# Levels of Evidence Are Not the Whole Story

*Susan E. Nelson MD, MPH[1]; Unni G. Narayanan MD, MSc, FRCSC[2]; Matthew R. Schmitz, MD[3]; Scott D. McKa,y MD[4]; The POSNA Evidence-Based Practice Committee**

*[1]Department of Orthopaedic Surgery, University of Rochester Medical Center, Golisano Children's Hospital, Rochester, NY; [2]Division of Orthopaedic Surgery & Child Health Evaluative Sciences, The Hospital for Sick Children, University of Toronto, Toronto, ON; [3]Department of Orthopaedics, San Antonio Military Medical Center, Ft. Sam, Houston, TX; [4]Department of Orthopedics, Baylor College of Medicine, Texas Children's Hospital, Houston, TX*

## Introduction

A 1992 article in *The Journal of the American Medical Association* (*JAMA*) described Evidence-Based Medicine (EBM) as a "paradigm shift"; one that would require new skills of physicians and new areas of teaching for residents to be able to find, appraise, and integrate findings into clinical practice to improve outcomes and patient care.[1] Almost 30 years of sifting through the ever growing body of evidence for any topic and being able to identify quality research that may help guide clinical decision making remains challenging.

To aid surgeons with deciphering the quality of the growing body of literature, the *Journal of Bone and Joint Surgery* (*JBJS*), among others, introduced Levels of Evidence (LOE) to the journal in 2003. The goal of this addition was to clarify research questions and help surgeons contextualize quality when deciding whether to apply results to their clinical dilemmas.[2] The caveat was made at the time that this was to be a guide and critical appraisal was needed for in-depth assessment. The most recent LOE guide for *JBJS* is adapted from the 2011 Oxford Centre for Evidence-Based Medicine Guide[3] which also highlights that levels of evidence may be changed based on several factors. Effect size, study quality, imprecision, or indirectness of the study question may all lead to downgrading the LOE.

The purpose of this article is to define LOE commonly used in the orthopaedic literature and to highlight that LOE alone is not always sufficient for assessing the quality of the evidence presented.

## What Are Levels of Evidence?

Levels of evidence are classification systems that use a hierarchal structure to indicate where the research in question may fall in regard to the strength of the recommendations. Finding the highest LOE available may help physicians make clinical decisions with confidence. LOE as a structure for examining research appeared in the 1970s and 80s. For example, a 1979 report by the Canadian Task Force examined evidence regarding period health examination[4] and provided a classification for examining the evidence with three levels.

Sackett then described levels of evidence to assess antithrombotic therapy in 1989[5] and similar to the Canadian Task Force, placed randomized control trials (RCT) at the highest level. However, this work acknowledged with the classification of a Level II that not all RCTs are equal when it comes to quality of evidence. Both of these early LOE classifications place research design with less risk of bias at the top. However, these early systems were more simplistic than contemporary LOE classification.

Contemporary classifications for LOE include subdivisions for different categories of research. The current LOE hierarchy that is used by *JBJS*[2] is also used by the *Journal of Pediatric Orthopaedics* (*JPO*) and categorizes research into diagnostic, therapeutic, prognostic,

