Evidence-based medicine is “the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients.” Although this form of medical practice is often contrasted with decision making on the basis of opinion and experience (a phenomenon sometimes termed “eminence-based medicine”), it is important to note that the ideal practice of evidence-based medicine involves “integrating individual clinical expertise with the best available external clinical evidence from systematic research.”

Although the term “evidence-based medicine” did not appear in the medical literature until 1992, it has since become the dominant paradigm in the practice of clinical medicine. Orthopaedic surgery has been no exception, as evidence-based medicine has generally been welcomed by practicing orthopaedists.

Along with this increased emphasis on evidence in clinical decision making, there has developed a need for the critical analysis of scientific research. Studies may be subject to bias, confounding, and other shortcomings, which may limit the extent to which the data may be used to guide the clinical care of patients. Therefore, hierarchical level of evidence classification systems have been developed to aid in the assessment of research quality.

The movement toward evidence-based medicine has also created a need for mechanisms to synthesize the available evidence and distill it into a form that can be used effectively by practicing physicians. The sheer volume of the scientific literature, coupled with ever-present time constraints, means that it is no longer feasible for physicians to critically review all studies published in one’s field. Clinical practice guidelines, which provide recommendations for the management of specific conditions based on a systematic review of the best available evidence, are well placed to satisfy this growing need.

**Levels of Evidence**

**History and Purpose**

The concept of grading studies on the basis of their methodology was first proposed in a 1986 article on the use of antithrombotic agents. The Journal of Bone and Joint Surgery: American volume was the first orthopaedic publication to adopt the classification system, as it began assigning levels of evidence to all clinical studies published in the Journal in 2003. In the adoption of this classification system, the Journal’s goals were threefold: to familiarize authors, reviewers, and readers with the concept of levels of evidence; to improve orthopaedic studies via the explicit articulation of a primary research question; and to place clinical research studies into context for the reader.

Over the past few years, several other orthopaedic journals have adopted similar level of evidence classification systems, including Clinical Orthopaedics and Related Research, Arthroscopy, The American Journal of Sports Medicine, and The Journal of Hand Surgery. The American Academy of Orthopaedic Surgeons (AAOS) followed suit by adopting its own level of evidence classification system in 2005 (Table 1).

**Assigning Levels of Evidence**

There are three basic steps in the assignment of a level of evidence to a particular clinical study: determine the primary research question, establish the study type, and assign a level of evidence.
Table 1

Levels of Evidence for Primary Research Question

<table>
<thead>
<tr>
<th>Types of Studiesa</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Therapeutic Studies</td>
<td>Prognostic Studies</td>
<td>Diagnostic Studies</td>
<td>Economic and Decision Analyses</td>
</tr>
<tr>
<td>Investigating the results of treatment</td>
<td>Investigating the effect of a patient characteristic on the outcome of disease</td>
<td>Investigating a diagnostic test</td>
<td>Developing an economic model or decision model</td>
</tr>
<tr>
<td>Level I</td>
<td>High-quality randomized trial with statistically significant difference but narrow CIs</td>
<td>High-quality prospective study (all patients were enrolled at the same point in their disease with ≥80% follow-up of enrolled patients)</td>
<td>Testing of previously developed diagnostic criteria on consecutive patients (with universally applied reference “gold” standard)</td>
</tr>
<tr>
<td></td>
<td>Systematic reviewb of level I RCTs (and study results were homogeneous)</td>
<td>Systematic reviewb of level I studies</td>
<td>Systematic reviewb of level II studies</td>
</tr>
<tr>
<td>Level II</td>
<td>Lesser-quality RCT (eg, &lt;80% follow-up, no blinding, or improper randomization)</td>
<td>Retrospectivec study Untreated controls from an RCT</td>
<td>Development of diagnostic criteria on consecutive patients (with universally applied reference “gold” standard)</td>
</tr>
<tr>
<td></td>
<td>Prospectivec comparative studyd</td>
<td>Lesser-quality prospective study (eg, patients enrolled at different points in their disease or &lt;80% follow-up)</td>
<td>Systematic reviewb of level II studies</td>
</tr>
<tr>
<td></td>
<td>Systematic reviewb of level II studies or level I studies with inconsistent results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level III</td>
<td>Case-control studye</td>
<td>Case-control studye</td>
<td>Study of nonconsecutive patients; without consistently applied reference “gold” standard</td>
</tr>
<tr>
<td></td>
<td>Retrospectivef comparative studye</td>
<td>Systematic reviewb of level III studies</td>
<td>Systematic reviewb of level III studies</td>
</tr>
<tr>
<td>Level IV</td>
<td>Case series</td>
<td>Case series</td>
<td>Case-control study Poor reference standard</td>
</tr>
<tr>
<td>Level V</td>
<td>Expert opinion</td>
<td>Expert opinion</td>
<td>Expert opinion</td>
</tr>
</tbody>
</table>

