

# Clinical Utility of Likelihood Ratios

From the Departments of Emergency Medicine, Medicine, Epidemiology, and Social Medicine, Albert Einstein College of Medicine, Bronx, NY.

Received for publication April 18, 1997. Revisions received July 21 and August 25, 1997. Accepted for publication September 8, 1997.

Copyright © 1998 by the American College of Emergency Physicians.

**E John Gallagher, MD**

Test-performance characteristics can be derived from a simple 2×2 table displaying the dichotomous relationship between a positive or negative test result and the presence or absence of a target disorder. Sensitivity and specificity, including a summary display of their reciprocal relationship as a receiver operating characteristics curve, are relatively stable test characteristics. Unfortunately, they represent an inversion of customary clinical logic and fail to tell us precisely what we wish to know. Predictive values, on the other hand, provide us with the requisite information but—because they are vulnerable to variation in disease prevalence—are numerically unstable. Likelihood ratios (LRs), in contrast, combine the stability of sensitivity and specificity to provide an omnibus index of test performance far more useful than its constituent parts. Application of Bayes' theorem to LRs produces the following summary equation: Clinically estimated pretest odds of disease×LR=Posttest odds of disease. This simple equation illustrates a concordance between the mathematical properties of likelihood ratios and the central strategy underlying diagnostic testing: the revision of disease probability.

[Gallagher EJ: Clinical utility of likelihood ratios. *Ann Emerg Med* March 1998;31:391-397.]

## INTRODUCTION

The performance characteristics of diagnostic tests are reported in a sometimes-bewildering variety of configurations. Because both the target disorder (present/absent) and the test results (positive/negative) are often expressed dichotomously, the relationship between the two can be summarized in the form of the familiar 2×2 table.

Although a 2×2 table may be constructed in any number of ways, Table 1 represents the conventional format, with disease status and test results expressed as column and row headers, respectively. The location of each of the four cells in relation to the table and to one another is also standard.

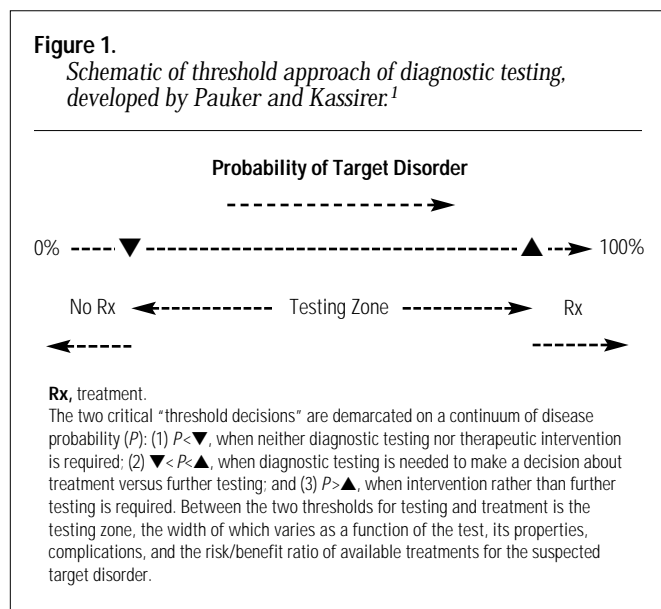
THE THRESHOLD APPROACH TO DIAGNOSTIC TESTING

The clinician’s educated “guess” about the prior probability that a particular patient has a given target disorder plays a crucial role, not only in the interpretation of test results but in the more fundamental decision to perform any diagnostic tests at all. The threshold approach, described more than 15 years ago by Pauker and Kassirer,<sup>1</sup> is a useful heuristic that places the clinician’s “index of suspicion” on a continuum of probability from 0 to 100%. At some point nearer the upper end of the scale, the probability of disease becomes high enough to cross the “treatment threshold” (▲ in Figure 1), and therapeutic intervention is undertaken without further testing. At some corresponding point toward the lower end of the scale, the probability of a particular target disorder becomes low enough that it drops beneath the “testing threshold” (▼ in Figure 1). At this juncture, the physician decides that the working diagnosis has been reasonably excluded from consideration, and no further testing is necessary.

