Fundamentals of Clinical Research for Radiologists

Susan Weinstein[1]
Nancy A. Obuchowski[2]
Michael L. Lieber[2]

Clinical Evaluation of Diagnostic Tests

The evaluation of the accuracy of diagnostic tests and the appropriate interpretation of test results are the focus of much of radiology and its research. In this article, we first will review the basic definitions of diagnostic test accuracy, including a brief introduction to receiver operating characteristic (ROC) curves. Then we will evaluate how diagnostic tests can be used to address clinical questions such as "Should this patient undergo this diagnostic test?" and, after ordering the test and seeing the test result, "What is the likelihood that this patient has the disease?" We will finish with a discussion of some important concepts for designing research studies that estimate or compare diagnostic test accuracy.

Defining Diagnostic Test Accuracy

Sensitivity and Specificity

There are two basic measures of the inherent accuracy of a diagnostic test: sensitivity and specificity. They are equally important, and one should never be reported without the other. Sensitivity is the probability of a positive test result (that is, the test indicates the presence of disease) for a patient with the disease. Specificity, on the other hand, is the probability of a negative test result (that is, the test does not indicate the presence of disease) for a patient without the disease. We use the term "disease" here loosely to mean the condition (e.g., breast cancer, deep venous thrombosis, intracranial aneurysm) that the diagnostic test is supposed to detect. We calculate the test's specificity based on patients without this condition, but these patients often have other diseases.

Table 1 summarizes the definitions of sensitivity and specificity [1]. The table rows give the results of the diagnostic test, as either positive for the disease of interest or negative for the disease of interest. The columns indicate the true disease status, as either disease present or disease absent. True-positives (TPs) are those patients with the disease who test positive. True-negatives (TNs) are those without the disease who test negative. False-negatives (FNs) are those with the disease but the test falsely indicates the disease is *not* present. False-positives (FPs) are those without the disease but the test falsely indicates the presence of disease. Sensitivity, then, is the probability of a TP among patients with the disease (TPs + FNs). Specificity is the probability of a TN among patients without the disease (TNs + FPs).

Consider the following example. Carpenter et al. [2] evaluated the diagnostic accuracy of MR venography (MRV) to detect deep venous thrombosis (DVT). They performed MRV in a group of 85 patients who presented with clinical symptoms of DVT. The patients also underwent contrast venography, which is an invasive procedure considered to provide an unequivocal diagnosis for DVT (the so-called "gold standard test" or "standard of reference"). Of a total of 101 venous systems evaluated, 27 had DVT by contrast venography. All 27 cases were detected on MRV; thus, the sensitivity of MRV was 27/27, or 100%. Of 74 venous systems without DVT, as confirmed by contrast venography, three tested positive on MRV (that is, three FPs). The specificity of MRV was 71/74, or 96% specificity (Table 2).

Combining Multiple Tests

Few diagnostic tests are both highly sensitive and highly specific. For this reason, patients sometimes are diagnosed using two or more tests. These tests may be performed ei-

Series editors: Nancy Obuchowski, C. Craig Blackmore, Steven Karlik, and Caroline Reinhold.

This is the 13th in the series designed by the American College of Radiology (ACR), the Canadian Association of Radiologists, and the *American Journal of Roentgenology*. The series, which will ultimately comprise 22 articles, is designed to progressively educate radiologists in the methodologies of rigorous research, from the most basic principles to a level of considerable sophistication. The articles are intended to complement interactive software that permits the user to work with what he or she has learned, which is available on the ACR Web site (www.acr.org).

Project coordinator: G. Scott Gazelle, Chair, ACR Commission on Research and Technology Assessment.

Staff coordinator: Jonathan H. Sunshine, Senior Director for Research, ACR.

[1]Department of Radiology, University of Pennsylvania Medical Center, Philadelphia, PA 19104. Address correspondence to S. Weinstein.

[2]Departments of Biostatistics and Epidemiology and Radiology, The Cleveland Clinic Foundation, Cleveland, OH 44195.

