10 Assessing the Performance and Validity of Diagnostic Tests and Screening Programs

David C. Miller, MD, MPH, Rodney L. Dunn, MS, and John T. Wei, MD, MS

CONTENTS

Although surgery is primarily a therapeutic intervention, surgeons also play a pivotal role in the initial evaluation and diagnosis of surgical disease. Indeed, recent scientific and technologic advances (e.g., molecular markers of disease) have considerably expanded the catalog of diagnostic tests available to contemporary surgeons. At the same time, many established (and widespread) screening programs (e.g., mammography, colonoscopy, prostate-specific antigen [PSA]) are designed to detect conditions that are treated primarily with surgical interventions. Moreover, given the substantial morbidity that may accompany surgical intervention, it is imperative that surgeons critically assess the value of a diagnostic test before using its results as the basis for intervention.

In this context, it is essential for surgeons to understand fundamental concepts related to the evaluation of clinical test performance, and for surgical investigators to be skilled in the interpretation of measures of test validity. Whether the test in question is from the patient history, physical examination, a laboratory test, or an imaging study, surgeons must be able to answer the question: How useful is this test for distinguishing diseased from disease-free individuals? *(1)*.

In this chapter, we will cover basic concepts related to the assessment of clinical test performance. We will introduce several statistical methods for assessing the validity of diagnostic tests including sensitivity, specificity, positive and negative predictive values, likelihood ratios and receiver operating characteristic curves (Appendix 1). To highlight their appropriate clinical application, the various measures of validity will be covered separately for tests with categorical (dichotomous) versus continuous results.

From: *Clinical Research for Surgeons*
Edited by: D. F. Penson and J. T. Wei © Humana Press Inc., Totowa, NJ

In addition, this chapter will address some of the most salient issues related to the selection, implementation, and evaluation of screening tests and programs. The rationale for disease screening efforts, as well as various risks and benefits associated with such programs, will be discussed. Finally, several sources of potential bias associated with screening programs, including lead-time bias and length-bias sampling, will be covered to provide a comprehensive framework for assessing the value and validity of a screening program.

1. ASSESSING THE VALIDITY OF DIAGNOSTIC TESTS

1.1. Sensitivity, Specificity, and Accuracy

The validity of a test refers to its ability to measure what it is purported to measure; in most clinical situations, this involves the ability of a diagnostic test to distinguish between individuals with and without a particular disease. Two principal measures of test validity are sensitivity and specificity. In general terms, sensitivity may be characterized as the degree to which a particular test correctly identifies diseased individuals; in contrast, specificity reflects the capacity of the test to distinguish individuals that are free of disease *(1)*. In statistics, sensitivity is defined as the proportion of diseased individuals with a positive test result; specificity, on the other hand, is the proportion of disease-free individuals with a negative test result. A complementary measure of the validity of a given test is its accuracy, which can be defined as the proportion of all tests results (both positive and negative) that are concordant with true health status.

An important caveat with regard to assessing the validity of a diagnostic is that, to assess the performance of a particular test, there must be a "gold standard" test available for comparison. In other words, a different and established test must be available that reliably and precisely differentiates individuals with and without a given disease. In many cases the gold standard may be the pathologic findings from an invasive procedure such as tissue biopsy or extirpative surgery. Alternatively, the gold standard may be based on an objective or subjective set of clinical findings, such as the National Institutes of Health/ National Institute of Diabetes and Digestive and Kidney criteria for the diagnosis of interstitial cystitis *(2–4)*. Thus, to properly assess the validity (sensitivity and specificity) of a diagnostic test, the investigator should identify and make use of an existing gold standard. Without a widely accepted gold standard for comparison, evaluations of test performance may be difficult.

1.2. How to Evaluate Tests With Categorical (Dichotomous) Results

A useful way to conceptualize the concepts of sensitivity and specificity is to start by examining a 2 × 2 table for a scenario involving a dichotomous disease state (i.e., disease present or disease absent) and a dichotomous test outcome (i.e. test positive or test negative) (Table 1). It should be mentioned that an ideal test would have both a sensitivity and specificity of 100%. Examining Table 1, such a test would classify subjects into only two outcome groups: individuals with the disease that have a positive test result (*true positives*, the upper left cell [a]) and individuals without the disease that have a negative test result (*true negatives*, the lower right cell [d]). In the clinical setting, there are no tests that perform at this ideal level. In fact, the outcomes of most tests include positive results in disease-free individuals (*false positives*, the upper right cell [b]) and negative results in people with that actually have the disease (*false negatives*, the lower left cell [c]). Based

Table 1
Standard Table for Comparison of Test Results With Actual Disease Status

	Disease Present	*Disease Absent*	
Test Positive	**a** (true positives)	**b** (false positives)	**a+b**
Test Negative	**c** (false negatives)	**d** (true negatives)	**c+d**
	a+c	**b+d**	**a+b+c+d**

Sensitivity = **a**/(**a**+**c**)
Specificity = **d**/(**b**+**d**)
Accuracy = **a+c**/(**a**+**b**+**c**+**d**)
Positive predictive value (PPV) = **a**/(**a**+**b**)
Negative predictive value (NPV) = **d**/(**c**+**d**)

on these four possible outcomes, this standardized 2 × 2 table can be used to further illustrate the calculation of sensitivity and specificity.

