Level of Evidence Gap in Orthopedic Research

KEITH D. BALDWIN, MD, MSPT, MPH; JOSEPH BERNSTEIN, MD; JAIMO AHN, MD, PHD; SCOTT D. MCKAY, MD; WUDBHAV N. SANKAR, MD

abstract

Full article available online at Healio.com/Orthopedics. Search: 20120822-31

Level of evidence is the most widely used metric for the quality of a publication, but instances exist in which a Level I study is neither feasible nor desirable. The goal of this study was to evaluate the level of evidence gap in current orthopedic research, which the authors defined as the disparity between the level of evidence that would be required to optimally answer the primary research question and the level of evidence that was actually used. Five orthopedic surgeons (K.D.B., J.B., J.A., S.D.M., W.N.S.) evaluated blinded articles from the first 6 months of 2010 in the *Journal of Bone and Joint Surgery (American Volume)* (JBJS-Am), classifying the study type and design and extracting a primary research question from each article. Each evaluator then defined the study type and method, along with the level of evidence that would ideally be used to address the primary research question. The level of evidence gap was then calculated by subtracting the actual level of evidence of the manuscript from the level of evidence of the idealized study. Of the 64 JBJS-Am manuscripts eligible for analysis, the average level of evidence was between Level II and III (mean, 2.73). The average level of evidence gap was 1.06 compared with the JBJS-Am–designated level of evidence and 1.28 compared with the evaluators' assessment. Because not all questions require Level I studies, level of evidence alone may not be the best metric for the quality of orthopedic surgery literature. Instead, the authors' concept of a level of evidence gap may be a better tool for assessing the state of orthopedic research publications.

Drs Baldwin and Sankar are from the Department of Orthopaedic Surgery, Children's Hospital of Philadelphia, and Drs Bernstein and Ahn are from the Department of Orthopaedic Surgery, Hospital of the University of Pennsylvania, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania; and Dr McKay is from the Department of Orthopaedic Surgery, Texas Children's Hospital, Baylor College of Medicine, Houston, Texas.

Drs Baldwin, Bernstein, Ahn, McKay, and Sankar have no relevant financial relationships to disclose.

Correspondence should be addressed to: Wudbhav N. Sankar, MD, Children's Hospital of Philadelphia, 34th and Civic Center Blvd, 2nd Floor Wood Bldg, Philadelphia, PA 19104 (sankarw@email.chop.edu).

doi: 10.3928/01477447-20120822-31

Although the practice of medicine has been based on empirical knowledge since at least the time of Galen, the term *evidence-based medicine* is more recent, appearing for the first time in a 1992 article.[1] The modern concept of evidence-based medicine centers on the notion that physicians' decisions should rely not merely on evidence in general, but on evidence of the highest quality. To that end, hierarchies based on levels of evidence were devised by the US Preventative Services Task Force, the National Health Service in Great Britain, and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group internationally.[2,3] The evidence hierarchy ranges from meta-analyses of high-quality prospective, randomized, controlled trials (Level I), through case-control studies (Level III), case series and case reports (Level IV), to the lowest form, expert opinion (Level V).[1]

This hierarchy follows the approach of Wright,[2] who noted that "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." In turn, orthopedic surgery journals have been evaluated according to the relative prevalence of studies with higher levels of evidence.[4-6] Obremskey et al[7] reviewed 382 clinical articles published in a 6-month period in 9 orthopedic journals and rated them according to the *Journal of Bone and Joint Surgery (American Volume)* (JBJS-Am) level of evidence scheme, ranging from I to IV. They found that 11.3% of the sample were Level I studies, whereas 58.1% were Level IV studies. They concluded that their work "exposes a well-known weakness in the orthopaedic literature, which is its tendency to contain retrospective studies with a lower level of evidence."[7]

Implicitly, Obremskey et al[7] claim that all studies should be Level I; why else would a preponderance of retrospective studies be a "weakness," and not simply a feature? However, instances exist in which a Level I study is neither feasible nor desirable. Smith and Pell[8] make that clear in their parody, "Parachute Use to Prevent Death and Major Trauma Related to Gravitational Challenge: Systematic Review of Randomised Controlled Trials."[8] Studies reporting lower levels of evidence can nonetheless aptly alter practice. For example, the Level IV survey study of Wong and Williams,[9] reporting axillary nerve injuries after thermal capsulorrhaphy of the shoulder, might reasonably dissuade surgeons from using this technique.

Accordingly, the state of orthopedic surgery literature may not be accurately represented by the average level of evidence of its articles, but rather the deficit, if any, between the methods used in published studies and the methods that should have been used. Therefore, the authors introduce the concept of the level of evidence gap: the difference between the level of evidence that should have been used to answer the primary research question of the study and the level of evidence that was actually used in the study.

The research question addressed in the current study was, "What is the current level of evidence gap in orthopedic surgery clinical research papers?" This metric can serve as a better assessment of the state of the orthopedic literature, as well as a benchmark for future studies assessing trends within the field.