RCT = randomized controlled trial; CI = confidence interval
a A complete assessment of quality of individual studies requires critical appraisal of all aspects of the study design.
b A combination of results from two or more prior studies.
c Studies provided consistent results.
d Study was started before the first patient enrolled.
e Patients treated one way (eg, cemented hip arthroplasty) compared with a group of patients treated another way (eg, uncemented hip arthroplasty) at the same institution.
f The study was started after the first patient enrolled.
g Patients identified for the study based on their outcome, called “cases” (eg, failed total arthroplasty), are compared with those patients who did not have outcome, called “controls” (eg, successful total hip arthroplasty).
h Patients treated one way with no comparison group of patients treated another way.


It is important to note that levels of evidence are applied to the primary research question of a given study only. Although a study may have multiple research questions, only one should be designated as primary. The primary research question of a given study should be specified by the authors, preferably at the time of study inception.

Once the primary research question has been determined, the study type must be established. In the classification scheme used by the AAOS as well as several orthopaedic journals, including The Journal of Bone and Joint Surgery, studies are categorized as therapeutic, prognostic, diagnostic, or economic/decision analysis (discussed in detail in the next sections). Although nearly all clinical studies can be placed into one of these categories, the categorization scheme is not exhaustive and it is possible for a particular study to not fit neatly into any one of the above categories. Only after the pri-
mary research question has been determined and the study type has been established can a specific level of evidence be assigned (Table 1).

**Therapeutic Studies**

Therapeutic studies investigate the effect of treatment on the outcome of disease and represent the most common type of study in the orthopaedic literature.16-18 Level I therapeutic studies are high-quality randomized controlled trials (RCTs), which are generally considered to represent the best possible evidence available. To be considered a high-quality study, an RCT must satisfy several criteria. It must be appropriately powered, either by detecting a significant difference (in the case of a “positive” trial) or documenting sufficient power (in the case of a “negative” trial). High-quality RCTs must use an appropriate randomization technique, in which allocation of the next study participant cannot be determined by members of the research team before the patient receives his or her treatment allocation. Rates of follow-up must be high – generally above 80%. Whenever possible, patients, caregivers, and researchers should be blinded to the treatment assignment. This is by no means an exhaustive list, and there are several other characteristics that must be fulfilled for a trial to be considered of high quality. However, one can reliably assume that studies not fulfilling the above criteria will generally not be considered level I evidence.

Level II therapeutic studies include lesser-quality RCTs (as previously discussed), as well as prospective comparative studies. Comparative studies (also known as cohort studies) involve the comparison of one group of patients treated in a particular way with another group of patients treated in another way. For example, a study comparing the outcomes of patients with intertrochanteric hip fractures treated with a sliding hip screw or a cephalomedullary device would be considered a comparative (or cohort) study. Although the distinction between prospective and retrospective can sometimes be confusing,15 AAOS considers prospective investigations to be that in which the study was initiated (the research question was posed) before the first patient was enrolled or treated.

Level III therapeutic studies include retrospective comparative studies (as previously discussed), as well as case-control studies. Case-control studies, which are retrospective, involve the comparison of one group of patients who have a particular outcome with another group of patients who do not have the outcome of interest. These “case” and “control” groups are compared to each other on the basis of characteristics plausibly associated with the outcome of interest. For example, a comparison of children who developed slipped capital femoral epiphysis to a similar group of children who did not develop this condition would be considered to be a case-control study. Such a comparison could be made on the basis of risk factors, such as obesity or sex.