Recall that sensitivity is defined as the proportion of individuals with a disease that have a positive test result. From Table 1, the total number of diseased individuals is represented by the sum of cells a and c; the number of positive test results for this group is represented in cell a. Thus, for this standard 2 × 2 table, sensitivity is defined as:

$$\text{Sensitivity} = a/(a+c) \tag{1}$$

Similarly, specificity refers to the proportion of disease-free individuals (b+d) that have a negative test result (d) and is, therefore, represented by the following formula:

$$\text{Specificity} = d/(b+d) \tag{2}$$

It should also be noted that for a test with dichotomous results, the accuracy of the test is calculated based on the following formula:

$$\text{Accuracy} = (a+d)\ (a+b+c+d) \tag{3}$$

In a recent publication, Staib and associates used these calculations to evaluate the validity of a newly available diagnostic imaging modality. Specifically, the authors examined the ability of ^{18}F-fluorodeoxyglucose positron emission tomography (FDG-PET) to detect recurrent colorectal cancer in patients who had previously undergone surgical resection with curative intent. In this study, the diagnostic gold standard for recurrent cancer was either histologic confirmation via tissue biopsy or clinical progression of the presumably malignant site identified by FDG-PET *(5)*. The relevant results from this study are summarized in Table 2. Among the 58 patients with recurrent colorectal cancer, as documented by the gold standard described previously, 57 had increased tracer uptake on an FDG-PET scan (interpreted as a positive result). Therefore, the sensitivity of the FDG-PET scan was reported as 57/58 = 98.2%. In terms of specificity, negative FDG-PET results were observed in 38/42 men without recurrent cancer, indicating a specificity for this test of 90.5% (Table 2). The accuracy of FDG-PET imaging for detecting a recurrence was (57+38)/(57+4+1+38) = 95%. Based on these results, the authors concluded that FDG-PET had reasonable validity and may be a useful adjunct to conventional imaging studies in patients with colorectal cancer *(5)*.

Table 2
Validity of ^{18}F-Fluorodeoxyglucose Positron Emission Tomography (FDG-PET) for Detecting Recurrent Colorectal Cancer

	Recurrent Colorectal Cancer	*No Recurrent Colorectal Cancer*	
FDG-PET positive	57	4	61
FDG-PET negative	1	38	39
	58	42	100

Sensitivity = 57/58 = 98.2%
Specificity = 38/42 = 90.5%
Accuracy = (57+38)/(57+4+1+38) = 95%
PPV = 57/61 = 93.4%
NPV = 38/39 = 97.4%
Data from Staib et al. *(5)*.

1.2.1. Positive Predictive Value and Negative Predictive Value

Although sensitivity and specificity are useful measures for evaluating test validity, they are less helpful from a clinical standpoint where disease status is typically unknown and surgeons are faced with assessing the likelihood of disease given a particular test result. It is in this clinical context that understanding and applying the concepts of the positive predictive value (PPV) and negative predictive value (NPV) of a diagnostic test is essential. In general, the PPV (or NPV) helps clinicians answer the following question: "Given that this test is positive (or negative), what is the probability that this patient actually has (or does not have) the disease?" Similar to sensitivity and specificity, an ideal test would have both a PPV and NPV of 100%; however, tests with such optimal performance characteristics are exceedingly rare in clinical practice.

Turning our attention back to Table 1, the PPV of a test is defined as the proportion of individuals with positive tests that actually have the disease:

$$\mathrm{PPV} = a/(a+b) \tag{4}$$

Correspondingly, the NPV is defined as the proportion of individuals with a negative test result that are actually disease-free:

$$\mathrm{NPV} = d/(c+d) \tag{5}$$

In more general terms, the PPV is the probability that someone with a positive test result actually has the disease. The NPV describes how likely it is that a patient with a negative test result is truly unaffected. Based on these definitions, a general principle is that the number of false-positive and false-negative tests will affect the PPV and NPV, respectively. The study from Staib and colleagues (Table 2) can also serve as a useful example for calculating PPV and NPV. Specifically, the PPV of FDG-PET for detecting recurrent cancer was 57/61 = 93.4%; the corresponding NPV was 38/39 = 97.4% *(5)*.

An important caveat with regard to PPV and NPV is that the predictive value of a test may vary based on several factors, including disease prevalence in the community or study sample and the specificity and sensitivity of a particular test *(1)*. An example from the literature is useful to illustrate this concept *(6)*. Lachs and colleagues evaluated the

Table 3
Urine Dipstick Example Illustrating the Relationship Between Disease Prevalence and Predictive Value Data from Lachs et al. *(6)*

A: UTI Prevalence 7% (Low Prior Probability)

	Urine Culture Positive	*Urine Culture Negative*	
Dipstick positive	10	53	63
Dipstick negative	8	188	196
	18	241	259

Positive predictive value = 10/63 = 16%
Negative predictive value = 188/196 = 96%

B: UTI Prevalence 52% (High Prior Probability)

	Urine Culture Positive	*Urine Culture Negative*	
Dipstick positive	49	29	78
Dipstick negative	4	21	25
	53	50	103

PPV = 49/78 = 63%
NPV = 21/25 = 84%

performance of the rapid dipstick test for urinary tract infections (UTI) in two groups of patients that differed in their prior probability of UTI. The investigators defined patients at high-risk for UTI as those with a high proportion of symptoms (dysuria, urgency, frequency, hematuria, fever) and signs (abdominal and costovertebral angle tenderness) consistent with UTI. Conversely, the same signs and symptoms were significantly less frequent among patients classified as having a low prior probability of infection. As expected, the actual prevalence of UTI, based on urine culture as the diagnostic gold standard, was different for the two groups, with 52% (53/103) of the high-risk patients having a culture-proven UTI vs only 7% (18/259) of low-risk patients *(6)* (Table 3). Based on Table 3, in the sample with a prevalence of 7%, 18 women are affected with a UTI and 241 women are disease-free. However, 63 women in this sample have a positive result on their urine dipstick test, and only 10 of these were true positives. Therefore, in this low prevalence sample, the PPV of a urine dipstick test is only 10/(10+53) = 16% *(6)*.

