and they are now widely employed. They have been used in cohort studies and audits and in national joint replacement registries, including those in England, New Zealand and Sweden. They have also been applied to other disorders of the joints and in surgical management other than arthroplasty.

A number of issues have been raised and various groups have made modifications to the scoring system which has resulted in some confusion. We therefore now review their use and recommend ways in which they should be best employed. If the scores are used in a standard fashion, data from different series can be more easily compared which is one of the main reasons for developing and using such systems.

Wording of questions and response categories

While the response rates for the Oxford hip and knee scores are generally higher than those for many other measures of health status, we and others initially found that some patients had difficulty in answering specific questions. For example, in question 4 in the knee score, where patients are asked how long they are able to walk before pain becomes severe, the extreme response ‘not at all’ was sometimes wrongly inferred by the patient to mean the opposite of that intended. For clarification, we made small amendments to the response categories for two items. The recommended format for questions and their response categories can be found on the Patient-Reported Health Instruments website.

A different problem is typified by question 7 which regards kneeling. This causes particular difficulty when patients have been told not to kneel. The item reads “Could you kneel down and get up again afterwards?” The word ‘could’ is in bold type, suggesting that patients should answer this item hypothetically if they have been told not to kneel. This is our recommendation, which can be communicated to the patient. A similar approach should be taken for the question about stairs by patients living in bungalows and about housework by those who have housekeepers or live in residential care. If, after clarification, an item is still unanswered, it should be dealt with as missing data.

It has been further suggested that the Oxford scores should be extended by adding two questions on the need for walking aids and, in the case of the hip, sexual problems. We think that this would be more detrimental than helpful for three reasons. First, the questions should be relevant to most patients, and these issues were not found to be generally important in the initial interviews from which the Oxford hip score
resulted. Secondly, a new questionnaire would no longer be comparable with the old and thirdly, it would mean that further work on validation would have to be undertaken. It is wrong to imagine that there will ever be a perfect questionnaire which suits all people at all times.

**Scoring**

When the Oxford hip and knee score systems were originally devised they were made as simple as possible in order to encourage their use. Thus, each question was scored from 1 to 5, with 1 representing the best outcome/least symptoms. The scores from each question were added so that the overall figure lies between 12 and 60, with 12 being the best outcome. Subsequently, many surgeons have found this scoring to be unintuitive and have modified it, particularly when using the knee score. These changes have led to considerable confusion. For example, the use of a system whereby each question has been scored between 0 and 4, with 4 being the best outcome, produces overall scores running from 0 to 48, with 48 being the best outcome. The 60 to 12 system may be converted to the 0 to 48 score and vice versa by subtracting the score from 60. It has also been suggested that the score could be modified to range between 0 and 100 with either 100 or 0 being the best. Although it is possible to convert between one system and another it would simplify the situation considerably if a standard form was used. Our view is that both the hip and knee scores should be used from 0 to 48, with 48 being the best. The method used should always be clearly stated.

**Comparison with other scoring systems**

The Oxford scores were designed to be joint-specific in order to increase their sensitivity to the outcome of the joint replacement as far as possible and to be influenced as little as possible by other comorbidities. Although the scores are influenced by pathology, such as strokes and back problems elsewhere, they seem to be influenced to a less extent than is the case for other patient-reported outcome measures, used in this context. This feature of a score – its specificity – influences its responsiveness or sensitivity to change, which is the most important aspect in relation to prospective outcome studies. The Oxford hip and knee scores have been shown to have particularly high responsiveness.

Since the Oxford hip and knee scores have been evaluated independently and found to be the best and most reliable systems for the assessment of hip and knee replacement, respectively, there is some justification for using these scores in isolation. However, if it is important to compare the improvement resulting from hip or knee procedures with those occurring at other sites, it is sensible to use a general health measure, such as the SF-12 as well as the Oxford score. If information on health economics is needed the EuroQol is valuable. If specific clinical and surgical data, such as range of movement, are required then a formal clinical assessment would be necessary.