Level IV therapeutic studies, which represent the most common type of study in the orthopaedic literature,16-19 are case series of patients treated in one way without any comparison group of patients treated in another way. Although the evidence provided by this type of study is relatively low in the level of evidence hierarchy, it is important to emphasize that these studies do have their place in the orthopaedic literature. If a level IV case series is well executed, it can provide important information to guide patient care. A well-designed case series is one that has 100% of patients with the same diagnosis, strict inclusion and exclusion criteria, a standard treatment protocol, patient follow-up at specified time intervals, well-defined outcome measures that include clinical parameters, and validated patient-derived instruments for functional assessment.17

Expert opinion, without the support of clinical data, is considered to be level V evidence. This is true for all study types, including therapeutic studies.

The terms “systematic review” and “meta-analysis” are often assumed to be interchangeable, but their definitions do differ slightly. Whereas a systematic review is a comprehensive literature search to identify studies appropriate for answering a particular clinical question, a meta-analysis is a statistical method of combining the data provided by these studies. Combining multiple studies into a single meta-analysis may address problems of small sample size and insufficient power, but it will not alter the level of evidence because meta-analyses are assigned levels of evidence based on the quality of the studies used in the meta-analysis.

**Nontherapeutic Studies**

Prognostic studies, which represent the second most common type of study in the orthopaedic literature,16-18 investigate the effect of a patient characteristic on the outcome of disease. Differentiating between therapeutic and prognostic studies can be difficult because both examine the effects of factors with the potential to influence the outcome of disease. James Wright, MD, MPH, Associate Editor for Evidence-Based Orthopaedics at The Journal of Bone and Joint Surgery: American volume, suggests considering whether the factor of interest can be randomized or not: “If [a factor] can be randomly allocated, one is dealing with a therapeutic study.”15 For example, an investigation of physical therapy on outcome after proximal humerus fracture would be considered a therapeutic study, whereas an investigation of the effect of age following this same injury would be considered to be a prognostic study. The criteria for assigning levels of evidence to prognostic studies are detailed in Table 1.

Diagnostic studies are the third most common type of study in the orthopaedic literature16-18 and evaluate the performance of tests designed to detect the presence or absence of a particular condition. Of central importance in the evaluation of a diagnostic test is the “gold standard,” a second diagnostic test generally regarded to provide the most definitive evidence for or against the presence of a particular condition. For example, a study evaluating the ability of physical examination to
detect a meniscal tear could use arthroscopy as the gold standard. For a diagnostic study to be considered of high quality, there must be a gold standard that is universally applied to all cases. (In the previous example, all study participants would have to undergo arthroscopy for it to be considered a high-quality diagnostic study.) Another feature of high-quality diagnostic studies is the inclusion of consecutive patients, as detailed in Table 1.

The final study type included in the AAOS level of evidence classification system is economic and decision analyses. This study type is relatively uncommon in the orthopaedic literature, as it has been found to represent less than 1% of clinical studies published in orthopaedic journals. Details on the criteria used to assign levels of evidence to these types of studies are described in Table 1.

Inter-rater Reliability in the Assignment of Levels of Evidence

For the level of evidence classification system to be most useful, there must be consistency of grade assignment between different raters. Recently, there have been several studies assessing inter-rater agreement in the assignment of levels of evidence to orthopaedic research, and inter-rater reliability has been found to range widely.16-18,20,21

One of the factors that appears to influence inter-rater reliability is the training and experience of the raters who are submitting level of evidence grades. For example, in a study of individuals reviewing manuscripts for The Journal of Bone and Joint Surgery: American volume, the authors found that reviewers with training in epidemiology demonstrated near-perfect agreement (intraclass correlation coefficient [ICC] of 0.99 to 1.00, versus 0.61 to 0.75 for reviewers not trained in epidemiology).16 Similarly, the authors of another study found that agreement in the assignment of level of evidence was higher among experienced reviewers (practicing orthopaedic surgeons) than between experienced and inexperienced reviewers (orthopaedic residents and medical students) (kappa 0.75 versus 0.62).17 In addition, the ability of orthopaedic residents to identify the level of evidence of 10 blinded articles from The Journal of Bone and Joint Surgery: American volume, was assessed. The mean percentage correct was found to be only 29.5% (41.3% after an educational intervention).21

