Evidence-based Medicine: Critical Appraisal of the Literature (Critical Appraisal Tools)

Marc A. Raslich, MD,*
Gary M. Onady, MD, PhD*

Author Disclosure
Drs Raslich and Onady did not disclose any financial relationships relevant to this article.

To view Educational Tutorial Reference Links for evidence-based medicine, visit pedsinreview.org and click on the title of this article.

Introduction

Teaching the incorporation of evidence-based medicine (EBM) into clinical decision-making is the major goal of this series. The first article defined and introduced the process of EBM. (1) The second two articles in the series introduced the first two steps and tools used in the EBM process, which are to develop an answerable question (2) and then to conduct an evidence-based search. (3) Before integrating this knowledge into clinical decision-making, the found information must be tested for validity. This third step in the EBM process is referred to commonly as critical appraisal. Critical appraisal forms the bridge between finding relevant data and applying the information to clinical practice.

How do we ensure that we have found the best answer to our clinical question? When using the medical literature to answer our questions, sometimes we can rely on others to do the background work for us. For example, a previous article in this series (3) refers to secondary sources, including synopses such as *AAP Grand Rounds* and syntheses such as the *Cochrane Database*. Secondary sources refer to publications that review research articles independently and appraise them for evidence. These resources can be useful but cannot be the sole source of information. Secondary resources do not always address specific clinical questions.

Therefore, the ability to evaluate the medical literature personally and judge its value independent of assessments made by others is essential. Critical appraisal provides the skill for evaluating the literature and reaffirming the quality of the originally structured answerable question. (4) This process enables physicians to recognize potential problems with the evidence, allowing use of the results in making an informed decision or deciding that the data are of insufficient quality to draw any useful conclusions. (5)

Thousands of studies are published each year. Given the sheer quantity of data, a credible answer to a specific clinical question is likely to exist, but a large quantity of information either is not credible or not applicable toward a specific patient's care. In fact, as little as 10% of original research and review articles provide good evidence that is ready for application in clinical practice. (6) Therefore, having the skills and tools to interpret studies and select *useful* information from clinical research becomes important. Little benefit comes from collecting evidence if it cannot be interpreted properly. Fortunately, many textbooks and articles on EBM address this potential skill gap by describing techniques for assessing the methodologic quality of research evidence. (7)(8)

Critical assessment of an article highlighted earlier in the series is used to illustrate whether a reported clinical prediction algorithm can be applied to the specific patient who has the following clinical presentation:

A 12-year-old boy is evaluated for a limp. On physical examination, he demonstrates minimal weight bearing, localized right lateral thigh tenderness, and a fever of 102°F (38.5°C). Both the erythrocyte sedimentation rate and white blood cell count are elevated significantly.

The article by Kocher and associates, (9) discovered by searching reports of clinical manifestations similar to those of this patient, needs to be checked for relevance and validity before applying this information to this specific patient for clinical decision-making. The investigation, which was found using the search terms "weight bearing" and "fever," involved study of the records of 282 children who had an acutely irritable hip, looking at a variety of parameters. An algorithm predicted the probability that a child would have septic arthritis. When the algorithm is applied to the patient described previously, the probability is 99.6% that the child has septic arthritis of the hip. The

*Medicine-Pediatrics Program, Wright State University, Dayton, Ohio.

