Levels of evidence and the orthopaedic surgeon

D. Tovey, G. Bognolo
From London, England

Readers of medical journals should ask themselves two questions on reading a piece of original research. Firstly, what are the direction and magnitude of any specified effect? Secondly, how much can I believe it? In attempting to help clinicians address the second of these questions, the American Volume of The Journal of Bone and Joint Surgery requires all contributors to specify the level of evidence for each submission. The intention to familiarise surgeons, for whom evidence-based medicine terminology might not be their first language, with the concept of grading studies, is laudable.

The adoption of the principles of ‘evidence-based’ medicine with respect to decisions about surgical procedures is not universally accepted. Surgeons are not alone in challenging the assumption that randomised, controlled trials are always the best way of answering a clinical question.1 These objections are often made on the grounds of practicality and the difficulty of applying evidence-based approaches to treatments which rely heavily on individual expertise and the clinical environment. Surgeons may also be uncomfortable with carrying out an unfamiliar procedure, and feel that having to do so would be unethical.

Methodological difficulties exist in designing surgical trials. Blinding is difficult and often impracticable, but solutions can be found. One proposed approach is to conduct expertise-based, randomised trials, where patients are randomised to a procedure, and then allocated to the surgeon with the best skill to perform it.2 This may provide patients with the best expert in carrying out the procedure, and allow evaluation of a technique performed at its best.

A further obstacle to understanding and acceptance is an over-rigid approach to evidence hierarchies. If the systematic review of high-quality randomised, controlled trials (RCTs) is always the best approach, then everything else must, by definition, be inferior. Expert opinion, on this scale, is barely above useless. However, it is clearly undesirable, in the quest for understanding, to undervalue anything that does not come in the form of an RCT. In truth ‘lesser’ forms of evidence provide many insights, hail many breakthroughs and warn of impending disasters in ways which would simply be impossible with RCTs which are constructed to demonstrate benefits. The case series which highlights the rare unanticipated harm is crucial, even if it does take a subsequent randomised trial to test the association conclusively.

In most circumstances, however RCTs are feasible and are the most reliable way of comparing alternative treatments. Perhaps the most infamous example in support of this was the over-enthusiastic promotion of hormone replacement therapy (HRT) for symptoms of the menopause in women based on observational data. This evidence which appeared to show that HRT was associated with favourable cardiovascular outcomes, proved to be misleading following an RCT. The consequence of the adoption of the earlier evidence has been a painful lesson for clinicians and more importantly, for many patients.3

Some subtle inconsistencies among the different systems for grading evidence make their interpretation difficult for occasional users. Further confusion may arise when assessing prognostic, diagnostic and economic studies, where some of the frameworks used for grading treatment may not be appropriate. The system used by the American Volume of The Journal of Bone and Joint Surgery is based on the Oxford Centre for Evidence-based Medicine Levels of Evidence (May 2001) framework.4 It uses letters, numbers and numerals to quantify the strength and quality of the evidence. This is similar to other systems, including those of the Scottish Intercollegiate Guidelines Network (SIGN), the American Heart Association (AHA) and the American Collect of Chest Physicians (ACCP). The rating of the quality of evidence of individual studies is problematical and the various systems, includ-
ing that of the Oxford Centre, recommend that they should be based on a systematic review of the literature rather than on single studies.

An alternative approach is proposed by the GRADE working group,\(^5\) an informal network which has been collaborating over several years to develop a grading system which accommodates both the quality of evidence and the strength of effect in a simple, comprehensible, form with a symbol-based solution. Under the GRADE system, studies are given an initial rating (randomised trial: high; observational study: low; any other form of evidence: very low) and this is amended both up and down by further consideration of other parameters.\(^5\) Methodological failings, a small sample size and inconsistency of results downgrade the quality rating whilst factors such as a strong evidence of association or a dose-response gradient can improve the ranking. Evaluations by GRADE are based on systematic reviews, and there is uncertainty about how effective this system may be to assess individual studies.\(^6\)

Systems for grading the quality of evidence are important in determining the level of confidence that can be applied to published studies. The models which are presently available may sometimes be confusing and inconsistent in their conclusions. Interpreting grading systems in an inflexible way - RCT good, everything else bad - is over simplistic. Perhaps it is better to think of the hierarchy as one of progression: lower levels of evidence playing their part by generating hypotheses, randomised trials coming into their own by testing them. To the extent that publishing the grading systems enables readers to understand the place of an article on this journey, they are helpful. If the message goes out that nothing but an RCT is worth reading, it is surely wrong.

References