Materials and Methods

Five orthopedic surgeons (K.D.B., J.B., J.A., S.D.M., W.N.S.) reviewed 67 consecutive articles from the first 6 months of 2010 in JBJS-Am and identified those clinical research manuscripts with a discrete and identifiable research question. Basic science articles, systematic reviews, economic analyses, technique articles, and case series were excluded.

A redacted copy of each article was sent to the evaluators, blinding them to the designated study type (ie, therapeutic, prognostic, or diagnostic) and level of evidence (I-IV) as assigned by the journal. The evaluators were asked to classify each article by type and study design (eg, case series; case-control studies; and randomized, controlled trials). They were then asked to assign a level of evidence to the study and extract the primary research question of the manuscript.

Because this extracted primary research question was not amenable to aggregation due to its qualitative nature, the evaluators then met after these data were collected to agree on a consensus primary research question. Unless the wording proposed was exactly the same, 2 evaluators' versions of the question, even if ostensibly similar, could not be combined; hence, the evaluators met to arrive at a consensus. The evaluators were then asked to define the study type and method, along with the level of evidence that would ideally be used to best address the consensus primary research question.

The evaluators' classification regarding article type, study design, and level of evidence were compared with those published in JBJS-Am. Intraclass correlation coefficients and 95% confidence intervals (CIs) for single measures were calculated to establish agreement between evaluators on the actual level of evidence, the type of study, the actual study design used, the ideal study design that should have been used, and the ideal level of evidence. These were then all compared to the actual JBJS-Am rating of level of evidence as a quality control and validation method for the evaluators' ratings.

A level of evidence gap was then defined for each manuscript by subtracting the level of evidence of the manuscript from the evaluators' consensus ideal level of evidence. This gap was calculated twice: first using the level of evidence assigned to the manuscript by the evaluators and then using the level of evidence assigned by JBJS-Am.

A total of 10 items were collected for each article under consideration:

1. Study type published by JBJS-Am.
2. Study design published by JBJS-Am.
3. Level of evidence published by JBJS-Am.
4. Study type defined by evaluators.
5. Study design defined by evaluators.
6. Level of evidence defined by evaluators.
7. Research question defined by evaluators.
8. Level of evidence of ideal study type for the research question as defined by evaluators.
9. Level of evidence gap: ideal level of evidence minus published level of evidence.
10. Level of evidence gap: ideal level of evidence minus evaluator-defined level of evidence.

All statistical analyses were performed with SPSS version 16.0 software (SPSS, Inc, Chicago, Illinois).

Results

Sixty-four JBJS-Am manuscripts were eligible for analysis. The distribution of study types among the 64 manuscripts is shown in Table 1. The interobserver agreement on type of study was .693 (95% CI, .599-.781). The agreement rate between the consensus-assigned study type and that assigned by JBJS-Am was .956 (95% CI, .928-.974).

Seventeen studies were classified as case series, 10 as case-control studies, 9 as prospective cohort studies, 18 as retrospective cohort studies, and 9 as randomized, controlled trials. One study was diagnostic and did not fit neatly into any of the groups. At the time of the study, it was not JBJS-Am policy to explicitly state the study design other than its type and level of evidence. The evaluators' agreement on the actual study method as published was .814 (95% CI, .746-.872). Their consensus on the level of evidence of the published papers compared with the grades assigned by JBJS-Am is shown in Table 2. Interobserver agreement of the assessment of the actual level of evidence of a study was .662 (95% CI, .562-.757). When comparing the evaluators' assigned level of evidence to the JBJS-Am level of evidence, the overall agreement was .857 (95% CI, .771-.912). For the primary research question of the 64 studies analyzed, the evaluators' consensus was that 46 studies would ideally require Level I evidence, 8 could be addressed with Level II, and 2 and 8 needed Levels III and IV, respectively.

Although the evaluators' overall agreement was .857 (95% CI, .771-.912) when comparing the group's assigned level of evidence to the JBJS-Am level of evidence, when more than 2 evaluators could not agree on the study type (ie, therapeutic, diagnostic, or prognostic), the agreement on level of evidence was .245 (95% CI, .027-.583), compared with .730 (95% CI, .630-.818) when fewer than 2 disagreements existed. The evaluators' agreement on the ideal level of evidence for a given study was .413 (95% CI, .297-.540).

The average level of evidence gap (the difference between the ideal level of evidence and the actual level of evidence) was 1.06 compared with the JBJS-Am–designated level of evidence and 1.28 compared with the evaluators' assessment.

Discussion

Since its adoption by orthopedic surgery journals, the level of evidence designation has become a widely used metric to gauge the quality of orthopedic research. In turn, orthopedic surgery researchers have been criticized for publishing too few Level I studies.[7] However, because not all questions are amenable to Level I studies, the relative preponderance of lower-level studies may not correctly describe the frequency with which orthopedic researchers use inadequate means, and in turn may not accurately represent the quality of the orthopedic surgery literature. Therefore, the current authors introduce the concept of level of evidence gap to quantify the true deficit.

Table 1
Assigned Study Types

Study Type	n	
	Evaluators	JBJS-Am
Therapeutic	38	37
Prognostic	25	23
Diagnostic	1	0
Not classified	0	4

Abbreviation: JBJS-Am, Journal of Bone and Joint Surgery (American Volume).