The width of the testing zone shown in Figure 1 is a function of three variables: (1) test properties, (2) risk of excess morbidity/mortality attributable to the test relative to that of the target disorder, and (3) risk/benefit ratio of available therapies for the working diagnosis.<sup>1</sup>

CALCULATION OF CONVENTIONAL TEST PROPERTIES

Traditionally, markers of test performance have been divided into two groups, depending on whether one focuses on the rows or columns of the standard 2×2 table (Table 1).



Rows

Working across the rows yields predictive values, which vary markedly as a function of disease prevalence, or prior probability of disease. The positive predictive value (PPV) of a test is expressed as  $(TP/[TP+FP])$ ; the negative predictive value  $(NPV)=(TN/[TN+FN])$ . The overall prevalence of disease in the population displayed in any given table can be expressed simply as  $([TP+FN]/[TP+FP+FN+TN])$ .

Columns

Working down the columns provides estimates of sensitivity— $(TP/[TP+FN])$ —and specificity— $(TN/[TN+FP])$ —which are true “test properties” in the sense that they are relatively—though not absolutely—dependent of variation in disease prevalence.<sup>2</sup> Sensitivity is also referred to as the true-positive rate (TPR) because it represents the proportion of patients with the disease who have a positive test result. Similarly, specificity represents the proportion of patients without the disease who have a negative result and is therefore known as the true-negative rate (TNR). Both sensitivity and specificity have complements— $(1-\text{sensitivity})$  and  $(1-\text{specificity})$ . These represent the false-negative rate, or FNR  $(FN/[FN+TP])$  and the false-positive rate, or FPR  $(FP/[FP+TN])$ , respectively. Thus the FNR is the complement of the TPR or sensitivity, and the FPR is the complement of the TNR or specificity.

SOME PROBLEMS WITH CLINICAL APPLICATION OF TRADITIONAL TEST PROPERTIES

Predictive values

The vulnerability of predictive values to shifts in disease prevalence severely limits their usefulness. As disease prevalence increases in a population, the PPV increases, and, reciprocally, the NPV decreases. Similarly, as prevalence decreases, the reverse occurs: NPV increases and PPV diminishes.

Because of this variation, one must adjust predictive values for prior probability by examining the difference between the two for the “predictive increment”—that is, the difference between pretest probability (prevalence) and posttest

**Table 1.**  
Standard table for derivation of test-performance characteristics.

Test Result	Disease Present	Disease Absent	Row Totals
Positive	TP	FP	TP+FP
Negative	FN	TN	FN+TN
Column totals	TP+FN	FP+TN	TP+FP+FN+TN

probability of disease presence or absence. For example, the prevalence of coronary heart disease is so low in women younger than 30 years (say, 1%) that a negative ECG stress test does not have much opportunity to improve on prior probability. Thus the NPV of a stress test in this population will be at least 99%, but the negative predictive increment cannot exceed 1%, making the test of very little clinical value in this setting.

**Sensitivity/specificity**

The instability of predictive values is one of the principal reasons that test properties are commonly expressed as sensitivity and specificity. Although these performance characteristics have the virtue of remaining relatively constant among different patient populations, what they actually tell us is how likely a test result is to be positive or negative, given that a patient does or does not have the target disorder for which one is testing. There is a paradoxical inversion of customary clinical logic intrinsic to this definition<sup>3</sup> because knowledge of whether the patient had the illness would clearly obviate the need for a diagnostic test in the first place.

On the other hand predictive values tell us what we really want to know clinically: Given that a test result is positive or negative, what is the probability that the patient does or does not have the illness for which one is testing? Unfortunately, as noted above, these values do not travel well from one patient population to another.

