ROC Curve Analysis in Assessing The Usefulness Of Noninvasive Vascular Tests*

Kenneth Ouriel,** M.D., Anne E. McDonnell,** B.S., and Christopher K. Zarins,** M.D.

ABSTRACT  Diagnostic tests are imperfect tools, and occasionally healthy individuals will be classified as diseased, while diseased individuals will frequently not be detected as abnormal. The frequency of these errors defines the diagnostic value of the test. Receiver operating characteristic (ROC) curve analysis provides a statistical means for evaluating diagnostic tests. A cutoff point for determining an abnormal test result may be selected using the principles of cost-benefit analysis in conjunction with the ROC curve, and the diagnostic usefulness of two tests may be compared by statistically testing the difference between their corresponding ROC curves. The use of ROC curve analysis was illustrated by comparing two forms of stress testing, treadmill exercise and postocclusive reactive hyperemia (PORH), in the diagnosis of arterial occlusive disease. Treadmill exercise was of greater diagnostic value (p < .02), and we now employ treadmill exercise rather than PORH when stress testing is desired.

The evaluation of diagnostic tests must be an ongoing process in the noninvasive vascular laboratory. There are numerous methods for assessing a diagnostic test, and all involve comparisons between the results of the test and the results of another test, a so-called "gold-standard" with well proven validity. Perhaps the simplest measure of diagnostic decision quality is the "accuracy" of the test, expressed as the percentage of test results that agree with the results obtained by the gold standard. However, accuracy can be misleading because it depends on the prevalence of a disease in the population under study. Therefore, "sensitivity" and "specificity" are required, since they are independent of disease prevalence. These two indices are more meaningful than accuracy, but they do not provide a unique description of diagnostic performance because they depend on the selection of an arbitrary decision threshold, or cutoff point, between results that are to be considered abnormal and normal. The Receiver Operating Characteristic (ROC) curve avoids these limitations by providing a description of all combinations of true positive and false positive frequencies, rather than the frequencies at only one specific decision threshold.1,2

The purpose of this study was to illustrate the use of ROC curve analysis using an example: the comparison between treadmill exercise (TE) and postocclusive reactive hyperemia (PORH) for the diagnosis of peripheral arterial occlusive disease.

Materials and Methods

During the three year period ending July, 1980, 218 consecutive patients with the clinical diagnosis of lower extremity peripheral vascular disease were evaluated by stress testing in the vascular noninvasive laboratory at the University of Chicago. Patients deemed suitable for treadmill exercise were tested in this manner (100 patients), while patients unfit for exercise (previous MI, angina, arrhythmia, amputation, CVA) underwent PORH testing (118 patients). Arteriographic evaluation of the aorta, iliac, and lower limb vessels was done in 70 of the patients (33%). Sensitivity and specificity were calculated and ROC curve analysis was performed using the data from these 70 patients.

In addition, we studied 25 normal controls with TE and PORH. In an effort to assure disease free vascu-
lature, all control subjects were less than 30 years of age and without history of smoking, diabetes, or trauma.

Treadmill exercise was performed at 1.5 mph with a 7% grade. Patients were instructed to walk as long as possible, to a maximum of ten minutes. At the termination of exercise, ankle and brachial Doppler systolic pressure measurements were made, and the minimum poststress ankle index was calculated as the lowest ankle/brachial pressure ratio following the termination of stress. Ankle and brachial pressures

---

* Presented at the Fourth Annual Scientific Session of the Society of Non-Invasive Vascular Technology, Dallas, Texas, June 8-10, 1981.
** From the Department of Surgery, The University of Chicago, Chicago, Illinois.
were measured during reactive hyperemia after four minutes of calf compression using a below knee cuff inflated to 200 mmHg. Serial measurements were made following release of the cuff and the minimum poststress ankle index was determined.

The ROC curves were constructed by choosing seven different decision thresholds, or cutoff values. For each cutoff point, the percent false positive (x) and percent true positive (y) ratios were calculated, yielding seven (x,y) pairs. A computer program was implemented to fit an ROC curve to these seven points. One curve was plotted for each test to be evaluated, and the higher curve represented the better test. Using a test for the significance of observed difference between ROC curves, a probability value was generated for rejecting the null hypothesis that the tests were of equivalent diagnostic value.3

Results

The poststress minimum ankle indices after TE and PORH are illustrated in Figure 1. The mean index after TE was significantly less in the diseased limbs (0.36 ± .02) than in the normal limbs (1.01 ± .01, p < .001). PORH also produced a lower poststress minimum ankle index in the diseased limbs (0.33 ± .02) than in the normals (0.81 ± .02, p < .001).

The ROC curves for TE and PORH were constructed by plotting the percent false positive results versus the percent true positive results (Fig. 2). When the assumptions were made that the added cost of a false negative test result was equal to the added cost of a false positive test result, and that the prevalence of disease in the population undergoing stress testing was 80 percent, the most appropriate poststress minimum index in defining the lower limit of normal was 0.85 for TE and 0.68 for PORH. Using these cutoff points, the percentage of abnormal limbs documented by arteriography that had abnormal stress tests was calculated, and a sensitivity of 99 percent for TE and 92 percent for PORH was obtained. When the number of control limbs with poststress indices below the cutoff points were determined, the specificity of TE was 95 percent and PORH, 91 percent. The differences between the ROC curves were statistically compared, and it was found that TE was diagnostically more useful than PORH in discriminating the normal from the diseased state (X² = 8.05, p < .02).

