Curtis P. Langlotz, MD, PhD

Index terms:
Diagnostic radiology
Statistical analysis

Published online
10.1148/radiol.2281011106
Radiology 2003; 228:3–9

Abbreviation:
ROC = receiver operating characteristic

Fundamental Measures of Diagnostic Examination Performance: Usefulness for Clinical Decision Making and Research[1]

Measures of diagnostic accuracy, such as sensitivity, specificity, predictive values, and receiver operating characteristic curves, can often seem like abstract mathematic concepts that have a minimal relationship with clinical decision making or clinical research. The purpose of this article is to provide definitions and examples of these concepts that illustrate their usefulness in specific clinical decision-making tasks. In particular, nine principles are provided to guide the use of these concepts in daily radiology practice, in interpreting clinical literature, and in designing clinical research studies. An understanding of these principles and of the measures of diagnostic accuracy to which they apply is vital to the appropriate evaluation and use of diagnostic imaging examinations.

The bulk of the radiology literature concerns the assessment of examination performance, which is sometimes referred to as diagnostic accuracy. Despite the proliferation of such research on examination performance, it is still difficult to assess new imaging technologies, in part because such initial assessments are not always performed with an eye for how the results will be used clinically (1). The goal of this article is to describe nine fundamental principles (Appendix) to help answer specific clinical questions by using the radiology literature.

Consider the following clinical scenario: A referring physician calls you about the findings of a diagnostic mammogram that you interpreted yesterday. In the upper outer quadrant of the left breast you identified a cluster of suspicious microcalcifications—not the kind that suggests definite cancer but rather that which indicates the need for a more definitive work-up. The referring physician relays to you the patient's desire to explore the possibility of breast magnetic resonance (MR) imaging.

In this article, I will use this clinical example to illustrate the basic concepts of examination performance. To supplement previously published introductory material (2–4), I will relate the nine fundamental principles to the specific clinical scenario just described to illustrate the strengths and weaknesses of using them for clinical decision making and clinical research. I plan to answer the following questions in the course of this discussion: Which descriptors of an examination are the best intrinsic measures of performance? Which are the most clinically important? What are the limitations of sensitivity and specificity in the assessment of diagnostic examinations? What are receiver operating characteristic (ROC) curves, and why is their clinical usefulness limited? Why are predictive values more clinically relevant, and what are the pitfalls associated with using them? The ability of radiologists to understand the answers to these questions is critical to improving the application of the radiology literature to clinical practice.

TWO-BY-TWO CONTINGENCY TABLE: A SIMPLE AND UNIVERSAL TOOL

One of the most intuitive methods for the analysis of diagnostic examinations is the two-by-two table. This simple device can be jotted on the back of an envelope yet is quite

[1] From the Departments of Radiology and Epidemiology, University of Pennsylvania, 3400 Spruce St, Philadelphia, PA 19104. Received June 25, 2001; revision requested August 2; revision received May 3, 2002; accepted May 15. **Address correspondence to** the author (e-mail: *langlotz@rad.upenn.edu*).

versatile and powerful, both in the analysis of a diagnostic examination and in increasing our understanding of examination performance.

Simplifying Assumptions

The use of two-by-two tables (a more general term is *contingency tables*) requires certain simplifying assumptions and prerequisites. The first assumption that I will make is that the examination in question must be compared with a reference-standard examination—that is, one with results that yield the truth with regard to the presence or absence of the disease. In the past, this reference standard has commonly been called a "gold standard"—a term that is falling out of favor, perhaps because of the recognition that even some of the best reference standards are imperfect. For example, even clinical diagnoses supplemented by the results of the most effective histopathologic analyses are fallible (5).

A second major assumption that I will make is that the examination result must be considered either positive or negative. This is perhaps the least appealing assumption with regard to a two-by-two table, because many examinations have continuous result values, such as the degree of stenosis in a vessel or the attenuation of a liver lesion. As we will see later, this is one of the first assumptions that I will discard when advanced concepts such as ROC curves are discussed (6–8). The final assumption is that one assesses examination performance with respect to the presence or absence of a single disease and not several diseases.

Example Use of a Two-by-Two Table

Table 1 is a prototypical form of a two-by-two table. Across the top, we see two center columns, one for all cases (or patients) in which the disease is truly present (D+) and the other for all cases in which the disease is truly absent (D−). In the far left column of the table, we see the two possible examination results: positive, indicating disease presence, and negative, indicating disease absence. This table summarizes the relationship between the examination result and the reference-standard examination and defines four distinct table cells (ie, true-positive, false-positive, true-negative, and false-negative examination results). In the first row (T+), we see that a positive examination result can be either true-positive or false-positive, depending on whether the disease is present or absent, respectively. The second row (T−) shows that a negative examination result can be either false-negative or true-negative, again depending on whether the disease is present or absent, respectively.