AJR 2005;184:14–19

0361–803X/05/18405–14

TABLE 1 Defining Sensitivity and Specificity

Test	Disease	
	Present	Absent
+	True-positive (TP)	False-positive (FP)
–	False-negative (FN)	True-negative (TN)

Note.—Sensitivity = TPs/(TPs + FNs), specificity = TNs/(TNs + FPs).

TABLE 2 Sensitivity and Specificity of MRV in 101 Venous Systems

MRV	Deep Venous Thrombosis	
	Present	Absent
+	27	3
–	0	71

Note.—MRV = magnetic resonance venography.

ther in parallel (i.e., at the same time and interpreted together) or in series (i.e., the results of the first test determine whether the second test is performed at all) [3]. The latter has the advantage of avoiding unnecessary tests, but the disadvantage of potentially delaying treatment for diseased patients by lengthening the diagnostic testing period.

Tests can be interpreted in parallel in two ways. The first, called "the OR rule," yields a positive diagnosis if either test (let's assume there are two tests) is positive and a negative diagnosis if both tests are negative. That is, if test A and test B are both negative, then the combined result is negative, but if either or both are positive, then the combined result is positive.

The second rule, called "the AND rule," yields a positive diagnosis only if both tests are positive and a negative diagnosis if either test is negative. That is, if test A and test B are both positive, then the combined result is positive, but if either or both are negative, then the combined result is negative.

Let us denote the sensitivities of the two tests by SE_a and SE_b, and their specificities by SP_a and SP_b. To calculate the sensitivity of the combined test in parallel using the OR rule, the formula is: $SE_a + SE_b - (SE_a \times SE_b)$. Specificity under the OR rule is simply $SP_a \times SP_b$. Conversely, to calculate sensitivity using the AND rule, the formula is: $SE_a \times SE_b$, while specificity under the AND rule is $SP_a + SP_b - (SP_a \times SP_b)$.

Under the OR rule, the sensitivity of the combined result is higher than that of either test alone, but the combined specificity is lower than that of either test. With the AND rule, this is reversed: The specificity of the combined result is higher than either test alone, but the combined sensitivity is lower than that of either test.

Serial testing is an alternative to parallel testing that is particularly cost-efficient when screening for rare conditions and often is used when the second test is expensive and/or risky. Under the OR rule, if the first test is positive, the diagnosis is positive; otherwise, the second test is performed. If the second test is positive after a negative first test, then the diagnosis also is positive; otherwise, the diagnosis is negative. The OR rule, then, leads to a higher overall sensitivity than either test by itself. With the AND rule, if the first test is positive, the second test is performed. If the second test is positive, the diagnosis is positive; otherwise, the diagnosis is negative. The AND rule, then, leads to a higher overall specificity than either test by itself.

To calculate sensitivity of the combined test using serial testing with the OR rule, the formula is: $SE_a + (1 - SE_a) \times SE_b$. Specificity under the OR rule is simply $SP_a \times SP_b$. Conversely, to calculate sensitivity using the AND rule, the formula is: $SE_a \times SE_b$, while specificity under the AND rule is $SP_a + (1 - SP_a) \times SP_b$.

ROC Curves

While some tests provide dichotomous results (that is, positive or negative), other tests yield results that are numeric values (for example, attenuation of a lesion on CT) or ordered categories (for example, BI-RADS scoring used in mammography). Consider CT attenuation as a diagnostic test for distinguishing papillary renal cell carcinomas from other types of renal masses [4]. In Table 3, the ratio of tumor enhancement to normal kidney enhancement (T–K ratio) of 10 masses is listed.

How do we calculate the basic measures of accuracy, that is, sensitivity and specificity, for T–K ratio as a diagnostic test for papillary masses? We shall consider each unique T–K ratio value as a "cutoff," or "decision threshold" and calculate the sensitivity and specificity associated with each cutoff. Masses with T–K ratio values greater than or equal to the cutoff are called "negative" for papillary lesions and masses with T–K ratio values less than the cutoff are called "positive" for papillary lesions. In Table 3, the third and fourth columns give the calculated sensitivity and specificity, respectively, using the T–K ratio value in column 2 as the cutoff. Note that as the value of the cutoff increases, the specificity decreases while the sensitivity increases.