**ROC curves**

The problem with using a single 2x2 table to generate sensitivity and specificity is that the results are vulnerable to the "single cutoff trap."<sup>4</sup> Superimposition of a dichotomous cutoff point onto a continuous distribution of data demotes quantitative information to a relatively qualitative status. By essentially ignoring degrees of positivity or negativity of test results, the information contained in the gray-scale spectrum of diagnostic testing is lost. Recognition of this problem led to the notion of adjusting the threshold for test positivity up or down, depending on the clinical circumstances. This yielded a series of tables with different sensitivity/specificity pairs, varying as a function of each table's threshold for declaring a test positive. On a graph of sensitivity versus specificity, each threshold would define a unique combination of sensitivity and specificity, representing a single point on a curvilinear plot.

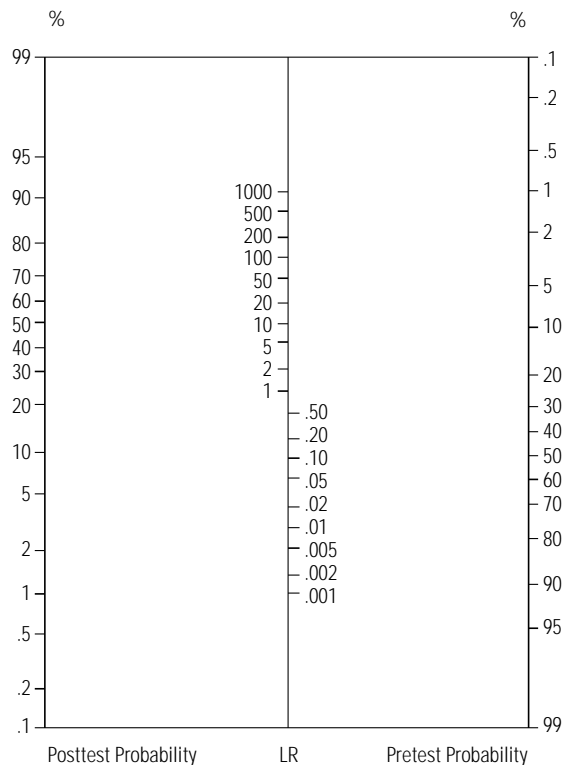
ROC curves also readily lend themselves to display of the performance characteristics of test results reported as continuous data. By convention, the ROC curve displays sensitivity (the TPR) on the vertical axis against the complement of specificity (1-specificity or the FPR) on the horizontal axis. The ROC curve then demonstrates the characteristic

reciprocal relationship between sensitivity and specificity, expressed as a tradeoff between the TPR and FPR. This configuration of the curve also facilitates calculation of the area beneath it as a summary index of overall test performance.

The acronym ROC stands for receiver operating characteristic and was initially used to measure skills among radar operators by determining each individual's capacity to distinguish signal from noise on their screens. An alternative way to think about ROC curves is to regard the TPR on the vertical axis (sensitivity) as the "signal" and the FPR on the horizontal axis (1-specificity) as the "noise."

Although ROC curves went a long way toward avoiding the single-cutoff trap and giving clinicians a better feel for

**Figure 2.** Nonogram for use of LRs to convert pretest probabilities (right vertical axis) into posttest probabilities (left vertical axis).



Extension of a straight line, connecting the pretest probability found on the right vertical axis with the LR for a diagnostic test, determines the posttest probability, as displayed on the left vertical axis. (Adapted with permission from Fagan TJ: Nomogram for Bayes' theorem [letter]. *N Engl J Med* vol 293, p 257. Copyright © 1975 by the Massachusetts Medical Society.)

the utility of diagnostic tests, they were limited by the same problems of direct clinical applicability that restricted use of sensitivity and specificity. Ultimately, ROCs have found more use as descriptors of overall test performance through calculation of the area under the curve than in the interpretation of diagnostic tests in individual patients at the bedside.