Data in the D+ column show how the examination performs (ie, yields results that indicate the true presence or true absence of a given disease) in patients who have the disease in question. Data in the D− column show how the examination performs in patients who do not have the disease (ie, who are "healthy" with respect to the disease in question). The total numbers of patients who actually do and do not have the disease according to the reference-standard examination results are listed at the bottom of the D+ and D− columns, respectively.

The datum in the first row at the far right (TP + FP) is the total number of patients who have positive examination results; the datum in the second row at the far right (FN + TN) is the total number of patients who have negative examination results. The overall total *(N)* is the total number of patients who participated in the study of examination performance.

The example data in Table 2 are interim data from an experiment to evaluate the accuracy of breast MR imaging in patients with clinically or mammographically suspicious lesions. Like the patient with suspicious microcalcifications who is considering undergoing MR imaging in the hypothetical scenario described earlier, all patients in this experiment had suspicious lesions and were about to undergo open excisional biopsy. Prior to biopsy, each woman underwent dynamic contrast material–enhanced MR imaging of the breast. The results of histopathologic examination of the specimen obtained at subsequent excisional biopsy were used as the reference standard for disease. (A more detailed description of the experimental methodology and a more recent report of the data are published elsewhere [9].) As shown in Table 2, a total of 182 women were enrolled in the study at the time the table was constructed. Seventy-four of these women had cancer, and 108 did not. There were a total of 99 positive examination results: 71 were true-positive and 28 false- positive. The 83 negative examination results comprised three false-negative and 80 true-negative results.

In the following sections, I describe the important quantitative measures of examination performance that can be computed from a two-by-two table.

TABLE 1
Shorthand Two-by-Two Table Describing Diagnostic Examination Performance

Examination Result	D+	D−	Total
T+	TP	FP	TP + FP
T−	FN	TN	FN + TN
Total	TP + FN	FP + TN	*N*

Note.—D+ = all cases or patients in which disease is truly present (ie, according to reference-standard examination results), D− = all cases or patients in which disease is truly absent, FN = number of cases or patients with false-negative examination results, FP = number of cases or patients with false-positive examination results, *N* = overall total number of cases or patients, T+ = positive examination result, T− = negative examination result, TN = number of cases or patients with true-negative examination results, TP = number of cases or patients with true-positive examination results.

TABLE 2
Patient Data in Experiment to Study Breast MR Imaging

MR Imaging Result	Malignant	Benign	Totals
Positive	71	28	99
Negative	3	80	83
Total	74	108	182

Note.—Data are numbers of women with malignant or benign breast tumors.

SENSITIVITY AND SPECIFICITY: INTRINSIC MEASURES OF EXAMINATION PERFORMANCE

Sensitivity: Examination Performance in Patients with the Disease in Question

Principle 1: Sensitivity is a measure of how a diagnostic examination performs in a population of patients who have the disease in question. The value can be defined as how often the examination will enable detection of the disease when it is present: TP/(TP + FN).

Given the data in two-by-two Table 2, sensitivity is computed by using the numbers in the "Malignant" column. Of the 74 women who actually had cancer, 71 had a positive MR imaging result. Thus, the sensitivity of breast MR imaging in the sample of women undergoing breast biopsy was 96%—that is, 71 of 74 women with cancer were identified by using MR imaging.

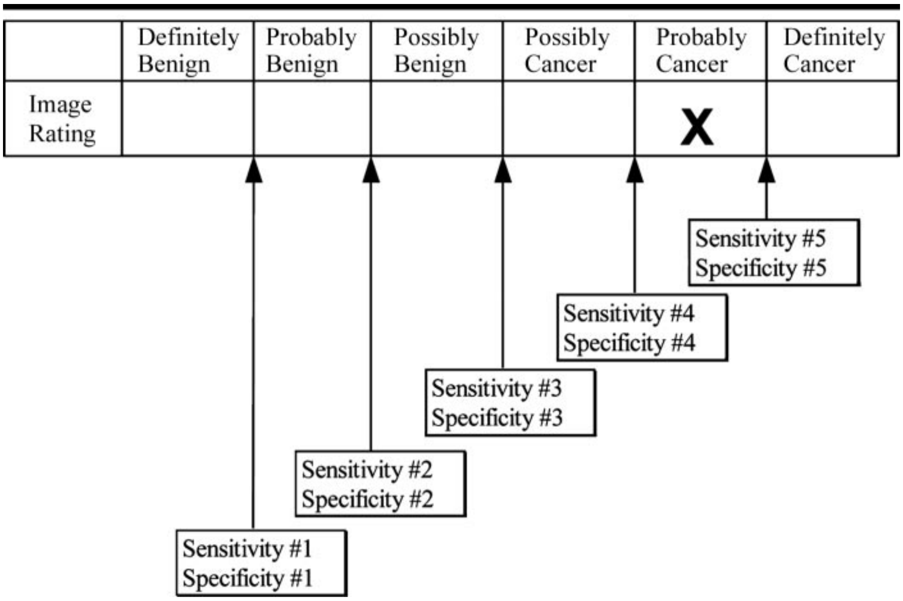


Figure 1. Six-category scale for rating the presence or absence of breast cancer. By varying a cutoff for rating categories, one can create five two-by-two tables of data from which sensitivity and specificity values can be calculated (sensitivity and specificity *#1–#5*).