TABLE 3 T–K Ratio Values of 5 Papillary and 5 Nonpapillary Renal Masses

Cell Type	T–K Ratio	Sensitivity	Specificity	FPR
PRCC	0.05	0.0	1.0	0.0
PRCC	0.11	0.2	1.0	0.0
Other	0.20	0.4	1.0	0.0
PRCC	0.22	0.4	0.8	0.2
PRCC	0.25	0.6	0.8	0.2
Other	0.29	0.8	0.8	0.2
Other	0.38	0.8	0.6	0.4
PRCC	0.43	0.8	0.4	0.6
Other	0.56	1.0	0.4	0.6
Other	0.66	1.0	0.2	0.8

Note.—PRCC = papillary renal cell carcinoma, FPR = false-positive rate, or 1 – specificity.

In Figure 1, we have plotted the 10 pairs of sensitivity and specificity calculated in Table 3. The *y*-axis is the sensitivity and the *x*-axis is 1 minus the specificity, or the false-positive rate (FPR). Connecting these points with line segments, we have constructed an ROC curve [5]. A test with an ROC curve that lies near the "chance diagonal line" in Figure 1 has no ability, beyond mere guessing, at distinguishing between patients with and without the disease. In contrast, a test with an ROC curve that passes near the upper left corner (that is, near 100% sensitivity and 0% FPR [100% specificity]) is nearly perfect at distinguishing disease from no disease. T–K ratio has moderate accuracy, with its ROC curve falling between these two extremes.

Suppose now that an investigator proposes the ratio of the attenuation of the mass to the attenuation of the abdominal aorta (T–A ratio) as a new diagnostic test for papillary lesions. This investigator, however, arbitrarily chooses a single cutoff and reports only the sensitivity and specificity at that cutoff. Figure 2 illustrates this single point (labeled A) in relation to the ROC curve for T–K ratio. We might be tempted to conclude that T–K ratio is superior to T–A ratio because point A falls below the ROC curve for T–K ratio. There

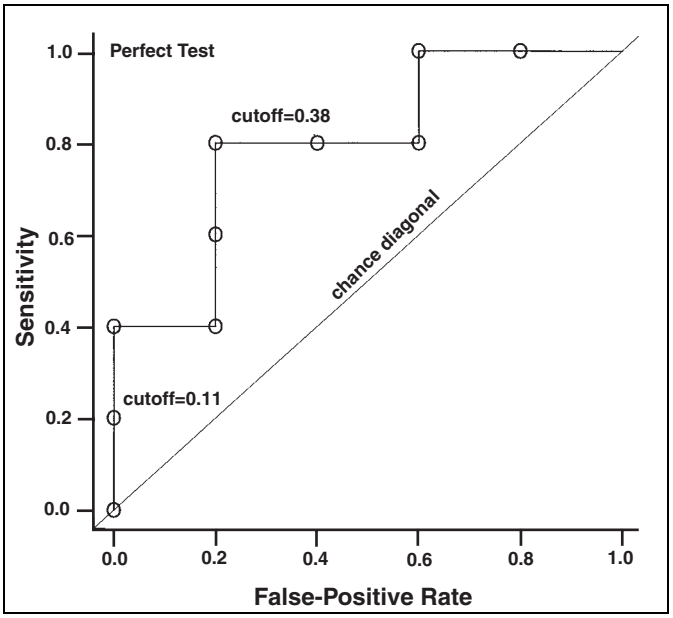

Fig. 1.—10 pairs of sensitivity and specificity as calculated in Table 3. The *y*-axis is the sensitivity and the *x*-axis is 1 minus the specificity, or the false-positive rate (FPR). Receiver operating characteristic (ROC) curve is created by connecting points with line segments.

