

E. John Gallagher, MD

From the Department of Emergency Medicine, Albert Einstein College of Medicine, Bronx, NY.

The Problem With Sensitivity and Specificity...

See related article, p. 292.

[*Ann Emerg Med.* 2003;42:298-303.]

The goal of all diagnostic tests is the revision of disease probability. The power of a test to accomplish this goal is reflected in its performance characteristics. Unfortunately, these characteristics are expressed in a wide and unnecessarily confusing array of formats. To untangle the various descriptors of test performance, one can begin by dividing all diagnostic tests into 2 groups: (1) those that produce dichotomous outcomes and (2) those that produce continuous or ordinal outcomes.

If the outcome of a test is reported dichotomously (ie, positive or negative), its performance characteristics can be expressed in at least 3 ways: (1) sensitivity and specificity, (2) positive and negative predictive values, or (3) positive and negative likelihood ratios.

If the test outcome is reported as a continuous variable (eg, age, temperature, serum sodium, leukocyte count, troponin level) or on an ordinal scale (eg, a 5-point scoring system, rating the probability of appendicitis on computed tomography as very low, low, intermediate, high, or very high), the performance of the test may be expressed in at least 2 ways: (1) graphically, as a receiver operating characteristics (ROC) curve (which incorporates sensitivity and specificity at multiple thresholds), or (2) as interval likelihood ratios.

Of these several kinds of test performance characteristics, only likelihood ratios are indifferent to the for-

Copyright © 2003 by the American College of Emergency Physicians.

0196-0644/2003/\$30.00 + 0
doi:10.1067/mem.2003.273

mat of the test output, managing dichotomous, ordinal, and continuous data with equal facility. In a well-written article in this month's *Annals*, Brown and Reeves¹ argue persuasively that interval likelihood ratios are a particularly useful way of interpreting certain kinds of diagnostic tests. This commentary, written to accompany Brown and Reeves' work, constitutes an effort to compare these various test performance characteristics with the explicit intent of persuading readers that likelihood ratios, whether expressed as positive, negative, or interval likelihood ratios, are the most clinically useful tools currently available for the formulation of diagnostic testing strategies in everyday emergency practice.

SENSITIVITY AND SPECIFICITY: CONCEPTUALLY COUNTERINTUITIVE

Sensitivity and specificity are the most commonly used measures of diagnostic test performance. Their fundamental flaw is their failure to tell us what we wish to know. In fact, they seem configured to tell us very much the opposite by answering the wrong question: Given that a patient does or does not have a particular disease, respectively, what is the probability that a test result is positive (sensitivity) or negative (specificity)? This constitutes a confusing reversal of customary clinical logic, because knowledge of the patient's disease status would presumably preclude a diagnostic test aimed at detection of an illness the patient is already known to have.

The counterintuitive formulation of the definitions of sensitivity (proportion of patients with the disease who have a positive test result) and specificity (proportion of patients without the disease who have a negative test result) causes many young clinicians, particularly when first introduced to these concepts, to wonder how one can use a sensitive test (in which the denominator consists of all individuals with the target disorder) to exclude from further consideration the very disease that, by definition, all of them have. Similarly, how one can use a specific test (in which the denominator consists only of individuals who do not have the target disorder) to confirm the presence of a disease which, again

by definition, none of them has? The answer is that we set aside the clinically confusing conceptual definitions of sensitivity and specificity and make use of their mathematic properties. In the case of sensitivity (true positives/[true positives+false negatives]), we rely on the fact that a test with a very high sensitivity approaching 100% is likely to have a very low number of false negatives. A negative result generated by a test known to have few false negatives is, therefore, likely to be a true negative test result, thus excluding the target disorder. Analogous to this, in using specificity (true negatives/[true negatives+false positives]), we rely on the fact that a test with a very high specificity approaching 100% is likely to have a low number of false positives. A positive test result generated by a test known to have very few false positives is therefore likely to be a true positive test result, thus confirming the presence of the target disorder.

All of this, however, is a very contorted, roundabout way of obtaining a test result. Furthermore, this mathematic strategy rapidly unravels as one moves away from tests with near-perfect sensitivity or specificity (>95%) into the lower range of performance where most diagnostic tests reside. Likelihood ratios, in contrast, cannot only be calculated readily and directly from sensitivity and specificity, but have several additional important advantages over sensitivity and specificity in simplifying clinical problem solving.