Specificity: Examination Performance in Patients without the Disease in Question

Principle 2: Specificity is a measure of how a diagnostic examination performs in a population of patients who do not have the disease (ie, healthy subjects)—in other words, a value of the ability of an examination to yield an appropriately negative result in these patients. Specificity can be defined as how often a healthy patient will have a normal examination result: TN/(FP + TN).

In the example scenario, specificity is calculated by using the numbers in the "Benign" column of two-by-two Table 2. Of the 108 women who had benign lesions, 80 had negative MR imaging results. Thus, the specificity of breast MR imaging in the sample of women undergoing breast biopsy was 74%—that is, 80 of 108 women without cancer were identified by using MR imaging.

Relative Importance of Sensitivity and Specificity

How can sensitivity and specificity values be used directly to determine whether an examination might be useful in a specific clinical situation? Which value is more important? A quantitative analysis of these questions (4) is beyond the scope of this article. However, here are two qualitative rules of thumb, which together make up principle 3: A sensitive examination is more valuable in situations where false-negative results are more undesirable than false-positive results. A specific examination is more valuable in situations where false-positive results are more undesirable than false-negative results.

For example, with regard to a woman with a suspicious breast mass, we must consider how we would feel if we were to miss a cancer owing to a false-negative examination. Because we would regret this outcome, we place appropriate emphasis on developing and enhancing the sensitivity of breast MR imaging to avoid missing cancers that may progress during the follow-up interval after a false-negative MR imaging examination. We would also feel uncomfortable about referring a patient for excisional biopsy of a benign lesion, but perhaps less so, since this result would occur even if MR imaging was never performed. Consequently, this principle leads us to the conclusion that sensitivity is more important than specificity with respect to breast MR imaging in this clinical setting. Because the main potentially beneficial role of breast MR imaging in this clinical setting is to allow some women without cancer to avoid excisional biopsy, this principle also highlights the greater importance of sensitivity compared with specificity in this case. Pauker and Kassirer (10) provide a quantitative discussion of how this principle functions.

Limitations

Sensitivity and specificity are important because they are diagnostic examination descriptors that do not vary greatly among patient populations. A detailed analysis of the limitations of these measures is described elsewhere (11). Let us return to the woman with a suspicious lesion on the mammogram. She wants to know whether breast MR imaging might help her. Now that we have computed the sensitivity and specificity of MR imaging by using the data in the two-by-two table, we can convey to her the following: "If you have cancer, the chance that your MR imaging examination will be positive is 96%. If you don't have cancer, the chance that your MR imaging examination will be negative is 74%." Statements of this kind are often difficult for patients and health care providers to incorporate into their clinical reasoning. Thus, a key weakness of sensitivity and specificity values is that they do not yield information about a diagnostic examination in a form that is immediately relevant to a specific clinical decision-making task. Therefore, while the diagnostic imaging literature may contain a great deal of information about the measured sensitivity and specificity of a given examination, it often contains few data that help us assess the optimal clinical role of the examination (12).

Principle 4: The sensitivity and specificity of a diagnostic examination are related to one another. An additional important weakness of sensitivity and specificity is that these two measures cannot always be used to rank the accuracy of two examinations (or two radiologists). This weakness is particularly evident when one examination has a higher sensitivity but a lower specificity than another. The reason that examination comparison difficulties often arise with regard to sensitivity and specificity is that these two values are inherently related: You cannot evaluate one without the other. As sensitivity increases, specificity tends to decrease, and vice versa. We see this phenomenon every day when two colleagues interpret the same images differently.

Consider, for example, how two radiologists decide whether congestive heart failure is depicted on a chest radiograph. One reader may use strict criteria for the presence of congestive heart failure and thus interpret fewer findings as positive, facilitating decreased sensitivity and increased specificity. The other reader may use much more flexible criteria and thus interpret more image findings as positive for congestive heart failure, facilitating increased sensitivity but decreased specificity.