Using the same urine dipstick test in the sample of women with a higher prevalence of UTI (52%) (Table 3), we see that among the 78 women with positive dipstick tests, 49 are true positives and 29 are false positives; the resulting PPV is 49/78 = 63% *(6)*. Therefore, as the prevalence of disease in the sample being tested increases, the PPV of the test increases as well. Likewise, as the prevalence of a particular disease *decreases*, the NPV *increases* (although, given the rarity of many diseases, this tends to be less dramatic than the association between prevalence and PPV). This correlation between prevalence and predictive value is an important and consistent principle that should be kept in mind when considering the potential applications for a clinical test. Furthermore,

Table 4
Urine Dipstick Example Illustrating the Relationship
Between Test Specificity and Predictive Value

A: UTI Prevalence 7%

	Urine Culture Positive	*Urine Culture Negative*	
Dipstick positive	10	53	63
Dipstick negative	8	188	196
	18	241	259

Sensitivity = 56%
Specificity = 78%
PPV = 10/63 = 16%
NPV = 188/196 = 96%
Data from Lachs et al *(6)*.

B: UTI Prevalence 7%

	Urine Culture Positive	*Urine Culture Negative*	
Dipstick positive	10	12	22
Dipstick negative	8	229	237
	18	241	259

Sensitivity = 56%
Specificity = 95%
PPV = 10/22 = 45%
NPV = 229/237 = 97%
Data based on the results for a hypothetical urine dipstick test applied to the sample for **A** (*see* text) *(6)*.

this relationship provides the rationale for selective implementation of screening tests in populations that are at increased risk for a particular disease *(1,4)*.

Independent of the effect of disease prevalence, changes in the specificity, and, to a lesser degree, the sensitivity, of a particular test will also affect its predictive value. This principle is illustrated with a hypothetical example based on the study from Lachs and associates (Table 4). Suppose that a new rapid urine dipstick test was developed and found to have an improved specificity (but identical sensitivity) when compared with available tests. Suppose also that a subsequent study was undertaken to compare the predictive value of this new urine dipstick with the "conventional" dipstick test employed by Lachs et al. To control for the effect of disease prevalence on predictive value, the two dipstick tests were applied only in low-risk sample of patients (UTI prevalence = 7%). As determined by Lachs et al, the specificity of the "conventional" dipstick test in this sample is 78%; in contrast, the (hypothetical) specificity of the newly available dipstick in the same population is 95% (Table 4). The sensitivity of both tests is 56%. From Table 4, we see that a change in the specificity from 78% to 95% substantially decreases the number of false-positive test results (53 with the "conventional" dipstick vs 12 with the "improved" dipstick). Consequent to this improved specificity, there is a simultaneous improvement in the PPV of the rapid dipstick test from 16% to 45% (Table 4). The key principle in this example is that changes in the specificity of a diagnostic test tend to have a dramatic effect on the predictive values of the test, with increases in specificity increasing the PPV and vice versa. The PPV and NPV of a test will also increase concurrently with increases in the sensitivity of a particular test; however, the effect of sensitivity on predictive value is modest for low prevalence conditions.

Although their derivations are beyond the scope of this introductory chapter, the previously described relationships between predictive value, prevalence, sensitivity and specificity may also be summarized by the following equations (based on Bayes theorem):

$$\text{PPV} = \frac{(\text{sensitivity})(\text{prevalence})}{[(\text{sensitivity})(\text{prevalence}) + (1 - \text{specificity})(1 - \text{prevalence})]} \tag{6}$$

$$\text{NPV} = \frac{(\text{sensitivity})(\text{prevalence})}{[(1 - \text{sensitivity})(\text{prevalence}) + (\text{specificity})(1 - \text{prevalence})]} \tag{7}$$

Based on Equation 6, it is clear that as sensitivity, specificity or prevalence increase, PPV will increase correspondingly. Similar to PPV, increases in NPV will occur in concert with increases in specificity and sensitivity; however, increases in disease prevalence will actually be associated with a lower NPV (Table 4).

1.2.2. Likelihood Ratios

Another method for describing the performance of a diagnostic test is the likelihood ratio (LR). The use of LRs is increasingly common in the medical literature, and a basic understanding of their derivation is useful for clinical researchers in the surgical disciplines. In general, the LR indicates how much a particular test result raises (or lowers) the pretest probability of the disease of interest and provides an alternative method for determining the PPV and NPV. Furthermore, an important advantage of LRs is that, to determine the PPV and NPV, a clinician must only remember one number for a particular test (the LR) rather than having to recall both the sensitivity and specificity. Furthermore, the availability of validated nomograms has greatly enhanced the clinical value and application of this measure of test performance.

A positive LR is defined quantitatively as the probability of a positive test result in patients with the disease of interest divided by the probability of that test result in disease-free individuals *(7)*. Conversely, a negative LR is derived from the probability of a negative test result among healthy individuals divided by the probability of the same result among those affected with the disease of interest. To illustrate this point further, consider the following equations:

$$\begin{array}{l}\text{LR for a}\\ \text{positive test}\end{array} = \frac{\text{Probability (+ test) among diseased individuals}}{\text{probability (+ test) among disease-free individuals}} \tag{8}$$

$$\begin{array}{l}\text{LR for a}\\ \text{negative test}\end{array} = \frac{\text{Probability (– test) among disease-free individuals}}{\text{probability (– test) among diseased individuals}} \tag{9}$$

Recalling our definitions of sensitivity and specificity, equivalent equations for the LR of a positive and negative test, respectively, are:

$$\begin{array}{l}\text{LR for a}\\ \text{positive test}\end{array} = \frac{\text{Sensitivity (true-positive “rate”)}}{1 - \text{specificity (false-positive “rate”)}} \tag{10}$$

$$\text{LR for a negative test} = \frac{\text{Specificity (true-negative “rate”)}}{1 - \text{sensitivity (false-negative “rate”)}} \qquad (11)$$

As previously mentioned, the clinical value of a LR is based on the fact that this information can be combined with pre-test assessment of disease probability to calculate the posttest probability of disease (PPVs or NPVs) *(7)*. Indeed, the LR specifies how much a particular test result increases or decreases the pretest probability of the disease of interest. In practice, the pretest probability of disease is typically estimated by the clinician based on the patient's history and physical examination, as well as adjunctive epidemiologic data and personal experience.