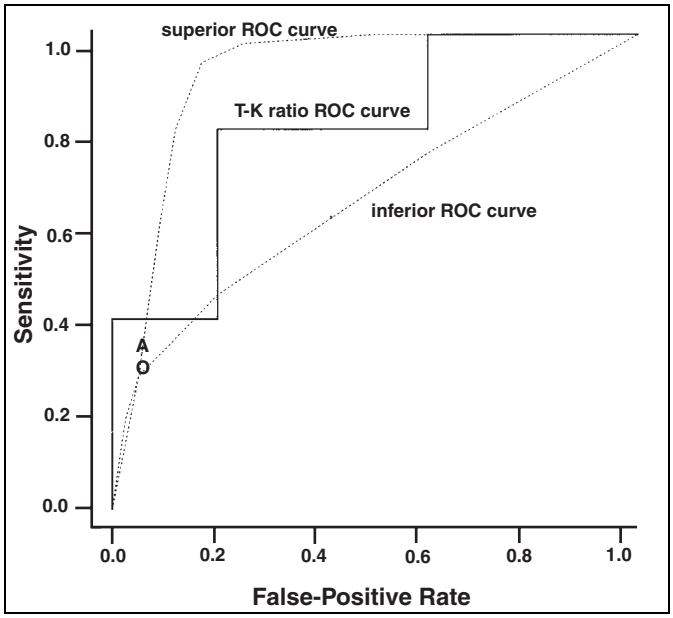

Fig. 2.—Single cutoff point (labeled A) in relation to the receiver operating characteristic (ROC) curve for T–K (tumor enhancement to normal kidney enhancement) ratio.

are, however, an infinite number of ROC curves that could pass through point A, two of which are depicted by dashed curves in Figure 2. Some of these ROC curves could be superior to the ROC curve for T–K ratio for most FPRs and others inferior. Based on the single sensitivity and specificity reported by the investigator, we cannot determine if the T–A ratio is superior or inferior in relation to the T–K ratio. However, if we had been given the ROC curves of both the T–A and T–K ratio, then we could compare these two diagnostic tests and determine, for any range of FPRs, which test is preferred.

This example illustrates the importance of ROC curves and why they have become the state-of-the-art method for describing the diagnostic accuracy of a test. In a future module in this series Obuchowski [6] provides a detailed account of ROC curves, including constructing smooth ROC curves, estimating various summary measures of accuracy derived from them, finding the optimal cutoff on the ROC curve for a particular clinical application, and identifying available software.

Interpretation of Diagnostic Tests

Calculating the Positive and Negative Predictive Values

Clinicians are faced each day with the challenge of deciding appropriate management for patients, based at least in part on the results of less than perfect diagnostic tests. These clinicians need answers to the following questions. "What is the likelihood that this patient has the disease when the test result is positive?" and "What is the likelihood that this patient does *not* have the disease when the test result is negative?" The answers to these questions are known as the positive and negative predictive values, respectively. We illustrate these with the following example.

The lemon sign has been described as an important indicator of spina bifida. Nyberg et al. [7] describe the sensitivity and specificity of the lemon sign in the detection of spina bifida in a high-risk population (elevated material serum α-fetoprotein level, suspected hydrocephalus or neural tube defect, or family history of neural tube defect). A portion of their data is summarized in Table 4.

Spina bifida occurred in 6.1% (14/229) of the sample, that is, sample prevalence was 6.1%. The lemon sign was seen in 92.9% (13/14) of the fetuses with spina bifida (92.9% sensitivity), and was absent in 98.6% (212/215) of the fetuses without spina bifida (98.6% specificity).

We also can calculate the positive and negative predictive values of the lemon sign from the available data. The positive predictive value (PPV) is the probability that the fetus has spina bifida when the lemon sign is present. The PPV is calculated as follows:

$$\text{PPV} = \text{TP} / (\text{TP} + \text{FP}) = 13 / (13 + 3) \times 100\% = 81.3\% \quad (1)$$

The PPV differs from sensitivity. While the PPV tells us the probability of a fetus with spina bifida following detection of the lemon sign (that probability is 0.813, or 81.3%), the sensitivity tells us the probability that the lemon sign will be present among fetuses with spina bifida (probability is 0.929, or 92.9%). PPV helps the clinician decide how to treat the patient after the diagnostic test comes back positive. Sensitivity, on the other hand, is a property of the diagnostic test and helps the clinician decide which test to use.