ROC CURVES

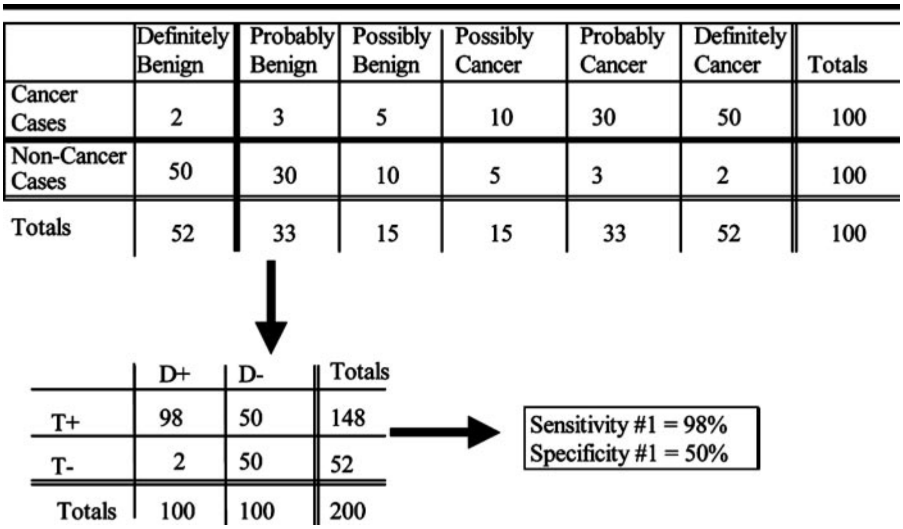
Comprehensive Comparisons among Diagnostic Examinations

Principle 5: ROC curves provide a method to compare diagnostic examination accuracy independently of the diagnostic criteria (ie, strict or flexible) used. When one examination is more sensitive but another is more specific, how do we decide which examination provides better diagnostic information? Or, are the accuracies of the two examinations really similar, with the exception that one involves the use of stricter criteria for a positive result? ROC curves are important and useful because they can answer these questions by explicitly representing the inherent relationship between the sensitivity and specificity of an examination. As such, ROC curves are designed to illustrate the overall information yielded by an imaging examination, regardless of the criteria a reader uses to interpret the images. Therefore, ROC curves specifically address situations in which examinations cannot be compared on the basis of sensitivity and specificity alone. A detailed discussion of ROC methodology is published elsewhere (7).

For the discussion of ROC curves, I will “relax” one of the assumptions made earlier—that of a two-value (ie, positive or negative) examination result. Instead, the readers of the images generated in the two examinations will be allowed to specify their results on a scale. Figure 1 is an illustration of a six-point rating scale for imaging-based identification of breast cancer. When using this scale, the reader of breast images is asked to specify the interpretation in terms of one of six finding categories: definitely cancer, probably cancer, possibly cancer, possibly benign, probably benign, or definitely benign. One then tabulates the ratings by using the two-by-six table shown in Figure 2.

The rating scale and the table produced by using it provide multiple opportunities to measure sensitivity and specificity. For example, we can assume that the examination is positive only when “definitely cancer” is selected and is negative otherwise. Next, we can assume that the probably cancer and definitely cancer ratings both represent positive examination results and that the remaining ratings represent negative results. When the two-by-six table is collapsed in this manner, a new two-by-two table is formed comprising higher sensitivity and lower specificity values than the first two-by-two table (because less strict imaging criteria were used and more image findings were rated as positive).

We can repeat this process five times, concluding with a two-by-two table such as that shown in Figure 2 (bottom table), in which only the definitely benign ratings are considered to represent negative results and the remaining ratings are considered to represent positive results. This approach would result in low specificity and high sensitivity. Figure 3 shows a plot of all five sensitivity-specificity pairs that can be derived from the two-by-six table data in Figure 2 and the ROC curve defined by these points.



	Definitely Benign	Probably Benign	Possibly Benign	Possibly Cancer	Probably Cancer	Definitely Cancer	Totals
Cancer Cases	2	3	5	10	30	50	100
Non-Cancer Cases	50	30	10	5	3	2	100
Totals	52	33	15	15	33	52	100

	D+	D-	Totals
T+	98	50	148
T-	2	50	52
Totals	100	100	200

Sensitivity #1 = 98%
Specificity #1 = 50%

Figure 2. Sample two-by-six table showing the results of an ROC study of breast cancer identification in 200 patients. The two-by-two table at the bottom can be created by setting a cutoff between the ratings of definitely benign and probably benign. This cutoff corresponds to sensitivity *#1* and specificity *#1* in Figure 1. Sensitivity and specificity values are calculated by using the two-by-two table data. As expected, use of the more flexible criteria leads to high sensitivity but low specificity.

