Evidence-Based Medicine, Part 3. An Introduction to Critical Appraisal of Articles on Diagnosis

Damon A. Schranz, DO
Michael A. Dunn, OMS III, MBA

This article provides an introductory step-by-step process to appraise an article on diagnosis. The authors introduce these principles using a systematic approach and case-based format. The process of assessing the validity of an article on diagnosis, determining its importance, and applying it to an individual patient is reviewed. The concepts of study population homogeneity, reference and criterion standards, and completeness are discussed to help physicians determine an article’s validity. Instruction on calculating prevalence, sensitivity, specificity, and positive and negative predictive values and likelihood ratios is provided and applied to a hypothetical clinical scenario. Study generalizability and the role of patient values, expectations, and concerns are also addressed. The skills learned from appraising an article on diagnosis in the manner outlined provides a solid basis for life-long learning and improved patient care.

(Editor’s note: This article is part 3 of a six-article series intended to introduce the principles of evidence-based medicine (EBM) to busy clinicians, physician residents, and medical students. Because the application of EBM is a career-long process, further training is needed beyond the information provided within this article and series. A foundation of knowledge about research methods is critical in understanding EBM; however, such details, though introduced, are beyond the scope of this series.)

J Am Osteopath Assoc. 2007;107:304-309

Evidence-Based Medicine, Part 3. An Introduction to Critical Appraisal of Articles on Diagnosis

Every medical school graduate is taught how to assess and diagnose a patient’s condition. A diagnostic test and its results are important tools that help guide physicians to the appropriate diagnosis by revealing the likelihood of whether or not a patient has a specific condition.1 Results of the best diagnostic tests remove all doubt that a patient has (or does not have) an identifiable disease or disorder. However, not all diagnostic tests are equal in their ability to differentiate the presence, absence, or severity of a particular disease or condition present in a patient. Therefore, clinicians need a method for selecting the best test to meet a particular patient’s needs.2 Evidence-based medicine (EBM), the practice of appraising the literature in a time-efficient manner to answer a clinical question about, and for, the patient,3 is such a method.

In this article, we present a strategy for busy clinicians, physician residents, and medical students to critically assess the medical literature on diagnosis. In-depth details of research methods are beyond the scope of this introductory series on EBM. Readers are encouraged to seek further training on these topics with supplemental learning opportunities and continuing medical education. Finally, the clinical scenario described has been simplified to provide readers with an illustrative example for the general concepts introduced.

Searching the Evidence

To find an article that is appropriate to review for the purpose of better establishing patient diagnosis, physicians can approach searching the evidence in two ways. In general, physicians who practice EBM search the evidence for an article that contains the information sought. However, physicians in the habit of summarizing articles relevant to their practice can first refer to their clinically-appraised topics (CATs) when faced with a clinical question.

Critically Appraised Topics

Similar to the index card method of recording researched information, CATs are a personal method of documenting the results of any article in medical literature for a specific clinical problem.3 These records are simply summaries of a study and its results that a physician can create for later retrieval, review, and reuse (Figure 1). The most thorough CATs consist of the article title, the clinical “bottom line,” the clinical question, a summary of the results, comments, the date the study was published, and any relevant citations.3 A more detailed description of these components is available in Figure 2.4 Physicians may choose to share their CATs with colleagues, in which case physicians should also include their name or initials as the CAT appraiser.

A CAT is not a systematic review and should not be considered a practice guideline because the information found in it may not be authoritative.3 However, physicians will begin...
to refine and improve their EBM skills after summarizing varying clinical issues in this fashion.3

Systematic Reviews vs Individual Articles
When searching the evidence for a clinically relevant article on diagnosis, systematic reviews and meta-analyses are the most authoritative types of reports.3 These studies, which critically appraise and summarize multiple similar studies concerning a common medical problem, are not as numerous as individual articles. However, such reviews are only as good as the individual studies they include. A physician must be vigilant in critically assessing a systematic review or meta-analysis before putting its recommendations into practice. For guidelines on how to appraise such review articles, a handbook is available on The Cochrane Collaboration Web site (http://www.cochrane.org/resources/handbook/Handbook426Sept2006.pdf).

