Criterion Validity (Accuracy)

Criterion validity (accuracy) refers to how well a measure, such as the classification of some test results indicating, for example, having a disease or not, matches a phenomenon that the test is intended to capture, such as the actual sickness or health of examinees. Two distinct aspects of accuracy are *sensitivity* and *specificity*. Table below gives definitions of sensitivity, specificity, and other key terms relevant to measuring the accuracy of a test/classification model. It also shows the quantitative relationships among the terms.

Terms Relevant to Measuring the Accuracy of a Test

The table below shows the four possible combinations of actual true sick status and test results.

	True Condition		
Test Result	Positive (sick)	Negative (healthy)	Total
Positive (testing sick)	a (true positive)	b (false positive)	a+b
Negative (testing healthy)	c (false negative)	d (true negative)	c+d
Total (N)	a+c	b+d	a+b+c+d

Sensitivity - The proportion of truly positive (sick) cases that give positive results on the test (a/[a+c])., i.e. the conditional probability of a true-positive test or the true-positive proportion (TP).

False negative probability - The proportion of truly positive cases that give negative results on the test (c/[a+c]). i.e. the conditional probability of a false-negative test (FN) and is the complement of *sensitivity.*

Specificity - The proportion of truly negative (healthy) cases that give negative results on the test (d/[b+d]), i.e. the conditional probability of a true-negative test (TN).

False positive probability - The proportion of truly negative cases that give positive results on the test (b/[b+d]), i.e. the conditional probability of a false-positive test (FP) and is the complement of *specificity.*

Three terms use test results as a reference point and reveal how well the test results indicate the true conditions (see text for further discussion).

Positive predictive value (PPV) - The predictive value of a positive test, that is, the percentage of positive tests that are correct (a / [a+b]).

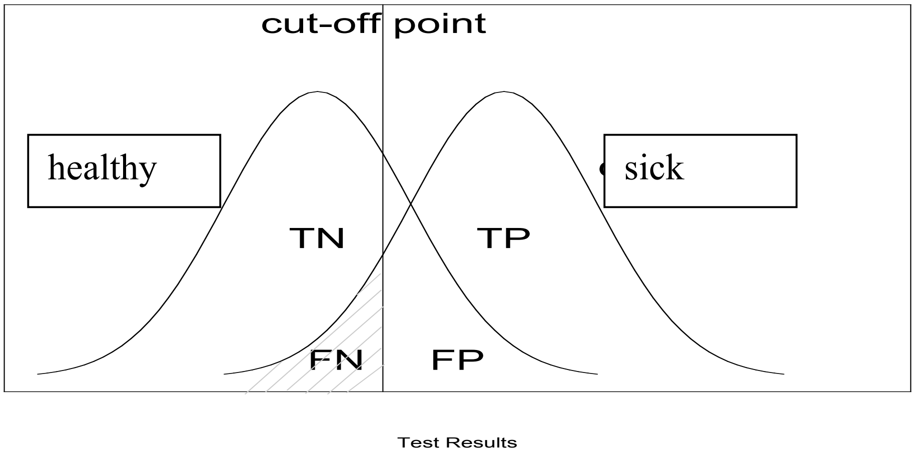
Negative predictive value(NPV) - The predictive value of a negative test, that is, the percentage of negative tests that are correct (d / [c+d]).

False positive index(FPI) - Number of false positives for each true positive (b/a).

The Receiver Operating Characteristic Curve (ROC)

The Receiver Operating Characteristic Curve (ROC) is a standard technique for summarizing classifier performance over a range of tradeoffs between true positive (TP) and false positive (FP) error rates (Sweets, 1988). ROC curves and their analyses are based on statistical decision theory; they were originally developed for electronic signal–detection theory. They have been applied in many medical and nonmedical endeavors, including studies of human perception and decision making.

The position of the ROC on the graph reflects the accuracy of the diagnostic test, independent of any decision threshold(s) that may be used. It covers all possible thresholds, with one point on the curve reflecting the performance of the diagnostic test for each possible threshold (cut-off point), expressed in terms of the proportions of true and false positive and negative results for each threshold. (see Figure below). The ROC of random guessing lies on the diagonal line. The ROC of a perfect diagnostic technique is a point at the upper left corner of the graph, where the TP proportion is 1.0 and the FP proportion is 0.



The Area Under the Curve (AUC), also refered to as index of accuracy (A), is an accepted traditional performance metric for a ROC curve, and it is the proportion of the unit area of the graph that lies under the ROC. Its possible range is from 0.50 at the “chance” diagonal to 1.0 for perfection. The curve would be higher for diagnostic techniques that provide greater separations of the distributions (i.e., higher accuracy) and lower for techniques that provide lesser separations (i.e., lower accuracy). A=0.8 can be interpreted to mean that a randomly selected individual from the positive group has a test value larger than that for a randomly chosen individual from the negative group in 80 percent of the time The area can be estimated with parametric or nonparametric techniques. Parametric methodology refers to inference (MLEs) based on the bivariate normal distribution When this assumption is true, the MLE is unbiased. Nonparametric refers to inference based on the trapezoidal rule (which is equal to the Wilcoxon estimate of the area under the ROC curve). Nonparametric estimates of the area under the ROC curve (AUC) tend to underestimate the "smooth curve" area (i.e., parametric estimates), but this bias is negligible for continuous data.