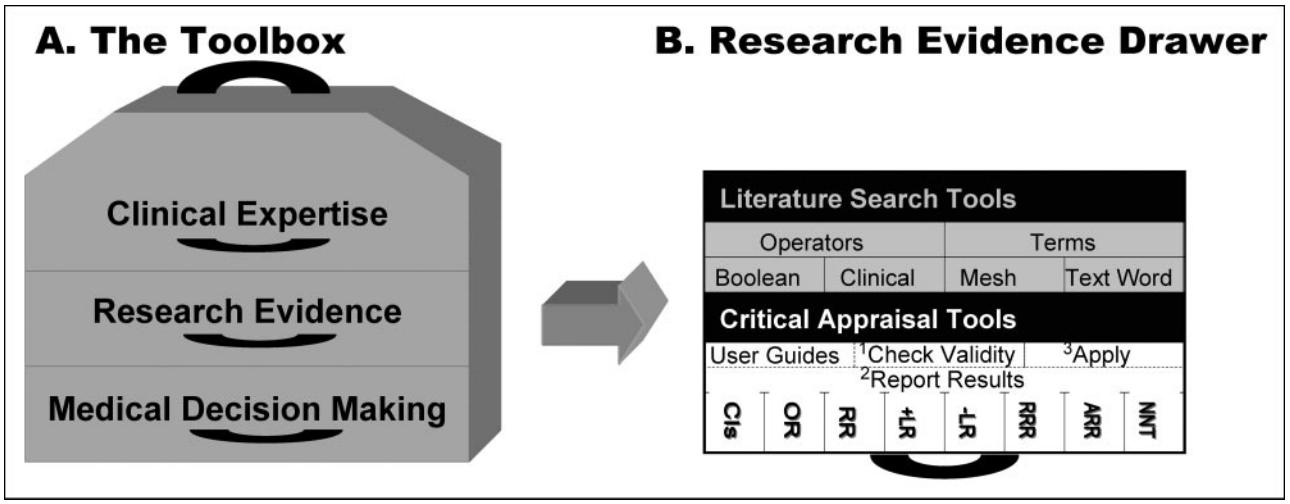

Figure. The EBM Toolbox. Mesh=medical subject heading, CIs=confidence intervals, OR=odds ratio, RR=relative risk, LR=likelihood ratio, RRR=relative risk reduction, ARR=absolute risk reduction, NNT=numbers needed to treat

question is whether this study and algorithm can be trusted as a tool for treating this specific patient.

Digging Deeper

The tools found in the EBM Toolbox were introduced in the second article of this series. The focus now is directed toward the front part of the second drawer, as indicated in the Figure. Three tasks are listed in that part of the toolbox: 1) evaluate basic study design to check **validity**, 2) analyze and summarize **results** for use in patient care decisions, and 3) **apply** these results to the specific clinical scenario. Table 1 lists the specific elements of an article that should be examined when determining validity, evaluating results, and deciding how applicable the data are to a specific clinical dilemma.

No one master checklist can be used for all types of articles. The items that help determine validity for an article focusing on therapy are different from those that govern study design for an article on diagnosis or prognosis. Fortunately, the evidence that pediatricians require to answer most clinical questions is limited to a few types. For interested readers, a summary of more detailed critical evaluation guides has been published to help physicians perform their own critical reviews. (7)

Checklists should make the critical appraisal process structured, explicit, and straightforward. It is important to remember that this is not a simple "yes" or "no" process. Errors or shortcomings in study design almost certainly are uncovered with close scrutiny. The reviewer must determine how accurately the results represent the truth and how much impact the degree of validity has on clinical decisions. For example, if an article fails to meet any of the standards outlined in the checklists, it is a good idea to search for other data to answer the relevant question. On the other hand, if the study meets all of the basic requirements in the checklists, the results should aid in making an accurate clinical decision. Explaining each of the terms in the checklists is beyond the scope of this article, but additional tools may be necessary to help refine the statistical results extracted from Table 1. Such tools are included in the online edition of this article as educational links for readers who seek further detail.

Checking Validity: The Problem of Bias

The appraisal process begins with evaluating the study design to ensure adequate validity. Validity hierarchy has been described (10) and is outlined in Table 2, which is based on the concept that some study types are better suited than others to measure the effects being studied. However, deficiencies in validity may be influenced by subtleties related to study design or methodologic flaws that may influence results further.

Threats to validity are termed bias. These biases or systematic errors can lead to false conclusions – an extremely undesirable result in nearly any field, but particularly when the conclusions influence medical decisions. A multitude of biases have been described that can be reduced to three broad categories (Table 3).

The items provided in Table 1 focus on the authors' efforts to minimize biased results through appropriate study design. Each category has different criteria. For example, with therapeutic studies, randomization is an important method of avoiding selection bias. For the Kocher and associates article, the relevant items in Table

