Evidence-Based Medicine: An Introduction

Noah D. Cohen, VMD, MPH, PhD, Diplomate ACVIM

Basing clinical decisions on the best available evidence should improve patient care. Understanding the strengths and limitations of evidence-based medicine is essential for making such decisions. Author’s address: Department of Large Animal Clinical Sciences, College of Veterinary Medicine and Biomedical Sciences, Texas A&M University, College Station, Texas 77843-4475; e-mail: ncohen@cvm.tamu.edu. © 2009 AAEP.

1. Introduction
The term evidence-based medicine (EBM) is increasingly used in the medical literature. In 1993, a search using the term yielded 6 citations, whereas a similar search in April 2009 yielded 34,742 citations. The concepts of EBM were developed by clinical epidemiologists >20 yr ago. This burgeoning of EBM in human medicine has been extended to veterinary medicine, including equine medicine. An excellent presentation on how to apply EBM in practice was presented at last year’s (2008) annual convention of the AAEP. It seems that the terminology has at once generated considerable enthusiasm, confusion, and skepticism. The purpose of this presentation will be to review the underlying principles of EBM and to discuss some of the limitations of EBM for equine practice.

2. What Is EBM and Why Is It Important?
EBM has been described as “the integration of the best research evidence with our clinical expertise and our patient’s unique values and circumstances.” EBM refers not only to this laudable intention, but also to the methodology developed for accomplishing it. A fundamental principle of EBM is that, whenever possible, evidence for our clinical activities should be derived from well-designed studies of patients with spontaneous disease. The rationale for emphasizing patient-centered research is 2-fold. First, well-designed studies of spontaneous disease have greater relevance to clinical practice than do experimentally induced disorders, because the latter are rarely good mimics of their natural counterparts. Furthermore, diagnostic procedures or therapeutic interventions often will perform differently when applied to patients than when applied to animals with experimentally-induced disease. Traditionally, equine clinical practice has been based primarily on anecdotes, opinions, and an understanding of the pathogenesis and mechanisms of disease derived from experimental studies of horses or other animals rather than on patient-based studies. Explanations for the primacy of experimentation in equine medicine include practical considerations, including the fact that funding amounts and sources for patient-based studies are quite limited. Second, equine practice has been traditionally empiric as well as experimental. Veterinary practitioners have relied extensively on the knowledge, wisdom, and experiences of experts, along with intuition and logical deduction. EBM arose partly in response to widespread recognition that that the impressions
and judgments of experts can be wrong and that logical deduction can be inaccurate.

3. How Does one Practice EBM?
The methodology described for practicing EBM consists of the following five steps: (1) specifying a clinical question; (2) searching for evidence to answer the question; (3) critically appraising the evidence gathered; (4) applying the answer to the question to one’s patient(s); and (5) auditing the outcome of applying the results (i.e., how well did the EBM-derived answer work?). A brief overview of these five steps is provided.

Asking the Clinical Question(s)
The first step in the EBM approach is to define the question(s) to be answered for the clinical circumstances of the patient of interest. Although this seems simple, there are a number of complexities. A question asked at too superficial of a level (e.g., “What causes forelimb lameness in a horse?”) may yield numerous irrelevant reports from searching, whereas a question that is too focused may yield too few results (e.g., “What are the results of randomized, controlled clinical trials for intra-articular management of caudal heel pain of the forelimb in Quarter Horses used for team roping?”). Any clinical encounter will likely generate multiple questions regarding signalment and history, clinical examination findings, etiology and differential diagnoses, diagnosis, treatment, prognosis, prevention, and client experiences. Attempting to answer all questions relevant to a given patient simultaneously will be unrewarding for the clinician, client, and patient. Thus, it is important to prioritize one’s questions. The most important aspect of posing clinical questions is likely to be as a reminder to persistently inquire about the evidence on which one’s clinical actions are based: what is really known about any of the plans for one’s patient? This does not mean disregarding expert opinion, but it does mean abandoning the assumption that expert opinion is always best.

