The use of outcome scores in surgery of the shoulder

The pursuit of ‘best practice’, health economic planning, the increasing awareness and expectations of patients, pressure from politicians and the media, and the emergence of league tables for surgeons are some of the reasons why orthopaedic surgeons are encouraged to adopt evidence-based strategies for managing their patients. Levels of evidence have been devised which allow publications to be ranked or given a grade of recommendation.\(^1,2\) The highest levels are assigned to well-designed, randomised, controlled trials and systematic reviews of such trials.

Lower levels are offered by cohort studies in which patients are compared with a control group treated at the same time and in the same institution. Such studies are ranked higher than randomised trials of poor quality, retrospective cohort studies or case-control studies. Individual case series and poorly designed cohort studies are lower still while the final level is expert opinions without critical appraisal and descriptive studies or reports from expert committees (Table I).\(^3-5\) Proper studies require good design and the use of validated outcome measures. We have carried out a systematic review of the use of outcome scores and research methods in surgery of the shoulder to establish whether the literature provides suitable evidence on which to establish best practice.

Review of the literature

A systematic review was undertaken of all articles relating to the shoulder published in the Journal of Shoulder and Elbow Surgery, the Journal of Bone and Joint Surgery [Br] and the Journal of Bone and Joint Surgery [Am] between January 1992 and December 2002. After manual searching, all papers which documented any form of clinical outcome were included for more detailed review.\(^3-5\) Those relating to anatomy, pathology, biomechanics, engineering design or technical aspects which did not involve a clinical outcome were excluded. Each paper chosen was placed into one of 16 broad categories according to its subspecialty. The exact number of patients studied as well as the minimum, maximum and mean periods of follow-up were recorded. A ‘grade of recommendation’ and ‘level of evidence’ were assigned to each paper in accordance with the standards shown in Table I. All criteria used to describe a clinical outcome were recorded, whether in the form of observations such as power or range of movement, or by the use of a recognised scoring system. Each paper was reviewed to ascertain whether a description of the outcome method used and the reasons for its selection were included in the text. In particular, we looked for details of the original group of patients on which any outcome score was based. An outcome score was regarded as appropriate if it was used unmodified for a validated group of patients.

Results

We reviewed 1106 articles relating to surgery on the shoulder. Of these, 496 were excluded on the basis of non-clinical content. The remaining 610 underwent more detailed review. There were 198 case reports and 379 cohort studies, the latter including 19 RCTs, but no systematic reviews (Table I). The mean sample size was 42 (1 to 1063). The overall mean follow-up was 27 months (1 to 540) with a minimum of 12 (1 to 540) and a maximum of 68 months (1 to 540). A formal outcome was described in 569 (93.3%) articles. Of these, 271 (47.6%) used clinical assessment, 217 (38.1%) an outcome score and 81 (14.2%) both. A total of 44 different outcome scores were encountered, 22 clinician-based (50.0%), 21 patient-based (47.7%) and one clinician- and patient-based (2.3%). Of 439 applications of an outcome score, 266 (60.6%) were clinician-based, 105 (23.9%) patient-based and 68 (15.5%) clinician- and patient-based. Trends in the use of the different types of score are shown in Figure 1. Of 298 articles using outcome scores, 126 (42.3%) described the details of the score within the text, but only eight (2.7%) made clear the reasons for the choice of the particular score.
Closer scrutiny of the use of clinical assessment in 352 articles showed a mean of 2.3 observations (1 to 6) per article. Those used were range of movement (208), pain (202), function/activities of daily living (129), power (88), radiological appearance (83), patient satisfaction (67) and stability (47). In the 298 articles using a formal outcome score a mean of 1.5 outcomes (1 to 6) was used per article. Overall, of the 439 applications of an outcome score 282 (64.2%) were regarded as being appropriate (Fig. 2). All formal outcome measures identified during the course of this review are listed.