**Use of the scores**

While the scores were designed to be primary measures of outcome in randomised, controlled trials, they have been much more widely used in cohort studies and audits. It has become apparent that one of the biggest determinants of outcome after a joint replacement is the pre-operative score. If the treatment of different cohorts of patients is being studied in a non-randomised setting, it is essential that both the pre-operative and post-operative scores are obtained. The change in the score should be analysed in addition to the post-operative score. Likewise, if a multivariate analysis of outcomes is undertaken, this should take into account the pre-operative score. After joint replacement, most improvements in function and in the Oxford scores occur within the first year. It is therefore not unreasonable to assess the outcome at one year.

Given that the score at one year is related to the pre-operative score when using the scores for audit purposes, it is useful to know approximately what outcome would be expected after a total joint replacement given a particular pre-operative score. A publication based on data from a national audit has reported the mean Oxford hip score before and after total hip replacement (THR), presented by 10% bands (deciles) of the pre-operative score. Table I summarises these data converted to the 0 to 48 scoring system. A similar analysis has been undertaken on data from a large multicentre randomised, controlled trial of total knee replacement (TKR) and is presented in Table II. It should, however, be noted that, like all outcome scores, the absolute score tends to decrease with age. Therefore, in elderly patients, a ‘normal’ score may be somewhat less than 48.

Although we have achieved very high rates of response when the Oxford hip and knee scores have been sent to patients, other authors have not been as successful. Practical approaches to maximising rates of response have been recommended elsewhere. The scores include using carefully-worded covering letters, sending reminders with second copies of the questionnaire; using prepaid reply envelopes and contacting patients by telephone. When patients have bilateral joint problems we favour giving two questionnaires, one for each side, rather than modifying the questionnaires to include both sides.

Surgeons are attracted to systems of categorisation, grouping the results according to whether they are considered to be excellent, good, fair or poor, rather than simply using a score. However, the cut-off points in categories are always very approximate and are likely to vary from one population to another. Work is currently in progress to produce categories based on large international data sets. Until this is available we believe that surgeons should avoid categorisation. If, in spite of this, surgeons are still keen to categorise, then they can use published cut-off points. For
example, Kalairajah et al.29 developed categories for the Oxford hip score based on the Harris hip score,30 which when translated to the 0 to 48 system suggested that a score of more than 41 was excellent, 34 to 41 good, 27 to 33 fair and less than 27 poor.

Use other than in joint replacement

The Oxford scores were designed specifically for the assessment of joint replacement but they have also been used for evaluating other interventions such as pharmacological treatment,31 after osteotomy32 or rehabilitation33 and the treatment of conditions such as fractures.34 Although the scoring systems have not been validated for these uses,35 we see no reason why they should not be used in these situations. However presentation of some evidence that appropriate measurement properties have been maintained, such as by the use of Cronbach’s alpha,36 should be made in such circumstances. It is nevertheless very useful for comparative purposes that the same scoring systems are used across related areas of treatment.

Other languages

The Oxford scores are now being used worldwide and have been translated into a number of different languages. If data from the Oxford scores are to be used comparatively in different languages then the translation has to be standardised as described by Haverkamp et al.37 This should include forward and backward translation, plus an assessment of the measurement properties of the translated score. Translations of the Oxford hip and knee scores which have used these methods include Dutch,8,37 Swedish,16 Thai38 and Chinese39 versions.

Missing data

A common problem with questionnaires includes the incomplete responses that some patients provide. If, after repeated attempts to obtain complete data from an individual, only one or two questions have been left unanswered, it is feasible to enter the mean value representing all of their other responses in order to fill the gaps. An alternative computerised method of imputing values

### Table I. The mean Oxford hip scores pre-operatively and at 12 months after hip replacement and by deciles of the pre-operative score. Based on data from Hajat et al.22

