Understanding evidence-based medicine: A primer

J. Kell Williams, MD
Tampa, Fla

Evidenced-based medicine is the concept of formalizing the scientific approach to the practice of medicine for identification of "evidence" to support our clinical decisions. It requires an understanding of critical appraisal and the basic epidemiologic principles of study design, point estimates, relative risk, odds ratios, confidence intervals, bias, and confounding. By using this information, clinicians can categorize evidence, assess causality, and make evidence-based recommendations. Evidence-based medicine allows analysis of complicated material so that we can make the best possible clinical decisions for the populations we serve. (Am J Obstet Gynecol 2001;185:275–78.)

Key words: Evidence-based medicine

As the science of medicine advances, decision making based on clinical experience and opinion are bowing to a more evidence-based approach. Archie Cochrane first formally suggested the idea when he published in 1972 an article entitled *Effectiveness and Efficacy: Random Reflections on Health Services*, that advanced the concept that health care should be evaluated on the basis of scientific evidence rather than clinical opinion.[1] Out of this was formed the Cochrane Collaboration, with multiple centers in North America and Europe. The Collaboration "prepares, maintains, and promotes the accessability of systematic reviews of the effects of health care interventions."[2]

The mostly widely accepted definition of evidence-based medicine (EBM) is that of David Sackett, MD (McMaster University, Hamilton, Ontario). 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,"[3] a definition requiring practitioners to synthesize information regarding populations and, using clinical expertise, relate this evidence to the individual. In 1998, the American College of Obstetricians and Gynecologists (ACOG) published a manuscript on applying the precepts of EBM to clinical practice.[4]

There are common misconceptions about EBM. Many think of EBM as "cookbook" medicine that is based solely on protocols. However, a compilation of rules or guidelines may or may not reflect EBM principles. The managed care industry has consistently touted EBM as cost saving, an assertion not uniformly true. In fact, the application of EBM may often justify increased utilization of available resources. In addition, the practice of EBM is not restricted to conclusions based on randomized controlled trials (RCTs) and large meta-analyses but in fact uses all available data.

From the Department of Obstetrics and Gynecology, University of South Florida College of Medicine.
Presented at the Sixty-third Annual Meeting of the South Atlantic Association of Obstetricians and Gynecologists, Hot Springs, Va, January 20-23, 2001.
Reprint requests: J. Kell Williams, MD, Department of Obstetrics and Gynecology, University of South Florida College of Medicine, 4 Columbia Drive, Suite 529, Tampa, FL 33606. E-mail: jkwillia@hsc.usf.edu

0002-9378/2001 $35.00 + 0 ***6/6/116740***
doi:10.1067/mob.2001.116740

Critical appraisal

Grimes[5] suggests that understanding EBM is understanding critical appraisal, which is defined as the systematic gathering and synthesizing of the best available evidence using the acumen of the clinician.[5, 6] Clinicians can thus understand study limitations, recognize bias, extract information, and reach appropriate conclusions. Critical appraisal is based on well-established epidemiologic principles. Although they are not the only kind of study used in critical appraisal, clinicians rely on epidemiologic studies because they are interested primarily in the health outcomes of specific patients.

Epidemiology

The narrow definition of epidemiology is the study of the distribution of diseases in a population. The broader definition—an inductive science of biologic inferences derived from observations—is more appropriate when relating EBM principles to clinical decision making. The goal of epidemiology is to identify exposures, make observations, and estimate outcome relationships in order to determine if a causal relationship exists. Classically these studies have been quite good at the "who," the "what," and the "where." The application of EBM principles helps us to get to the "why."

The basis for the evidence

Analytic and descriptive are the two fundamental types of epidemiologic studies. Analytic studies attempt to asso-

ciate a given exposure to (or dissociate from) a risk of a particular disease or outcome. In the hierarchy of analytical studies, experimental designs carry more weight than observation ones for EBM purposes. Experimental design, such as the RCT, is considered the *gold standard* because the investigator controls the exposure of the factor of interest. However, because of the inherent limitations, impracticality and costs of RCTs, the majority of studies have been of observational design, notably cohort or case-control studies.