Searching for Evidence
Once a clinical question has been posed, one needs to find the best available evidence using a review of the literature. The search process should be as comprehensive as possible. New information is generated faster than it can be included in textbooks, and authors of review articles have beliefs, biases, and perceptions that influence their interpretations and recommendations. Thus, relying exclusively or even predomately on book chapters or review articles is not an appropriate strategy for gathering the best available information. Relevant clinical research reports (particularly patient-based research reports) are a preferred source of evidence. Acquiring these reports is easily accomplished at institutions with appropriate licenses for digital/electronic publications. Moreover, reference librarians may be available for assistance. If one lacks such licenses, it may still be possible to access resources through institutions, such as one’s alma mater or a local college or university. High-quality information can be accessed via the Internet, using resources such as PubMed (http://www.ncbi.nlm.gov/PubMed/) or BioMed Central (http://www.biomedicalcentral.com). These sources (and others, including links at www.AAEP.org) are freely available, often in abstracted form; however, access to full articles may be limited. Relying solely or predominately on subscriptions to individual journals is an inefficient means for acquiring clinically relevant information. Too many journals with too many articles that have clinically important information for equine practice are published each month for one to keep abreast. Moreover, one may miss critical information by limiting oneself to a routine list of journals. For example, few equine practitioners subscribe to or regularly peruse the journal Genomics, but all would have some interest in reading about the mutation that causes polysaccharide storage myopathy.

Critical Appraisal
Critical appraisal is arguably the most important step of EBM for veterinarians interested in practicing EBM. The process of critical appraisal is the basis for establishing the relative merit of information for basing our clinical actions. A hierarchy of evidence has been proposed for EBM that places a premium on information derived from patient-based (i.e., epidemiologic) studies. Studies derived from experimental models are considered weak in terms of clinical relevance, even when they involve in vivo experiments among the species of interest. This means that an understanding of the relative merits of various epidemiologic study designs must be familiar to those interested in practicing EBM. Information about epidemiologic study designs and their strengths and limitations in the context of equine medicine has been summarized recently.

Briefly, epidemiologic study designs can be defined as either experimental or observational. The term randomized, controlled clinical trial (RCT) is a synonym for an experimental epidemiologic study of patients. In RCTs, investigators control the assignment of the exposure (often a treatment) to which the patients are assigned. It is preferable that treatment assignment be randomized because the process helps to render the treatment groups as similar as possible for both measured and unmeasured factors that might be associated with the outcome of interest. It should be noted that randomization does not guarantee that significant differences between groups will be eliminated. Randomly assigned groups can differ simply by chance, and the potential for differences between groups to occur is inversely related to sample size. The RCT is considered the highest form of evidence for an individual study because of the advantages that result from randomly assigning patients, spec-
ifying primary study endpoints a priori, and blinding of patients and clinicians to the primary study outcomes. It should be noted that the term “randomized clinical trial” or “clinical trial” is often misused in equine medicine to refer to experimental studies involving research horses rather than clinic patients. True RCTs are rare in equine medicine. Thus, the best source of evidence in equine medicine is derived from observational epidemiologic studies. Observational study designs include cohort, case-control, and cross-sectional designs. In a cohort study, one first determines exposure status (such as horses that did or did not receive a treatment) and then one follows the horses forward in time to determine disease incidence; the incidence of disease can be compared among groups with differing exposures. In a case-control study, one first determines disease status (either a case or a control) and then one looks back historically at exposure status to determine whether the odds of disease are greater for certain exposures. In a cross-sectional study, one simultaneously determines both the disease status and the exposure status of the individuals under study.

Critical appraisal of studies involves three processes: (1) determining whether study results are valid; (2) assessing the clinical importance of valid studies; and (3) assessing the relevance to our practice. Results of studies that are not valid are biased. In this context, the term bias refers to a systematic error in study results resulting from shortcomings with the design, methods of data collection, or analysis of the data. Such systematic error is distinct from random error that results from imprecision inherent to the device(s) used for collecting data. Biases in patient-based studies fall into three categories: selection bias, information bias, and confounding bias. Studies can be assessed for validity/bias on the basis of either the design used for the study or on the basis if the type of clinical activity being addressed (i.e., studies designed to address questions regarding diagnosis, treatment, prognosis, or harm). Only the latter will be addressed herein.

