Understanding Levels of Evidence for Scientific Communication

Abstract: Foot & Ankle Specialist (FAS) is adding a “level of evidence” rating to all new submissions. This is important as evidence-based medicine in practice is increasingly important. The rationale for adding this rating to FAS publications is multifaceted. It will encourage researchers to design better projects and become adept at critically evaluating research. It will put research publications into appropriate context, and it will guide the profession in general to a higher level of scientific analysis. Four types of studies will be used at FAS: therapeutic, prognostic, diagnostic, and economic/decision making. Within each of these types of studies, levels of evidence between I and V can be assigned. A description of each type and explanation of how to assign evidence level is given. This may assist authors and readers as they work to design projects and critically evaluate literature.

Keywords: level of evidence; recommendation; guide; design; rationale

History
The era of evidence-based guidelines for scientific evaluation was first introduced by the members of the American College of Chest Physicians in 1986. They devised a “level of evidence” table consisting of 5 levels, related to “grades of recommendations,” grades A to C. These initial recommendations were updated and published in March 2009 (University of Oxford Centre for Evidence Based Medicine Web site [www.cebm.net]; see Table 1). Musculoskeletal science entered the era of evidence-based guidelines in 2003 with the Journal of Bone and Joint Surgery (American) adopting a level-of-evidence rating to all clinical articles using a modified version of the Oxford Centre’s scale.

Rationale
Our goal for implementing a level-of-evidence rating to clinical research publications in Foot & Ankle Specialist is to improve the scientific process by influencing the profession in 3 core areas.

1. Improve the focus of the authors as they develop their research protocol. Implementing the level-of-evidence rating will encourage and require physicians to become familiar with the nomenclature and methodology behind the rating. A better understanding will facilitate the process to clearly identify the primary research question, as well as classify and clarify the methodology used to answer their question.

2. Ensure that data are put into appropriate context so that unintended and unsupported conclusions are not made. Studies that are well designed, have a comparative group, and minimize the opportunity for bias through strict methodology will generally result in higher levels of evidence. This higher level of evidence correlates into stronger grades of recommendation, and clinicians can more easily evaluate how the results should influence their practice.

3. Guide the profession toward a higher level of scientific analysis. The current level of understanding of levels of evidence is evolving and can be significantly improved with training. Grading of levels of evidence between surgeons familiar or unfamiliar with

DOI: 10.1177/1938640010375184. From the Ripon Medical Center Advanced Orthopaedics, Ripon, Wisconsin (JGD), and Orthopedic Foot and Ankle Center, Westerville, Ohio (GCB). Address correspondence to J. George DeVries, DPM, AACFAS, Ripon Medical Center Advanced Orthopaedics, 933 Newbury St, PO Box 390, Ripon, WI, 54971; e-mail: ofacresearch@orthofootankle.com.

For reprints and permissions queries, please visit SAGE’s Web site at http://www.sagepub.com/journalsPermissions.nav.

Copyright © 2010 The Author(s)
### Table 1.
Levels of Evidence for Primary Research Question