Discussion
The proposal that clinical outcome in orthopaedic surgery could be analysed systematically so that patients would receive increased benefits from their treatment was first introduced by Codman et al\(^2\) in the second decade of the 20th century, and is the basis of his concept of the “End Result”. Unfortunately, his peers did not share his enthusiasm. Codman’s frustration culminated at a meeting on January 6, 1915 in which he ridiculed his colleagues and members of the hospital board, portraying them in a large cartoon as an ostrich burying its head in the sand and choosing to ignore what was happening around it. Codman’s career declined thereafter and he died in relative anonymity. Systematic reviews of randomised, controlled trials offer the maximum levels of evidence upon which clinical decisions can be based. No such reviews were found in the course of this investigation. Although 19 randomised, con-
trolled trials (3.1%) were identified, 538 papers (88.2%) described case series offering low levels of evidence. The undertaking of a randomised, controlled trial for a surgical procedure is costly and time-consuming. Nevertheless, increased use of cohort or case-control studies would considerably improve the level of evidence available.

The use of validated outcome scores allows comparisons to be made between studies. If scores are modified or used on inappropriate groups of patients, such comparisons are flawed. The European Society for Surgery of the Shoulder and Elbow and the Japanese Orthopaedic Association have each given guidance on the preferred use of outcome scores. However, such recommendations are not uniformly accepted. Our review has shown that study cohorts are generally small, periods of follow-up short and levels of evidence low. The overall pattern of the application of an outcome score is highly variable and at times inappropriate. We have identified changes made to outcome scores, often without proper testing of the modification and without justification. For example, the Neer rating was initially used to assess the outcome of displaced fractures of the proximal humerus, but was modified to assess total shoulder arthroplasty and, more recently, repair of the rotator cuff, although its formal statistical validation for use with these differing groups has not been undertaken.

The score of Constant and Murley is widely used, but large variations occur in how it is formulated. Pain is often measured using separate visual analogue scales, the methods of measuring power vary and, most importantly, the fact that scores should be normalised for age and gender is selectively ignored. The application of objective clinical assessment of pain, range of movement, power and stability are acceptable means of measuring outcome. However, the means by which such assessments are measured and documented and the number of such criteria used in studies is variable. Scores may be patient-based such as the Oxford shoulder score, clinician-based as the Constant-Murley score or a combination of both as in the modified American Shoulder and Elbow Surgeons form. There are condition-specific scores such as the Oxford shoulder instability score and non-condition-specific scores such as the simple shoulder test. In recent years there has been a proliferation of patient-based outcome scores recognising the benefits of such scores compared with clinician-based assessments. The latter are susceptible to bias and error, and may not represent the view of the patient. Patient-based scores are designed for use in clinical trials and are valid for comparing and aggregating cohort studies. Their use will directly improve levels of evidence. Despite the trend to move away from the application of clinician-based outcome scores, our review has shown that in practice the magnitude of this shift is highly variable. Over the last decade the use of clinician-based scores has remained high. An overall understanding of the initial population upon which scores were first based is lacking. Newer scores such as the shoulder pain and disability index (SPADI) were initially based on a cohort of 37 male patients with shoulder pain which was either musculoskeletal, neurogenic or of unknown aetiology. The patient self-reporting section of the modified American Shoulder and Elbow Surgeons assessment form (M-ASES) has undergone validation. However, this was based on only 63 patients, 25 of whom had impingement, but only one had undergone hemiarthroplasty and two had tears of the rotator cuff. The use of outcome scores on cohorts for which they have not been validated casts doubt on the validity of the results.

Conclusion
Forty-four different outcome scores were encountered in the course of this review, many being applied inappropriately. There is a trend towards the increased use of validated patient-based scores, but many have not been properly tested for validity, repeatability and sensitivity to change. Scores are not valid when used in a modified form and their use should be discouraged. Levels of evidence were generally low, with 88.2% of level 4, and with only a small number of RCTs. Improvement in the design of the studies and the use of appropriately validated outcome scores would substantially increase the levels of evidence on which to base best practice in surgery of the shoulder.

Supplementary material
A table showing the list of outcome scores identified in the course of this review is available with the electronic version of this article, on our website at www jbjs org uk.

We wish to thank Mrs Pat Deele, Academic Secretary to Professor A J Carr, for her assistance in collating the review of the literature.

References