<table>
<thead>
<tr>
<th>Pre-operative Oxford hip score</th>
<th>Decile cut-off points</th>
<th>Decile cut-off points</th>
<th>Pre-operatively</th>
<th>12 months post-operatively</th>
<th>Change in score between 0 and 12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band</td>
<td>60 to 12 scale</td>
<td>0 to 48 scale</td>
<td>Mean SD Number</td>
<td>Mean SD Number</td>
<td>Mean SD</td>
</tr>
<tr>
<td>1 Best</td>
<td>12 to 34</td>
<td>26 to 48</td>
<td>29.7 3.4 279</td>
<td>41.2 6.9 202</td>
<td>11.5 6.9</td>
</tr>
<tr>
<td>2</td>
<td>35 to 38</td>
<td>22 to 25</td>
<td>23.3 1.1 346</td>
<td>41.5 0.64 266</td>
<td>18.2 6.4</td>
</tr>
<tr>
<td>3</td>
<td>39 to 41</td>
<td>19 to 21</td>
<td>19.9 0.8 392</td>
<td>40.2 7.6 291</td>
<td>20.3 7.7</td>
</tr>
<tr>
<td>4</td>
<td>42 to 43</td>
<td>17 to 18</td>
<td>17.5 0.5 289</td>
<td>39.4 7.6 226</td>
<td>22.0 7.5</td>
</tr>
<tr>
<td>5</td>
<td>44 to 45</td>
<td>15 to 16</td>
<td>15.5 0.5 322</td>
<td>38.1 8.2 237</td>
<td>22.6 7.2</td>
</tr>
<tr>
<td>6</td>
<td>46 to 47</td>
<td>13 to 14</td>
<td>13.5 0.5 361</td>
<td>37.1 9.3 255</td>
<td>23.5 9.4</td>
</tr>
<tr>
<td>7</td>
<td>47 to 50</td>
<td>10 to 12</td>
<td>11.0 0.8 541</td>
<td>36.0 9.6 386</td>
<td>25.0 9.6</td>
</tr>
<tr>
<td>8</td>
<td>51 to 52</td>
<td>8 to 9</td>
<td>8.5 0.6 338</td>
<td>36.6 9.6 236</td>
<td>28.1 9.5</td>
</tr>
<tr>
<td>9</td>
<td>53 to 54</td>
<td>6 to 7</td>
<td>6.5 0.5 273</td>
<td>33.6 10.4 175</td>
<td>27.1 10.4</td>
</tr>
<tr>
<td>10 Worst</td>
<td>55 to 60</td>
<td>0 to 5</td>
<td>3.8 1.3 326</td>
<td>31.1 10.7 205</td>
<td>27.3 10.6</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>14.7 7.3 3466</td>
<td>37.6 9.2 2479</td>
<td>22.6 9.8</td>
</tr>
</tbody>
</table>

### Table II. The mean Oxford knee scores (using the 0 to 48 scoring method) pre-operatively and at 12 months after knee replacement, by 10% bands (deciles) of the pre-operative score. Data from the Knee Arthroplasty Trial.24,25