Most recently, the feasibility of assigning levels of evidence to abstracts presented at the AAOS Annual Meeting was examined. In particular, the levels of evidence assigned by authors, volunteer graders (with access to the abstract only), and session moderators (with access to the full paper) were examined. Agreement ranged from slight to moderate (kappa 0.16 to 0.46).20 However, this study did not consider the experience or epidemiologic training of the raters in question. The study also found that authors tended to grade the level of evidence of their own work more favorably than did other graders,20 which is a finding deserving of further investigation.

Current Levels of Evidence in the Orthopaedic Literature

Recent studies have been fairly consistent in describing the current status of the orthopaedic literature. The most common study type is therapeutic (69% to 71%), followed by prognostic (20% to 25%), diagnostic (6% to 9%), and economic/decision analysis (0% to 0.5%).16,17 With regard to level of evidence, level IV studies comprise 54% to 58% of the orthopaedic literature, whereas level I evidence accounts for only 11% to 16%.16,18 Upon closer examination of the level I data, many of these studies appear to have shortcomings.22,23

However, recent studies provide reason to be optimistic regarding levels of evidence in the orthopaedic literature. In a 2009 study, 551 articles in The Journal of Bone and Joint Surgery: American volume published over the past 30 years were examined. The percentage of clinical studies providing level I evidence increased from 4% to 21% between 1975 and 2005, while the percentage of level IV studies decreased from 81% to 48% over this same period.24 An earlier study found that the number of randomized trials published in orthopaedic journals increased substantially between 1968 and 199925 (Figure 1). Similar trends have been observed within the subspecialty field of sports medicine, as a 2005 study documented significant increases in the number of randomized controlled trials and prospective cohort studies coupled with a significant decrease in the number of case series and descriptive studies between 1991-1993 and 2001-2003.26

In recent years, some researchers have sought to investigate the relationship between levels of evidence and other factors in orthopaedic research. Manuscripts
submitted to The Journal of Bone and Joint Surgery: American volume were studied; studies with a higher level of evidence were more likely to be accepted for publication. Articles published in three prominent general orthopaedics journals also were examined, and it was found that studies with a higher level of evidence were also more likely to be cited following publication. The relationship between level of evidence and declared funding support was examined in a recent study. It was determined that industry-funded studies were more likely to provide level IV evidence than were studies funded by not-for-profit sources. This last association has not been documented previously and is certainly deserving of further investigation.

Using Levels of Evidence in Orthopaedic Practice

The levels of evidence classification system provides a rapid assessment of study quality, which may help readers to quickly place studies into context. Because studies of higher level of evidence have greater methodologic safeguards against bias, they may provide better information to guide physicians in their care of patients. However, it is important to emphasize that a level of evidence rating does not tell the whole story with regard to a particular study. Assignment of a particular level of evidence represents an assessment of the study design as reported by the authors but does not address other factors such as quality of data gathering or interpretation. This is emphasized in the first footnote in Table 1, which states that “a complete assessment of [the] quality of individual studies requires critical appraisal of all aspects of the study design.” Levels of evidence cannot be used in a blind manner; orthopaedic surgeons must always consider all aspects of a given study, and integrate these data with their individual clinical expertise.

Clinical Practice Guidelines

History and Purpose

When medical decision making is based largely on opinion, wide variation in clinical practice is expected. Although variability certainly persists within the evidence-based model, large discrepancies in practice patterns invite investigation. It was in this context that the Institute of Medicine first proposed clinical practice guidelines, defined as “systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances.”

Since they were first proposed in 1990, the number of such guidelines has grown rapidly. By August of 2010, for example, there were nearly 2,500 guidelines listed in the National Guideline Clearinghouse. Over the past few years, AAOS has begun to publish clinical practice guidelines on a variety of topics within orthopaedic surgery. As of August 2010 there were nine such guidelines that had been approved by AAOS, with several more in various stages of the development process.