**Levels of Evidence for Primary Research Question**

| | Types of Studies | | | |
|---|---|---|---|---|
| | Therapeutic Studies—Investigating the Results of Treatment | Prognostic Studies—Investigating the Outcome of Disease | Diagnostic Studies—Investigating a Diagnostic Test | Economic and Decision Analyses—Developing an Economic or Decision Model |
| Level I | 1. Randomized controlled trial<br>a. Significant difference<br>b. No significant difference but narrow confidence intervals<br>2. Systematic review[2] of Level-I randomized controlled trials (studies were homogeneous) | 1. Prospective study[1]<br>2. Systematic review[2] of Level-I studies | 1. Testing of previously developed diagnostic criteria in series of consecutive patients (with universally applied reference "gold" standard)<br>2. Systematic review[2] of Level-I studies | 1. Clinically sensible costs and alternatives; values obtained from many studies; multiway sensitivity analyses<br>2. Systematic review[2] of Level-I studies |
| Level II | 1. Prospective cohort study[3]<br>2. Poor-quality randomized controlled trial (e.g., <80% follow-up)<br>3. Systematic review[2]<br>a. Level-II studies<br>b. nonhomogeneous Level-I studies | 1. Retrospective study[4]<br>2. Study of untreated controls from a previous randomized controlled trial<br>3. Systematic review[2] of Level-II studies | 1. Development of diagnostic criteria on basis of consecutive patients (with universally applied reference "gold" standard)<br>2. Systematic review[2] of Level-II studies | 1. Clinically sensible costs and alternatives; values obtained from limited studies; multiway sensitivity analyses<br>2. Systematic review[2] of Level-II studies |
| Level III | 1. Case-control study[5]<br>2. Retrospective cohort study[4]<br>3. Systematic review[2] of Level-III studies | | 1. Study of nonconsecutive patients (no consistently applied reference "gold" standard)<br>2. Systematic review[2] of Level-III studies | 1. Limited alternatives and costs; poor estimates<br>2. Systematic review[2] of Level-III studies |
| Level IV | Case series (no, or historical, control group) | Case series | 1. Case-control study<br>2. Poor reference standard | No sensitivity analyses |
| Level V | Expert opinion | Expert opinion | Expert opinion | Expert opinion |

1. All patients were enrolled at the same point in their disease course (inception cohort) with ≥80% follow-up of enrolled patients.
2. A study of results from two or more previous studies.
3. Patients were compared with a control group of patients treated at the same time and institution.
4. The study was initiated after treatment was performed.
5. Patients with a particular outcome ("cases" with, for example, a failed total arthroplasty) were compared with those who did not have the outcome ("controls" with, for example, a total hip arthroplasty that did not fail).

***Table 1.*** *Contemporary Levels of Evidence used by JBJS and JPO. In: Wright, J.G., M.F. Swiontkowski, and J.D. Heckman, Introducing levels of evidence to the journal. J Bone Joint Surg Am, 2003. 85(1): p. 1-3.*

and economic studies, followed by assignment of the LOE (Table 1). The Oxford Evidenced-Based Working group highlights that LOE guidelines should be easy to use for the busy clinician, but no guideline can be used without judgement and careful consideration.[6] The Working Groups' most recent modifications of the LOE table is intended to increase the facility of interpreting LOE in clinical context.[3]

Keeping in mind that individual judgement is paramount in assessing the quality of evidence, we must look beyond just LOE but also consider the grade of the evidence. There are formalized systems for grading evidence that help clinicians determine how confident they can be in recommendations based on the best available evidence, and they are often used when making clinical practice guidelines. Whether evidence is graded strong or weak may depend not only on the LOE but also on effect size, as well as individual or population circumstances. Although the RCT is placed at the top of the hierarchy in LOE classification systems, not all RCTs provide strong evidence. Similarly, observational study de-

signs may be categorized as lower LOE but may provide higher grades of evidence. Sometimes observational studies may provide the strongest evidence available.

The following examples will highlight well-known research with various study designs and LOE as well as explore some of the critical appraisal needed to decide if the research presents best evidence and if recommendations can be applied in practice. To begin, we highlight two landmark orthopaedic papers, both multicenter randomized control trials.

## Starting High

*Should We Fix Clavicle Fractures?*

The Canadian Orthopaedic Trauma Society performed a multicenter randomized trial where they compared operative and nonoperative treatment of clavicular fractures. Although prospective in nature, if one were to simply read the abstract it would lead you to believe that all clavicular fractures should be treated operatively with improved functional outcomes and decreased rates of malunion and nonunion. But there are weaknesses to the study which deserve mentioning. Fifteen out of 65 patients originally randomized to the non-operative treatment group were lost to follow up (23%), with an additional patient deceased; they did not follow one-quarter of their participants in that arm out for the year. This has the potential to jeopardize the results with the differences in patient-reported outcomes compared to the operative treatment group that with 95% follow up at one year. In addition, eligibility criteria required patients to be between 16 and 60 years old for this trial, and it is not reported how many patients were in the adolescent/young adult age group between ages 16-19, thus the results may not be generalizable to adolescent patient populations.[7]