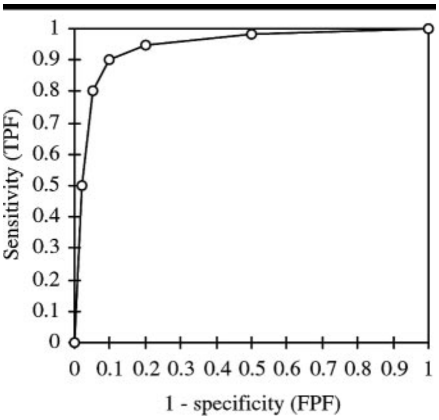


Figure 3. Sample ROC curve. This curve is a plot of sensitivity versus (1 − specificity). The 0,0 point and 1,1 point are included by default to represent the situation in which all images are considered to be either negative or positive, respectively. *FPF* = false-positive fraction, *TPF* = true-positive fraction.

Clinical Limitations

Several methods to quantitatively assess an examination on the basis of the ROC curve that it yields have been designed. The most popular method is to measure the area under the ROC curve, or the A_z. In general, the larger the area under the ROC curve, the better the diagnostic examination. Despite the advantages of these measurements of area under the ROC curve for research and analysis, they do not provide information that is useful to the clinician or patient in clinical decision making. Thus, the value of the area under the ROC curve has no intrinsic clinical meaning. Also, there is no cutoff above or below which one can be certain of the usefulness a diagnostic examination.

Should a patient with an abnormal mammogram be satisfied if she is told that the area under the ROC curve for breast MR imaging is 0.83? Although the area under the ROC curve is helpful for comparing two examinations, it has limited usefulness in facilitating the clinical decisions of patients and referring physicians.

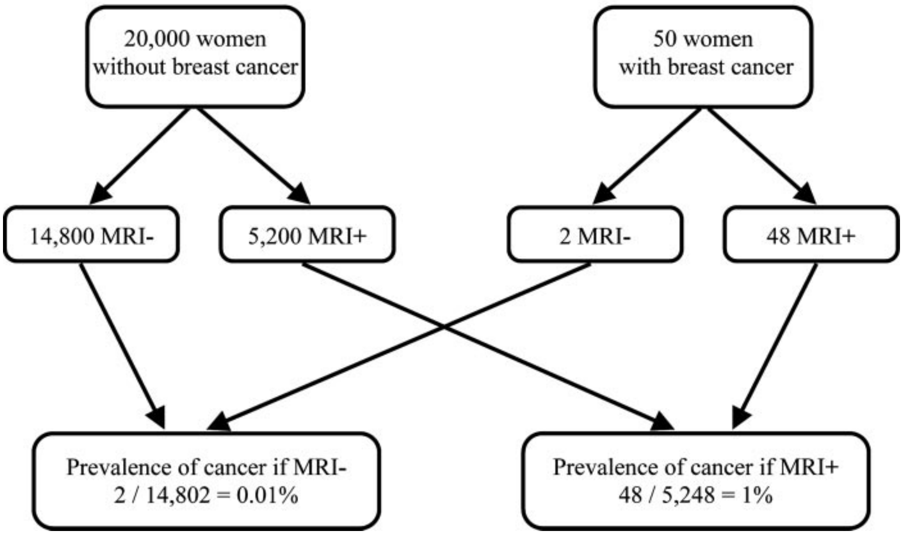

20,000 women without breast cancer
50 women with breast cancer
14,800 MRI-
5,200 MRI+
2 MRI-
48 MRI+
Prevalence of cancer if MRI- 2 / 14,802 = 0.01%
Prevalence of cancer if MRI+ 48 / 5,248 = 1%

Figure 4. Flowchart depicts a simulated population of 20,050 low-risk asymptomatic women who might be screened with breast MR imaging. The assumed prevalence of cancer in this screening population of 50 women with and 20,000 women without cancer is 0.25%. *MRI*+ and *MRI*− = positive and negative MR imaging results, respectively.

PREDICTIVE VALUES

Measuring Postexamination Likelihood of Disease

Because sensitivity, specificity, and ROC curves provide an incomplete picture of the clinical usefulness of an imaging examination, I will now shift back to the original two-by-two table for breast MR imaging (Table 2) to examine two additional measurements that have much greater clinical relevance and intuitive appeal: positive and negative predictive values. These quantities emphasize an important principle of diagnostic examinations—principle 6: A diagnostic examination causes a change in our belief about the likelihood that disease is truly present.

For example, the data in two-by-two Table 2 indicate that the probability of cancer in the women with suspicious mammograms was 41% (74 of 182 women). Thus, simply on the basis of her referral for excisional biopsy (without knowing the specific mammographic appearance of the lesion), the chance of cancer in the woman in the hypothetical scenario is about 41%. How can MR imaging help modify this likelihood to benefit the patient? The predictive values can help answer this question.