<table>
<thead>
<tr>
<th>Pre-operative Oxford knee score</th>
<th>Decile cut-off points</th>
<th>Decile cut-off points</th>
<th>Pre-operatively</th>
<th>12 months post-operatively</th>
<th>Change in score between 0 and 12 months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band</td>
<td>60 to 12 scale</td>
<td>0 to 48 scale</td>
<td>Mean SD Number</td>
<td>Mean SD Number</td>
<td>Mean SD</td>
</tr>
<tr>
<td>1 Best</td>
<td>12 to 32</td>
<td>28 to 48</td>
<td>31.7 3.8 224</td>
<td>39.3 8.0 186</td>
<td>7.6 8.3</td>
</tr>
<tr>
<td>2</td>
<td>33 to 35</td>
<td>25 to 27</td>
<td>26.0 0.8 192</td>
<td>36.7 8.1 151</td>
<td>10.7 8.1</td>
</tr>
<tr>
<td>3</td>
<td>36 to 38</td>
<td>22 to 24</td>
<td>22.9 0.8 255</td>
<td>36.6 8.3 206</td>
<td>13.7 8.3</td>
</tr>
<tr>
<td>4</td>
<td>39 to 40</td>
<td>20 to 21</td>
<td>20.5 0.5 195</td>
<td>35.7 8.7 155</td>
<td>15.2 8.8</td>
</tr>
<tr>
<td>5</td>
<td>41 to 42</td>
<td>18 to 19</td>
<td>18.5 0.5 226</td>
<td>35.1 9.1 186</td>
<td>16.6 9.1</td>
</tr>
<tr>
<td>6</td>
<td>43 to 44</td>
<td>16 to 17</td>
<td>16.5 0.5 192</td>
<td>34.9 9.5 150</td>
<td>18.4 9.5</td>
</tr>
<tr>
<td>7</td>
<td>45 to 46</td>
<td>14 to 15</td>
<td>14.5 0.5 235</td>
<td>34.3 9.1 174</td>
<td>19.8 9.1</td>
</tr>
<tr>
<td>8</td>
<td>47 to 48</td>
<td>12 to 13</td>
<td>12.6 0.5 189</td>
<td>31.8 9.9 144</td>
<td>19.2 9.8</td>
</tr>
<tr>
<td>9</td>
<td>49 to 51</td>
<td>9 to 11</td>
<td>10.1 0.8 223</td>
<td>28.3 11.0 165</td>
<td>18.2 11.0</td>
</tr>
<tr>
<td>10 Worst</td>
<td>52 to 60</td>
<td>0 to 8</td>
<td>6.1 1.9 224</td>
<td>27.5 12.1 155</td>
<td>21.4 11.7</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>18.0 7.5 2165</td>
<td>34.2 10.0 1672</td>
<td>15.9 10.2</td>
</tr>
</tbody>
</table>
which can be applied to many questionnaires has been described by Jenkinson et al. If more than two questions are unanswered, the overall score should not be calculated. If patients indicate two answers for one question the worst response is adopted.

Statistical issues
By the time the patients receive a joint replacement, their symptoms are likely to be fairly severe, whereas generally, most patients 12 months post-operatively have only very mild symptoms, if any. Hence, data from the Oxford scores obtained at these different times tend to be skewed in either direction. It could therefore be argued that the use of transformations or non-parametric statistics for analyses involving absolute scores are sensible. However, we have found that parametric statistics have generally been satisfactory in most analyses. Analysis using change scores is less difficult since they tend to be more normally distributed.

While it is simple to determine the statistical significance of changes in the quality of life as expressed by the Oxford hip or knee scores, it can be harder to determine the real clinical or subjective meaning of these changes. There are a number of approaches for assessing the smallest amount of change on a measure that is likely to be of importance, which includes the minimal clinically important difference (MCID). This is the smallest change in score which patients perceive as meaningful and which would cause clinicians to consider a change in management. Work is in progress to produce MCID estimates for the Oxford hip and knee scores. Until these are available an approximation to the MCID can be obtained, based on the observation that for many patient-reported outcome measures the MCID is about half the SD of change. In joint replacement studies, the SD of the Oxford change scores tended to be about 10, as seen in the totals in Tables I and II, but can be as low as 6. Therefore the MCID would be expected to be between 3 and 5 points. However, a study of THR found that at one year the mean score after a posterior approach was about 2 points better than that after a Hardinge approach. This suggests that the MCID may actually be lower than 3 points.

Information about clinically important differences is needed in power calculations to determine the size of a study. It is essential that power calculations are undertaken before a study begins. For comparative studies of joint replacements using the Oxford scores as the primary outcome measure, power calculations are likely to suggest that a minimum of 100 patients, and usually many more, are required.

Now that the involvement of the patients in outcome assessment is becoming more widely established, it is particularly important to achieve standardisation in the use of instruments. We trust that this paper will assist in this.

We are most grateful to the Knee Arthroplasty Trial team for providing data on the outcome of knee replacement. We are also grateful to Barbara Marks for help with the manuscript and Cathy Jenkins and Jo Copp for insights into the collection of data.

No benefits in any form have been received or will be received from a commercial party related directly or indirectly to the subject of this article.

References