## LIKELIHOOD RATIOS

A better approach to diagnostic testing than use of predictive values, sensitivity/specificity, or ROC curves is to use likelihood ratios (LRs). LRs are defined as the likelihood that a particular test result would be found in a patient with the target disorder, relative to the likelihood of that same test result occurring in a patient without the target disorder. LRs combine the conventional test properties of sensitivity and specificity into a summary index that is far more useful clinically than its constituent parts. Once determined, an LR can be incorporated directly into the calculation of posttest probability by employing the following formulation of Bayes' theorem:  $LR \times \text{Pretest odds of disease} = \text{Posttest odds of disease}$ .

### Odds and probabilities

Although odds and probabilities are mathematically different, they are conceptually similar, and are related as follows:

Odds of a disease = Probability of a disease / (1 - Probability of a disease), or  $O = P / (1 - P)$ . Conversely,  $P = O / (O + 1)$ . Thus if the odds are 1:1 (an "even bet"),  $P = 1 / (1 + 1) = 1/2 = .5$  or 50% probability of winning; conversely, if probability is .5 or 50%, odds =  $.5 / (1 - .5) = .5 / .5 = 1:1$ .

When used in betting, odds are traditionally expressed as a ratio of two whole numbers separated by a colon. However, for purposes of diagnostic testing, odds are easier to use if they are expressed simply as whole numbers or decimals, without the use of the colon. Because most clinicians think in terms of prior probabilities expressed as percentages, odds will be derived from, and converted back to, probabilities for decisionmaking. At no point in this process is it necessary to express or think about odds using the colon.

Because the quantitative use of LR requires a fair amount of arithmetic, there are two practical choices: (1) Make a rough approximation by using the rule of thumb that pretest probability divided by one minus itself ( $1 - P$ ) calculates pretest odds, then, after multiplying by the LR, the posttest odds divided by itself plus one ( $1 + O$ ) converts back to posttest probability; or (2) use the nomogram shown in Figure 2. By extending a straightedge from any pretest probability

through any LR, the point of intersection of the line with the posttest probability axis defines the value of the corresponding posterior probability.<sup>5</sup> By treating the LR as an axis, one can obtain a sense of the variation of the posttest probability as a function of shifts in pretest probability. Similarly, by treating the pretest probability as an axis, one can observe the variation in posttest probability as a function of changes in the LR.

### Positive and negative LRs

LRs are often divided into positive and negative LRs, expressed as follows:

LR of a positive test =  $TPR / FPR = \text{sensitivity} / (1 - \text{specificity})$ .

LR of a negative test =  $FNR / TNR = (1 - \text{sensitivity}) / \text{specificity}$ .

The formal definition of a positive LR would simply be a special case of the general definition of LRs given above: A positive LR is defined as the likelihood that a positive test result would be found in a patient with the target disorder, compared with the likelihood of a positive test result occurring in a patient without the target disorder. The definition of a negative LR is the likelihood that a negative test result would be found in a patient with the target disorder, compared with the likelihood of a negative test result occurring in a patient without the target disorder.

Because LRs can be defined for multiple threshold values of a given test result, dividing them into positive and negative LRs is not mathematically or clinically essential. However, it is a common convention congruent with clinical logic in the sense that one is usually driving toward either confirming ("ruling in") or excluding ("ruling out") a particular diagnosis. Under these conditions, the positive and negative signs aid in the choice of a diagnostic test whose properties are oriented in the direction in which the clinician is headed.

Similar to other commonly used ratios in clinical epidemiology, such as the odds ratio (OR) or the risk ratio (RR), likelihood ratios range from 0 to  $\infty$ , with a null point at 1 (representing "even odds," or a valueless test). LRs, like ORs and RRs, do not drop below 0. Consequently a negative LR still refers to a positive number. As the positive LR becomes larger, and the negative LR becomes smaller, both increase their power to revise pretest probability in proportion to the extent to which they have diverged from the null point of 1. In examining Figure 2, as is characteristic of an odds scale, a positive LR of 10 is just as far from the null point—and therefore just as clinically significant—as its reciprocal, represented by a negative LR of .10.

