

THE JOURNAL OF BONE & JOINT SURGERY

JB&JS

This is an enhanced PDF from The Journal of Bone and Joint Surgery

The PDF of the article you requested follows this cover page.

A Practical Guide to Assigning Levels of Evidence

James G. Wright

J Bone Joint Surg Am. 2007;89:1128-1130. doi:10.2106/JBJS.F.01380

Reprints and Permissions

Click here to [order reprints or request permission](#) to use material from this article, or locate the article citation on jbjs.org and click on the [Reprints and Permissions] link.

Publisher Information

The Journal of Bone and Joint Surgery
20 Pickering Street, Needham, MA 02492-3157
www.jbjs.org

THE ORTHOPAEDIC FORUM



A Practical Guide to Assigning Levels of Evidence

By James G. Wright, MD, MPH, FRCSC

Evidence-based medicine uses the best available evidence to make decisions with patients. The highest-quality evidence is well-designed randomized trials. The most compelling example of the power of trials comes from pediatric oncology. The improvement in the survival rate of children with cancer from 10% to 90% has been attributed almost exclusively to multiple randomized trials¹. Since January 2003, every clinical article published in *The Journal of Bone and Joint Surgery* has been assigned a level-of-evidence rating². Levels of evidence provide a concise and simple appraisal of study quality. The essence of levels of evidence is that, in general, controlled studies are better than uncontrolled studies, prospective studies are better than retrospective studies, and randomized studies are better than nonrandomized studies. Levels of evidence have multiple purposes. First, levels of evidence provide the readers of *The Journal* with a rapid appraisal of study quality. Although a complete critical appraisal is required to determine study quality³, readers generally find higher levels of evidence more compel-

ling. Second, levels of evidence for multiple studies evaluating a clinical question can be summarized as a “grade of recommendation.” Grades of recommendations, such as A, B, C, or I, provide an overall appraisal of the quality of literature for or against a treatment recommendation⁴. Third, levels of evidence can be used to develop practice guidelines and performance measures. In an era of pay-for-performance, whereby evidence-based interventions receive higher reimbursement^{5,6}, understanding levels of evidence is important for surgeons.

In considering levels of evidence, surgeons need to know that the evaluation is reliable and valid. Several studies have shown that the assignment of levels of evidence is reliable^{7,8}. The first table explaining the level-of-evidence ratings, published in the January 2003 issue of *The Journal*², was revised in July 2004 to further simplify it by eliminating subcategories and providing more explicit criteria for the categorizations to further improve the reliability (see Instructions to Authors). The level-of-evidence ratings have also been shown,

as one measure of validity, to be correlated with the citation index⁸. The assignment of levels of evidence, however, is not always straightforward. A clear understanding of how to assign a level-of-evidence rating would be useful for authors of submitted articles, readers of this and other journals, participants in journal clubs, and clinical groups responsible for developing practice guidelines and performance measures. Furthermore, as level-of-evidence ratings have been provided for articles published in *Arthroscopy*, *The Journal of Hand Surgery*, and *Clinical Orthopaedics and Related Research*, numerous queries have arisen regarding the assignment of levels of evidence.

As the Associate Editor for Evidence-Based Orthopaedics, I have reviewed and assigned a level-of-evidence rating to every clinical article published in *The Journal* since January 2003. The purpose of this article is to provide a practical guide to assigning level-of-evidence ratings to the orthopaedic clinical literature. The three steps in assigning a level-of-evidence rating are determining the primary re-

Disclosure: In support of their research for or preparation of this work, the author received, in any one year, outside funding or grants in excess of \$10,000 from the R.B. Salter Chair in Surgical Research. Neither he nor a member of his immediate family received payments or other benefits or a commitment or agreement to provide such benefits from a commercial entity. No commercial entity paid or directed, or agreed to pay or direct, any benefits to any research fund, foundation, division, center, clinical practice, or other charitable or nonprofit organization with which the author, or a member of his immediate family, is affiliated or associated.

search question, establishing the study type, and assigning a level.

Determining the primary research question. A level of evidence is applied only to the primary research question of a study. Authors should specify the primary research question of their study. Authors, however, are not always clear about the primary research question, and some studies may have multiple purposes. Journal reviewers and editors have a role to ensure that the study purpose and hypothesis are explicit.