Table 1. **Medical Literature Critical Evaluation Guides**

Focus of Study	Clinical Manifestations	Differential Diagnosis	Diagnostic Testing	Therapy/Prevention	Prognosis	Harm/Etiology	Reviews
Study Validity	☐ Diagnosis made using explicit and credible criteria	☐ Well-defined clinical problem	☐ Independent comparison to gold standard	☐ Assignment of patients randomized	☐ Defined, representative sample of patients	☐ Clearly defined groups of patients similar in all ways except exposure	☐ Clearly stated, focused question
	☐ Criteria independent of clinical manifestation under study	☐ Final diagnostic criteria explicit and credible	☐ Diagnostic standard applied to all patients	☐ Follow-up complete and sufficiently long	☐ Patients at common point in disease	☐ Assessment of outcomes either objective or blind to exposure	☐ Thorough search
	☐ Patients represent full spectrum of clinical presentation	☐ Diagnostic evaluation complete and applied consistently	☐ Testing methods used are described to permit replication	☐ Patients analyzed in groups to which they were randomized (intention to treat)	☐ Follow-up complete and sufficiently long	☐ Follow-up complete and sufficiently long	☐ Explicit inclusion and exclusion criteria
	☐ Manifestations methodically sought and fully classified	☐ Patients represent full spectrum of clinical presentation		☐ Patients and clinicians blind to treatment	☐ Outcome criteria applied blindly	☐ Results satisfy basic criteria for causation	☐ Reviewed studies of good quality
		☐ Follow-up complete and sufficiently long		☐ Groups similar at start of trial	☐ Adjustment for important prognostic factors ☐ Validation in separate patients		☐ Consistent results from study to study (homogeneity)
Statistical Results	☐ Frequency of each clinical finding listed	☐ Probabilities for each diagnosis	☐ Likelihood Ratios (LR) Large effect if +LR > 10; −LR < 0.2 Moderate effect if +LR = 10 to 5; −LR = 0.5 to 0.2 Minimal effect if +LR = 0 to 1; −LR = 1 to 0.5	☐ Relative risk reduction	☐ Survival curves (likelihood of outcome over time, eg, Kaplan-Meier)	☐ Relative risk for randomized or cohort study (prospective)	☐ Detailed data summarized
	☐ Confidence intervals reveal precise frequency estimates	☐ Confidence intervals reveal precise estimate	☐ Confidence intervals acceptable	☐ Absolute risk reduction and number needed to treat	☐ Confidence intervals reveal precise prognostic estimate	☐ Odds ratio for case-control (retrospective)	☐ Relative risk or odds ratio reported
	☐ Findings correlated to course of disease			☐ Confidence intervals acceptable		☐ Confidence intervals acceptable	☐ Confidence intervals acceptable
Applicability	☐ Study population similar to our own	☐ Study population similar to our own	☐ Study population similar to our own	☐ No important disease or patient differences that will affect treatment response	☐ Study population similar to our own	☐ Study population similar to our own	☐ Study population similar to our own
	☐ No significant comorbid conditions or event rates that will change frequency	☐ Disease possibilities have not changed since study completion	☐ Test is available, affordable, and accurate in our setting	☐ Treatment is feasible	☐ Results will lead directly to therapy selection	☐ Potential benefits outweigh potential harms	☐ Treatment is feasible
	☐ Disease manifestations have not changed since study		☐ Disease possibilities have not changed since study completion	☐ Potential benefits outweigh potential harms	☐ Results useful for counseling patient or family	☐ Comparison to available alternative therapy	☐ Potential benefits outweigh potential harms
			☐ Test/treatment thresholds are affected	☐ Consider family's values and expectations			☐ Consider family's values and expectations

Table 2. Hierarchy of Evidence for Study Design/ Study Type

Systematic reviews and meta-analyses	Higher Validity
Randomized, controlled trials with definitive results	
Randomized, controlled trials with nondefinitive results	
Cohort studies	
Case-control studies	
Cross-sectional studies	
Case reviews and reports	Lower Validity

1 are found in the column headed "Clinical Manifestations." The methods section of that article documents that the four criteria most important in reducing bias and ensuring validity for this type of study are satisfied:

- The study made its final diagnoses using explicit and credible criteria
- The diagnostic criteria were independent of the clinical manifestation under study
- The patients under study represented the full spectrum of clinical presentation
- The manifestations were sought methodically and classified fully

Because all criteria were met with regard to study validity, bias is minimized, and the process now can proceed to assessment of the results.

Reporting Results

The next step involves assessing and summarizing results for use in patient care decisions. A few areas of importance are common to all research results. Because it is not possible to study all patients who have the disease or symptom of interest, studies generally are conducted using a smaller group or sample of patients having that particular disorder. Even if a study is appropriately designed and bias is minimized, the results from one sample of patients may misrepresent the results that would be found in the population due solely to chance.

A coin flip can be used to illustrate the effect of chance on study results. Theoretically, if a coin is flipped 10 times, five heads and five tails should be observed. However, it would not be surprising if the results were seven heads and three tails. This divergence from the "true" probability of 50% heads and 50% tails is due to chance or random variation. Chance can occur at any step in a clinical study and never can be eliminated totally when assessing results.

Statistics can be used to estimate the extent to which chance accounts for the results of a particular study. Results can be categorized as significant with quantification of the probability that chance alone accounts for the findings. If results are "statistically significant," the likelihood that the results are due to chance alone is at an acceptably low level. Two values commonly used for quantification of chance are the *P* value and confidence interval (CI). By consensus, a *P* value of 0.05 or less is considered statistically significant, meaning there is a 5% or less probability that the reported results are due to chance alone.