Table 2
Assigned Levels of Evidence

Level of Evidence	n	
	Evaluators	JBJS-Am
I	4	11
II	25	25
III	19	11
IV	16	13
Not assigned	0	4

Abbreviation: JBJS-Am, Journal of Bone and Joint Surgery (American Volume).

In this study, 64 JBJS-Am articles were reviewed. The overall level of evidence was between Level II and III (mean, 2.73), implying a deficit of close to 2 relative to the presumed ideal of Level I. However, the data suggest that the gap was closer to 1. The evaluators' agreement suffered substantially when 2 or more evaluators disagreed on the type of study. This was largely due to a quirk in the grading of levels of evidence based on study type: a data set representing a case series, collected to answer the question, "How do patients fare after a given treatment?" could be labeled as Level II if the study were considered prognostic or Level IV if it were considered therapeutic. Wolf et al[10] also demonstrated differences in the accuracy of level of evidence grading by orthopedic residents based on study type.

The current authors' level of evidence gap is an imperfect measurement. Although evaluators were in substantial agreement regarding the actual level of evidence and other study design questions, their agreement on the ideal level of evidence was moderate.[11] This suggests that the evaluators' assessments varied by their interpretation of the research question, and reasonable and experienced evaluators may disagree about this.

The strengths of this study include the blinding of evaluators to levels of evidence, the use of consensus questions, and the aggregation of multiple raters. The evaluators' near-perfect agreement with the JBJS-Am levels of evidence and study types provides internal consistency to the authors' methods and lends credibility of rating to the rest of the metrics measured. Weaknesses of the study include the fact that manuscripts from the recently published literature were assessed; therefore, the authors could not assure complete blinding because the evaluators may have recalled the assigned level of evidence prior to initiation of this study. However, this study was performed 1 year after the articles were published, so clear memory of the assigned level of evidence is unlikely.

The more significant limitation of the level of evidence gap is that it relies on an idiosyncratic definition of the ideal level of evidence for a given question. An article reporting axillary nerve injuries after thermal capsulorrhaphy of the shoulder is a Level IV study. That study has a level of evidence gap of 0 if the evaluators assert that the question is, "Can thermal capsulorrhaphy of the shoulder injure the axillary nerve?" but a gap of 3 if the question is defined as, "What is the true rate of neurological complications after thermal capsulorrhaphy?"

Another important limitation of the level of evidence gap concept is that, in the name of keeping this gap small, authors may be encouraged to ask poorer questions. If a journal were inclined to keep its gap to a minimum, it may favor Level III for questions that could be answered with Level III studies instead of selecting a good Level II study that addresses a Level I question. However, any consideration of levels of evidence may skew the tastes of editors.

Overall, an assessment of the level of evidence gap may be an important metric for assessing the state of the literature in a given specialty. In addition, the level of evidence gap concept may be useful as an educational tool for orthopedic trainees—who have been shown by Wolf et al[10] to have limited accuracy in determining the level of evidence of a given study—because it calls specific attention to the primary research question and to the methodology used to answer this question. Moreover, attention to the level of evidence gap may improve the quality of published papers by reminding authors to temper their conclusions in light of the quality of the evidence they used. Answering the question at hand should be the goal of all researchers, and attention to the level of evidence gap, not the absolute level of evidence, will help assure that.

References

1. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA. 1992; 268(17):2420-2425.
2. Wright JG. A practical guide to assigning levels of evidence. *J Bone Joint Surg Am.* 2007; 89(5):1128-1130.
3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *J Clin Epidemiol.* 2011; 64(4):383-394.
4. Hanzlik S, Mahabir RC, Baynosa RC, Khiabani KT. Levels of evidence in research published in *The Journal of Bone and Joint Surgery (American Volume)* over the last thirty years. *J Bone Joint Surg Am.* 2009; 91(2):425-428.
5. Cashin MS, Kelley SP, Douziech JR, Varghese RA, Hamilton QP, Mulpuri K. The levels of evidence in pediatric orthopaedic journals: where are we now? *J Pediatr Orthop.* 2011; 31(6):721-725.
6. Barske HL, Baumhauer J. Quality of research and level of evidence in foot and ankle publications. *Foot Ankle Int.* 2012; 33(1):1-6.
7. Obremskey WT, Pappas N, Attallah-Wasif E, Tornetta P III, Bhandari M. Level of evidence in orthopaedic journals. *J Bone Joint Surg Am.* 2005; 87(12):2632-2638.
8. Smith GCS, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ.* 2003; 327(7429):1459-1461.
9. Wong KL, Williams GR. Complications of thermal capsulorrhaphy of the shoulder. *J Bone Joint Surg Am.* 2001; 83(suppl 2 pt 2):151-155.
10. Wolf JM, Athwal GS, Hoang BH, Mehta S, Williams AE, Owens BD. Knowledge of levels of evidence criteria in orthopedic residents. *Orthopedics.* 2009; 32(7):494-497.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33(1):159-174.