In the absence of a systematic review or meta-analysis, individual articles are often the only source of new information available to clinicians. Assessing these individual articles (Figure 3) is the focus of this paper.

Validity of Articles on Diagnosis
To ascertain the validity of an individual article, physicians need to determine not only if the study’s results and conclusions were accurately deduced but also if the methods used to arrive at the conclusions were free of error and bias. This is the most crucial step in evaluating an article. If its validity is questionable, the article’s results cannot be confidently interpreted.2,5,6 Physicians may use the following questions3 to help them determine an article’s validity:

- Was there an independent and blind comparison to a reference standard?
  A reference standard is a method of defining the presence or absence of the disease or condition in question.7 To determine whether a diagnostic test is effective, a reference standard is needed for comparison.8 If a reference standard is not used in the study, the benefit of the diagnostic test cannot be ascertained. In addition, not all reference standards are equal or subjective.9 For example, reference standards for psychiatric disorders may not be clear-cut and subjective, and other standards, such as biopsies, rely on expert interpretation. The best reference standard to evaluate the effectiveness of a diagnostic test is the criterion standard, which is considered the diagnostic model for identifying a specific disease or condition.3

The study’s data collection and analysis must be carefully planned and executed to ensure that unconscious (or conscious) biases are maximally reduced.3 In other words, in clinical investigations, those who perform tests and those who interpret the results should be independent of one another. Both groups of researchers should be blinded to the diagnostic and reference standard test results.

- Was the diagnostic test evaluated in subjects similar to patients seen in practice?
  Because physicians practice in a wide range of geographic areas and within various medical specialties, the patients they treat have distinct characteristics. For a study to be applicable to a physician’s patient, the study’s subjects need to have similar baseline characteristics. A physician who evaluates the applicability of an article in this way maximizes the likelihood that a study’s results can be generalized to his or her patient.

- Was the reference standard obtained regardless of the diagnostic test’s result?

Figure 1. Clinical scenario.

A 58-year-old man visits your primary care clinic for the first time in 17 years. He states that, although he feels fine and has no complaints, he is concerned that his blood pressure may be elevated. He relays to you that, while shopping at a supermarket last week, he had his blood pressure measured by the store’s blood pressure machine. He was alarmed to see a reading of 182/104 mm Hg and wanted to verify that his blood pressure was indeed that high.

During a routine physical examination you confirm that he is hypertensive but in otherwise good physical health. Naturally, your first reaction is to help the patient relieve his hypertension, but you are unsure about the appropriate goal.

You remember generating a critically appraised topic (CAT) from a study that demonstrated a marked reduction in major cardiovascular events when diastolic targets of 80 mm Hg were achieved. After reviewing this CAT, you realize that the benefits to the lower diastolic pressure applied only to patients with diabetes. However, the CAT also states that the US Preventative Service Task Force recommends screening hypertensive patients for type 2 diabetes mellitus.

You discuss the finding of the study with your patient along with the pros and cons of screening for diabetes. He agrees to think it over and return in a week for a follow-up visit. After writing a prescription to initiate treatment of his hypertension, you make a note to review the literature regarding diagnostic tests for type 2 diabetes mellitus in asymptomatic patients.
Assessment of a diagnostic test to a reference standard (preferably the criterion standard) requires that both tests are performed and their effectiveness compared, which should not be an issue if the comparison study is truly independent and blinded. One exception to the rule is a negative noninvasive diagnostic test result coupled with an invasive or risky reference standard. In this situation, the investigators would be hesitant to perform the invasive reference standard if the noninvasive diagnostic test results were negative. Studies can be designed to reduce this risk by creating, for example, a method to screen persons who do not have the target disorder, thus eliminating the need to verify the noninvasive negative result with an invasive test. However, a study should be viewed with suspicion if it does not independently perform the reference standard test and diagnostic test on every participant, even if the reference standard was considered invasive or risky.