Principle 7: The positive predictive value indicates the likelihood of disease given a positive examination. Positive predictive value is defined as the probability of disease in a patient whose examination result is abnormal: TP/(TP + FP). Thus, we can compute the positive predictive value by using only the numbers in the first row ("Positive") of two-by-two Table 2. A total of 99 patients had positive MR imaging results. Seventy-one of these patients actually had cancer. Thus, the positive predictive value of breast MR imaging is 72%—in other words, 71 of the 99 women with positive MR imaging results had cancer. These values are sometimes referred to as "posttest likelihoods" or "posttest probabilities of disease," because the predictive value simply reflects the probability of disease after the examination result is known. The positive predictive value tells us, as expected, that a positive breast MR imaging result increases the probability of disease from 41% to 72%.

Principle 8: The negative predictive value indicates the likelihood of no disease given a negative examination. The negative predictive value is the negative analog of the positive predictive value. The negative predictive value can be defined as the probability that disease is absent in a patient whose examination result is negative: TN/(FN + TN). Thus, the negative predictive value can be computed solely by using the values in the second row ("Negative") of two-by-two Table 2. A total of 83 patients in the sample had negative MR imaging results. Eighty of these patients actually had benign lesions; there were three false-negative results. Thus, the negative predictive value of breast MR imaging was 96%—in other words, 80 of the 83 women with negative MR imaging results did not have cancer. (Note: It is coincidence that the sensitivity and negative predictive value are approximately equivalent in this case.) The probability of disease after a negative examination is simply 100% minus the negative predictive value, or 4% in this case. This computation tells us—as expected—that a negative examination decreases the probability of disease in this case from 41% to 4%.

The clinical usefulness of the predictive values is best illustrated by the first question that the patient in our hypothetical scenario might ask after she has undergone MR imaging: "Do I have cancer?," which in the uncertain world of medicine, can be translated as "How likely is it that I have cancer?" The predictive values, in contrast to sensitivity and specificity, answer this question and therefore are helpful in the clinical decision-making process. Knowing that a negative MR imaging result decreases the chance of cancer to 4% raises several additional questions for the clinician and patient to consider: Is a 4% likelihood low enough that excisional biopsy could be deferred during a short period of follow-up? Is it worth it to trade the potential harm of tumor progression during short-interval follow-up for the potential benefit of not undergoing excisional biopsy? These trade-offs can be considered explicitly by using decision analysis (13) but are routinely considered implicitly by referring physicians, patients, and other medical decision makers.

Limitations

Although predictive values have substantial clinical usefulness, a discussion of their weaknesses is warranted. The most important weakness is the dependence of predictive values on the preexamination probability, or the prevalence of disease in the imaged population. As emphasized earlier, a diagnostic examination causes a change in our belief about the likelihood of disease. And, as expected, the higher the preexamination probability of disease, the higher the postexamination probability of disease. Thus, predictive values are directly dependent on the population in which the given examination is performed.

Consider the following example, which illustrates this dependence: Since breast MR imaging may depict some cancers that were missed with mammography, why not use MR imaging as a screening tool to detect cancer in low-risk asymptomatic women? This approach has some intuitive appeal since breast MR imaging is a highly sensitive examination and might depict a greater number of cancers than would be detected with screening mammography. To illustrate the implications of this approach, a simulation of what would occur if a group of low-risk asymptomatic women were screened with MR imaging is provided (Fig 4). To approximate the prevalence of cancer in a low-risk screening population and to simplify the calculations, I will use 0.25% as the prevalence of breast cancer. For simplicity, I will consider a screening population of 20,050 women, 50 of whom have occult cancer.

Since we have established that MR imaging is 96% sensitive, 48 of the 50 women who actually have cancer will have positive MR imaging examinations (50 × .96 = 48). The other two women will have false-negative MR imaging examinations and undetected cancer, just as if they had never undergone screening MR imaging. With the 74% specificity for breast MR imaging computed earlier, 14,800 women will have normal MR imaging examinations (20,000 × .74 = 14,800). The remaining 5,200 women will have false-positive examinations. Table 3 is a two-by-two table containing these data.

Because the true disease status of these women will be unknown at the time of examination, clinical inferences must be drawn from the examination results and predictive values. According to principle 8, the negative predictive value indicates the clinical implications of a negative examination. Since there are 14,802 women with negative examinations, only two of whom have cancer, the negative predictive value is 0.01% (two of 14,802 women). This value represents a decrease from 0.25% and has no real clinical importance in a screening population; however, it does have some potential reassurance value.

There are 5,248 women with positive examinations in the simulation, and 48 of them actually have cancer, so the likelihood of cancer is approximately 1.00% (48 of 5,248 women). Thus, a positive MR imaging examination increases the likelihood of cancer from 0.25% to 1.00%. This group of women represents a clinical problem, however: Are we willing to perform 100 biopsies to find one cancer? Probably not. Should these women be followed up with a special more intensive regimen? Perhaps, but there are lingering questions regarding the cost-effectiveness of this program, which would likely cost tens of millions of dollars and lead to a substantial increase in the number of negative excisional biopsies.