The corollary to the PPV is the negative predictive value (NPV), that is, the probability that spina bifida will not be present when the lemon sign is absent. The NPV is calculated as follows:

$$\text{NPV} = \text{TN} / (\text{TN} + \text{FN}) = 212 / (212 + 1) \times 100\% = 99.5\% \quad (2)$$

If the lemon sign is absent, there is a 99.5% chance that the fetus will not have spina bifida. The NPV is different from the test's specificity. Specificity tells us the probability that the lemon sign will be absent among fetuses without spina bifida (that probability is 0.986, or 98.6%).

TABLE 4 Lemon Sign Versus Spinal Cord Defect in Fetuses Prior to 24 Weeks

Lemon Sign	Spina Bifida	No Spina Bifida	Total
+	13	3	16
–	1	212	213
Total	14	215	229

Note.—SE = 92.9%, SP = 98.6%, PPV = 81.3%, NPV = 99.99%, prevalence = 6.1%.

TABLE 5 The PPV of the Lemon Sign in the General Population

Lemon Sign	Spina Bifida	No Spina Bifida	Total
+	9	140	149
–	1	9,850	9,851
Total	10	9,990	10,000

Note.—SE = 90.0%, SP = 98.6%, PPV = 6.0%, NPV = 99.99%, prevalence = 0.1%.

The PPV and NPV can also be calculated from Bayes' theorem. Bayes' theorem allows us to compute the PPV and NPV from estimates of the test's sensitivity and specificity, and the probability of the disease before the test is applied. The latter is referred to as the pretest probability and is based on the patient's previous medical history, previous and recent exposures, current signs and symptoms, and results of other screening and diagnostic tests performed. When this information is unknown or when calculating the PPV or NPV for a population, the prevalence of the disease in the population is used as the pretest probability. The PPV and NPV, then, are called posttest probabilities (also, revised or posterior probabilities), and represent the probability of the disease after the test result is known.

Let p denote the pretest probability of disease, and SE and SP the sensitivity and specificity of the diagnostic test. Recalling the expression for a conditional probability (see module 10 [8]),

$$\text{PPV} = \text{P(disease}|+\text{test)} = [\text{SE} \times p] / [\text{SE} \times p + (1 - \text{SP}) \times (1 - p)] \quad (3)$$

$$\text{NPV} = \text{P(no disease}|-\text{test)} = [\text{SP} \times (1 - p)] / [\text{SP} \times (1 - p) + (1 - \text{SE}) \times p] \quad (4)$$

Thus, the posttest probability of disease for any patient can be calculated if one knows the accuracy of the test and the patient's pretest probability of disease.

The PPV and NPV can vary markedly, depending on the patient's pretest probability, or prevalence of disease in the population. In the Nyberg et al. [7] study the prevalence rate of spina bifida in their high risk sample was 6.1%. In the general population, however, the prevalence of spina bifida is much less, about 0.1%. Filly [9] studied the predictive ability of the lemon sign in the general population. He assumed that the sensitivity of the lemon sign was 90.0% and the specificity was 98.6% (very similar to that in Nyberg's small study, 92.9% and 98.6%, respectively). In a sample of 10,000 fetuses from a low-risk population (see Table 5), Filly showed that the positive predictive value is only 6%. This is in contrast to the PPV of 81.3% in the Nyberg study. The drastic difference in PPVs is due to the different prevalence rates of spina bifida in the two samples, 6.1% in Nyberg's and 0.1% in Filly's. Thus, while a high-risk fetus with a lemon sign may have an 81% chance of having spina bifida, "a low risk fetus with a lemon sign has a 94% chance of being *perfectly normal*" [9]. This example illustrates the importance of reporting the pretest probability or prevalence rate of disease whenever one presents a PPV or NPV.

Rationale for Ordering a Diagnostic Test

The previous section described how clinicians can use the results of a diagnostic test to plan a patient's management. Let's back up a bit in the clinical decision-making process and look at the rationale for ordering a diagnostic test.