In general, LRs greater than 1 indicate that the test result increases the probability that a patient has the disease of interest. Conversely, LRs less than 1 decrease the probability of the target disorder *(8)*. A LR equal to 1 indicates that the pretest and posttest probabilities of disease are equivalent. Some authorities define likelihood ratios ≥5 or ≤0.2 as being associated with moderate to large shifts in pretest to posttest probability (and therefore having a greater impact on clinical decision making).

In a recent article, McCormick and colleagues applied this concept to the diagnostic evaluation of orthopedic trauma patients *(9)*. In this study, the authors evaluated the accuracy of four different physical exam maneuvers for diagnosing posterior pelvic ring injuries in patients with traumatic pelvic fractures. For each physical examination modality, sensitivity and specificity for the detection of posterior ring injury was determined based on comparison with computed tomography findings (considered the diagnostic gold standard) *(9)*. One of the examination modalities assessed was posterior pelvic palpation, which involves careful palpation of the sacrum and bilateral sacroiliac joints; this diagnostic maneuver was considered positive when local tenderness was noted on examination. When compared with computed tomography scan results, the sensitivity and specificity of posterior pelvic palpation were 98% and 94%, respectively *(9)*. Based on Equation 10, the authors determined that the positive LR for posterior pelvic palpation (for the diagnosis of posterior ring injuries) was 16.3, indicating that this physical examination finding is 16 times more likely to be present in a patient with a posterior ring injury than one without such a lesion. Based on these results, the authors concluded that the positive findings on posterior palpation provide strong evidence in favor of a posterior ring injury and that this test can, therefore, be used to refine and guide the subsequent radiologic evaluation of patients with traumatic pelvic injuries *(9)*. Indeed, applying this concept further, a LR of 16.3 for pain on posterior palpation means that even if the pre-examination probability of a posterior ring fracture is fairly low (based, perhaps, on patient history and mechanism of injury), the presence of this physical exam finding generates a large, and potentially conclusive, change from pre-test to post-test probability of a posterior ring injury *(8, 9)*.

The mechanics by which LRs are used to translate from pretest to posttest disease probability are fairly complex and require a brief review of the concept of the odds of a disease. Statistically, the odds of an event (such as the presence of a disease) may be defined as follows:

$$\text{Disease odds} = \text{disease probability}/1 - \text{disease probability} \qquad (12)$$

After calculating the pretest odds, this statistic may be combined with the LR to calculate the posttest odds of disease (which are much more useful to a clinician than the pretest odds). For a positive test result, the following equation illustrates this point:

$$\text{Posttest disease odds} = \text{pretest disease odds} * \text{positive LR} \tag{13}$$

The posttest disease probability (PPV) may then be determined as follows:

$$\text{Posttest disease probability (PPV)} = \frac{\text{posttest disease odds}}{1 + \text{posttest disease odds}} \tag{14}$$

It should also be noted that the posttest disease probability is mathematically equivalent to the positive predictive value for the diagnostic test. Similar calculations can be performed for negative test results, based on the corresponding negative LR. Recognizing the relative complexity and time requirements of such calculations, sophisticated nomograms have been developed that allow clinicians to move rapidly from pretest (based on clinical data and disease prevalence) to posttest disease probability, thereby facilitating clinical decision making and broadening the applicability of this measure of test performance *(8, 10)*.

2. HOW TO EVALUATE TESTS WITH CONTINUOUS RESULTS

Until now, we have focused on tests with only two possible outcomes (positive or negative). In surgical practice, however, clinicians frequently order and interpret diagnostic tests (e.g., PSA, carcinoembryonic antigen) that have continuous outcomes. In this context, there is no concrete positive or negative test result; rather, a threshold level must be established for the test such that values above this threshold are considered positive and those below the threshold are considered negative. In truth, the choice of cutoff levels can have important implications with regard to the performance of tests with continuous outcome values.

PSA, an important tumor marker for patients with prostate cancer, is an example of a test with continuous outcomes that is widely used in clinical practice. Indeed, the application of PSA as a diagnostic test for prostate cancer serves as a useful illustration of the effects of changes in cutoff levels on the performance of a diagnostic test. Consider, for example, the data in the attached PSA screening dataset, which summarizes serum PSA levels and cancer status for 100 men undergoing screening for adenocarcinoma of the prostate (Table 5). Overall, 40 men have biopsy-confirmed prostate cancer, whereas 60 patients had no evidence of cancer in their biopsy specimen. However, there is no precise PSA threshold that unequivocally separates men with and without prostate cancer; instead, there is overlap of diseased and nondiseased individuals at most levels of PSA. Nonetheless, in clinical practice, a PSA cutoff must be defined such that individuals with values above this level can be referred for additional testing (i.e., transrectal ultrasound-guided prostate biopsy), whereas those with PSA values below the threshold are spared further workup.