*To BrAIST or Not to BrAIST: Bracing for Adolescent Idiopathic Scoliosis (AIS)*

The BrAIST trial[8] was designed as a multicenter randomized control trial to provide the highest level of evidence. The goal was to improve on previous pitfalls in studies on bracing for AIS and determine whether bracing prevents progression of high-risk curves to surgical range. When examining the trial for the quality of the evidence, importantly there was a priori determination of effect size, a well-defined question and outcome, randomization, objective assessment of brace compliance, a control comparison group, and blinded curve measurement system. Despite the heterogeneity of the AIS population being studied, the experimental design aimed to minimize bias. However, the number of patients that accepted randomization was lower than the investigators had anticipated, citing a strong treatment preference as the reason for declining randomization. It is possible that patients and their families were influenced by their own research or unconscious influence of their surgeon's preferences. A preference arm was added to the study protocol to increase enrollment and results were analyzed in the primary analysis including both preference and randomized subjects together, with the randomized cohort analyzed separately. The trial thus presents Level I and Level II evidence. To reduce the influence of bias introduced due to the self-selected preference arm, a propensity adjusted analysis was used; bias may remain despite statistical adjustment. Even as originally designed, the trial was not blinded, a practical limitation of many surgical randomized trials. With the refinement of classifications of skeletal maturity that is ongoing the inclusion criteria for skeletal maturity based on the Risser 0-2 may not have captured the entire at-risk population. Despite these limitations and the deviation from the purely randomized design, the trail provides evidence to help guide decision making and counsel families regarding treatment options. This research frames the discussion with families in ways that were not previously possible.

## Looking Lower

While higher levels of evidence are often desired, there are many questions that will likely never be answered by prospective comparative research. The parody article, "Parachute use to prevent death and major trauma

related to gravitational challenge: systematic review of randomized controlled trials" makes this point well.[9]

Additionally, long-term follow-up and natural history studies are typically presented as single-center retrospective studies. This is still valuable information, especially if there are multiple centers reporting separate studies.

*For example:*
In 2007, a 45-year follow up of DDH patients treated at ages 18 months to 5 years with innominate osteotomy and open reduction showed a 54% survival rate. All of these patients were treated with traction 2-3 weeks prior to surgery, a practice which is not performed frequently in these times. Also, at least 20% of hips had subsequent procedures for subluxation, dislocation, or dysplasia.[10]

In 2020, these same patients were compared with a cohort of DDH patients from another center treated with closed reduction at 18 months to 5 years, also with 40+ years of follow-up. This brought the analysis of the paper up to a level 3 study (retrospective study with comparison group). Comparisons were difficult as nearly 20% of these hips also underwent further surgery. In both studies, outcome and survival decreased after age 40, but the closed reduction group had a higher drop off. Nevertheless, up until age 40, the closed reduction and open reduction cohorts had similar survival rates of around 50%.[11]

Despite the confounding variable of further surgery, selection bias, and other imperfections, these two retrospective studies provide the best information we have on the long-term outcome of DDH treatment in children ages 18 months to 5 years.

Lower level evidence studies affect our everyday orthopaedic knowledge. A 2013 study of the most frequently cited pediatric orthopaedic papers found that 72% were level 4 evidence. Consider these examples of low-level evidence papers that provide us with valuable findings:

- Important descriptions of uncommon conditions—TRASH elbow fractures[12]
- Innovative techniques—Ganz description of surgical dislocation of the hip[13]
- Warnings of significant complications - Wong & Williams description of thermal capsulorrhaphy[14]

## Conclusion

Levels of Evidence were introduced as a way to stratify research projects based on chance of bias. Higher level (level 1) is generally desired over lower level evidence (level 4). However, level 1 evidence is not flawless and often not required to answer clinical questions. Not all clinical quandaries will be able to be examined through the lens of a meta-analysis of level 1 evidence or through level 1 evidence at all. The reader is encouraged to implement fundamentals of literature review and critical thinking to determine whether the methods of the study answer the research question and if the outcomes apply to their specific clinical scenario.