CIs are more informative than a single result because they provide a range of plausible values for the answer of interest. This interval represents the range of values that is believed to encompass the "true" value with a defined probability. The defined value typically is at the 95% level. For example, the number of patients who have acute otitis media and need to be treated for one patient to benefit clinically is 17 (95% CI, 13 to 22). This means that the "true" number needed to treat has a 95% probability of lying between 13 and 22. The CI also provides a measure of the precision of the estimate, with wider intervals indicating lower precision and narrow intervals indicating greater precision.

Both the CI and the *P* value describe the same statistical significance. If the *P* value is greater than 0.05 or the CI contains a value corresponding to "no effect" (sometimes expressed as a relative risk of 1 or a treatment difference of 0), the results can be considered nonsignif-

Table 3. Types of Bias Influencing Study Validity

Bias	Description
Selection bias	When comparisons are made between groups of patients that differ in ways other than those factors under study that could affect the outcome of the study
Measurement bias	When methods used for measurement are not applied consistently between groups under study
Confounding bias	When an associated factor other than the one under study is confused with or altering the true results

Table 4. Common Variables Used to Describe Results From Clinical Research

Therapeutic Results

$$\text{Relative Risk Reduction (rRR)} = \frac{\text{Control event rate (CER)} - \text{Experimental event rate (EER)}}{\text{Control event rate (CER)}}$$

$$\text{Absolute Risk Reduction (ARR)} = \text{CER} - \text{EER}$$

$$\text{Number Needed to Treat (NNT)} = 1/\text{ARR}$$

Diagnostic Results

Sensitivity: Proportion of persons *with* the condition who *test positive* for the condition $= \frac{a}{a+c}$*

Specificity: Proportion of persons *without* a condition who *test negative* for the condition $= \frac{d}{b+d}$*

$$\text{Positive Likelihood Ratio (LR+)}: \frac{\text{Probability of a positive test if the disease is present}}{\text{Probability of a positive test if the disease is absent}} = \frac{sensitivity}{1-specificity}$$

$$\text{Negative Likelihood Ratio (LR−)}: \frac{\text{Probability of a negative test if the disease is present}}{\text{Probability of a negative test if the disease is absent}} = \frac{1-sensitivity}{specificity}$$

*See above

	Disorder present	Disorder absent
Result positive	a	b
Result negative	c	d

Cause Results

$$\text{Relative Risk (RR)} = \frac{\text{Probability of the outcome if the risk factor is present}}{\text{Probability of the outcome if the risk factor is absent}} \text{ (Cohort, prospective)}$$

$$\text{Odds Ratio (OR)} = \frac{\text{Odds of having risk factor if condition is present}}{\text{Odds of having risk factor if the condition is not present}} \text{ (Case-control, retrospective)}$$

icant. Verifying that the result is within the range expressed in the CI or that the *P* value is less than 0.05 indicates that it is appropriate to continue with the appraisal process. On the other hand, if the results lie outside the CI or the *P* value is greater than 0.05, it is appropriate to disregard the results and search elsewhere for answers to the clinical question.

Once the level of chance is determined as acceptable and the results are statistically significant, the focus is directed toward interpreting the results. Pediatricians must understand some statistical terms to interpret studies properly. Only a few of the more clinically useful terms are noted in Table 1 under the "statistical results" portion. Definitions for some of these terms are provided in Table 4. Further definitions and suggestions for appropriate use of the skills identified in the EBM Toolbox are available. (8) Reviewing criteria in the article by Kocher and associates, as listed under the "Statistical Results" section of Table 1 for articles on clinical manifestations, (11) indicates that the following three criteria have been met within the article:

- The frequencies for each clinical finding were apparent
- The findings correlated with the course of each disorder
- Confidence intervals were provided and did not include 0.

Thus, the article by Kocher and associates satisfies the most important validity criteria with regard to study design (thereby minimizing potential sources of bias) and appropriately reports the findings for clinical use (precisely quantifying the extent that chance affected the reported results). Now the question is asked, "Are these results useful for the specific patient?"

Applicability

The next step is to ensure that the valid data are applicable to the specific clinical situation. Technically, application is not part of the critical review process. However, because we are doing the review to treat a real person, the applicability of the study becomes critical in answering the question by assuring that a study that has good validity matches closely and, therefore, can be applied directly to a specific individual. As described by Dans and colleagues, (12) the issues regarding applicability can be organized around three areas: biologic issues, social and economic issues, and epidemiologic issues (Table 5).