Discussion

Diagnostic tests are imperfect tools, and occasionally healthy individuals will be classified as diseased, while diseased individuals will frequently not be detected as abnormal. The frequency of these errors determines the diagnostic usefulness of the test. In order to critically evaluate a diagnostic test it is useful to arrange a 2 x 2 decision matrix such as Table 1. Four possible diagnostic outcomes are displayed in the matrix: the number of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) results.4

Accuracy is the simplest measure of diagnostic value, and it is calculated as the total number of correct tests (TP + TN) divided by the total number of tests results (TP + TN + FP + FN), but its usefulness is limited by the fact that it is highly dependent on the prevalence of disease in the population under study. For example, suppose PORH is used and a normal test result is defined as one with a minimum poststress ankle index of greater than 0.20. Virtually every truly normal limb will have a minimum poststress index above 0.20, however, many diseased limbs will also have indices above this level. When the test is used as a screening procedure in a population with a low prevalence of disease, the prior probability of encountering a truly diseased limb is low.
The number of false negative results will be high when compared to the total number of results. The apparent accuracy of the test will be high. In contrast, when the test is used in a population where the prevalence of disease is high, the number of truly diseased limbs with indices greater than 0.20 (false negative results) will be large, revealing the inaccuracy of using PORH with a cutoff of 0.20.

The limitations of accuracy as a measure of decision performance requires the introduction of the concepts of sensitivity and specificity. Sensitivity is a measure of the ability of a test to discover disease when disease is present and is equal to TP divided by (TP + FN). Specificity is an index of the ability of a test to define the absence of disease when disease is not present and is equal to TN divided by (TN + FP). Sensitivity and specificity are more meaningful than accuracy because they are not dependent on disease prevalence. Although using PORH with a cutoff of 0.20 as a screening procedure is associated with a high degree of accuracy, the uselessness of this testing paradigm would have been uncovered through the determination of sensitivity.

Sensitivity and specificity are more useful parameters than accuracy, but they rely on arbitrarily chosen decision thresholds. Consider the example of PORH when an index of 0.20 was used as the decision threshold. The sensitivity of this test is high and the specificity is low. However, if a decision threshold of 1.05 were used, the sensitivity would be low and the specificity high. Clearly, by varying the decision threshold from low levels to high levels, an infinite number of sensitivities and specificities will be generated. The ROC curve illustrates this continuum of sensitivity/specificity trade-offs by plotting the way in which the true positive ratio (sensitivity) and the false positive ratio (1-specificity) change as the decision threshold is increased. Since a good test is one with high true positive ratios for the corresponding false positive ratios, the higher ROC curve represents the better test. While it is difficult to determine statistical significance between two sensitivities or specificities, a test has been developed to compare two ROC curves, calculate a chi-square value and generate a probability value for rejecting the null hypothesis that the tests are of equal diagnostic value.

Once an ROC curve is generated, one may arrive at a rational cutoff point to use in future decisions regarding the presence of disease using a cost-benefit analysis (Fig. 3). If it is more costly to improperly classify diseased individuals as normal than it is to classify normals as diseased, a strict threshold low on the ROC curve is appropriate. If the reverse is true, one should use a lax threshold and a point high on the curve. The disease prevalence is also of importance when picking a cutoff point. As illustrated by the example of PORH, in a population with a high prevalence of disease a lax threshold is desired in order to maximize the number of correct diagnoses. Conversely, if the diagnostic test is to be used in a population with a low disease prevalence, for example, in a screening program, then a strict threshold is appropriate.

Once the disease prevalence and the added costs of diagnostic errors have been estimated, one may determine the optimal cutoff point by finding the point...
on the ROC curve where the slope is equal to:

$$\frac{1 - \text{prevalence of disease}}{\text{prevalence of disease}} \times \frac{\text{added cost of a false positive result}}{\text{added cost of a false negative result}}$$

The costs may be based on mortality, morbidity, or finances. In many cases, one does not have accurate estimates of the health or monetary costs associated with errors in diagnosis. In these instances it is appropriate to assume that the cost of a false positive result equals the cost of a false negative result, as was done in the comparison of TE and PORH.

In summary, we believe that vascular laboratory tests are best assessed through the use of ROC curves. In this way one may rationally choose a decision threshold to use in separating normal from abnormal, and two diagnostic tests may be compared and a probability calculated for selecting the more useful test. These concepts have been illustrated by comparing treadmill exercise and PORH in the diagnosis of peripheral vascular disease. Appropriate cutoff points for defining the lower limit of normal were selected based on principals of decision analysis, and treadmill exercise was found to be of greater diagnostic use than PORH. We now employ TE rather than PORH when stress testing is desired, except in patients with contraindications for treadmill exercise.

References