As opposed to analytical studies, descriptive studies merely report noncomparative results, such as incidence, prevalence, and mortality rates of diseases. Most epidemiologists do not consider a descriptive study to be evidence-based science because such a study does not lead to the goal of determining cause and effect.

In an RCT, individuals are assigned randomly to a group, usually by computer or card draw. At the top of the hierarchy of study design is the RCT with a placebo control group that is not exposed to the factor of interest. In double-blind RCTs neither the investigator nor the subject knows who is in which group. The investigator assesses the various factors prospectively throughout the duration of the study and determination of outcomes. Because they are cumbersome and costly, RCTs are most commonly short-term studies. When critically appraising RCTs, Dolan suggests the following questions be asked: Were patients truly randomized? Was the sample size large enough to detect a clinically meaningful difference in outcome? Were all patients accounted for? Were the investigators, clinicians, and/or patients blinded?[7]

If the interest is in an outcome that may develop after years of exposure, such as heart disease or cancer, it may not be practical to perform a randomized study that is sufficiently large and lasts long enough to provide the answers. For this reason, most studies that examine long-term health outcomes are either cohort or case-control studies. In 1996 reports of observational studies were most common among the articles published in *Obstetrics & Gynecology*.[8] In the classic cohort study design, subjects are categorized with respect to their exposure/treatment status and are followed prospectively. The investigator measures various factors and evaluates outcomes. These studies are also known as *follow-up* or *longitudinal*. Before the development of computerized databases, cohort was often synonymous with prospective. Today, investigators can readily perform retrospective cohort studies such as those published by the Kaiser-Permanente group (example[9]). In critically appraising a cohort study one should ask: Was exposure status clearly defined? How was the outcome assessed? What was the follow-up time and was it adequate to assess the outcome of interest? What was the loss to follow-up?[7]

Case-control studies are the traditional retrospective observational studies. Subjects who have a particular medical condition are interviewed about past exposure to the factor of interest. In a hospital-based case-control study, the controls are similar subjects who are in the hospital at the same time and do not have the particular condition. In a population-based control study, the controls are similar to subjects within a certain geographic area. By using this methodology, the investigators quantitate associations between various exposures and the outcome of interest. For example, did the cases use a certain medication more often than controls, and if so, which medication, for how long, and when. Case-control studies can be performed more quickly and less expensively than cohort studies. Questions to ask case-control studies are: How were the cases and controls chosen? What about recall bias? Are there factors not addressed by the investigators that could be related to both the exposure and outcome?[7]

A previously controversial use of statistics is meta-analysis, which is simply an analytical technique that combines data from many studies to get more power and precision. Results of studies are often imprecise because of the small number of subjects. The combining of many small studies results in larger numbers of subjects and a more precise estimate of whether or not a causal relationship exists. Early meta-analyses merely averaged the results of various studies and were justifiably rejected by epidemiologists. The more legitimate approach to meta-analysis extracts the data subsets in the studies, combines the data, and provides a summary estimate. In the best and most powerful meta-analyses, individual investigators agree to pool their original raw data, as if the pooled data had come from the same place at the same time and in the same manner, and recalculate the results. Clinicians are rapidly accepting this latter method because it overcomes the shortcoming of earlier aggregate analyses.[10]

In observational studies, the magnitude or strength of an observed association. Point estimates are expressed in cohort studies as relative risk (RR) and in case-control studies as odds ratio (OR). The RR is the risk of disease in the exposed group divided by the risk of disease in the unexposed group. The OR is the odds of exposure in the case group divided by the odds of exposure in the control group. Both provide a measure of the association between the disease and exposure. An RR or OR of >1.0 indicates the exposure is associated with an increased risk of the outcome; a value <1.0 indicates that the exposure is associated with a reduced risk of the outcome. The further the RR or OR is from 1.0 in either direction, the stronger the relationship and the more likely it is real. When the RR or OR is 1.0 or close to 1.0, there is little or no association between the exposure and the outcome.

Study results are generally considered statistically significant if the *P* value is <.05. Confidence intervals (95%) are calculated for individual point estimates. Results are statistically significant if the CI does not overlap 1.0. In

addition, a narrow CI indicates greater precision. Small studies are often imprecise and have wide confidence intervals. On the other hand, large studies often have a very narrow confidence interval indicating greater precision.