PREDICTIVE VALUES: NUMERICALLY UNSTABLE

In contrast to sensitivity and specificity, predictive values provide dichotomous results that do tell us what we wish to know: Given a positive or negative test, what is the probability, respectively, that the patient does or does not have a particular disease? Unfortunately, predictive values, whether positive or negative, provide test performance characteristics that fluctuate markedly as a function of the prevalence of the target disorder. As disease prevalence increases in a population, the positive predictive value of a particular test increases, and reciprocally, the negative predictive value of the test

decreases. As prevalence decreases, the opposite occurs: the negative predictive value of the same test increases while the positive predictive value decreases. Thus, the very same diagnostic tests appear to perform better or worse simply because the prevalence of the target disorder changes. Because of this vulnerability to variation in disease prevalence (or prior or pretest probability), predictive values are numerically unstable and do not travel well from one population of patients to another.

In contrast to predictive values, sensitivity and specificity, likelihood ratios are reproducible test performance characteristics that tend to remain highly stable across populations, unless both disease severity and prevalence happen to shift concurrently.²

ROC CURVES: CLINICALLY CUMBERSOME

ROC curves offer a graphical display of the relationship between sensitivity and specificity for diagnostic tests with continuous or ordinal outcomes. In so doing, ROC curves avoid the loss of information that naturally occurs when a single dichotomous cut-point is superimposed on test results that contain more than simple binary (positive/negative) data. By taking into account degrees of positivity and negativity, the information contained in the gray-scale spectrum of diagnostic testing that would otherwise be lost by forcing a continuum of data into binary categories can be reclaimed. As Brown and Reeves¹ point out, the data distortion, which is particularly marked at or near the threshold dividing positive from negative test results, can be attenuated by slicing the output into more than 2 contiguous categories.

For reasons of mathematic convenience, ROC curves plot sensitivity (the true positive rate) on the vertical axis against the complement of specificity ($[1 - \text{specificity}]$ or the false positive rate) on the horizontal axis. The tradeoff between the true positive and false positive rates reflects the reciprocal relationship between sensitivity and specificity expressed as the ability of a given test to distinguish “noise” from “signal plus noise.”³

ROC curves have not found much application in the clinical arena. This is not only because they suffer from

the same limitations as sensitivity and specificity, but also because needing a curve for interpretation of test results in a clinical setting is unwieldy. In the future, this could be resolved by clinical decision support systems that display test-specific ROC curves on screen to accompany results and facilitate their interpretation. However, as noted by Tandberg et al,⁴ this technology would be put to better use by applying it to generalized likelihood ratios linked directly to Bayes' theorem. At present, ROC curves are more useful as descriptors of overall test performance, reflected by the area under the curve (AUC), with a maximum of 1.00 describing a perfect test and an AUC of 0.50 describing a valueless test. Although the meaning of the AUC is often difficult to grasp intuitively, it is a useful concept. Using the ROC curve displayed in Brown and Reeves' Figure 2 as an example,^{1,5} the AUC of 0.87 can be interpreted as follows: If a pair of patients, one who suffered a congestive heart failure–related cardiac event within the ensuing 6 months and one who did not, were selected at random from the study cohort, there would be an 87% probability that the patient destined to have a cardiac event had a higher B-type natriuretic peptide level at the time of study entry than the patient who was not slated to have a cardiac event.⁶

CLINICAL APPLICATION OF LIKELIHOOD RATIOS

Likelihood ratios express the likelihood that a particular diagnostic test result is likely to occur in a patient with (as opposed to a patient without) the target disorder. The “particular diagnostic test result” might include: (1) the presence or absence of a specific sign such as an S_3 gallop; (2) a reading of a V/Q scan, expressed on an ordinal scale as normal, very low probability, low probability, intermediate probability, and high probability; or (3) a laboratory value, such as B-type natriuretic peptide, expressed as continuous data in nanograms per milliliter.

In practice, one simply asks the question: Given the likelihood ratio of this test result, what are the odds that the patient has the disease at which the test is targeted?

The answer to this question will be the simple arithmetic product of the clinician's best guess at pretest odds multiplied by the likelihood ratio of the test result. Thus, likelihood ratios ranging from more than 1 to infinity increase disease likelihood or probability, likelihood ratios ranging from more than 0 to less than 1 decrease disease likelihood or probability, and likelihood ratios at or around 1 leave the posttest likelihood or probability of disease unchanged from pretest estimates. Likelihood ratios are always positive numbers.

