Outcome measures for orthopaedic interventions on the hip

E. Ashby, M. P. W. Grocott, F. S. Haddad

From University College London Hospital, London, England

Orthopaedic interventions are assessed on their outcome. Outcome scores are used for research purposes to compare prostheses, surgical techniques, methods of fixation and types of peri-operative care. They are also used for audit to allow comparison between surgeons, departments, institutions and countries. Outcome after surgery can be defined in many different ways such as mortality, morbidity, clinical findings, radiological findings, post-operative complications, rates of re-operation, pain, length of hospital stay and health-related quality-of-life.

Generic short-term clinical outcomes address overall morbidity and mortality related to surgery. Mortality at a variety of time points, the duration of post-operative stay and the incidence of specific complications fall within this group. Morbidity and complications are poorly and inconsistently recorded. The post-operative morbidity survey is a recently validated measure of post-operative morbidity which focuses on indicators of dysfunction of organ systems that can be easily obtained. It can be used to record general complications such as deep-vein thrombosis, pulmonary embolism, wound problems and infection in the immediate post-operative period. It also assesses mobility and can identify orthopaedic-specific complications such as fracture and dislocation.

Orthopaedic interventions are used to evaluate the effect of operative interventions. They are used for audit and research. Knowledge of these measures is becoming increasingly important with league tables comparing surgeons and hospitals being made accessible to the profession and the general public.

Several types of tool are available to describe outcome after hip surgery such as generic quality-of-life questionnaires, disease-specific quality-of-life questionnaires, hip-specific outcome measures and general short-term clinical measures. We provide an overview of the outcome measures commonly used to evaluate hip interventions.

Validation of outcome measures

For an outcome measure to be meaningful, it must be psychometrically evaluated and shown to be reliable, valid and sensitive to change. Tests of reliability answer the question: Is the test measuring something, for example an underlying concept such as mobility, in a reproducible fashion? Internal consistency describes whether a survey measures a single underlying concept and can be tested using Cronbach’s alpha. Reproducibility (or stability) defines whether the questionnaire yields the same results in repeated trials under the same conditions. Paired sets of data can be compared using the kappa coefficient.
Tests of validity assess whether the survey measures what it is proposed to measure. There are many types of validity: content validity, criterion validity and construct validity. Content validity examines whether items in a questionnaire cover the intended topics. It is a subjective measure and cannot be evaluated statistically. Criterion validity examines how a new measure relates to an established ‘gold standard’ in the field. It can be measured by the Pearson correlation coefficient between the score for the questionnaire and the ‘gold standard’. Construct validity assesses if a single underlying entity is being measured and is assessed using correlation coefficients between scale scores.

Sensitivity to change or responsiveness indicates whether the survey is able to detect clinically significant changes. It is assessed by comparing outcome scores before and after an intervention and is defined as the difference between the mean pre- and post-operative scores divided by the standard deviation of the pre-operative scores.

A detailed description of the development and validation of a new patient-reported outcome score, the Oxford elbow score has been presented in this journal.8

**Generic quality-of-life outcome measures**

Generic outcome measures aim to assess all dimensions of health-related quality-of-life. The World Health Organisation Quality-of-Life group has recommended that five dimensions should be assessed in any generic quality-of-life survey: physical health, psychological health, social relationship perceptions, function and well-being.9 Generic outcome measures are used across a wide range of medical and surgical specialties. Commonly-used measures are the medical outcomes study 36-item short-form health survey (SF-36),10 the medical outcomes study 12-item short form health survey (SF-12),11 the Nottingham health profile12 and the European quality-of-life 5-dimension (EuroQol) questionnaire.13

The **SF-36 score.** This is a 36-item questionnaire which explores health over the previous four weeks.10 Each question has a choice of between two and six answers on a Likert-type scale and each answer is scored between 0 (worst health) and 100 (best health). The questions cover eight health concepts: bodily pain, physical functioning, role limitations due to physical health, general health, mental health, vitality, social functioning and role limitations due to emotional health.