This example of screening MR imaging illustrates clearly that the predictive values for breast MR imaging are vastly different in a screening population with a much smaller prevalence of disease. Likewise, the values of clinical usefulness of breast MR imaging as a screening examination are in stark contrast to the analogous measures of MR imaging performed in women with mammographically suspicious lesions. This contrast illustrates a weakness of predictive values: They vary according to the population in which the given examination is performed. Although the predictive values for breast MR imaging performed in women with suspicious mammograms are appealing, the predictive values for this examination performed in a screening population suggest that it has little value for asymptomatic women with low breast cancer risk. Another realistic clinical example of this phenomenon is described elsewhere (14).

Despite these limitations, predictive values can be determined mathematically from sensitivity, specificity, and prevalence data. Because sensitivity and specificity values are often published, a clinician can compute the predictive values for a particular population of interest by using the prevalence of disease in that population and the sensitivity and specificity values provided in the literature.

LIKELIHOOD RATIO

Quantifying Changes in Disease Likelihood

Principle 9: Likelihood ratios enable calculation of the postexamination probability of disease from the preexamination probability of disease. A brief description of the likelihood ratio (15,16) is relevant here because this measurement is not affected by disease prevalence and can yield clinically useful information. Likelihood ratio is defined as the probability that a person with a disease will have a particular examination result divided by the probability that a person with no disease will have that same result. Positive likelihood ratio (LR+), sometimes expressed as λ, is defined as the likelihood, or probability, that a person with a disease will have a positive examination divided by the likelihood that a person with no disease will have a positive examination: LR+ = sensitivity/(1 − specificity). Negative likelihood ratio (LR−) is defined as the probability that a person with a disease will have a negative examination divided by the probability that a person without the disease will have a negative examination: LR− = (1 − sensitivity)/specificity.

TABLE 3
Patient Data in a Hypothetical Group of Women Undergoing Screening Breast MR Imaging

MR Imaging Result	Malignant	Benign	Total
Positive	48	5,200	5,248
Negative	2	14,800	14,802
Total	50	20,000	20,050

Note.—Data are numbers of women with malignant or benign breast tumors.

To illustrate the use of likelihood ratios, consider the breast MR imaging example: The sensitivity of breast MR imaging is 96%, and the specificity is 74%. Therefore, the positive likelihood ratio is calculated by dividing the sensitivity by (1 − specificity). Thus, LR+ = 0.96/(1 − 0.74) = 3.7. The negative likelihood ratio is calculated by dividing (1 − sensitivity) by the specificity. Thus, LR− = (1 − 0.96)/0.74 = 0.055.

Once the likelihood ratios have been calculated, they can be used to calculate the postexamination probability of disease given the preexamination probability of disease, or $P(\mathrm{D}+|\mathrm{T}+)$. For this calculation, one first must convert probabilities of disease to odds of disease. Odds of disease, Odds(D+), is defined as the probability that disease is present, $p(\mathrm{D}+)$, divided by the probability that disease is absent, $p(\mathrm{D}-)$: Odds(D+) = $p(\mathrm{D}+)/p(\mathrm{D}-)$.

To compute the postexamination probability of disease given a positive examination result for any preexamination probability value and examination result, multiply the preexamination odds of disease, Odds(D+), by the positive likelihood ratio for the examination result, LR+, to obtain the postexamination odds, Odds(D+|T+). In other words, the postexamination odds of having a given disease given a positive examination result is equal to the positive likelihood ratio multiplied by the preexamination odds of the disease: Odds(D+|T+) = LR+ · Odds(D+).

Finally, to determine the postexamination probability of disease, $P(\mathrm{D}+|\mathrm{T}+)$, convert the postexamination odds back to a postexamination probability value. The postexamination probability of disease given the examination result can be computed from the postexamination odds as follows:

$$P(\mathrm{D}+|\mathrm{T}+) = \frac{\mathrm{Odds}(\mathrm{D}+|\mathrm{T}+)}{1+\mathrm{Odds}(\mathrm{D}+|\mathrm{T}+)}$$

With these three formulas, the likelihood ratio can be used to determine the postexamination probability of disease (or predictive value) from any preexamination probability of disease.

Limitations of the Likelihood Ratio

The likelihood ratio has several properties that limit its usefulness in describing diagnostic examinations. First, it functions only on the basis of the odds of disease rather than the more intuitive probability of disease. Accordingly, the likelihood ratio is best considered on a logarithmic scale: Likelihood ratios of less than 1.0 indicate that the examination result will decrease disease likelihood, and ratios of greater than 1.0 indicate that the examination result will increase disease likelihood. To many, it is not obvious that a likelihood ratio of 4.0 increases the likelihood of disease to the same degree that a likelihood ratio of 0.25 decreases the likelihood. Furthermore, it is sometimes counterintuitive that the same likelihood ratio causes different absolute changes in probability, depending on the preexamination probability. Despite these weaknesses, the likelihood ratio is probably underused in the radiology literature today as a measure of examination performance.