**Study Results**

Now that a diagnostic article of interest is found and is deemed to have merit, one can evaluate its results to determine its general usefulness (Figure 4). Although this step of the appraisal process for articles on diagnosis appears intimidating, it only requires basic mathematic and statistical skills. With practice, these invaluable calculations will become second nature.

### Does the diagnostic test help determine who has the target disorder?

Research articles present information to emphasize the authors’ point of interest. Although this focus may be different from the reader’s particular interest, the information sought can usually be found within the article. To determine the diagnostic discrimination of a test, or the statistical assessment of how a diagnostic test compares with a reference standard,
With a high sensitivity and specificity, the test result, or the negative predictive value, was 97%. Therefore, within the study’s population, a positive diagnostic test result shifted the pretest odds of having type 2 diabetes mellitus from 11% to 43% (posttest), which is clinically significant.

Sensitivity, specificity, and positive (LR+) and negative (LR−) likelihood ratios are additional parameters to help physicians determine the usefulness of a test’s diagnostic abilities. Sensitivity is defined as the proportion of true positives (e.g., patients who test positive for a disease as measured by both the criterion or reference standard and the diagnostic test) of a study population. Specificity is the proportion true negatives (e.g., patients who test negative for a disease as measured by both the criterion or reference standard and the diagnostic test) of a study population. These parameters can be used to calculate the diagnostic test’s LR+ and LR−, which are the probabilities of getting a positive or negative test result if the patient has the condition compared with the probability of getting the result if the patient does not have the condition.

According to the Table, the LR+, the ratio of the true positive rate to the false positive rate, means that a positive test result would be 6.25 times as likely in someone with type 2 diabetes mellitus as in someone without type 2 diabetes mellitus. Likewise, in the referenced study, the LR−, the ratio of the false negative rate to true negative rate, a negative test result would be 0.28 times as likely in someone with type 2 diabetes mellitus as in someone without type 2 diabetes mellitus.

How can a diagnosis be determined?
An interesting and useful feature of high sensitivity and specificity values is that they can help rule in or rule out a diagnosis, respectively. Mnemonic devices can be used to help one remember how to use specificity and sensitivity to make a clinical decision.

- With a high sensitivity (Sn), a negative (N) result effectively rules out the diagnosis (SnNout)
- With a high specificity (Sp), a positive (P) result effectively rules in the diagnosis (SpPin)

For example, a positive result on a rapid streptococcal antigen test rules in (SpPin) the diagnosis of streptococcal pharyngitis, and a negative D-dimer test result effectively rules out (SnNout) the diagnosis of deep venous thrombosis (Figure 5).

Practical Use
Now that the article has been reviewed for its validity and relevance to the physician’s patient and it is determined to have significant clinical applicability, one still needs to answer a fundamental question: Can these results benefit the patient? If a physician cannot confidently answer “yes,” the article must be placed aside and a new search started. The potential for “wasted time” is the main factor behind why physicians often do not apply this step. However, the real waste of time—
not to mention a potential for harm—would result from implementing results that cannot be expected to help the patient or that are unrealistic to apply in the clinical setting.

Is the diagnostic test available and affordable in the physician’s clinical setting?
The diagnostic test must be available to a physician before he or she can order it. In addition, the diagnostic test must be affordable to patients or covered by their health insurance. Applying the right diagnostic tool at the appropriate time assists one’s efforts in reducing healthcare costs by reducing the number of unnecessary tests.

How can the physician determine a specific patient’s pretest probability of having the target disorder?
One method for determining a patient’s pretest probability of having the target disorder has already been discussed: using the study’s inherent disease prevalence. This inherent prevalence, however, is appropriate only if the physician’s patient is similar to those in the study’s population. Other means of determining a patient’s pretest probability include the physician’s clinical experience, regional and national statistics, and studies specifically developed to determine pretest probabilities for the target disorder. All of these methods have merit and should be considered. The one that is chosen should be based on available data and their applicability to the particular patient.