The tradeoff between the TPR and FPR contained in the positive LR is also reflected in the ROC curves mentioned above. The numerator (TPR or sensitivity) and denominator (FPR or  $[1 - \text{specificity}]$ ) of the positive LR represent the

vertical and horizontal axes of the ROC curve, respectively. Although the positive LR and ROC curves are closely related, they serve different functions. As noted above, ROC curves permit comparison of diagnostic tests over the entire range of their performance; LRs facilitate application of diagnostic test results to revision of disease probability in individual patients.

#### Interpretation of LRs

In general, a positive LR of 1 to 2, or a negative LR of .5 to 1, alters disease probability by a small and clinically insignificant degree. In contrast, positive LRs greater than 10, or negative LRs less than .1, may have a very substantial impact on clinical decisionmaking through meaningful revision of disease probability.<sup>6</sup> Positive LRs of 2 to 10, or negative LRs of .5 to .1, reside in between and may occasionally be clinically important. Thus a diagnostic test with a negative LR of .1 is as useful as a diagnostic test with positive LR of 10, depending on whether the clinician is focused on driving the posttest probability above the treatment or below the testing threshold.<sup>1</sup>

#### COMPARISON OF LR<sub>s</sub> WITH OTHER MEASURES OF TEST PERFORMANCE

LRs have unique properties that give them a clinical edge over predictive values, sensitivity/specificity, and ROC curves:

(1) Like predictive values, LRs provide information that can be applied directly to the clinical problem at hand with the use of Bayes' theorem. However, unlike predictive values, or any test property that is "horizontally derived"—that is, calculated across the rows of a table—LRs, which are "vertically derived," possess stability in the face of shifting disease prevalence.

(2) Sensitivity and specificity, the most commonly used measures of test performance, are also vertically derived and therefore have the advantage noted above of relative stability in the face of alterations in disease prevalence. However, they may become unstable if disease severity in the study population shifts concurrently with a change in prevalence.<sup>2</sup> LRs appear to be less prone to instability under similar circumstances, in part because they incorporate both sensitivity and specificity into their vertical derivation.

An additional, and more important, advantage that LRs have over sensitivity and specificity is their direct applicability to a clinical problem. In this respect, LRs are similar to predictive values, but without the vulnerability to variation in disease prevalence. In tests with either high sensitivity or specificity, one can qualitatively apply the test result to an estimate of prior probability and obtain useful clinical information. For example, if a test has a 98% sensitivity,

application of a negative test result to a patient with a low (10%) prior probability of disease will generally drop that diagnosis below the threshold of further consideration. However, when one begins to work with tests that have more commonly encountered performance characteristics, such as 90% specificity and 85% sensitivity, intuitive translation of the information derived from a positive or negative test becomes considerably more difficult. It is under these circumstances that a quantitative analysis is more helpful, and LRs, unlike sensitivity and specificity, can provide that.

(3) As noted earlier, ROCs were devised to offset the problem of lost information that occurs when test results expressed as continuous data are aggregated into dichotomous categories. Although ROCs are traditionally expressed in a graphical format, the varying combinations of sensitivity and specificity occurring at different thresholds of test positivity could easily be stratified and calculated. However, stratification of sensitivity and specificity, which is commonly done with LRs, does not deal with the fundamental difficulty that occurs when attempting to apply a test with only modest sensitivity or specificity directly to the estimated prior probability of disease in a given patient. This will remain a problem that only worsens as sensitivity and specificity drift away from near-perfect performance characteristics toward more realistic ones. This in fact represents an intrinsic limitation of sensitivity and specificity, whether the test results are stratified, expressed in ROC format, or expressed in a traditional dichotomous configuration.