The guidelines can help the busy pediatrician make decisions regarding applicability and protect against making erroneous generalizations of study results, while not being overly conservative and discarding useful clinical data for use in the patient's care.

If there are no important disease or patient differences to affect treatment response, no patient or practitioner compliance problems, and no significant comorbid conditions or expected target event rates to change treatment efficiency, the study results can be applied confidently to the individual patient. If, however, there are clinically important issues of relevance that cannot be resolved, clinicians should consider the results as not applicable to the particular decision and seek other data.

Applying the "Applicable" criteria in Table 1 to the study by Kocher and associates shows:

- No significant comorbid conditions or event rates that will change frequency
- Disease manifestations have not changed since study completion

The age range in the study is appropriately narrow and includes patients the same age as ours within the demographic population. However, because the hospital setting in which this study was conducted does not reflect our ambulatory outpatient presentation and the differential diagnostic spectrum may vary, the box regarding study population is not checked. No concerning comorbid conditions or event rates in the patient population create significant differences between our patient and the study population, and the differential diagnosis considered for the presenting clinical manifestations has changed little over time. Overall, the study meets most of the salient issues regarding applicability to our patient.

The three areas delineated in Table 1 have shown that the results of the article by Kocher and associates are valid and applicable. We can use the 99.6% probability of septic arthritis predicted with their algorithm with good confidence.

Table 5. Guides on Applicability

Biologic Issues
☐ Pathophysiologic differences in the illness that could lead to a different response
☐ Patient differences that could lead to a different response
Social and Economic Issues
☐ Patient who has compliance issues or beliefs that may affect response
☐ Practitioner who has compliance issues or beliefs that may affect response
Epidemiologic Issues
☐ Patient has comorbid conditions that may change beneficial or harmful outcomes
☐ If untreated, the patient's risk of adverse events alters the treatment efficiency

Conclusion

One of the many challenges clinicians face today is providing medical care that incorporates valid, current information. Credibility of clinical results varies from study to study, reinforcing the need to use an approach that sorts out valid results and determines those findings applicable to individual clinical practice. Critical appraisal provides a process of evaluating research data via an objective and structured approach to decide on its validity and applicability.

Clinical decisions should be based on clinical experience and the family's beliefs, but must incorporate the most current, valid evidence available. Most busy clinicians do not have hours to spend critiquing an article, so the tools and techniques described herein can provide a brief and efficient approach to the literature to extract the information necessary to make evidence-based decisions.

How do we apply this valid, relevant data in our clinical scenario? The answer to this question relies on the integration of newly found EBM knowledge within medical decision making, which will be the focus of the next article in this series.

References

1. Onady GM, Raslich MA. Evidence-based medicine for the pediatrician. *Pediatr Rev*. 2002;23:318–322

2. Onady GM, Raslich MA. Evidence-based medicine: asking the answerable question (question templates as tools). *Pediatr Rev*. 2003; 24:265–268

3. Onady GM, Raslich MA. Evidence-based medicine: searching literature and databases for clinical evidence. *Pediatr Rev*. 2004;25: 358–363

4. Scott JN, Markert RJ. Relationship between critical thinking skills and success in preclinical courses. *Acad Med*. 1994;69:920–924

5. Grimes DA, Bachicha JA, Learman LA. Teaching critical appraisal to medical students in obstetrics and gynecology. *Obstet Gynecol*. 1998;92:877–882

6. Haynes RB. Where's the meat in clinical journals? *ACP Journal Club*. 1993;119:A22–A23

7. Guyatt GH, Rennie D, eds. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice*. Chicago, Ill: American Medical Association Press; 2002

8. Straus SE, Richardson WS, Glasziou P, Haynes RB. *Evidence-based Medicine, How to Practice and Teach EBM*. 3rd ed. Edinburgh, Scotland: Churchill Livingstone; 2005

9. Kocher MD, Zurakowski D, Kasser JR. Differentiating between septic arthritis and transient synovitis in the hip in children: an evidence-based clinical prediction algorithm. *J Bone Joint Surg Am*. 1999;81:1662–1670

10. Guyatt GH, Sackett DL, Sinclair JC, et al. Users' Guides to the Medical Literature: IX. A method for grading health care recommendations. *JAMA*. 1995;274:1800–1804

11. Richardson WS, Wilson MC, Williams JR Jr, Moyer VA, Naylor CD. Users' Guides to the Medical Literature XXIV. How to use an article on the clinical manifestations of disease. *JAMA*. 2000; 284:869–875

12. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' Guides to the Medical Literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. *JAMA*. 1998;279:545–549