The primary research question can be found in several locations. The abstract, intended to be a concise summary of the article, is the first potential source for the primary research question. The abstract through virtue of brevity often states and provides the most succinct description of the primary research aim. If the research question is not explicit in the abstract, the next step is to review the article introduction. The introduction of the article frames the research study and usually ends with an explicit research question. If the primary research question is still not clear, the next step is to review the major conclusions of the article. The conclusions should correspond with what the authors thought would have been, if they had specified, the primary research question. In the case of multiple conclusions, one should select what appears to be the major conclusion (or if there is no clear major conclusion, the first conclusion). Finally, in those rare cases in which the primary research question remains unclear, one should review the results and evaluate where the majority of the analyses have been directed to determine the primary research question.

Establishing the study type. Once the primary research question is established, the next step is to determine the study type. Levels of evidence can be assigned to the following four study types: therapeutic, prognostic, diagnostic, and economic decision analyses. The distinction between prognostic and therapeutic studies provides the most confusion because both evaluate the effect of one or more factor(s) on the de-

velopment or outcome of disease.

Therapeutic studies evaluate the effect of treatment on the outcome of disease, whereas prognostic studies evaluate the effect of patient characteristics on the outcome of disease. A pragmatic test to differentiate between therapeutic and prognostic studies is to consider whether the factor could be randomly allocated. If it can be randomly allocated, one is dealing with a therapeutic study. On the other hand, patient age or fracture type cannot be randomly allocated to two groups of patients, and therefore any study evaluating the effect of patient age or fracture severity on outcome would be a prognostic study. As another example, the investigation of the stage of disease or the presence of blood coagulation factors, which cannot be randomly allocated, on the outcome of Legg-Calvé-Perthes disease would be a prognostic study.

The remaining two study types are diagnostic and economic or decision analyses. Diagnostic studies specifically evaluate whether the results of a test are related to the presence or absence of a disease. For example, a physical examination may be used to detect shoulder instability when the definitive test or so-called gold standard is arthroscopy⁹. Economic and decision analyses are modeling exercises and relatively easy to identify. Economic analyses evaluate and compare the costs of care. For example, a study may determine the relative cost-effectiveness of operative and nonoperative treatment for calcaneal fractures¹⁰. Decision analyses evaluate and compare the outcomes of care, such as determining what the preferred treatment is for slipped capital femoral epiphysis¹¹. Finally, some study types, such as those describing the development of new measures and reliability or validity studies, are not included in the current version of the levels-of-evidence table.

Assigning a level. Once the study type has been chosen for the primary research question, the next step is to assign a level of evidence within each of the four study types. Therapeutic type-

I studies are defined as high-quality positive or negative randomized clinical trials.

“Positive” trials have significant differences between groups in one or more important outcome in favor of one treatment. “Negative” trials do not report a significant difference in favor of one treatment. A critical requirement of a high-quality Level-I “negative” study is sufficient power. If a negative randomized trial has insufficient power, then the rating is Level II. Sample size determination, usually found in the methods section, should be performed before the onset of the trial. Among the parameters required for a sample size determination is a prespecified clinically significant difference. The clinically significant difference is that difference above which the investigators believe that clinicians would be led to adopt a new therapy and below which the clinicians would conclude that the difference between two treatments was not clinically meaningful. If the trial does not demonstrate a significant difference, the confidence interval around the primary outcome should not include the clinically significant difference. If the confidence interval does not include the clinically significant difference, then the trial has a narrow confidence interval and is rated as Level I. If the confidence intervals for the difference between the two treatment groups (including the clinically important difference) are wide, then the study is a Level-II randomized clinical trial. If a sample size calculation was not included in the methods, then one should look for a power analysis. If a post hoc power analysis identifies power of <0.8, then the study is identified as Level II.

Although determination of a “high-quality” randomized clinical trial requires a complete critical appraisal of all elements of the study design¹², the critical characteristics of a lesser-quality trial (and therefore warranting a Level-II designation) include poor randomization technique, such as randomization by days of the week or hospital record number (from which

investigators can easily determine the randomization assignment); less than 80% follow-up; or evaluators who are unblinded to treatment assignment.

Meta-analyses are summaries of multiple studies. The level of evidence assigned to a meta-analysis is based on the literature used in the meta-analysis. If all Level-I studies are used, then the meta-analysis is Level I. However, if the meta-analysis is based on Level-IV studies, then it is Level IV.

An important criterion for differentiating between Level-II and Level-III studies is whether the study is prospective or retrospective. This is a confusing distinction because these terms have been and can be used in many ways¹³. Furthermore, different aspects of the study can be prospective or retrospective. For example, the decision to perform the study could have occurred long after patients received treatment and therefore be retrospective, whereas the collection of the outcomes of treatment could have occurred prospectively. For the *JBJS* levels of evidence, the term *prospective* is used for studies in which the study question was articulated before the first patient was enrolled and therefore before any patient data were collected. All studies that are not prospective are retrospective.