The eight health concepts can be grouped into two higher-order clusters known as the physical component summary (calculated from the bodily pain, physical functioning, role limitations due to physical health and general health scores) and the mental component summary (calculated from the mental health, vitality, social functioning and role limitations due to emotional health scores). The former is most responsive to treatments which alter physical symptoms such as hip replacement.14

The SF-36 score was developed in American English but a United Kingdom English version is available. It takes five to ten minutes to complete and is suitable for self-administration or administration by a computer or by an interviewer. Scores are aggregated without standardisation or weighing. The SF-36 score is one of the most evaluated generic questionnaires and is known to be valid and consistent,15,16 sensitive14 and reproducible.10 In patients undergoing total hip replacement, the SF-36 score has been shown to be valid17 and reliable.18 However, it does have minor ‘floor’ and ‘ceiling’ effects.19,20 ‘Floor’ effect refers to data that cannot be less than a defined minimum number. Any deterioration in this group will not be detected. ‘Ceiling’ effect refers to the opposite situation where data cannot be greater than a defined maximum number. In this group, improvement will not be detected.

The **SF-12 score.** This is a short form of the SF-36 score and consists of 12 of the 36 questions, but if the sample size is sufficiently large, the SF-12 can produce profiles of the eight SF-36 health concepts. The SF-12 scoring algorithms are weighted and a computer program is available for scoring.

The advantages of the SF-12 over the SF-36 score are that it improves efficiency and lowers cost. The main disadvantage is that it has less construct validity and sensitivity thus producing less precise scores for the eight-scale health profile.11

The **Nottingham health profile questionnaire.** This is a self-administered questionnaire which takes five to ten minutes to complete. It was developed in English and consists of two parts. Part I contains 38 ‘yes/no’ items covering six dimensions: pain, physical mobility, emotional reactions, energy, social isolation and sleep. Part II has seven ‘yes/no’ questions concerning problems of daily living. It has been shown to be internally consistent, valid, reproducible and sensitive.21 No psychometric analysis has been performed on patients undergoing hip replacement.

The Nottingham health profile has one major disadvantage when compared with the SF-36 score concerning the dichotomous ‘yes/no’ response format. This restricted response format explores only ill health whereas the SF-36 score with its multiple-response options can detect positive as well as negative states of health. This produces higher ceiling effects in all dimensions for the Nottingham health profile when compared with the SF-36 score.22 They both have equal minor floor effects.

The **EuroQol questionnaire.** This is self-administered and takes approximately five minutes to complete. The first part contains 15 questions which explore five health dimensions: mobility, self-care, usual activities, pain and depression. The three possible replies are: ‘no problem’, ‘moderate problem’ or ‘extreme problem’. The second part examines the patient’s perception of their overall health and contains a 100-point visual analogue scale.

The questionnaire is known to be both valid23 and reliable,24 but EuroQol suffers from the same ceiling effects as the Nottingham health profile because of the restricted response format. In patients undergoing hip replacement test-retest reliability has been shown22 and there is evidence of construct validity and responsiveness.26
Disease-specific quality-of-life outcome measures
These provide patient-centred information about a particular disease. This allows comparison of different surgical and medical treatment options for that disease. Quality-of-life outcome measures commonly used to assess arthritis of the hip are the Western Ontario and McMaster Universities (WOMAC) osteoarthritis index and the arthritis impact measurement scales. Although these measures are specific for arthritis, they can be used to assess any joint and any intervention.

The WOMAC osteoarthritis index. This was developed in Canadian English for patients with osteoarthritis. The original version has undergone several refinements and WOMAC 3.1 has been the standard form for several years. It is self-administered and contains 24 questions covering three dimensions: pain, stiffness and physical function. The standard version uses a 48-hour timeframe. The WOMAC index is available in a five-point Likert, 100 mm visual analogue and 11-point numerical rating format. It is valid, reliable, responsive, easy to complete and simple to score. Most clinical studies use the Likert and visual analogue versions of WOMAC 3.1. The index has been extensively evaluated in patients undergoing hip replacement and has been shown to be responsive, have high internal consistency and acceptable test-retest reliability.