<table>
<thead>
<tr>
<th>Types of Studies</th>
<th>Economic and Decision Analyses—Developing an Economic or Decision Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Therapeutic Studies—Investigating the Results of Treatment</strong></td>
<td><strong>Prognostic Studies—Investigating the Effect of a Patient Characteristic on the Outcome of Disease</strong></td>
</tr>
<tr>
<td><strong>Level I</strong></td>
<td><strong>Diagnostic Studies—Investigating a Diagnostic Test</strong></td>
</tr>
<tr>
<td>• High-quality randomized controlled trial with statistically significant difference or no statistically significant difference but narrow confidence intervals</td>
<td>• Testing of previously developed diagnostic criteria in series of consecutive patients (with universally applied reference gold standard)</td>
</tr>
<tr>
<td>• Systematic review(^b) of level I randomized controlled trials (and study results were homogeneous(^c))</td>
<td>• Systematic review(^b) of level I studies</td>
</tr>
<tr>
<td><strong>Level II</strong></td>
<td><strong>Sensible costs and alternatives; values obtained from many studies; multiway sensitivity analyses</strong></td>
</tr>
<tr>
<td>• Lesser quality randomized controlled trial (eg, &lt;80% follow-up, no blinding, or improper randomization)</td>
<td>• Development of diagnostic criteria on the basis of consecutive patients (with universally applied reference gold standard)</td>
</tr>
<tr>
<td>• Prospective(^e) comparative study(^g)</td>
<td>• Systematic review(^b) of level II studies</td>
</tr>
<tr>
<td>• Systematic review(^b) of level II studies or level I studies with inconsistent results</td>
<td><strong>Level V</strong></td>
</tr>
<tr>
<td><strong>Level III</strong></td>
<td><strong>Sensible costs and alternatives; values obtained from limited studies; multiway sensitivity analyses</strong></td>
</tr>
<tr>
<td>• Case control study(^h)</td>
<td><strong>Level IV</strong></td>
</tr>
<tr>
<td>• Retrospective(^e) comparative study(^g)</td>
<td><strong>Expert opinion</strong></td>
</tr>
<tr>
<td>• Systematic review(^b) of level III studies</td>
<td><strong>Case series</strong></td>
</tr>
<tr>
<td><strong>Level IV</strong></td>
<td><strong>Case control study</strong></td>
</tr>
<tr>
<td><strong>Level V</strong></td>
<td><strong>Case series</strong></td>
</tr>
</tbody>
</table>

This chart was adapted from material published by the Centre for Evidence-Based Medicine, Oxford, United Kingdom. For more information, please see www.cebm.net.

\(^a\)A complete assessment of the quality of individual studies requires critical appraisal of all aspects of the study design.

\(^b\)A combination of results from 2 or more prior studies.

\(^c\)Studies provided consistent results.

\(^d\)Study was started before the first patient enrolled.

\(^e\)Patients treated one way (eg, with cemented hip arthroplasty) compared with patients treated another way (eg, with cementless hip arthroplasty) at the same institution.

\(^f\)Study was started after the first patient enrolled.

\(^g\)Patients identified for the study on the basis of their outcome (eg, failed total hip arthroplasty), called “cases,” are compared with those who did not have the outcome (eg, had a successful total hip arthroplasty), called “controls.”

\(^h\)Patients treated one way with no comparison group of patients treated another way.
evidence levels has been shown to have high levels of agreement when a description of the grading criteria is available. Overall, a more defined research process will benefit both authors and the editorial board as this will encourage researchers to improve the level of their research to boost the likelihood of eventual publication.

Explanation

The determination of the level of evidence for a clinical research study using the modified scale developed at the Journal of Bone and Joint Surgery (American) is a stepwise process. First, the type of study must be defined, then the level of evidence, then the specific subcategory within each level. This modified scale differs from the one put forth by the Oxford Centre by omitting the “Differential Diagnosis/Symptom Prevalence” type of study, and there were some changes within the various subcategories. This modified scale seems well suited for orthopaedic research questions and is somewhat simplified in its application. The 4 types of studies include therapeutic, prognostic, diagnostic, and economic/decision analysis. Therapeutic studies involve research questions aimed at determining the efficacy of various treatments or modalities on certain conditions or outcomes. The most commonly performed type of study undertaken in orthopaedic literature. Inherently, it may also be the easiest to comprehend as the variable that is being examined is controlled by the surgeon or allocation criteria. Prognostic studies, on the other hand, involve studying the outcome an uncontrolled variable will yield. The major difference between the 2 types of studies is researcher control on the variable being studied. To illustrate the difference, a therapeutic study may evaluate the type of fixation used in an ankle fracture and see how this affects the outcome, whereas a prognostic study may evaluate the fracture pattern of the ankle fracture and see how it affects the outcome. In the therapeutic study, the variable being studied is researcher controlled, compared to the prognostic study, in which the variable being studied is out of researcher control. Diagnostic studies evaluate the effectiveness of a certain test or modality on detecting the condition of interest. Economic studies model cost comparisons related to the variable being studied. Diagnostic studies will involve an investigator-controlled variable to check effectiveness, and economic studies may be investigator controlled or not. An example of a diagnostic study is to check the reliability of magnetic resonance imaging (MRI) to detect osteomyelitis in the Charcot foot, whereas an example of an investigator-controlled economic test is to determine the cost of screening all Charcot patients with foot ulcerations with an MRI. An example of an economic study not controlled by the researcher is to determine the overall cost to health care in diabetics with neuropathy compared to those with intact sensation.