In the simplest scenario (ignoring monetary costs, insurance reimbursement rates, etc.), there are three pieces of information that a clinician needs to determine whether a diagnostic test should or should not be ordered:

1. From the patient's previous medical history, previous and recent exposures, current signs and symptoms, and results of other screening and diagnostic tests performed, what is the probability that this patient has the disease (that is, the pretest probability)?

2. How accurate (sensitivity and specificity) is the diagnostic test being considered?

3. Could the results of this test affect the patient's management?

In the previous section, we saw how the pretest probability and the test's sensitivity and specificity fit into Bayes' theorem to tell us the posttest probability of disease. We also saw, even for a very accurate test, how the PPV can be quite low when the pretest probability is low. The clinician ordering a test needs to consider how the patient will be managed if the test result is negative versus if the test result is positive. If the probability of disease will still be low after a positive test, then the test may have no impact on the patient's management.

An example is screening for intracranial aneurysms in the general population. The prevalence of aneurysms is low, maybe 1%, in the general population. Even though magnetic resonance angiography (MRA) may have excellent accuracy, say 95% sensitivity and specificity, the PPV is still quite low, 0.16 (16%) from equation 3. Considering the nontrivial risks of invasive catheter angiography (which is the usual presurgical tool) [10], the clinician may decide that even after a positive MRA, the patient should not undergo catheter angiography. In this scenario, the clinician may decide not to order the MRA, given that its result, either positive or negative, will not impact the patient's management.

Designing Studies to Estimate and Compare Tests' Diagnostic Accuracy

As with all new medical devices, treatments, and procedures, the efficacy of diagnostic tests must be assessed in clinical studies. In the second module of this series Jarvik [11] described six levels of diagnostic efficacy. Here, we will focus on the second level, which is the stage at which investigators assess the diagnostic *accuracy* of a test.

Phases in the Assessment of Diagnostic Test Accuracy

There typically are three phases to the assessment of a diagnostic test's accuracy [3]. The first is the *exploratory phase*. It usually is the first clinical study performed to assess the efficacy of a new diagnostic test. These tend to be small, inexpensive studies, typically involving 10 to 50 patients with and without the disease of interest. The patients selected for the study samples often are cases with classical overt disease (for example, symptomatic lung cancer) and healthy volunteer controls. If the test results of these two populations do not differ, then it is not worth pursuing the diagnostic test further.

The second phase is the *challenge phase*. Here, we recognize that a diagnostic test's sensitivity and specificity can vary with the extent and stage of the disease, and the comorbidities present. Thus, in this phase we select patients with subtle, or early disease, and

TABLE 6 Common Features of Diagnostic Test Accuracy Studies

Feature	Explanation
Two samples of patients	One sample of patients with and one sample without the disease are needed to estimate both sensitivity and specificity.
Well-defined patient samples	Regardless of the sampling scheme used to obtain patients for the study, the characteristics of the study patients (e.g., age, gender, comorbidities, stage of disease) should be reported.
Well-defined diagnostic test	The diagnostic test must be clearly defined and applied in the same fashion to all study patients.
Gold standard/reference standard	The true disease status of each study patient must be determined by a test or procedure that is infallible, or nearly so.
Sample of interpreters	If the test relies on a trained observer to interpret it, then two or more such observers are needed to independently interpret the test [15].
Blinded interpretations	The gold standard should be conducted and interpreted blinded to the results of the diagnostic test, and the diagnostic test should be performed and interpreted blinded to the results of the gold standard.
Standard reporting of findings	The results of the study should be reported following published guidelines for the reporting of diagnostic test accuracy [16].

with comorbidities that could interfere with the diagnostic test [12]. For example, in a study to assess the ability of MRI to detect lung cancer, the study patients might include those with small nodules (3 cm), and patients with nodules and interstitial disease. The controls might have diseases in the same anatomic location as the disease of interest, for example, interstitial disease but no nodules. These studies often include competing diagnostic tests to compare their accuracies with the test under evaluation. ROC curves are most often used to assess and compare the tests. If the diagnostic test shows good accuracy, then it can be considered for the third phase of assessment.