When appraising an article that relates to a diagnostic test, it is critical to assess three elements: (1) the spectrum of disease represented among the patients studied; (2) whether the “gold standard” test was applied to patients irrespective of the results of the diagnostic test being evaluated; and (3) whether the “gold standard” was measured independently of the alternative test(s) being evaluated.

It is common for studies evaluating the performance of diagnostic tests to be assessed using severe forms of disease (e.g., necropsy-confirmed cases of equine protozoal myeloencephalitis) and horses free of signs of disease. Studies evaluating diagnostic tests should include cases that encompass the full spectrum of disease to which the test will be applied. Patients with mild and florid forms of the disease and those in early through late stages of disease should be evaluated; it also may be important to evaluate the test both among treated and untreated patients. In general, case-control studies are a weak source of evidence for evaluating diagnostic tests. Prospective studies of consecutively enrolled patients that undergo pre-specified criteria for diagnostic testing relative to a reference standard that is consistently applied are the best sources of evidence for evaluating diagnostic tests. Studies in which patients are not enrolled consecutively represent weaker evidence because there is potential for bias in which cases are selected for inclusion.

When a patient has a negative test, investigators may be inclined to forego testing with a reference standard, especially when that standard is more invasive than the test against which it is being compared. For example, to evaluate the diagnostic sensitivity and specificity of thoracic ultrasound for detecting subclinical *Rhodococcus equi* pneumonia using foals at a farm with endemic *R. equi* pneumonia, one might choose not to complete tracheobronchial aspiration to obtain samples for microbiologic culture and cytologic evaluation in foals that appear healthy and that lack abnormal findings of thoracic ultrasound examinations. Failure to perform such testing, however, introduces an important source of bias.

The diagnostic test and the reference standard should be assessed independently, and preferably results of other diagnostic tests should be unknown to those conducting a given test. One should be wary of studies in which the same individuals are performing all tests, particularly when any of the tests has categorical outcomes that involve subjective determinations (e.g., negative, weak positive, moderate positive, strong positive). Although blinding individuals might help reduce or eliminate bias when the same individuals are performing the test, the potential for the samples to be unmasked exists. A good understanding of the principles of sensitivity, specificity, predictive values, and likelihood ratios is essential for interpreting test results. This is because, if the study design is valid, one needs to be sure that the test can accurately distinguish patients with and without the disease. Although review of these topics is beyond the scope of this report, they are considered in many other places.

The reference standard for an individual study to evaluate treatment is the RCT. Because a study is an RCT does not mean by default that results should be accepted as valid. As with other study designs, bias can occur in RCTs. There are published criteria for evaluating RCTs. In brief, individual RCTs should be evaluated for
the following: (1) description of the randomization process; (2) whether allocation of treatment was concealed from those managing and evaluating patients; (3) the extent to which the study groups were similar at the time the study is initiated; (4) the extent to which patients were followed is important; (5) the duration of follow-up; (6) whether patients were followed up and monitored similarly irrespective of treatment group assignment; (7) whether data were analyzed according to the group to which patients were originally assigned (even if it is known that the patient was inadvertently assigned another treatment or failed to take their treatment); and (8) whether those administering treatment and monitoring patients were blind to which treatment was being given to each patient. There are numerous resources available that can help with assessing RCTs, including the Consolidated Standards for Reporting Clinical Trials (CONSORT; www.consort-statement.org).

Unfortunately, because of the relatively high costs and the amount of effort required for their conduct, valid RCTs are rarely available for evaluating most treatments used in equine medicine. Consequently, veterinarians often must rely on lesser forms of evidence. In the absence of RCTs, evidence from well-designed prospective (cohort) studies is preferred. Except for the absence of randomization, the principles of evaluating cohort studies for treatment are similar to those used for RCTs. Case-control studies are relatively weak sources of evidence for evaluating therapies because they are more subject to biases. Case-series and individual case reports should be viewed as preliminary and particularly subject to being misleading with respect to evaluating treatment effects.