(4) In addition to all of the above, LRs have the happy mathematical property of immediate and quantitative clinical utility through direct application of Bayes' theorem: Estimated pretest odds of disease  $\times$  LR = Posttest odds of disease. One can then treat the newly calculated posttest odds as new pretest odds and apply another LR from an additional "independent" diagnostic test through simple multiplication. This will generate a more refined posttest odds of the presence of the target disorder. Theoretically, the diagnostic tests must be independent to avoid overestimating posttest odds of disease. However, Sackett et al<sup>2</sup> maintain that in practice the phenomenon of convergence resulting from treating conditional probabilities as if they were independent of one another is not of great consequence, provided that one uses short chains of two or three diagnostic tests. Longer sequences may result in significant distortions of disease likelihood.<sup>2</sup>

Given that many diagnostic tests aimed at the same target disorder show a high degree of interrelation or collinearity, it seems prudent to regard the practice of "chaining" LRs with some skepticism. One logical, though untested, approach might be to chain two tests if their mechanism

of disease identification arises from measurement of different manifestations of the same pathologic process—for instance, measurement of myocardial ischemic damage using the ECG and enzymatic markers, or detection of pulmonary embolus, using an imaging modality (such as lung scan or magnetic resonance pulmonary angiography) and a measure of fresh clotting, such as the D-dimer assay.

(5) Finally, LRs may be applied not only to categorical but to continuous data through stratification. The LRs that emerge from aggregating continuous test results into contiguous strata at different thresholds are known as “interval likelihood ratios.” These can be very powerful in deriving a quantitative array of posttest odds of disease that vary with the LR of a given interval within which a particular test result falls.

A PRACTICAL EXAMPLE

Using sensitivities and specificities reported on several non-invasive diagnostic tests for thoracic aortic dissection,<sup>7</sup> we can derive the LRs shown in Table 2. Next we can focus on the direct application of LRs as stable test properties of each of the above imaging modalities, under assumed conditions of high (90%), intermediate (50%), and low (10%) pretest probability. Converting these three prior probabilities to pretest odds yields  $(.9/[1-.9])=9$ ,  $(.5/[1-.5])=1$ , and  $(.1/[1-.1])=.11$ , respectively. Because of an extremely high positive LR of 49 for the magnetic resonance imaging (MRI), the consequences of positive MRI in each of these settings yields posttest odds of about  $(9 \times 49)=441$ ,  $(1 \times 49)=49$ , and  $(.11 \times 49)=5.4$ ; converting these back to posttest probabilities results in  $(441/[1+441])= >99\%$ ,  $(49/[1+49])=98\%$ , and  $(5.4/[1+5.4])=84\%$ , respectively. The first two of these should be sufficient to drive decisionmaking well over the therapeutic threshold. A test with such an extraordinarily high positive LR is very powerful and can markedly revise disease probabilities, as is evident in this example, where positive MRI raised the absolute probability of aortic dissection by about 10%, 48%, and 74%, respectively, in patients with

high, intermediate, and low prior probabilities of aortic dissection.

Negative MRI under similar conditions of 90%, 50%, and 10% prior probability, given a negative LR of .02, yields posttest odds, respectively, of .18, .02, and less than .01; converting to posttest probability results in 15%, 2%, and less than 1%, thus revising downward the absolute probability of dissection by about 75%, 48%, and 10%, respectively, in patients with high, intermediate, and low prior probabilities of this diagnosis. For all but the patient with a very high prior probability, these results should be sufficient to drive the diagnosis of dissection below the testing threshold and turn the clinician’s attention to other diagnostic possibilities.

In practice, however, MRI is often unavailable or the patient with suspected dissection is too unstable to undergo the procedure. Under these conditions, the clinician may choose to order computed tomography (CT) of the thoracic aorta or transesophageal echocardiography (TEE), both of which have negative LRs comparable to that of MRI (.02 to .03). CT or TEE is therefore capable of revising pretest probability downward—although not upward—to an extent similar to that seen with MRI. Because of the excellent negative LRs, either of these would be a reasonable choice, particularly if one’s clinical sense of the prior probability of a dissection was near or below 50%. As shown in Table 2, the positive LR for CT is markedly lower than that for MRI, and TEE’s positive LR is lower still. This makes a positive test result more difficult to interpret, but it also offers an opportunity to combine the two tests in tandem to increase their aggregate diagnostic power.