Once the study type has been determined, the level of evidence should be determined. In therapeutic studies, the pinnacle of studies is the high-quality randomized controlled trial (RCT), and thus it is given level I evidence and may be given the greatest weight when determining the impact the results should have on clinical practice. However, poorly designed or executed RCTs will be dropped to level II. The factors that can affect a drop in evidence level for an RCT include insufficient power to predictably replicate the results, poor randomization, poor follow-up (generally <80%), or insufficient binding of evaluators or patients. Thus, poor methodology can predicate a drop in level of evidence.

A level II therapeutic study may also be a prospective cohort study. Prospective refers to a study in which the primary research question was designed and the methodological processes used to answer that question were established before any patient data were collected. This can get somewhat ambiguous in cases in which a previously treated group is now prospectively reevaluated for a specific outcome, but in the strictest sense, this too would fall under a retrospective study. A cohort is defined as when one group has gotten one treatment, compared to a different group that has gotten a different treatment. An example of a level II prospective cohort study is to prospectively gather validated outcomes scores on patients with stage 2 posterior tibialis tendon dysfunction and compare the results of patients treated with custom orthoses to those of patients treated with a specified surgical reconstruction. A level III therapeutic study would entail a retrospective cohort study and also a case control study, either prospective or retrospective. A case control study involves comparing the outcomes of 2 groups of patients, the “cases” and the “controls.” The “cases” in this scenario involve patients who have one outcome (the outcome of interest and generally the less common one) and the “controls” who have a different outcome. Once these 2 groups are established, they are compared for differences in the frequency of a specified therapy. An example of a therapeutic level III case control study is comparing patients with ankle nonunions (cases) to patients with solid ankle unions (controls) after arthrodesis and looking at the frequency in which bone morphogenetic protein was used.

Prognostic studies follow slightly different criteria in determining levels of evidence. As the variable being studied is not researcher controlled, it cannot be randomly assigned, and as such, a randomized controlled trial is inherently not possible. The difference in level I and level II prognostic studies is whether the study was undertaken in a prospective or retrospective manner. In the above example of a prognostic study evaluating ankle fracture pattern and the effect on patient outcomes, a level I study is designed and defined prior to any patient data being collected, whereas a level II study is undertaken in a retrospective manner. Similarly, a level II prognostic study can be undertaken by evaluating the untreated controls of a previous randomized controlled trial. Even though the patients were enrolled in a prospective manner, if the prognostic study was not designed prior to patient enrollment, the study is technically retrospective and
earn level II evidence. There is no level III prognostic study.

Diagnostic studies are stratified into level of evidence based on 3 factors: the use and acceptance of the studied test, the availability and use of a “gold-standard” test, and the elimination of selection bias by universal application on a consecutive set of patients. A level I diagnostic study evaluates a test or modality that has been in use and already has set criteria for evaluating the condition of interest and in all patients compares those results to a gold standard. A level II study is similar, except the test that is being compared to the gold standard is either being developed or there are no set criteria to evaluate the condition of interest. This can be a new diagnostic modality or an old modality used in a novel scenario. So a level I study is used to assess the reliability and usefulness of a test using predetermined criteria for that test or modality, whereas a level II study is used to develop the criteria that will be used for that test concerning the specific condition. The above example of osteomyelitis in the Charcot foot can be either level I or II. A level I study is when the sensitivity and specificity of an MRI, which already has defined criteria to determine osteomyelitis, is compared to a bone biopsy to determine osteomyelitis in all patients. A level II study is the use of skin thermometry to test for osteomyelitis in the Charcot foot, which has no accepted criteria in its usefulness to determine the presence of osteomyelitis, and compares the results universally to a bone biopsy. A diagnostic study will be dropped to level III evidence if the study was performed on a nonconsecutive set of patients or there was inconsistent comparison of results to the gold standard.