The most widely accepted cutoff for a normal PSA level is 4.0 ng/mL *(11)*. Based on this threshold, the PSA screening dataset (combined with Table 1 as a reference) can be used to estimate the sensitivity and specificity of PSA (as a diagnostic test for prostate cancer). In this example, the calculated sensitivity is 87.5% (35/40 cancers detected) and the specificity is 25% (PSA <4.0 for 15/60 men without prostate cancer). Some urologists contend that a PSA cutoff of 4.0 has an unacceptably low sensitivity and, therefore, application of this threshold fails to detect a significant number of men with important prostate cancers (in other words, this cutoff is associated with an unacceptably high false-negative rate) *(12, 13)*. As a result, some authorities have advocated a lowering of the

Table 5
Summary of Prostate-Specific Antigen Screening Dataset Format

Patient Number	*Prostate-Specific Antigen Level (mg/dL)*	*Cancer Status (0 = No cancer, 1 = Cancer)*
1	7.2	1
2	6.7	0
3	1.4	0
4	8.2	0
5	0.7	0
6	10	1
7	5.5	0
8	2.5	1
9	5.7	1
10	8.5	0
…	…	…
91	2	0
92	5.1	0
93	5.4	0
94	4.8	0
95	6.9	0
96	4.6	0
97	7.2	0
98	9.7	1
99	4.1	1
100	11.3	0

threshold for a positive result to 2.5 ng/mL *(12)*. In the PSA screening dataset, lowering the PSA threshold to 2.5 ng/mL would increase the sensitivity of this test to 95.0%; however, the specificity would decrease to 21.7% because of an increased number of false-positive test results. In this setting, we see that very few men with prostate cancer would be undiagnosed (2/40); however, a concurrent effect of changing this threshold is that a large number of men without prostate cancer (47/60) will now be, unnecessarily, subjected to additional invasive diagnostic tests (i.e., a prostate biopsy).

In contrast, an inverse effect is seen when a higher threshold is applied. For instance, if clinical practice was changed such that a higher PSA cutoff level (i.e., 10 ng/mL) was implemented, many men that actually have prostate cancer would not be referred for additional workup, and their cancer would likely remain undiagnosed. At the same time, however, very few disease-free men would be subjected to needless additional testing. In the PSA screening dataset, the net effect of choosing 10 ng/mL as the PSA cut point is a decrease in the sensitivity of this test to 25.0% (10/40 cancers detected), with a simultaneous increase in the specificity to 85.0% (PSA <10.0 for 51/60 men without prostate cancer). In fact, sensitivity and specificity will always vary in an inverse fashion when the "normal" threshold changes for a diagnostic test with continuous results (Table 6).

As illustrated by this example, the choice of cutoff levels can dramatically affect the performance (sensitivity, specificity, and accuracy) of a diagnostic test with continuous outcome values. In general, lowering the cut point will increase the sensitivity, while simultaneously decreasing the specificity. Conversely, raising the cutoff level will gen-

Table 6
Summary of the Effect of Different PSA Cut Points on Its Performance as a Diagnostic Test for Prostate Cancer (Based on the PSA Screening Dataset)

PSA Cut Point (ng/mL)	*Sensitivity (True-Positive "Rate")*	*Specificity*	*1 – Specificity (False-Positive "Rate")*	*# True Positives*	*# False Positives*
2.5	95.0%	21.7%	78.3%	38	47
4.0	87.5%	25.0%	75.0%	35	45
10.0	25.0%	85.0%	15.0%	10	9

erally improve specificity at the expense of sensitivity (Table 6). Clinically, the most salient effect of this principle is that changes in cutoff levels will result in a variable number of false-negative or false-positive test results (Table 6). Accordingly, the choice of an optimal threshold depends on the relative balance between the adverse effects of false positive versus false negative test results. In the case of PSA testing, regardless of the specific threshold applied, two groups of patients of patients will be identified: (1) those with "positive" results that will be referred for biopsy and (2) those with "negative" results that will be spared further testing. In this example, if a low PSA threshold is chosen (resulting in excellent sensitivity but many false positives), then many men will be referred for additional testing that is not only expensive, but also carries a risk of unnecessary morbidity. On the other hand, if a high threshold is chosen, many men that actually have prostate cancer will be inappropriately reassured and their (potentially curable) cancer may remain undetected. Ultimately, for continuous tests, the choice of a clinical threshold depends on the relative significance (e.g., morbidity, cost, availability of effective treatment) of false-positive and false-negative results for the disease of interest.

2.1. Optimizing the Diagnostic Threshold for Continuous Tests Using Receiver Operating Characteristic Curves

As described in the previous section, when test values are measured on a continuum, the sensitivity and specificity of a test will vary based on the position of the cutoff between "positive" and "negative" values. An efficient method for displaying the effects of different cut points on test performance is a receiver operating characteristic (ROC) curve. ROC curves were first developed and used in the engineering and communication fields; currently, they are widely employed as a valid and reliable approach to assessing and comparing the accuracy of various diagnostic tests *(14)*.

In the most general sense, an ROC curve is a plot of the true-positive rate (sensitivity) vs the false-positive rate (1-specificity) for a range of diagnostic test thresholds. The PSA Screening Dataset used earlier in this chapter can be reformulated to determine the true positive and false-positive rates for each of the previously mentioned cutoffs (Table 6). Plotting the true-positive rate vs the false-positive rate (for each PSA threshold) generates an ROC curve for PSA as a diagnostic test (Figure 1); this plot graphically demonstrates the tradeoff between sensitivity and specificity that results from changing the cut point of a diagnostic test. Specifically, as the PSA cut point shifts from 2.5 to 4 and then from 4 to 10, you can see the concurrent decrease in sensitivity and increase in specificity. It