Is the pre- to posttest probability shift valuable to the specific patient?
The purpose of performing a diagnostic test is to confirm or rule out a diagnosis. Therefore, the shift from pre- to posttest probability of the diagnostic test must be clinically useful; if it is not, the test result will not be valuable to the patient or the decision-making process.\textsuperscript{11}

The shift in pretest probability to the positive predictive value (or posttest probability) for a given diagnostic test is an

\begin{table}[!h]
\centering
\caption{Diagnostic Test Results of Type 2 Diabetes Mellitus Compared With the Criterion Standard (N=1471) and Statistical Assessment of the Data}
\begin{tabular}{|c|c|c|c|}
\hline
\textbf{Diagnostic Test Result (Blood Glucose, mg/dL)} & \textbf{Label} & \textbf{With Type 2 Diabetes Mellitus} & \textbf{Without Type 2 Diabetes Mellitus} \\
\hline
True positive (\geq 120) & a & 118 & \\
False positive (\geq 120) & b & 158 & \\
False negative (<120) & c & 39 & \\
True negative (<120) & d & 1156 & \\
\hline
\textbf{Totals} & & \textbf{157} & \textbf{1314} \\
\hline
\end{tabular}
\begin{tabular}{|c|c|c|}
\hline
\textbf{Statistical Assessment} & \textbf{Equation*} & \textbf{Equation With Data} & \textbf{Result} \\
\hline
\textbullet Prevalence & (a+c)/(a+b+c+d) & 157/1471 & 0.11 or 11\% \\
\textbullet Positive predictive value & a/(a+b) & 118/276 & 0.43 or 43\% \\
\textbullet Negative predictive value & d/(c+d) & 1156/1314 & 0.88 or 88\% \\
\textbullet Sensitivity & a/(a+c) & 118/157 & 0.75 or 75\% \\
\textbullet Specificity & d/(b+d) & 1156/1314 & 0.88 or 88\% \\
\textbullet Positive likelihood ratio & sensitivity/(1-specificity) & 0.75/0.12 & 6.25 \\
\textbullet Negative likelihood ratio & (1-sensitivity)/specificity & 0.25/0.88 & 0.28 \\
\hline
\end{tabular}
\textsuperscript{*}a=118; b=158; c=39; d=1156
\textsuperscript{11}
\end{table}

You create a table to consolidate the article’s findings that are related to your clinical question. This table assists your calculation of prevalence, pre- and posttest probabilities, sensitivity, specificity, and positive and negative likelihood ratios. In reviewing these statistics, you determine that the diagnostic test is clinically important.

\begin{figure}[!h]
\centering
\caption{Clinical scenario (continued)}
\end{figure}
effective discriminator for choosing between competing tests. Large LR+ values and small LR- values are indicative of significant shifts. For example, a diagnostic test that provides a LR+ of 1.0 will not shift the posttest probability at all. Therefore, it would be wasteful to perform the test because its results would not benefit the patient or the clinical decision-making process. On the other hand, a test with a LR+ of 10.0 would shift a pretest probability of 50% to a positive predictive value of 92%, which would be clinically useful.

In addition to the test’s pre- to posttest shift, one needs to consider the cost and invasiveness of the tests when choosing between competing diagnostic tests. When these competing elements are considered and balanced with the patient’s needs and informed consent, physicians can be confident that the best evidence is being applied in the most efficient and effective manner (Figure 6).

Conclusion
Although most clinicians are already incorporating EBM principles in their practices, often instinctively, some physicians may require a more organized approach to integrating this relatively new model of self-education. Improved comfort levels and true expertise in the practice of EBM are the result of additional education, repetition, and self-assessment. The principles of EBM allow physicians to stay informed while also improving the quality of the information communicated to patients during patient encounters. The systematic approach that is used to appraise an article on diagnosis is but one step in practicing EBM. Remember, the goal is always to provide the best care possible to patients—using one’s clinical expertise to address patient values and expectations for treatment.

References