Consider an elderly man with long-standing uncontrolled hypertension who complains of mild substernal and interscapular discomfort. No additional information is forthcoming from an otherwise unremarkable history, physical examination, ECG, and chest radiography. Therefore a conservative estimate of this patient’s prior probability of dissection might be on the order of about 20%. TEE is available and is requested because a negative LR of  $.03 \times$  pretest odds of  $(.2/[1-.2])=.25$  will produce, if negative, posttest odds of less than .01 and a posttest probability less than 1%. This should be low enough to reasonably exclude the diagnosis of aortic dissection. However, for purposes of this example, let us suppose that TEE in this patient identifies a probable type A (proximal) dissection. Applying the positive LR of 4 to the same prior probability of 20% (pretest odds, .25), yields posttest odds of 1, or a 50% posttest probability. Although this represents a 30% upward revision of disease probability, the patient’s chances of having a dissection are now about that of obtaining heads on a single coin toss.

**Table 2.**  
*Test-performance characteristics.*

Diagnostic Test	Sensitivity (%)	Specificity (%)	Positive LR	Negative LR
MRI	98	98	49	.02
CT	98	87	8	.02
TEE	98	77	4	.03

All numbers are rounded.

This is neither sufficient basis for moving directly to surgery without a change in the patient's condition nor grounds for excluding dissection as a clinical possibility. One is, in fact, somewhere between the two critical thresholds of testing and treating.<sup>1</sup> At this point, one rational strategy would be to order CT, treat the post-TEE odds as pre-CT odds, and apply the LR's associated with the CT to move the posttest probability in one direction or the other (Table 2). A negative result (negative LR, .02) will bring the posttest probability down to about  $(1 \times .02) = 2\%$ , which should be below the testing threshold and low enough to pursue other diagnostic possibilities.

Alternatively, if CT is positive (positive LR, 8), the posttest odds  $(1 \times 8) = 8$ , and the posttest probability increases to  $(8/[1+8]) = 89\%$ , which should be high enough to either move to the definitive test (MRI or angiography) or drive the clinician across the treatment threshold to the operating room, depending on the patient's clinical status. Although CT and TEE are probably not independent of one another, for serial multiplication of two odds the overestimate of posttest probability—so long as one keeps it in perspective—is probably acceptable in return for the increase in clinically useful information that this tandem testing strategy yields.<sup>2</sup>

Thus the real value of LR's lies in their ability to optimize the power of diagnostic tests to revise disease probability. LR's may be used to do this reliably, across heterogeneous patient populations. The central strategy is to combine the LR directly with the clinician's prior probability of disease (expressed in the form of pretest odds)—essentially applying Bayes' theorem by means of simple multiplication or a nomogram—to yield a posttest odds of disease that drives the probability of the target disorder either below the testing threshold or above the treatment threshold.

Reprint no. 47/1/87765

**Address for reprints:**

E John Gallagher, MD  
Emergency Department  
Albert Einstein College of Medicine  
Montefiore Medical Center  
Bronx, NY 10467  
718-920-7459  
Fax 718-798-6084  
E-mail jgallagh@montefiore.org

REFERENCES

1. Pauker SG, Kassirer JP: The threshold approach to clinical decision-making. *N Engl J Med* 1980;302:1109-1117.
2. Sackett DL, Haynes RB, Tugwell P: *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Toronto: Little, Brown, 1985.
3. Feinstein AR: *Clinical Epidemiology: The Architecture of Clinical Research*. Philadelphia: Saunders, 1985:419.
4. Young MJ, Fried LS, Eisenberg JM, et al: The single-cutoff trap: Implications for Bayesian analysis of stress electrocardiograms. *Med Decis Making* 1989;9:176-180.
5. Fagan TJ: Nomogram for Bayes' theorem [letter]. *N Engl J Med* 1975;293:257.
6. Jaeschke R, Guyatt GH, Sackett DL for the Evidence-Based Working Group. Users' guide to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and how will they help me in caring for my patients? *JAMA* 1994;271:703-707.
7. Nienaber CA, von Kodolitsch Y, Nicolas V, et al: The diagnosis of thoracic aortic dissection by noninvasive imaging procedures. *N Engl J Med* 1993;328:1-9.