Developing Clinical Practice Guidelines

In the selection of topics for guideline development, ideal candidates include common conditions with a solid evidence base but wide variation in practice patterns. In these cases, clinical practice guidelines have the best opportunity to achieve their objectives of decreasing practice variability, optimizing clinical outcomes, and promoting cost efficiency.

The backbone of any clinical practice guideline is the literature review, which proceeds in a systematic and carefully considered manner. Retrospective case reviews are excluded, and underpowered studies are not considered (unless used as part of a de novo meta-analysis).

---

Table 2

<table>
<thead>
<tr>
<th>Grade</th>
<th>Overall Quality of Evidence</th>
<th>Description of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Good-quality evidence</td>
<td>More than one level I study with consistent findings for or against recommending intervention</td>
</tr>
<tr>
<td>B</td>
<td>Fair-quality evidence</td>
<td>More than one level II or III study with consistent findings or a single Level I study for or against recommending intervention</td>
</tr>
<tr>
<td>C</td>
<td>Poor-quality evidence</td>
<td>More than one level IV or V study or a single level II or III study for or against recommending intervention</td>
</tr>
<tr>
<td>I</td>
<td>No evidence or conflicting evidence</td>
<td>There is insufficient or conflicting evidence not allowing a recommendation for or against intervention.</td>
</tr>
</tbody>
</table>

On the basis of the best available evidence, recommendations are made regarding the management of patients. Each recommendation carries a level of evidence derived from the data underlying the recommendation, as well as a grade of recommendation (Table 2). Table 3 presents a sample of recommendations from a recent AAOS Clinical Practice Guideline on the management of pediatric diaphyseal femur fractures. Once a guideline has been created, it must undergo peer review and finally a thorough approval process. The entire process of developing an AAOS clinical practice guideline is quite extensive and usually takes 12 to 18 months to complete.

### Concerns Regarding Clinical Practice Guidelines

Although clinical practice guidelines have generally been welcomed by the evidence-based medicine community, this reception has not been universal. Some argue, for example, that clinical practice guidelines erode physician autonomy and run the risk of transforming clinical practice into “cookbook medicine.” Others complain that existing guidelines are too comprehensive or too narrowly focused, and quickly become outdated. Still others voice fears that guidelines will be used to critique the treatment decisions of physicians in legal and pay-for-performance settings.

More recently, critics have pointed to the potential for bias in the recommendations made by clinical practice guideline authors. In a recent study of physicians who authored 44 clinical practice guidelines on common adult diseases, it was found that 87% of the authors had some form of interaction with the pharmaceutical industry, including 59% who had relationships with companies whose products were considered in the guideline they authored. However, in only two cases were these personal financial interactions specifically disclosed in the final published guideline. Although clinical practice guidelines are meant to be completely objective, the development process does involve subjective judgments where competing interests could come into play. AAOS has taken several steps to combat bias in the development of clinical practice guidelines, including requiring full conflict of interest disclosure from all authors and using well-defined, systematic processes that are transparent and reproducible.

### Using Clinical Practice Guidelines in Orthopaedic Practice

Like levels of evidence, clinical practice guidelines are not instruments to be used blindly. Guidelines provide recommendations that are to be carefully evaluated by each physician and integrated with his or her clinical expertise. The Evidence-Based Medicine Working Group, writing in the *Journal of the American Medical Association* (JAMA) in 1995, proposed several questions that physicians should consider when using a particular clinical practice guideline. In particular, clinicians are urged to evaluate the validity and content of the guideline recommendations, as well as their applicability to the patient in question. By following the steps outlined in this article, physicians have the best opportunity to use clinical practice guidelines in a man-

### Table 3

Sample of Recommendations From a Recent AAOS Clinical Practice Guideline on the Treatment of Pediatric Diaphyseal Femur Fractures