CONCLUSION

In this article, I describe nine principles (Appendix) that guide the evaluation and use of diagnostic imaging examinations in clinical practice. These principles are helpful when choosing measures to describe the capabilities of a diagnostic examination. As I have discussed, sensitivity and specificity are relatively invariant descriptors of examination accuracy. However, they have limited clinical usefulness and often cannot be used directly to compare two diagnostic examinations. ROC curves can be used to directly compare examinations independently of reader temperament or varying image interpretation criteria, but they yield little information that is useful to the clinical decision maker. Positive and negative predictive values yield useful information for clinical decision makers by facilitating explicit consideration of the trade-offs at hand, but they are intrinsically dependent on the preexamination likelihood of disease and therefore on the population of patients in whom the given examination is performed. Finally, the likelihood ratio can be used to calculate the postexamination likelihood of disease from the preexamination likelihood of disease, but the associated use of odds and the logarithmic scale are counterintuitive for some. An understanding of the described fundamental measures of examination performance and how they are clinically useful is vital to the appropriate evaluation and use of diagnostic imaging examinations.

APPENDIX

Nine principles that are helpful when choosing measures to describe the capabilities of a diagnostic examination:

1. Sensitivity is a measure of how a diagnostic examination performs in a population of patients who have the disease in question.
2. Specificity is a measure of how a diagnostic examination performs in a population of patients who do not have the disease in question (ie, healthy subjects).
3. A sensitive examination is more valuable in situations where false-negative results are more undesirable than false-positive results. A specific examination is more valuable in situations where false-positive results are more undesirable than false-negative results.
4. The sensitivity and specificity of a diagnostic examination are related to one another.
5. ROC curves provide a method to compare diagnostic examination accuracy independently of the diagnostic criteria (ie, strict or flexible) used.
6. A diagnostic examination causes a change in our belief about the likelihood that disease is truly present.
7. The positive predictive value indicates the likelihood of disease given a positive examination.
8. The negative predictive value indicates the likelihood of no disease given a negative examination.
9. Likelihood ratios enable calculation of the postexamination probability of disease from the preexamination probability of disease.

Acknowledgments: My special thanks to Harold Kundel, MD, Kimberly Applegate, MD, MS, and Philip Crewson, PhD, for their thoughtful comments on earlier versions of this manuscript.

References

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11:88–94.
2. Brismar J, Jacobsson B. Definition of terms used to judge the efficacy of diagnostic tests: a graphic approach. AJR Am J Roentgenol 1990; 155:621–623.
3. Black WC. How to evaluate the radiology literature. AJR Am J Roentgenol 1990; 154:17–22.
4. McNeil BJ, Keeler E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med 1975; 293:211–215.
5. Burton E, Troxclair D, Newman W. Autopsy diagnoses of malignant neoplasms: how often are clinical diagnoses incorrect? JAMA 1998; 280:1245–1248.
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
7. Metz CE. ROC methodology in radiologic imaging. Invest Radiol 1986; 21:720–733.
8. Brismar J. Understanding receiver-operating-characteristic curves: a graphic approach. AJR Am J Roentgenol 1991; 157: 1119–1121.
9. Nunes LW, Schnall MD, Siegelman E, et al. Diagnostic performance characteristics of architectural features revealed by high spatial-resolution MR imaging of the breast. AJR Am J Roentgenol 1997; 169: 409–415.
10. Pauker S, Kassirer J. The threshold approach to clinical decision making. N Engl J Med 1980; 302:1109–1117.
11. Ransohoff D, Feinstein A. Problems of spectrum bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 229:926–930.
12. Hillman BJ. Outcomes research and cost-effectiveness analysis for diagnostic imaging. Radiology 1994; 193:307–310.
13. Hrung J, Langlotz C, Orel S, Fox K, Schnall M, Schwartz J. Cost-effectiveness of magnetic resonance imaging and needle core biopsy in the pre-operative workup of suspicious breast lesions. Radiology 1999; 213:39–49.
14. Filly RA. The "lemon" sign: a clinical perspective. Radiology 1988; 167:573–575.
15. Thornbury JR, Fryback DG, Edwards W. Likelihood ratios as a measure of the diagnostic usefulness of excretory urogram information. Radiology 1975; 114: 561–565.
16. Black WC, Armstrong P. Communicating the significance of radiologic test results: the likelihood ratio. AJR Am J Roentgenol 1986; 147:1313–1318.