To apply likelihood ratios to clinical practice, one can begin by dividing pretest odds, which are typically based on clinical data, into 3 qualitative categories of low, intermediate, and high, corresponding to prior probabilities of 25%, 50%, and 75%. These can easily be converted to pretest odds of 0.3, 1, and 3. Then, making use of Brown and Reeves¹ Table 2 as an example, patients with low, intermediate, and high pretest probabilities of appendicitis by history and physical examination, who are found to have a WBC count greater than $15 \times 10^3/\mu\text{L}$ (this interval of the test has a likelihood ratio=7), will have a posttest odds of appendicitis of approximately 2 (calculated as $0.3 \times 7 = 2.1$), 7 (calculated as $1 \times 7 = 7$), and 21 (calculated as $3 \times 7 = 21$), respectively. From this information alone, a rational argument could be made for further observation of the patient with a low pretest probability of appendicitis, whose odds of appendicitis have now increased to only about 2:1; obtaining a computed tomographic scan on the patient with an intermediate pretest probability of appendicitis, whose odds of appendicitis have just increased to approximately 7:1; and going directly to surgery with the patient with a high pretest probability of appendicitis, whose odds of appendicitis have just increased to approximately 21:1. Before obtaining this simple test and applying its performance characteristics expressed as likelihood ratios to one's clinical impression, odds of appendicitis of 1:3, 1:1, and 3:1 would not have provided sufficient separation between patients with low, intermediate, and high pretest probabilities to drive further clinical decisionmaking.

POSITIVE, NEGATIVE, AND INTERVAL LIKELIHOOD RATIOS

Although it is not mathematically or clinically essential, likelihood ratios are often divided for convenience into positive and negative categories when used to report test performance characteristics that are expressed dichotomously. A positive likelihood ratio tells us how much the odds of a disease increase when a test is positive. Derived from the same 2×2 table as sensitivity and specificity, a positive likelihood ratio = sensitivity / (1 - specificity) = (true positive rate) / (false positive rate). Conversely, a negative likelihood ratio tells us how much the odds of a disease decrease when a test is negative: negative likelihood ratio = (1 - sensitivity) / specificity = (false negative rate) / (true negative rate). As noted earlier, all likelihood ratios, including negative likelihood ratios, are positive numbers.

Seen in a broader context, positive and negative likelihood ratios are nothing more than a particular kind of interval likelihood ratio that surfaces only when test results are dichotomized into positive and negative outcomes. Under these conditions, there are only 2 "intervals" (one interval that contains all the positive test results and a second interval that contains all the negative test results). As soon as one moves beyond a dichotomous set of test results to more than 2 intervals, use of positive and negative likelihood ratios no longer has any meaning, and likelihood ratios must instead be identified as interval likelihood ratios defined by the upper and lower limits of the range of values within which they fall.

If we again use Table 2 in Brown and Reeves¹ article to examine the association between leukocyte count and appendicitis, the interval likelihood ratios are displayed in 3 strata: A patient with a WBC count of $15 \times 10^3/\mu\text{L}$ or greater has a likelihood ratio of 7, which increases the odds of appendicitis by about sevenfold from its pretest status; a patient with a WBC count between $8 \times 10^3/\mu\text{L}$ and $15 \times 10^3/\mu\text{L}$ has a likelihood ratio of only 1.3, which is close enough to 1 that it will leave the pre- and posttest odds unaltered; and a patient with

a WBC count less than $8 \times 10^3/\mu\text{L}$ has a likelihood ratio of 0.2, which reduces the odds of appendicitis by about fivefold from its pretest status.

If, however, one combines the bottom 2 contiguous strata so that the table is collapsed into a 2×2 configuration with only 2 intervals, one row will now contain only patients with a WBC count of $15 \times 10^3/\mu\text{L}$ or greater, which can be classified, for purposes of this example, as a “positive” test for appendicitis. Similarly, the other row will now contain only patients with WBC counts less than $15 \times 10^3/\mu\text{L}$, which can be considered a “negative” test for appendicitis. The likelihood ratio for a positive test or WBC count of $15 \times 10^3/\mu\text{L}$ or greater retains its interval likelihood ratio of 7, which equals its positive likelihood ratio. This is because the contents of the interval are unchanged and are different only in that the interval has been relabeled as positive. However, the new likelihood ratio for the 2 combined lower strata, which now contains all patients with a WBC count less than $15 \times 10^3/\mu\text{L}$, is reduced to a negative likelihood ratio of 0.7, which will leave the pre- and posttest odds of appendicitis in this group essentially unchanged. As Brown and Reeves¹ note, such an aggregation of test results into categories of positive and negative, without taking into account the degree of positivity or negativity contained in interval likelihood ratios, substantially reduces the amount of information one can extract from the results of any diagnostic test.