When considered alone, relative risk can be misleading when it deals with uncommon events. Doubling the risk of a rare event has very little impact on the actual number of persons affected. A more realistic interpretation of risk is reporting attributable risk: the actual number of extra cases that occur because of an exposure. Subtracting the incidence of the outcome without exposure (the baseline incidence in the general unexposed population) from the incidence of the outcome after exposure results in attributable risk.

From an epidemiologic standpoint, bias is the systematic variation of measurements from the true values and refers to the features of the study design or execution that result in an incorrect answer. There are various types of bias. Selection bias, also called referral bias or treatment bias, is often seen in observational studies. An example of selection bias is healthy women choosing a particular form of therapy versus the selection of that therapy by less healthy women, resulting in an exaggerated assessment of a therapy's favorable impact.

Diagnostic or detection bias is also common and appears when subjects avail themselves of more examinations and testing, resulting in the discovery of more pathology. For example, if exposed/treated subjects in an observational study are more clinically aware of a certain disease entity and more likely to be examined, subtle changes are more likely to be noted. If the clinician is subconsciously more alert to these complaints and orders diagnostic testing more frequently, there will be more disease discovered and at an earlier stage. This type of bias was clearly identified in the report of the Collaborative Group on Hormonal Factors in Breast Cancer in 1996,[11] likely resulting in an overestimate of breast cancer risk in women taking oral contraceptives.

Recall bias is important in case-control studies. If a woman has been diagnosed with breast cancer, she will try to think of everything she has ever done that might have contributed to her condition, whereas a healthy woman may not recall these things.

A newly defined bias is "attrition of susceptibles" bias, which is particularly important in drug studies. It is also called the "healthy user effect" bias or the "recency of market introduction" bias. Basically**, it means that the individual who does well when exposed to a particular factor of interest will be more likely than others to continue the exposure. Those who experience significant negative effects will discontinue use and will likely not restart in the near future. In hormonal contraception studies, for example, this bias can make older, well-established contraceptive methods appear safer than they actually are when compared to newer methods because those at greater risk of complication have already discontinued the method. If clinicians believe a newer product is inherently safer than an older product, they may prescribe it to patients at who are at increased risk of complication, resulting in an exaggerated estimate of the new product's risk.

Confounding is different from bias in that it results from an internal factor in the subject that distorts the relative risk rather than from a factor of study design. When epidemiologists suspect a confounding factor they control for it in the statistical analysis. Possible sources of confounding include gender, age, tobacco use, access to care, and reproductive history.

Categorizing evidence

Epidemiologic principles allow us to categorize evidence. Several groups have made significant contributions to developing strategies for categorizing the quality of evidence, such as the Canadian Task Force on the Periodic Examination,[12] the U.S. Preventive Services Task Force[13] and the Cochrane Collaboration.[2] Level I evidence is obtained from at least one properly controlled randomized trial and is considered the *gold standard* of evidence.[3] Level II-1 evidence is derived from controlled trials without randomization. The most common evidence derived from observational studies of contraception is Level II-2, well-designed cohort or case-control studies. Key elements to look for when critically appraising the quality of Level II-2 evidence include a clearly identified comparison group, blinding of the investigators, a description of the methods sufficient to allow confidence that major biases and confounding have been avoided, and an analysis that is consistent with the study design. Level II-3 evidence includes studies with external control groups or ecological studies. Level III evidence is derived from reports of expert committees, not because it is weaker than levels I or II, but because it is often difficult to ascertain the scientific origin of the committee opinion.

It is important to define what is not scientific evidence. Opinions of respected authorities that are based on clinical experience or descriptive studies and case reports are not scientific evidence. Many clinical decisions are made where no directly relevant data exists. Clinicians must rely on opinions of peers or on their own experience. There are estimates that only 4% of therapeutic decisions are based on strong evidence from clinical studies, whereas 45% are based on minimal evidence from studies but strong clinical consensus. The remaining 51% of decisions are based on neither evidence nor consensus but on personal opinion.[14]

Assessing causality

The assessment of cause and effect starts with determining the statistical relationships. If there is no association (RR close to 1.0) there is no cause and effect. If there

is an association, assessing the magnitude and precision of the relationship is the next step. If there is an important, precise relative risk, then the criteria proposed by Sir Austin Bradford-Hill should be used to determine causality.[15] The Bradford-Hill Criteria for Causation, first published in 1968, provide guidelines to differentiate causality from association. In the United States, these criteria are often referred to as the Surgeon General's Criteria because they were initially used to interpret the evidence regarding cigarette smoking and lung cancer.[16] The Bradford-Hill Criteria for Causation include study design, strength of association, consistency, gradient, biologic plausibility, specificity, coherence, temporality and the existence of analogies.[15] These criteria are often presented in table form; however, they are best utilized as a general concept rather than as a checklist.