<table>
<thead>
<tr>
<th>Recommendation</th>
<th>Level of Evidence</th>
<th>Grade of Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. We recommend that children younger than 36 months with a diaphyseal femur fracture be evaluated for child abuse.</td>
<td>II</td>
<td>A</td>
</tr>
<tr>
<td>2. Treatment with a Pavlik harness or a spica cast are options for infants 6 months and younger with a diaphyseal femur fracture.</td>
<td>IV</td>
<td>C</td>
</tr>
<tr>
<td>3. We suggest early spica casting or traction with delayed spica casting for children age 6 months to 5 years with a diaphyseal femur fracture with less than 2 cm of shortening.</td>
<td>II</td>
<td>B</td>
</tr>
<tr>
<td>4. It is an option for physicians to use flexible intramedullary nailing to treat children age 5 to 11 years diagnosed with diaphyseal femur fractures.</td>
<td>III</td>
<td>C</td>
</tr>
<tr>
<td>5. Rigid trochanteric entry nailing, submuscular plating, and flexible intramedullary nailing are treatment options for children age 11 years to skeletal maturity diagnosed with diaphyseal femur fractures, but piriformis or near piriformis entry rigid nailing are not treatment options.</td>
<td>IV</td>
<td>C</td>
</tr>
<tr>
<td>6. We are unable to recommend for or against removal of surgical implants from asymptomatic patients after treatment of diaphyseal femur fractures.</td>
<td>IV</td>
<td>Inconclusive</td>
</tr>
</tbody>
</table>

Levels of evidence and clinical practice guidelines are tools of the evidence-based medicine movement that can help physicians provide better care for their patients. They do not represent “cookbook” instructions to be followed blindly, but rather instruments to be carefully evaluated and integrated with clinical expertise.

There is reason to be optimistic regarding evidence-based practice in the field of orthopaedic surgery. Levels of evidence are steadily increasing in the orthopaedic literature, and randomized trials are becoming more common. Several carefully researched clinical practice guidelines have recently been approved by AAOS, and others are under development. These advances have the potential to not only enhance the orthopaedic evidence base but also improve patient care.

Annotated References


In this survey of Dutch orthopaedic surgeons, evidence-based medicine was welcomed and clinical practice guidelines were perceived as the best means of proceeding from opinion-based medicine to evidence-based practice.


In this study of articles published in Spine between January and June 2003, 16.1% were level I, 22.3% level II, 8.0% level III, and 53.6% level IV. With regard to study type, 43.8% were therapeutic, 37.5% prognostic, 17.9% diagnostic, and 0.9% economic.


In this study of manuscripts submitted to The Journal of Bone and Joint Surgery: American Volume, studies of higher level of evidence were more likely to be accepted for publication.

In this study of abstracts accepted for presentation at the 2007 AAOS Annual Meeting, interrater reliability ranged from slight to moderate, and authors were found to grade the level of evidence of their own work more favorably than did others who graded the abstract.


In this study, orthopaedic residents were able to identify the level of evidence of blinded studies only 29.5% of the time before an educational intervention, and just 41.3% of the time posteducation.


In this study of articles published in The Journal of Bone and Joint Surgery: American volume between 1975 and 2005, the percentage of level I studies increased from 4% to 21% while the percentage of level IV studies decreased from 81% to 48%.


In this study of articles published in three general orthopaedics journals in 2002-2003, factors associated with an increased number of citations at 5 years were high level of evidence, large sample size, representation from multiple institutions, and self-reported disclosure of a conflict of interest.


In this study of articles published in J Bone Joint Surg Am between 2003 and 2007, studies funded by industry were significantly more likely to report level IV evidence as compared to studies funded by governments, foundations, or universities.


A table summarizing the characteristics of each level of evidence as it applies to therapeutic, prognostic, diagnostic, and economic / decision analysis studies is presented.


The National Guideline Clearinghouse is a repository of Clinical Practice Guidelines maintained by the Agency for Healthcare Research and Quality (AHRQ), a branch of the United States Department of Health and Human Services. Its stated mission is to “provide physicians and other health professionals, health care providers, health plans, integrated delivery systems, purchasers, and others an accessible mechanism for obtaining objective, detailed information on clinical practice guidelines and to further their dissemination, implementation, and use.”


A repository of clinical practice guidelines developed by the AAOS is presented.


The authors describe the shortcomings of clinical practice guidelines as they currently exist and argue that they should undergo major changes or be abandoned.


A summary of answers to frequently asked questions regarding clinical practice guidelines in orthopaedics.

This document, which makes recommendations regarding the management of pediatric diaphyseal femur fractures, represents a typical example of a Clinical Practice Guideline.