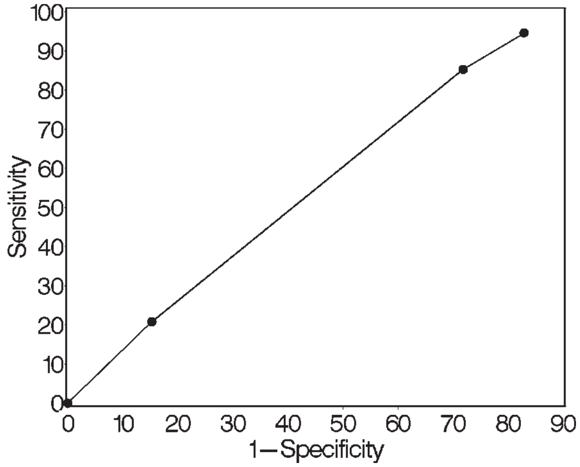

Figure 1: Receiver operating characteristic curve based on three prostate-specific antigen cut points (2.5, 4.0, 10.0 ng/mL) (from PSA Screening Dataset).

is important to recognize, however, that only three PSA cut points were used to generate the ROC curve in Figure 1; an idealized ROC curve for this example would be based on an infinite number of PSA thresholds and would have a (more typical) smoother appearance of the ROC curve in Figure 2.

There are several important caveats with regard to the interpretation of an ROC curve. First, the accuracy of diagnostic test can be assessed visually by examining the proximity of the ROC curve to the upper left-hand corner of the graph. An ROC curve for a "perfect" test would fill the entire area of the ROC space. Specifically, the closer the curve follows the upper left corner of the ROC space, the more accurate the test *(7)*. This makes sense because an ROC curve that approaches the upper left-hand corner of the graph reflects a test that achieves a high true-positive rate (sensitivity) while maintaining a low false-positive rate (1-specificity). Conversely, an ROC curve that approaches a 45° diagonal through the ROC space is a poorly performing test that does little to distinguish individuals with and without the disease of interest. In addition to visual inspection of an ROC curve, a more precise assessment of the accuracy of a test may be also obtained by measuring the area under the ROC curve.

As previously mentioned, the accuracy of a diagnostic test reflects how well the test distinguishes diseased from disease-free individuals. In the case of ROC curves, the most precise measurement of accuracy is the area under the curve; an area of 1 signifies a perfect test, while an area of 0.5 (represented by a 45° diagonal through the ROC space) indicates a poorly performing clinical test (e.g. the test performs no better than chance alone in terms of distinguishing between diseased and disease-free individuals). A useful way to conceptualize the meaning of this numeric value (area under an ROC curve) is to recognize that the area under the curve measures the discrimination of a particular test *(15)*. In other words, the area under the curve reflects the ability of a test to correctly classify individuals with and without the disease of interest. Continuing with our PSA example, consider a situation where the disease status is known for two different groups of men – one of the groups is comprised of men with prostate cancer (untreated) and the other group includes only men that are cancer-free. Suppose that one patient is randomly

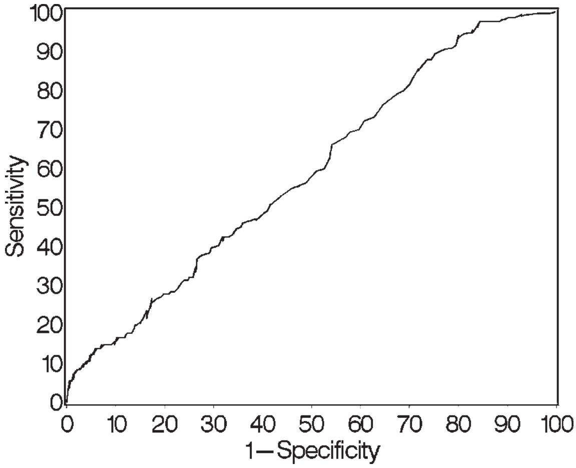

Figure 2: Idealized receiver operating characteristic curve–based PSA Screening Dataset.

selected from each group (e.g. one man with and one man without prostate cancer) and a PSA level is determined for each patient. If PSA is a useful diagnostic test, we presume that its value will be higher in the man with prostate cancer. Indeed, the area under the ROC curve (for PSA) is a numerical description of the percentage of times that this is true; more specifically, the area under the curve represents the percentage of randomly drawn pairs (cancer/cancer-free) for which the test of interest (i.e. PSA) correctly classifies the disease status of the two individuals in the random pair *(15)*.

Formal calculation of the area under an ROC curve is mathematically complex and almost exclusively performed by computer software. A comprehensive explanation of this methodology is beyond the scope of this chapter; however, suffice it to say that both non-parametric (trapezoidal rule) and parametric (maximum likelihood technique) techniques can be used to estimate both the area under the curve and its standard error *(15, 16)*. The point estimates for the area under the curve provide the basis for various statistical tests that assess whether or not two ROC curves are significantly different *(16)*. Although a detailed description is beyond the scope of this chapter, a common method for statistical comparison of ROC curves is to first calculate the area under each curve; the areas are then tested for statistically significant differences using a modification of the Wilcoxon rank-sum test *(7)*. A final caveat worth noting for ROC curves is that they are a function of disease prevalence like any other assessments of test performance such that using an identical assay, one can develop vastly different ROC curves in low prevalence and high prevalence populations.

3. SCREENING TESTS

No discussion of diagnostic test validity would be complete without considering the implications of test performance as they relate to the implementation and efficacy of disease screening programs. Screening tests (such a PSA, mammography and colonoscopy) are used to identify asymptomatic individuals with early-stage, potentially curable disease. In general, screening tests aim to classify individuals with regard to their probability of disease, rather than establishing a definitive diagnosis. The ultimate goal of screening is to alter the prognosis of a given condition by identifying patients in an early phase of the disease, thereby allowing the timely institution of effective therapy. For a screening program to be worthwhile and effective, the disease of interest (and screening test) must fulfill a number of criteria including: 1) the disease must be common and an important health problem; 2) the natural history of the disease should be well-defined and there should be an identifiable latent or presymptomatic stage; 3) if left untreated, the disease must be accompanied by significant morbidity or mortality; 4) there must be an accepted and effective treatment for patients with the disease and there must be some benefit, in terms of morbidity and/or mortality, when the disease is treated in the presymptomatic versus the symptomatic stage; 5) there must be a suitable screening test that is generally acceptable to the population; 6) the cost of screening (including diagnosis and treatment of diagnosed patients) must not be excessive relative to the overall costs of medical care; and 7) screening must be a continuous process and not a "one-time" event. For most widely available screening tests, including mammography, Pap smears and PSA testing, most, but not all, of these criteria are fulfilled *(1,4,17–25)*.