The arthritis impact measurement scale. This was developed in American English to measure outcome in patients with rheumatic disease. It has since been shown to be sensitive in patients with osteoarthritis. The arthritis impact measurement scale is a revised version and consists of 78 questions covering 12 scales: mobility, walking and bending, hand and finger function, arm function, self-care tasks, household tasks, social activity, support from family and friends, arthritis pain, work, level of tension and mood. The scores of each scale are normalised to give a value from 0 (good health) to 10 (poor health). This is self-administered and takes approximately 20 minutes to complete. It has been shown to be both valid and reliable.

A further version of the arthritis impact measurement scale has been produced specifically for patients undergoing hip replacement. This questionnaire consists of 57 items which are scored and weighted to produce four subscales: physiological function, self concept, role function and interdependence. Responsiveness, content and construct validity have been proved for this version of the system.

Hip-specific outcome measures
The outcome of total hip replacement (THR) was initially assessed by the surgeon using tools such as the Harris hip score (HHS) and Charnley score. Patients and surgeons often differ in their priorities and concerns and therefore hip-specific quality-of-life surveys were developed to elicit the patient’s perception of the outcome of surgery. Such surveys in common use today include the Oxford hip score (OHS) and the hip disability and osteoarthritis outcome score.

The HHS, Charnley score and the OHS were all developed to assess patients undergoing THR, irrespective of the underlying diagnosis. The hip disability and osteoarthritis outcome score can be used to assess any intervention on any hip pathology.

Harris hip score. This was developed in English in 1969. The assessment is performed by a surgeon and consists of eight questions and a physical examination. The questions are grouped into three categories: pain, function and level of activity, and the physical examination involves assessing the range of movement of the hip. Scores from each section are simply added together to make a maximum possible score of 100 (indicating the best possible outcome).

The original surgeon-reported measure was modified to create a patient-reported measure. This contains seven items: pain, support for walking, limping, walking distance, climbing stairs, putting on shoes and socks and sitting. Each item is reported using a Likert-type scale with between three and seven possible responses. This version of the HHS gives an overall score of 0 to 100 in which a lower score represents better health status. It has been found to be both valid and reliable in the assessment of the outcome of THR.

Charnley score. This was devised in 1972 and grades hip pain, mobility and walking on a six-point scale. The assessment is performed by the surgeon and lower scores indicate greater disability. The assessment is simple to perform but reflects the opinions of both the surgeon and the patient. There is no psychometric testing of the Charnley score supporting its use.

Oxford hip score. This was developed in English to assess disability in patients undergoing THR. The survey takes approximately five minutes to complete and assesses health over the previous four weeks. The measure contains 12 questions which assess pain and functional ability and each question has five Likert-type response choices. The overall score ranges from 12 to 60 with a higher score indicating greater disability.

The OHS has been used extensively in orthopaedics and is highly sensitive for patients undergoing primary and revision THR. It is internally consistent, reproducible and valid.

The hip disability and osteoarthritis outcome score. This is a 40-item questionnaire based on the knee disability and osteoarthritis outcome score. The questionnaire is self-administered and takes eight to ten minutes to complete. Each question has five possible answers which are scored from 0 to 4. The questions are grouped into five subscales: pain, other symptoms, activities of daily living, sport and hip-related quality of life. The score for each subscale is simply the sum of the individual question scores. The hip disability and osteoarthritis outcome score is then transformed on to a scale from 0 to 100 with 100 indicating the best possible outcome.
This scoring system contains all the WOMAC Likert 3.0 questions in an identical form and therefore can be used to calculate the WOMAC scores. It has been shown to be both valid and responsive.39

