Virtual Mentor
American Medical Association Journal of Ethics
January 2011, Volume 13, Number 1: 46-51.

POLICY FORUM

Rating Evidence in Medical Literature

Opeyemi O. Daramola, MD, and John S. Rhee, MD, MPH

A 24-year-old medical student comes to your clinic having had purulent rhinorrhea for 14 days, preceded by symptoms of upper respiratory infection. She reports having facial pain, frontal headache, nasal congestion, fever, and overall malaise. Nasal endoscopy reveals inflamed nasal mucosa with significant edema bilaterally. There is purulent rhinorrhea in the left middle meatus. Both cheeks are tender to the touch. You prescribe a 10-day course of amoxicillin and daily use of an intranasal steroid spray. She agrees with the use of amoxicillin but questions your nasal steroid recommendation. She proceeds to ask you about the effectiveness of intranasal steroids as adjunctive therapy and the strength of reported evidence supporting this recommendation.

An eager learner observing a seasoned physician will often probe the origin of the physician’s recommendation. Today’s patients are encouraged to seek more education about their health. Thus, they are not shy about questioning their physician’s recommendations. If the efficacy of an intervention has been established, how does it compare to available alternatives? How does one reach conclusions about the strength of relevant comparisons?

Evidence-Based Medicine

A concise and widely cited definition of evidence-based medicine (EBM) was formulated by David Sackett, one of its pioneers [1]. Sackett and colleagues define EBM as the “conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients” [1]. In practice, the provision of compassionate EBM reflects the integration of evidence from research, wisdom from clinical experience, and respect for the patient’s values and preferences, while recognizing existing circumstances [2, 3]. Most journals and specialty academies are dedicated to the continuous pursuit of high-quality studies and explicit grading recommendations in order to provide effective guidelines to physicians [4].

To understand the strength of guidelines and management strategies, one must be familiar with the different levels of evidence. The Oxford Centre for Evidence-Based Medicine (OCEBM) provides a popular scale for stratifying evidence from strongest to weakest on the basis of susceptibility to bias and the quality of the study design [5]. A modified and condensed version of the OCEBM scale is presented in table 1. A similar hierarchy is used by the U. S. Preventive Services Task Force in grading evidence [6, 7].

Table 1. Modified presentation of the Oxford Centre for Evidence-Based Medicine levels of evidence [5].

Grade of Recommendation	**Level of Evidence**	**Type of Study**
A	1a	SR (with homogeneity) of RCTs and of prospective cohort studies
	1b	Individual RCT with narrow confidence interval, prospective cohort study with good followup
	1c	All or none studies, all or none case series
B	2a	SR (with homogeneity) of cohort studies
	2b	Individual cohort study
	2c	Outcomes research, ecological studies
	3a	SR of case control studies, SR of 3b and better studies
	3b	Individual case control study, nonconsecutive cohort study
C	4	Case series/case report, poor quality cohort studies
D	5	Expert opinion, bench research

SR: systematic review; RCT: randomized controlled trial.

Randomized controlled trials (RCTs) are considered the gold standard in modern medicine for determining the efficacy of a treatment. Individual RCTs are level 1b evidence. Systematic reviews of homogenous RCTs are regarded as the highest level of evidence—level 1a. These systematic reviews consist of information synthesized from individual, well-designed RCTs where participants are similar and have equal chances of being assigned to an intervention group, a control group, or a placebo group. Systematic reviews of trials with blinded investigators and subjects (i.e., double-blinded RCTs) are even more desirable than reviews of non-double-blinded trials. These studies go through rigorous measures to eliminate bias, but they tend to be expensive and time-consuming.

In the case of our medical student, a literature search would reveal a published Cochrane Database systematic review of double-blinded RCTs. This review reported that intranasal corticosteroids (INCS) had been found to be effective as monotherapy or as adjunctive treatment when compared to placebo treatment for acute rhinosinusitis [8]. This review examined 475 studies but excluded 471. In the selected four studies, which had a robust total of 1,943 participants, those treated with INCS had earlier resolution or improvement of symptoms than those receiving a placebo. This systematic review selected high-quality, double-blinded placebo-controlled RCTs with homogenous design, clear reporting of outcomes, and an adequate number of subjects to establish clinical significance.

Cohort studies are considered level 2b evidence. In this design, a population (cohort) is defined according to the presence or absence of a variable that may potentially influence the occurrence of a specific disease. Cohort studies can be prospective or retrospective. In prospective cohort studies, people at risk for certain diseases are followed over time to investigate trends or risk factors in those who get the disease. Predictor variables are measured before outcomes occur. In retrospective cohort studies, the sample is defined and predictor variables are reported after the outcomes have occurred. Epidemiology studies that compare outcomes of people who had a certain exposure to unexposed subjects are examples of cohort studies.