In cases where an available screening test fulfills most of the above criteria, there are several potential benefits to screening programs. For instance, effective screening programs (coupled with appropriate follow-up testing and intervention) may improve the prognosis for treated cases. In addition, by detecting disease in its earliest (and presumably most treatable) stage, there is a potential for a reduction in treatment-related morbidity among screen-detected cases. Furthermore, assuming that an accurate test is available, screening programs can provide reassurance to individuals with a negative test result. Finally, when appropriately implemented, screening programs can serve as a cost-effective use of health resources *(17,19–21,23,25,26)*.

However, there are also several potential disadvantages that must be considered when assessing the relative merits of a screening test. First, screening efforts that employ a test with limited accuracy can result in unnecessary morbidity and anxiety for individuals with false positive results, as well as false reassurance for diseased patients that test negative *(17,27)*. Furthermore, there is often concern that screening programs are implemented in the absence of data that supports their ability to alter disease prognosis *(18)*. Indeed, the true effectiveness of a screening test can only be established by expensive and time-consuming randomized, controlled trials that are designed to evaluate meaningful end points such as morbidity and mortality. In the absence of such data, interpretation of the effectiveness of screening programs can be obscured by bias and confounding and, in fact, the question of whether or not current screening programs (including PSA testing) have been successful in altering the natural history of the disease or improving outcomes for patients remains controversial *(18,24)*. Another potential limitation of screening programs may be a lack of consensus regarding the optimal treatment of patients diagnosed with early disease of uncertain prognosis. Finally, the relative economic and human resources devoted to screening programs may be excessive when considered in the context of widespread population based screening efforts.

As mentioned previously, assessments of the relative value of screening programs may be limited by several sources of bias that frequently plague such evaluations. One source of bias that must be considered is patient-selection bias. Specifically, the results of screening programs may be biased by the presence of systematic differences between individuals that voluntarily participate in a screening test or program and those that choose not to participate. Factors that may contribute to selection bias include significant differences (between participants and nonparticipants) in the following characteristics: baseline health status and sociodemographic characteristics, history of screening, and distribution of risk factors that predict future incidence and mortality from the disease of interest. Once again, systematic differences (between participants and nonparticipants) in one or more of these areas may irreparably bias the interpretation of screening test effectiveness.

Two other sources of bias that often occur in the context of screening programs are lead-time bias and length-time bias. Lead time is defined as the period of time between diagnosis with a screening test and the time when the disease would have been otherwise diagnosed based on various signs and symptoms that prompt medical attention. For a given disease and screening test, the duration of lead time depends on both the biology of the disease and the ability of the screening test to truly detect early disease. Lead-time bias occurs if early diagnosis (screen-detection) results in patients living longer with a disease without ultimately affecting mortality because of the disease. With lead-time bias, the apparent improvement in survival occurs only because of a shift in the date of diagnosis, and intervention produces no real prolongation of life. When evaluating a screening program, avoidance of lead-time bias can be achieved by random assignment of individuals to screening and control groups. Furthermore, rather than comparing survival rates from the time of diagnosis, the effects of lead-time bias can also be reduced by comparing age- and disease-specific mortality rates among screened and control individuals, which are independent of the time since detection.

Length-bias sampling (or length-time bias) refers to the tendency of screening programs to preferentially detect more slowly progressive disease. This occurs because aggressive conditions (such as highly malignant tumors) typically produce symptoms early in the course of the disease and are, therefore, primarily identified by routine diagnostic procedures rather than screening tests. Length-time bias occurs when there is an impression of improved survival because of screening, based solely on the preferential detection of slowly progressive disease. Analogous to lead-time bias, length-time bias may be reduced by repeated screening examinations as often occur in an randomized, controlled trials. In sum, it is crucial to consider the potential for selection, lead-time, and length-time bias when assessing the value of any screening program.

4. CONCLUSIONS

This chapter describes the most salient issues relating to the validity of diagnostic tests and their application to screening programs. It is important to recognize that sensitivity and specificity are generally fixed for a test with a dichotomous outcome; in contrast, sensitivity and specificity will vary based on different cutoff levels for tests with continuous outcomes. NPV and PPV are arguably the most useful measures for clinicians, given that disease status is generally unknown prior to performance of a particular test. The PPV and NPV of a test may vary based on disease prevalence in the sample being studied, as well as changes in the specificity and sensitivity of a particular test. ROC curves are a useful method for further assessing the validity of tests with continuous outcomes. By and large, these statistics are determined by straightforward calculations and should be established for all diagnostic tests. An appreciation of these measures of test performance will allow the surgeon to critically assess the value of both proposed and established disease screening programs.