The third phase is the *advanced phase*. These studies often are multicenter studies involving large numbers of patients (100 or more). The patient sample should be representative of the target clinical population. For example, instead of selecting patients with known lung cancer and controls without cancer, we might recruit patients presenting to their primary care physician with a persistent cough or bloody sputum. Further testing and follow-up will determine which patients have lung cancer and which do not.

It is from this third phase where we obtain reliable estimates of a test's accuracy for the target clinical population. Estimates of accuracy from the exploratory phase usually are too optimistic because the "sickest of the sick" are compared with the "wellest of the well" [13]. In contrast, estimates of accuracy from the challenge phase often are too low because the patients are exceptionally difficult to diagnose.

Common Features of Diagnostic Test Accuracy Studies

The studies in the three phases differ in terms of their objectives, sampling of patients, and sample sizes. There are, however, some common features to all studies of diagnostic test accuracy, as summarized in Table 6. We elaborate here on a few important issues.

Studies of diagnostic test accuracy require both subjects with and without the disease of interest. If one of these populations is not represented in the study, then either sensitivity or specificity cannot be calculated. We stress that reporting one without reference to the other is uninformative and often misleading. The number of patients needed for diagnostic accuracy studies depends on the phase of the study, the clinical setting in which the test will be applied (for example, screening or diagnostic), and certain characteristics of the patients and test itself (for example, does the test require interpretation by human observers?). Statistical methods are available for determining the appropriate sample size for diagnostic accuracy studies [3, 14].

Studies of diagnostic test accuracy require a test or procedure for determining the true disease status of each patient. This test or procedure is called the "gold standard" (or "standard of reference," "reference standard," particularly when there is no perfect gold standard). The gold, or reference, standard must be conducted and interpreted blinded to the diagnostic test results to avoid bias. Common standards of reference in radiology studies are surgery, pathology results, and clinical follow-up. For example, in the study of Carpenter et al. [2] of the accuracy of MR venography for detecting deep venous thrombosis, contrast venography was used as the reference standard. Sometimes a study uses more than one type of reference standard. For example, in a study to assess the accuracy of mammography, patients with a suspicious lesion on mammography might undergo core biopsy and/or surgery, whereas patients with a negative mammogram would need to be followed for 2 years either to confirm that the patient was cancer free or to detect missed cancers on follow-up screenings. Note that when using different reference standards for patients with positive and negative test results, it is important that all the reference standards are infallible, or nearly so. One form of workup bias occurs when patients with one test result undergo a less rigorous reference standard than patients with a different test result [3].

Determining the appropriate reference standard for a study often is the most difficult part of designing a diagnostic accuracy study. Reference standards should be infallible, or nearly so. This is difficult, however, because even pathology is not infallible, as it is an interpretive field relying on subjective assessment from human observers with varying skill levels. One such example is the reader variability in pathologic interpretation of borderline intraductal breast carcinoma versus atypical ductal carcinoma. While some pathologists may interpret the lesion as intraductal cancer, others may interpret the same lesion as atypical ductal hyperplasia. While often we have to accept that a reference standard is not perfect, it is important that it be nearly infallible. If the reference standard is not nearly infallible, then *imperfect gold standard bias* can lead to unreliable and misleading estimates of accuracy. Zhou et al. [3] discuss in detail imperfect gold standard bias and possible solutions.

In other situations, no reference standard is available (for example, epilepsy) or it is unethical to subject patients to the reference standard because it poses a risk (for example, an invasive test such as catheter angiography). In these situations, we at least can correlate the test results to other tests' findings and to clinical outcome, even if we cannot report the test's sensitivity and specificity.

It is *never* an option to omit from the calculation of sensitivity and specificity those patients without a diagnosis confirmed by a reference standard. Such studies yield erroneous estimates of test accuracy due to a form of workup bias called *verification bias* [17, 18].

This is one of the most common types of bias in radiology studies [19] and is counterintuitive. Investigators often believe they are getting more reliable estimates of accuracy by excluding cases where the reference standard was not performed. If, however, the diagnostic test results were used in the decision of whether to perform the reference standard procedure, then verification bias most likely is present. For example, if the results of MR venography are used to determine which patients will undergo contrast venography, and if patients who did not undergo contrast venography are excluded from the calculations of the test's accuracy, then verification bias exists. Zhou et al. [3] discuss verification bias from a statistical standpoint and offer a variety of solutions.