Suppose you are counseling a 35-year-old woman whose husband is addicted to smoking tobacco about the risk of environmental tobacco smoke (ETS) on cardiovascular health. Because the deleterious effects of smoking tobacco are well-established, it would be unethical to perform a RCT to answer this question. An appropriate cohort study, such as one performed by Iribarren et al., would be the highest level of study that can be performed ethically and pragmatically to address the question in this scenario [9]. Iribarren et al. investigated the independent effect of exposure to environmental tobacco smoke (ETS) on the risk of stroke among 27,698 lifelong nonsmokers. They found that 20 hours or more a week of ETS exposure at home (compared to less than 1 hour a week) was associated with a 1.29-fold and a 1.50-fold increased risk of first ischemic stroke among men and women, respectively.

In matched-case control studies (level 3b evidence) investigators retrospectively evaluate two groups—one group with disease and the other without disease—with the intent of finding risk factors or trends. Subjects are matched for age, sex, and other demographics. For example, in a Swedish nationwide study, Lagergren et al. convincingly demonstrated that people who have weekly symptoms of esophageal reflux disease were eight times more likely to have adenocarcinoma of the esophagus than matched subjects without these symptoms [10]. In other words, these investigators looked for the prevalence of reflux (predictor variable) among subjects with confirmed esophageal adenocarcinoma (cases) and compared it to the prevalence of reflux symptoms in a sample of those who did not have adenocarcinoma of the esophagus (control).

A case report that provides information on the diagnosis, intervention, and outcome for a single individual is level 4 evidence. Case series—articles written about a series of patients with a specific diagnosis—are also regarded as level 4 evidence. Both case reports and case series describe characteristics of patients with certain diseases and may help identify questions for future research. These studies are ranked lower than other designs because of associated bias, lack of random sampling, the absence of controls or a comparison group, and heterogeneity of subjects. While these studies do not meet criteria necessary for achieving higher evidence level status, they are quite common in reporting outcomes in surgical specialties. Some diseases treated by surgical intervention (or nonintervention) do not lend themselves well to the higher level study designs previously mentioned. For example, performing sham surgeries for the sake of a controlled trial is ethically unacceptable. Systematic review of case series and case reports are helpful in identifying trends that lead to positive outcomes in diseases with high morbidity or that are treated surgically.

Grading Evidence in Medical Literature

Different specialty academies and journals have historically adopted unique systems to grade medical evidence and indicate the strength of disease-specific treatment guidelines [4, 6]. Grading systems arm physicians with information to help them make consistent, well-informed decisions and limit disparities in health care. Each system has its own shortcomings. A detailed explanation of the disadvantages of each system is beyond the scope of this article. (The reader is referred to a review paper by David Atkins et al., which appraised six prominent systems for grading levels of evidence [6]).

In 2002, the Agency for Healthcare Research and Quality (AHRQ) conducted a review of available methodologies for grading the strength of a body of scientific evidence [11]. This review identified three important characteristics to consider in assigning a grade to studies: quality, quantity, and consistency. Quality, as discussed above, refers to the methodologic rigor or extent to which bias was minimized in a study. Consistency refers to the similarities in design, population, outcome, and data analysis in studies attempting to answer the same question. Quantity refers to the number of subjects in individual studies and number of studies included in reviews. Seven systems fully addressed these key elements [11].

Grading of recommendations is useful when there is a need for a consensus guideline regarding the approach to a particular disease. Systematic reviews report the levels of evidence present in given studies and then assign grades to recommendations from these studies that reflect the strength of the intervention and likelihood of a successful outcome. The OCEBM system has grades of recommendations. Under this scheme, a grade A is a strong recommendation for or against an intervention. After critical appraisal, well-designed level 1a to 1c studies tend to result in grade A recommendations, level 2a to 3b studies result in grade B recommendations, and recommendations derived from level 4 studies are typically labelled grade C. Level 5 studies or “troubling,” “imprecise” studies at any level above 5 generate grade D recommendations (table 2). For example, recommendations from expert opinion without objective critical appraisal tend to be regarded as inconclusive and cannot be given a grade stronger than D.

Another popular grading system is the Strength of Recommendation Taxonomy (SORT) used by the journal of the American Academy of Family Physicians [4]. While the algorithms behind these systems are not identical, the outcomes are fundamentally similar. The simplified version in table 2 underrepresents the complexity of the system, and the reader is encouraged to peruse the algorithm behind these grading systems [4-6, 11].

Table 2. Similarities between the SORT and OCEBM grading systems.