REFERENCES

1. Gordis L. Epidemiology. 2nd ed. Philadelphia: W.B. Saunders Company, 2000.
2. Hanno PM, Landis JR, Matthews-Cook Y, Kusek J, Nyberg L Jr. The diagnosis of interstitial cystitis revisited: lessons learned from the National Institutes of Health Interstitial Cystitis Database study [comment]. J Urol 1999;161(2):553–557.
3. Kusek JW, Nyberg LM. The epidemiology of interstitial cystitis: is it time to expand our definition? [review]. Urology 2001;57(6:Suppl. 1):Suppl-9.
4. Hulley SB, Cummings SR, Browner WS, et al. Designing clinical research: an epidemiologic approach. 2nd ed. Baltimore: Lippincott Williams & Wilkins, 2001.
5. Staib L, Schirrmeister H, Reske SN, Beger HG. Is (18)F-fluorodeoxyglucose positron emission tomography in recurrent colorectal cancer a contribution to surgical decision making? Am J Surg 2000;180(1):1–5.
6. Lachs MS, Nachamkin I, Edelstein PH, et al. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection [comment]. Ann Intern Med 1992;117(2):135–140.
7. Dawson-Sanders B. TRG. Basic and clinical biostatistics. 2nd ed. Norwalk, CT: Appleton & Lange, 1994.
8. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271(9):703–707.
9. McCormick JP, Morgan SJ, Smith WR. Clinical effectiveness of the physical examination in diagnosis of posterior pelvic ring injuries. J Orthopaed Trauma 2003;17(4):257–261.
10. Fagan TJ. Letter: nomogram for Bayes theorem. N Engl J Med 1975;293(5):257.
11. Arcangeli CG, Ornstein DK, Keetch DW, Andriole GL. Prostate-specific antigen as a screening test for prostate cancer. The United States experience [review]. Urol Clin N Am 1997;24(2):299–306,.
12. Punglia RS, D'Amico AV, Catalona WJ, et al. Effect of verification bias on screening for prostate cancer by measurement of prostate-specific antigen [comment]. N Engl J Med 2003;349(4):335–342.
13. Catalona WJ, Smith DS, Ornstein DK. Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng/mL and benign prostate examination. Enhancement of specificity with free PSA measurements [comment]. JAMA 1997;277(18):1452–1455.
14. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer [review]. J Natl Cancer Inst 2003;95(7):511–515.
15. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29–36.
16. McNeil BJ, Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decision Making 1984;4(2):137–150.
17. Goldstein MM, Messing EM. Prostate and bladder cancer screening [review]. J Am Coll Surg 1998;186(1):63–74.
18. Harris R, Lohr KN. Screening for prostate cancer: an update of the evidence for the U.S. Preventive Services Task Force [review]. Ann Int Med 2002;137(11):917–929.
19. Lindfors KK, Rosenquist CJ. The cost-effectiveness of mammographic screening strategies [comment] [erratum appears in JAMA 1996;Jan 10;275(2):112]. JAMA 1995;274(11):881–884.
20. Mahadevia PJ, Fleisher LA, Frick KD, et al. Lung cancer screening with helical computed tomography in older adult smokers: a decision and cost-effectiveness analysis [comment]. JAMA 2003;289(3): 313–322.
21. Marks D, Thorogood M, Neil HA, Wonderling D, Humphries SE. Comparing costs and benefits over a 10 year period of strategies for familial hypercholesterolaemia screening. J Public Health Med 2003;25(1):47–52.
22. McGrath JS, Ponich TP, Gregor JC. Screening for colorectal cancer: the cost to find an advanced adenoma. Am J Gastroenterol 2002;97(11):2902–2907.
23. Pignone M, Saha S, Hoerger T, Mandelblatt J. Cost-effectiveness analyses of colorectal cancer screening: a systematic review for the U.S. Preventive Services Task Force [summary for patients in Ann Intern Med 2002;Jul 16:137(2):I38; PMID 12118986] [review]. Ann Intern Med 2002;137(2):96–104.
24. Smith DS, Catalona WJ, Herschman JD. Longitudinal screening for prostate cancer with prostate-specific antigen [comment]. JAMA 1996;276(16):1309–1315.

25. van Valkengoed IG, Postma MJ, Morre SA, et al. Cost effectiveness analysis of a population based screening programme for asymptomatic Chlamydia trachomatis infections in women by means of home obtained urine specimens.[comment]. Sex Transm Infect 2001;77(4):276–282.
26. McGrath JS, Ponich TP, Gregor JC. Screening for colorectal cancer: the cost to find an advanced adenoma. Am J Gastroenterol 2002;97(11):2902–2907.
27. Harris R, Lohr KN. Screening for prostate cancer: an update of the evidence for the U.S. Preventive Services Task Force [review]. Ann Intern Med 2002;137(11):917–929.

Appendix 1
Equations for the Assessment of Clinical Test Performance

$$\text{Sensitivity} = \frac{\text{number true positive test results}}{\text{number diseased individuals}}$$

$$\text{Specificity} = \frac{\text{number false positive test results}}{\text{number disease-free individuals}}$$

$$\text{Accuracy} = \frac{\text{(number true positive test results + number true negative test results)}}{\text{number disease-free individuals}}$$

$$\text{Positive predictive value} = \frac{\text{number true positives}}{\text{total number positive test results}}$$

$$\text{Negative predictive value} = \frac{\text{number true negatives}}{\text{total number negative test results}}$$

$$\begin{matrix}\text{LR for a}\\ \text{positive test}\end{matrix} = \frac{\text{probability (+ test) among diseased individuals}}{\text{probability (+ test) among disease-free individuals}}$$

or

$$\begin{matrix}\text{LR for a}\\ \text{positive test}\end{matrix} = \frac{\text{sensitivity (true-positive “rate”)}}{\text{1 – specificity (false-positive “rate”)}}$$

$$\begin{matrix}\text{LR for a}\\ \text{negative test}\end{matrix} = \frac{\text{probability (– test) among disease-free individuals}}{\text{probability (– test) among diseased individuals}}$$

or

$$\begin{matrix}\text{LR for a}\\ \text{negative test}\end{matrix} = \frac{\text{specificity (true-negative “rate”)}}{\text{1 – specificity (false-negative “rate”)}}$$