Economic studies are modeling studies to formulate decision-making standards, and because of this, the level of evidence assigned to the study is largely based on the availability and reliability of previous cost studies, the availability and sensibility of alternatives in the variable, and the statistical workup using a sensitivity analysis. A level I economic study will have alternatives in the treatment of a condition that have been used and advocated for the specific scenario. The costs of each of the alternatives will have ample previous studies and data that can be compiled together to get a reliable true cost estimate, and a multiway sensitivity analysis will be done to check the effect of the change of multiple parameters. The sensitivity analysis is performed because every model will have a certain level of uncertainty, and this test can be used to see how changing the value of certain parameters will affect the models’ outcomes. A 1-way sensitivity analysis will check how changing one parameter at a time will influence the outcome. An example of this is varying the cost of a certain intervention and seeing how this will affect the overall cost of the model. A multiway sensitivity analysis involves changing 2 or more key parameters (eg, cost and effectiveness) and showing how the various combination can have an effect on the model outcome. The range of values that should be checked can be based on values published in the literature, confidence intervals of the values, or extreme limits in a “best-case/worst-case” scenario, all based on what is most pertinent for that particular variable. An example of a level I economic study is the cost analysis of unstable Charcot ankle deformity reconstruction using external fixation compared to primary transfibial amputation. In this example, the alternatives compared are reasonable as both have been advocated and evaluated. The costs of the hospital stays and operating room times can be gathered from previous published works and from hospital data available from any hospital financial office. The costs of the hardware or the prosthesis can be found from industry, hospital, and prosthetics data. The effectiveness and complication rate can be pulled from previous published literature. Variables can then be run through a multiway sensitivity analysis comparing different values for cost in combination of different rates of efficacy. A level II economic study is when the values of various parameters are from a limited source and may be influenced by bias. A level III study is when there are limited alternative treatment parameters and if there are poor cost estimates available.

Level IV studies are the most commonly reported in orthopaedic literature, can be undertaken in any type of study, and can be prospective or retrospective. Therapeutic and prognostic level IV studies involve case series with no contemporary controls. Diagnostic level IV studies are case control studies, and the diagnostic test findings are examined after the outcome is known or in the case of having no gold-standard comparator. Economic studies are classified as level IV if there is no sensitivity analysis performed. Level V studies are expert opinion and can be undertaken in any type of study. Also, all bench or laboratory testing or papers written solely on economic theory are given level V evidence. This is because Sackett’s proposed levels of evidence are based on clinical trials, and all nonclinical trials are reduced to the lowest level of evidence.

Systematic reviews and meta-analyses are given levels based on the body of work used in the analysis. A systematic review deals with a focused question by combining the results of 2 or more studies, and a meta-analysis is a specific type of systematic review using the results of 2 or more studies, combining the data, and performing statistical analysis on these conglomerate data. In both types of literature, care must be taken to ensure rigorous and disciplined inclusion and analytic methodology to avoid bias. Therefore, a systematic review of 2 level I therapeutic RCTs will be given a level I level of evidence (provided it has sound methodology), and a meta-analysis of 10 level IV prognostic studies is level IV evidence.

Conclusion

As the age of evidence-based medicine continues to evolve and mature, so must researchers and publications. Surgeons have historically lagged behind our medical counterparts in regards to production of high evidence-level literature. Indeed, randomization of surgical interventions and the strict methodological application of procedures may not be ethical and may even undermine the efficacy of a technique. However, although
not all clinical quandaries are amenable to study via an RCT, its been suggested that up to 40% of the work performed in a surgical practice could be designed and carried out safely in an RCT. And in areas where RCTs are impossible or impractical, well-designed cohort and case control designs can be equally powerful tools in determining the evidence base of practice. Therefore, as surgeon-investigators, we should strive to always improve the quality of the research produced. This can be done by adding a control group to what would otherwise have been a level IV study or simply by using strict methodology and enrollment to produce quality and reliable level IV studies. All surgeons in practice have a responsibility to be using the best practice efforts for patients. This will continue to be driven by the wealth of literature available, so surgeons must be trained to effectively read and evaluate the research being published. Even when the level of evidence is available, it is important to realize this is simply a methodological grade given to an article that corresponds into a rough estimate of strength of recommendation. No one rating or level is definitive, and we must use care and understanding to relate it to the care of patients.

References