	Grading System	
	SORT*	**OCEBM****
A	Recommendation based on consistent and good quality patient-oriented evidence	Consistent level 1 studies
B	Recommendation based on inconsistent or limited-quality patient oriented evidence	Consistent level 2 or 3 studies or extrapolations from level 1 studies
C	Recommendation based on consensus, usual practice, disease-oriented evidence, case series for studies of treatment or screening, and/or opinion	Level 4 studies or extrapolations from level 2 or 3 studies
D		Level 5 evidence or troublingly inconsistent or inconclusive studies of any level

*SORT: Strength of Recommendation Taxonomy
**OCEBM: Oxford Centre for Evidence Based Medicine

Final Comment

One must be careful not to adopt an inflexible approach of only applying recommendations of greater strength. The practice of evidence-based medicine is not "cookbook" medicine, and therefore the basis for patient care decisions should not be restricted to randomized trials or meta-analyses [1, 12]. There are uncommon diseases and complex pathologies that cannot be investigated with study designs that achieve levels of evidence higher than 3 or 4.

In returning to our case illustration, let us assume that our medical student actually has chronic rhinosinusitis with nasal polyposis. She was counseled by a previous otolaryngologist that a surgical polypectomy may be performed to achieve better control of her disease. You perform a literature search and find level 3 and 4 evidence that supports polypectomy as an option. Although the level of evidence is not any higher than 3 or 4, surgery is not necessarily an inappropriate recommendation for the patient. As discussed earlier, study design limitations are inherent to some situations and therefore the physician must make a "conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients" [1].

References

1. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence-based medicine: what it is and what it isn't. *BMJ*. 1996;312(7023):71-72.
2. Dickersin K, Straus SE, Bero LA. Evidence based medicine: increasing, not dictating, choice. *BMJ*. 2007;334(Suppl 1):s10.
3. Burton MJ. Evidence-based medicine and otolaryngology-HNS: passing fashion or permanent solution. *Otolaryngol Head Neck Surg*. 2007;137(4 Suppl):S47-51.
4. Ebell MH, Siwek J, Weiss BD et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. *Am Fam Physician*. 2004;69(3):548-556.
5. Oxford Centre for Evidence-Based Medicine. Levels of evidence. http://www.cebm.net/index.aspx?o=1025. Accessed December 15, 2010.
6. Atkins D, Eccles M, Flottorp S, et. al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches; The GRADE Working Group. *BMC Health Serv Res*. 2004;4(1):38.
7. Hadorn DC, Baker D, Hodges JS, Hicks N. Rating the quality of evidence for clinical practice guidelines. *J Clin Epidemiol*. 1996;49(7):749-754.
8. Zalmanovici A, Yaphe J. Steroids for acute sinusitis. *Cochrane Database Syst Rev*. 2007;2:CD005149.
9. Iribarren C, Darbinian J, Klatsky AL, Friedman GD. Cohort study of exposure to environmental tobacco smoke and risk of first ischemic stroke and transient ischemic attack. *Neuroepidemiology*. 2004;23(1-2):38-44.
10. Lagergren J, Bergstrom R, Lindgren A, Nyren O. Symptomatic gastroesophageal reflux as a risk factor for esophageal adenocarcinoma. *N Engl J Med*. 1999;340(11):825-831.
11. West S, King V, Carey TS, et al. *Systems to Rate the Strength of Scientific Evidence*. Rockville, MD: Agency for Healthcare Research and Quality; 2002. Evidence Report/Technology Assessment series; no. 47.

12. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomized controlled trials. *BMJ*. 2003;327(7429):1459-1461.

Opeyemi O. Daramola, MD, is a third-year resident in the otolaryngology-head and neck surgery residency program in the Medical College of Wisconsin Affiliated Hospitals in Milwaukee. He completed his undergraduate education at Adams State College in Colorado and his medical studies at the University of Minnesota Medical School. He is working on research projects on hereditary hearing loss, outcomes in surgical management of pediatric vascular anomalies, and reporting of long-term outcomes of tracheotomy-related complications.

John S. Rhee, MD, MPH, is chief of the Division of Facial Plastic and Reconstructive Surgery and professor of otolaryngology and dermatology at the Medical College of Wisconsin in Milwaukee. Dr. Rhee was recently appointed the coordinator for research for the American Academy of Otolaryngology-Head and Neck Surgery Foundation (AAO-HNSF). He is the deputy editor in chief for the *Archives of Facial Plastic Surgery*, a senior examiner for the American Board of Otolaryngology, and a board member of the American Board of Facial Plastic and Reconstructive Surgery.

Related in VM

Evidence-Based Medicine: Can You Trust the Evidence? December 2010

The Role of Comparative Effectiveness Research in Developing Clinical Guidelines and Reimbursement Policies, January 2011

The viewpoints expressed on this site are those of the authors and do not necessarily reflect the views and policies of the AMA.