Another important distinction is between cohort and case-control studies. In a cohort study, one group of patients has a particular characteristic (or received a particular treatment) and another group of patients has a different characteristic (or received a different treatment). In a case-control study, patients are called “cases” on the basis of a particular outcome, such as failed hip replacement requiring revision, whereas “controls” refer to the patients who do not have the outcome, such as those with survival

of a hip replacement. In a case-control study, the two groups of patients, such as those with a total hip replacement that has failed or those with one that has not failed, are compared for the frequency of an intervention such as hip arthroplasty with or without cement. Cohort and case-control studies can be prospective if the study question is articulated before the first patient is enrolled, or they are retrospective if the study question is determined after the first patient is enrolled.

Case series are uncontrolled evaluations of the outcome of a group of patients treated in the same way, whether it be performed prospectively or retrospectively, and are assigned as Level IV.

Diagnostic studies have slightly different criteria for evaluation. The key concept for diagnostic studies is the availability and use of a so-called gold standard. The gold standard is the definitive diagnostic test, such as shoulder arthroscopy for clinical shoulder instability⁹. If the gold standard is applied inconsistently, such as only some patients receive shoulder arthroscopy or the study patients are selected or nonconsecutive, then the appropriate level is Level III. If the diagnostic study lacks controls by evaluating only patients with a disease, then the appropriate level is Level IV. For all types of studies, expert opinion constitutes Level-V evidence.

In conclusion, consistent criteria for the assignment of levels of evidence should improve the consistency of level assignment and increase surgeons’ understanding of levels of evidence. It is hoped that this summary provides useful advice and will generate discussion about future improvements in this discipline.

James G. Wright, MD, MPH, FRCSC
The Hospital for Sick Children, 555 University Avenue, Toronto, ON M5G 1X8, Canada.
E-mail address: james.wright@sickkids.ca

References

1. Wittes RE. Therapies for cancer in children — past successes, future challenges. *N Engl J Med*. 2003;348:747-9.
2. Wright JG, Swiontkowski MF, Heckman JD. Introducing levels of evidence to the journal. *J Bone Joint Surg Am*. 2003;85:1-3.
3. Spindler KP, Kuhn JE, Dunn W, Matthews CE, Harrell FE Jr, Dittus RS. Reading and reviewing the orthopaedic literature: a systematic, evidence-based medicine approach. *J Am Acad Orthop Surg*. 2005;13:220-9.
4. Wright JG, Einhorn TA, Heckman JD. Grades of recommendation. *J Bone Joint Surg Am*. 2005;87:1909-10.
5. Doran T, Fullwood C, Gravelle H, Reeves D, Kontopantelis E, Hiroeh U, Roland M. Pay-for-performance programs in family practices in the United Kingdom. *N Engl J Med*. 2006;355:375-84.
6. Rosenthal MB, Frank RG, Li Z, Epstein AM. Early experience with pay-for-performance: from concept to practice. *JAMA*. 2005;294:1788-93.
7. Bhandari M, Swiontkowski MF, Einhorn TA, Tornetta P 3rd, Schemitsch EH, Leece P, Sprague S, Wright JG. Interobserver agreement in the application of levels of evidence to scientific papers in the American volume of the *Journal of Bone and Joint Surgery*. *J Bone Joint Surg Am*. 2004;86:1717-20.
8. Obremsky WT, Pappas N, Attallah-Wasif E, Tornetta P 3rd, Bhandari M. Level of evidence in orthopaedic journals. *J Bone Joint Surg Am*. 2005;87:2632-8.
9. Farber AJ, Castillo R, Clough M, Bahk M, McFarland EG. Clinical assessment of three common tests for traumatic anterior shoulder instability. *J Bone Joint Surg Am*. 2006;88:1467-74.
10. Brauer CA, Manns BJ, Ko M, Donaldson C, Buckley R. An economic evaluation of operative compared with nonoperative management of displaced intra-articular calcaneal fractures. *J Bone Joint Surg Am*. 2005;87:2741-9.
11. Schultz WR, Weinstein JN, Weinstein SL, Smith BG. Prophylactic pinning of the contralateral hip in slipped capital femoral epiphysis: evaluation of long-term outcome for the contralateral hip with use of decision analysis. *J Bone Joint Surg Am*. 2002;84:1305-14.
12. Greenhalgh T. How to read a paper: the basics of evidence-based medicine. 3rd ed. Malden, MA: BMJ Books/Blackwell; 2006.
13. Feinstein AR. *Clinical epidemiology: the architecture of clinical research*. Philadelphia: W.B. Saunders; 1985.