## *POSNA Evidence-Based Practice Committee

*Scott D. McKay, MD (Chair); Tracey P. Bastrom, MA; Arvindera Ghag, MD, FRCSC, BC; Joseph Janicki, MD; Judson W. Karlen, MD; Indranil Kushare, MD; Ronald Lewis, MD; R. Justin Mistovich, MD; Unni G. Narayanan, MD, MSc, FRCSC; Susan Nelson, MD, MPH; Neeraj Patel, MD, MPH; William A. Phillips, MD; Jeffrey R. Sawyer, MD; Matthew R. Schmitz, MD; Kelly Vanderhave, MD; Maegen J. Wallace, MD, ATC*

## References

1. Guyatt, G., Cairns, J., Churchill, D. et al. *Evidence-based medicine. A new approach to teaching the practice of medicine.* JAMA, 1992. **268**(17): p. 2420-5.

2. Wright, J.G., M.F. Swiontkowski, and J.D. Heckman, *Introducing levels of evidence to the journal.* J Bone Joint Surg Am, 2003. **85**(1): p. 1-3.

3. Jeremy Howick, I.C., Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, and Hazel Thornton. *The Oxford 2011 Levels of Evidence*. 2011; Available from:

**https://www.cebm.ox.ac.uk/resources/levels-of-evidence/explanation-of-the-2011-ocebm-levels-of-evidence/**

4. Canadian Task Force on the Periodic Health Examination. *The periodic health examination.* Can Med Assoc J, 1979. **121**(9): p. 1193-254.

5. Sackett, D.L., *Rules of evidence and clinical recommendations on the use of antithrombotic agents.* Chest, 1989. **95**(2 Suppl): p. 2s-4s.

6. OCEBM Levels of Evidence Working Group. *Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document).* 2011; Available from: https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence.

7. Canadian Orthopaedic Trauma Society. *Nonoperative treatment compared with plate fixation of displaced midshaft clavicular fractures. A multicenter, randomized clinical trial.* J Bone Joint Surg Am, 2007. **89**(1): p. 1-10.

8. Weinstein, S.L., et al., *Effects of bracing in adolescents with idiopathic scoliosis.* N Engl J Med, 2013. **369**(16): p. 1512-21.

9. Smith, G., & Pell, J. *Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials.* BMJ, 2003.**327**(7429), 1459–1461.

10. Thomas, S., Wedge, J., Salter, R. *Outcome at Forty-five Years After Open Reduction and Innominate Osteotomy for Late-Presenting Developmental Dislocation of the Hip*. J Bone Joint Surg Am. 2007;**89**(11):2341-2350.

11. Scott, E., Dolan, L., & Weinstein, S. Closed Vs. *Open Reduction/Salter Innominate Osteotomy for Developmental Hip Dislocation After Age 18 Months: Comparative Survival at 45-Year Follow-up.*, J Bone Joint Surg Am, 2020. **102**(15), 1351–1357.

12. Waters, P., Beaty, J., & Kasser, J. *Elbow "TRASH" (The Radiographic Appearance Seemed Harmless) Lesions.* J Pediatr Orthop, 2010. 30 Suppl 2, S77–S81.

13. Ganz, R., Gill, T., Gautier, E., Ganz, K., Krügel, N., & Berlemann, U. *Surgical dislocation of the adult hip*. J Bone Joint Surg Br, 2001. **83**(8), 1119–1124.

14. Kong, K., & Williams, G. *Complications of Thermal Capsulorrhaphy of the Shoulder*. J Bone Joint Surg Am, 2001. **83**(2_suppl_2 Suppl 2), S151–155.