Summary

We conclude with a summary of five key points in the clinical evaluation of diagnostic tests:

1. Sensitivity and specificity always should be reported together.

2. ROC curves allow a comprehensive assessment and comparison of diagnostic test accuracy.

3. PPV and NPV cannot be interpreted correctly without knowing the prevalence of disease in the study sample.

4. Patients who did not undergo the reference standard procedure should never be omitted from studies of diagnostic test accuracy.

5. Published guidelines should be followed when reporting the findings from studies of diagnostic test accuracy.

References

1. Gehlbach SH. Interpretation: sensitivity, specificity, and predictive value. In: Gehlbach SH, ed. *Interpreting the medical literature.* New York: McGraw-Hill, 1993:129–139
2. Carpenter JP, Holland GA, Baum RA, Owen RS, Carpenter JT, Cope C. Magnetic resonance venography for the detection of deep venous thrombosis: comparison with contrast venography and duplex Doppler ultrasonography. *J Vasc Surg* 1993;18:734–741
3. Zhou XH, Obuchowski NA, McClish DK. *Statistical methods in diagnostic medicine.* New York: Wiley & Sons, 2002
4. Herts BR, Coll DM, Novick AC, Obuchowski N, Linnell G, Wirth SL, Baker ME. Enhancement characteristics of papillary renal neoplasms revealed on triphasic helical CT of the kidneys. *AJR* 2002;178:367–372
5. Metz CE. ROC methodology in radiological imaging. *Invest Radiol* 1986;21:720–733
6. Obuchowski NA. Receiver operating characteristic (ROC) analysis. *AJR* 2005(in press)
7. Nyberg DA, Mack LA, Hirsch J, Mahony BS. Abnormalities of fetal cranial contour in sonographic detection of spina bifida: evaluation of the "lemon" sign. *Radiology* 1988;167:387–392
8. Joseph L, Reinhold C. Introduction to probability theory and sampling distributions. *AJR* 2003;180:917–923
9. Filly RA. The "lemon" sign: a clinical perspective. *Radiology* 1988;167:573–575
10. Levey AS, Pauker SG, Kassirer JP, et al. Occult intracranial aneurysms in polycystic kidney disease: when is cerebral arteriography indicated? *N Engl J Med* 1983;308:986–994
11. Jarvik JG. The research framework. *AJR* 2001;176:873–877
12. Ransohoff DJ, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978;299:926–930
13. Sox Jr HC, Blatt MA, Higgins MC, Marton KI. *Medical decision making.* Boston: Butterworths-Heinemann, 1988
14. Beam CA. Strategies for improving power in diagnostic radiology research. *AJR* 1992;159:631–637
15. Obuchowski NA. How many observers in clinical studies of medical imaging? *AJR* 2004;182:867–869
16. Bossuyt PM, Reitsma JB, Bruns DE, et al. Toward complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Acad Radiol* 2003;10:664–669
17. Begg CB, McNeil BJ. Assessment of radiologic tests, control of bias, and other design considerations. *Radiology* 1988;167:565–569
18. Black WC. How to evaluate the radiology literature. *AJR* 1990;154:17–22
19. Reid MC, Lachs MS, Feinstein AR. Use of methodologic standards in diagnostic test research: getting better but still not good. *JAMA* 1995;274:645–651

The reader's attention is directed to earlier articles in the Fundamentals of Clinical Research series:

1. Introduction, which appeared in February 2001
2. The Research Framework, April 2001
3. Protocol, June 2001
4. Data Collection, October 2001
5. Population and Sample, November 2001
6. Statistically Engineering the Study for Success, July 2002
7. Screening for Preclinical Disease: Test and Disease Characteristics, October 2002
8. Exploring and Summarizing Radiologic Data, January 2003
9. Visualizing Radiologic Data, March 2003
10. Introduction to Probability Theory and Sampling Distributions, April 2003
11. Observational Studies in Radiology, November 2004
12. Randomized Controlled Trials, December 2004