Short-term clinical outcome measures
The short-term clinical impact of operations and the associated physiological disturbance can be described in different ways. Although mortality is important it is now so infrequent for most types of surgery that it is not useful as a comparative index. Length of hospital stay has been used as a marker of clinical outcome, but is recognised to be a compound measure of clinical status and the way in which a hospital functions.44

The post-operative morbidity survey1 is a measure of post-operative morbidity which focuses on indicators of dysfunction of organ systems that are easy to collect. Data for the post-operative morbidity survey (Table I) are collected from in-patients on selected post-operative days using observation charts, medication charts, patient notes, routine blood test results and direct patient questioning and observation. It requires no additional investigations. It can be used to assess short-term morbidity after any type of surgery and has been used in outcomes research45 and in effectiveness research.46 It has been shown to be reliable, valid and acceptable to patients.3

The post-operative morbidity survey identifies morbidity which is sufficient to delay discharge from hospital. Therefore, as well as providing useful research and audit data, it can assist with the planning of discharge and contribute to cost reduction.

Conclusions
Orthopaedic interventions are assessed on their outcome. There are several ways of assessing outcome including short-term post-operative measures and long-term patient-centred quality-of-life outcome measures. Since short-term post-operative surveys are only now evolving, it remains to be seen whether they will show any correlation with long-term outcome questionnaires. It is, therefore, advisable to use a combination of both measures to assess any operative intervention.

There are three main categories of long-term quality-of-life outcome measures, namely, generic, disease-specific and joint-specific. The most psychometrically evaluated generic measure is the SF-36 score, the most evaluated disease-specific measure is the WOMAC index and the most evaluated hip-specific measure is the OHS. Any hip intervention should be assessed by both a generic and specific outcome measure. They examine different aspects of the patient’s function and together provide more information than one questionnaire alone.

The choice of which generic and specific questionnaire to use depends on several factors including the aim of the study, the level of observation required, the funding available and the context of the study. Long-term follow-up is required, and therefore simple, short questionnaires may make long-term studies more feasible.

References

Table I. Criteria for a positive post-operative morbidity survey score (reproduced with permission from Elsevier3)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Criteria for positive result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pulmonary</td>
<td>Requires supplementary oxygen or ventilatory support</td>
</tr>
<tr>
<td>Infection</td>
<td>Currently on antibiotics or temperature &gt; 38°C in the last 24 hours</td>
</tr>
<tr>
<td>Renal</td>
<td>Oliguria (&lt; 500 ml/day), elevated creatinine (&gt; 30% pre-operative level), catheter in situ (for non-surgical</td>
</tr>
<tr>
<td></td>
<td>reason)</td>
</tr>
<tr>
<td>Gastrointestinal</td>
<td>Unable to tolerate enteral diet for any reason</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>Diagnostic tests or treatment within the last 24 hours for myocardial infarction, hypotension (requiring</td>
</tr>
<tr>
<td></td>
<td>pharmacological therapy or fluids &gt; 200 ml/hour, atrial/ventricular arrhythmia or cardiogenic pulmonary</td>
</tr>
<tr>
<td></td>
<td>oedema</td>
</tr>
<tr>
<td>Central nervous</td>
<td>Presence of new focal deficit, coma, confusion, delirium</td>
</tr>
<tr>
<td>complications</td>
<td>Wound dehiscence requiring surgical exploration or drainage of pus from operative wound with or without</td>
</tr>
<tr>
<td></td>
<td>isolation of organisms</td>
</tr>
<tr>
<td>Haematological</td>
<td>Requirement of blood transfusion, platelets, fresh-frozen plasma or cryoprecipitate within the last 24</td>
</tr>
<tr>
<td></td>
<td>hours</td>
</tr>
<tr>
<td>Pain</td>
<td>Wound pain requiring parenteral opioids or regional anaesthesia</td>
</tr>
</tbody>
</table>
OUTCOME MEASURES FOR ORTHOPAEDIC INTERVENTIONS ON THE HIP