ADVANTAGES OF LIKELIHOOD RATIOS

For the following reasons, likelihood ratios are a far more clinically useful approach to diagnostic testing than are sensitivity, specificity, predictive values, or ROC curves.

1. Unlike sensitivity and specificity, but similar to predictive values, likelihood ratios tell us what we wish to know and provide the information in an intuitive and straightforward format that runs parallel to the direction of diagnostic thinking: Given a particular test result, how likely is it that the patient has the disease?

2. Unlike predictive values, but similar to sensitivity and specificity, likelihood ratios are numerically stable and do not vary as a function of disease prevalence.

3. Likelihood ratios also incorporate the tradeoff between sensitivity and specificity seen in the ROC curve by using the 2 axes of the curvilinear plot (true positive rate versus the false positive rate) expressed in the form of a ratio to calculate interval likelihood ratios, as illustrated by Brown and Reeves.¹

4. Furthermore, as noted earlier, likelihood ratios are indifferent to the format in which the test results are expressed, whether as dichotomous, ordinal, or continuous data.

5. Finally, likelihood ratios plug directly into a simple formulation of Bayes’ theorem: pretest odds of disease \times likelihood ratio = posttest odds of disease.

In summary, as Brown and Reeves¹ have demonstrated in their excellent article, stratification by interval likelihood ratios can unearth a great deal of information buried in diagnostic tests that would otherwise be lost due to lumping. As shown here, positive and negative likelihood ratios are simply a unique case of interval likelihood ratios that occur when test results are reported with only 2 intervals (ie, dichotomously). Whether expressed as interval likelihood ratios or as positive or negative likelihood ratios, likelihood ratios are more intuitive than sensitivity and specificity, more numerically stable than predictive values, and more clinically applicable than ROC curves. Finally, the ability of a likelihood ratio to plug simply and directly into Bayes’ theorem illustrates the concordance between the mathematic properties of likelihood ratios and the central clinical strategy driving all diagnostic testing: the revision of disease probability.

Received for publication February 27, 2003. Revision received March 13, 2003. Accepted for publication March 18, 2003.

Reprints not available from the author.

Address for correspondence: E. John Gallagher, MD, Department of Emergency Medicine, Albert Einstein College of Medicine, Bronx, NY 10467; 718-920-7459, fax 718-798-0730; E-mail jgallagher@montefiore.org.

REFERENCES

1. Brown MD, Reeves MJ. Interval likelihood ratios: another advantage for the evidence-based diagnostician. *Ann Emerg Med.* 2003;42:292-297.
2. Sackett DL, Haynes RB, Tugwell P. *Clinical Epidemiology: A Basic Science for Clinical Medicine.* Toronto, Ontario, Canada: Little, Brown; 1985:123.
3. Green D, Swets J. *Signal Detection Theory and Psychophysics.* New York, NY: John Wiley and Sons; 1966:45-49.
4. Tandberg D, Deely JJ, O'Malley AJ. Generalized likelihood ratios for quantitative diagnostic test scores. *Am J Emerg Med.* 1997;15:694-699.
5. Harrison A, Morrison KL, Krishnaswamy P, et al. B-type natriuretic peptide predicts future cardiac events in patients presenting to the emergency department with dyspnea. *Ann Emerg Med.* 2002;39:131-138.
6. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology.* 1982;143:29-36.



IMPORTANT NOTICE TO CURRENT AND FORMER ABEM DIPLOMATES

- Emergency Medicine Continuous Certification (EMCC) will begin in 2004.
- All diplomates who want to maintain their certification with ABEM beyond the current expiration date must participate fully in the EMCC program.
- Effective 2004, the licensure requirement for all diplomates will change. Diplomates will be required to continuously maintain a current, active, valid, unrestricted, and unqualified license in at least one jurisdiction in the United States, its territories, or Canada, and in each jurisdiction in which they practice. Inactive medical licenses voluntarily held by physicians are in compliance with the *Policy on Medical Licensure*.
- Physicians eligible for ABEM recertification under current rules will maintain eligibility under EMCC. The written recertification examination as it currently exists will be offered for the last time on November 2, 2003.
- A special option will be available only from 2004-2006 for former diplomates to regain their diplomate status through participation in EMCC. Former diplomates must begin their participation in EMCC in 2004 to take advantage of this option.

A full description of EMCC including details of diplomates' participation requirements are available on the ABEM Web site (<http://www.abem.org>). Questions should be directed to:

American Board of Emergency Medicine
3000 Coolidge Road
East Lansing, MI 48823
Phone: 517-332-4800
E-mail: emcc@abem.org