For studies of prognosis, the prospective (cohort) design is considered best, although the case-control design may be useful for rare diseases or disorders for which follow-up must be very long. It is important that patients in cohort studies be included relatively early in the disease process, to avoid missing out on more severely affected patients that might die before being included. How individuals were identified for inclusion is critical for evaluating the potential for selection bias. As for clinical trials, it is important for cohort studies that the follow-up procedures are consistent among groups, that losses to follow-up are not excessive, that events of interest are not missed, and that the length of the study period is appropriate for the disease of interest. Evidence that those lost to follow-up were compared with the baseline population and that the groups appeared similar enhances the validity of a cohort study. As a guideline, the validity of study results should be interpreted with considerable caution when >20% of a cohort is lost to follow-up. Death is an outcome that is fairly objectively determined, but defining either the cause of death or other non-fatal outcomes (e.g., failure to return to racing, infertility) can be more subjective and less accurately determined. In evaluating prognostic studies, one should give greater credence to results of those in which clear definitions for objective assessments of outcomes are specified than those which use more subjective or less well-defined endpoints.

If a study is determined to be valid, one must assess the clinical importance of the study. Clinical importance is generally assessed with respect to the magnitude of the clinical effect/association that has been quantified. A number of measures of the strength of effect/association exist, such as relative risks, odds ratios, relative risk reduction, and risk differences. Another clinically useful measure of effect is the number needed to treat (NNT). The NNT is the number of patients that need to be treated during the study period to prevent one additional case of the disease. The NNT is the inverse of the absolute value of the absolute risk difference. For example, in an RCT of azithromycin for preventing R. equi pneumonia, the absolute risk difference through weaning between foals receiving azithromycin and those that did not was 16%, such that the estimated NNT was 6.25 (1/0.16). Thus, it was estimated that seven foals needed to be treated with azithromycin to prevent one new case of R. equi pneumonia developing from birth to weaning. To reflect the effects of random error inherent in measuring outcomes, confidence intervals should be calculated for all measures of clinical importance. For the R. equi prevention example, the 95% confidence interval for the estimated NNT was 4–12; thus, one can have 95% confidence that the true NNT for the period from birth to weaning for chemoprophylactic administration of azithromycin was between 4 and 12 foals.

If the results of a study are valid and clinically important, the next step is to assess the extent to which the patients studied are similar to the horse(s) to which the results will be applied. Generally, there will be considerable heterogeneity among study populations making it very difficult to have great confidence in results from most studies. Moreover, there are no formal guidelines or measures for implementing this aspect of assessing clinical importance.

In human medicine, there have been efforts to develop grading systems for evidence. Although similar systems may be developed for veterinary medicine, the need for application of these grading systems (such as formulating treatment policies or formal treatment guidelines) in equine medicine is not compelling.

Applying the Results
Applying the results of the best evidence is problematic for a number of reasons. In equine medicine, one is limited to relatively weak forms of evidence. This means that even our best evidence is not very good. In fact, there may not be evidence to suggest that applying results of EBM that are weak is superior to relying on expert opinion. Although EBM
calls for the integration of our clinical impressions and expertise with results of searching and appraising the literature, there is no prescribed method for how this integration is to be done, including how to resolve conflicts among evidence-based literature review and appraisal, expert opinion, and clinical impressions.

Assessing the Outcome of EBM
Outcomes assessments in equine medicine are fairly limited. Outcomes assessment by clinicians usually focuses on how the patient fared, but outcomes assessments can also include the perceptions of horse owners. The perceptions and desires of owners strongly influence their assessment of their horse’s clinical outcome. Their perceptions of their interactions with us may influence their assessment of the case’s outcome as much as objective measurements of health or disease. Cumulating and systematically evaluating these various outcomes assessments (patient health, owner perceptions, etc.), can help us to determine the extent to which our efforts benefit our patients and clients.

4. Limitations of EBM
There are a number of important limitations of EBM. These include the fact that the definitions underlying EBM (evidence, best, available, etc.) are not well defined; clear, widely accepted definitions are critical for reliability and validity in terminology or else the term may be variously invoked with differing meanings. Another important limitation is the fact that the methodology for and systematic evaluation of EBM are incomplete. Despite these limitations, there are at least two major advantages to embracing the concept of EBM in equine practice. First, practicing EBM will force us to think more carefully about what is known—or thought to be known—at all stages of case management. Second, awareness of EBM will help to move veterinarians toward seeking more information that is derived from clinical research involving patients (rather than experimental animal studies).

References