The use of Bradford-Hill Criteria enhances evidence-based recommendations. The U.S. Preventive Services Task Force used this process to determine whether screening tests prevented bad outcomes.[13] In addition, Bradford-Hill Criteria are used to aid in assessing most observational studies. Recommendations are categorized as: A, good evidence for cause and effect; B, fair evidence for cause and effect; C, insufficient evidence to make a recommendation; D, fair evidence against cause and effect; and E, good evidence against cause and effect. The categorizing of recommendations is very useful in explaining evidence against an association. A Level E recommendation, good evidence against a cause and effect, may be as close as we can get to "proving the null."

Comment

Critics of EBM claim that all clinical trials have inclusion and exclusion criteria that make their subjects not "real world." They also point out that subjects and investigators are human and, if given the opportunity, may unconsciously subvert randomization; that not all trials are correctly analyzed or reported; and that epidemiologic risk may not equate to individual risk. However, EBM, while never perfect, allows clinicians to synthesize complicated material so that the best-informed diagnostic and treatment decisions can be made.

REFERENCES

1. Cochrane AL. Effectiveness and efficacy: random reflections on health services. London: Nuffield Provincial Hospitals Trust, 1972.
2. Cochrane Collaboration website. Available at http://www.cochrane.org. Accessed March 23, 2001.
3. Sackett DL, Rosenberg WMC, Muir Gray JA, Richardson WS. Evidence based medicine: what it is and what it isn't. It's about integrating individual clinical expertise and the best external evidence. BMJ 1996;312:71-2.
4. American College of Obstetricians and Gynecologists. Reading the medical literature. Applying evidence to practice. ACOG Department of Practice Activities Washington, DC: ACOG; 1998.
5. Grimes DA, Bachicha JA, Learman LA. Teaching critical appraisal to medical students in obstetrics and gynecology. Obstet Gynecol 1998;92:877-82.
6. Scott JN, Markert RJ. Relationship between critical thinking skills and success in preclinical courses. Acad Med 1994;69:920-4.
7. Dolan MS. Interpreting the literature. Clin Obstet Gynecol 1998;41:307-14.
8. Funai EF. Obstetrics & Gynecology in 1996: marking the progress toward evidence-based medicine by classifying studies based on methodology. Obstet Gynecol 1997;90:1020-2.
9. Petitti DB, Sidney S, Bernstein A, Wolf S, Quesenberry C, Ziel H. Stroke in users of low-dose oral contraceptives, N Engl J Med 1996;335:8-15.
10. Goodman SN. Meta-analysis in health services research. In: Armerian HK, Shapiro S, editors. Epidemiology and health services research. New York: Oxford University Press; 1998. p. 229-42.
11. Breast cancer and hormonal contraception: further results. Collaborative Group on Hormonal Factors in breast cancer. Contraception 1996;54(suppl 3):S1-106.
12. Canadian Task Force on the Periodic Health Examination. The periodic health examination. CMAJ 1979;121:1193-254.
13. U.S. Preventive Services Task Force. Guide to Clinical Preventive Services. 2nd ed. Baltimore, MD: Williams & Wilkins; 1996. p. 861-2.
14. Field MJ, Lohr KN. Guidelines for clinical practice. Institute of Medicine. Washington, DC: National Academy Press; 1992. p. 34-9.
15. Bradford-Hill A. Principles of Medical Statistics. 9th ed. New York, NY: Oxford University Press; 1971. p. 309-23.
16. Koop CE, Luoto J. "The Health Consequences of Smoking: Cancer'" overview of a report of the Surgeon General. Public Health Rep 1982;97:318-24.