A Concise Introduction to Receiver Operating Characteristic (ROC) Curve Analysis:

ROC curve analysis is used widely in medicine as a method for evaluating the performance of diagnostic tests (3,5,6,10), but has been used recently in many agricultural applications (2,4,5,11,12). The ROC curve provides information regarding how often a test's predictions are correct, and provides a graphical method for evaluating and discriminating between different diagnostic tests or modifications of the same test (12). To perform the analysis, one starts with a diagnostic test that produces a range of values or test scores (T) in which a classification is decided. Decisions are arrived at by comparing the diagnostic's output to a threshold value (T_{thresh}). The data (i.e., individuals or test subjects) are then partitioned into two groups: the 'cases'—in which the disease is known to occur; and the 'controls'—in which disease was absent. Those test subjects with test scores above the threshold ($T>T_{thresh}$) are classified as diseased (D+), and those with test scores equal to or below the threshold ($T\#T_{thresh}$) are classified as not diseased (D-), irrespective of their true disease status.

For various reasons, diagnostic tests are not perfect predictors of disease. This can be depicted graphically as two overlapping distributions of threshold values; the cases and controls (Figure 1)(5,10). Thus, any decision threshold based on this test yields one of four possible decisions: 1) true positive (TP), in which a 'case' was correctly classified as diseased; 2) true negative (TN), in which a 'control' was correctly classified as not diseased; 3) false positive (FP), in which a 'control' was incorrectly classified as diseased; and 4) false negative (FN), in which a 'case' was incorrectly classified as not diseased (6,7).

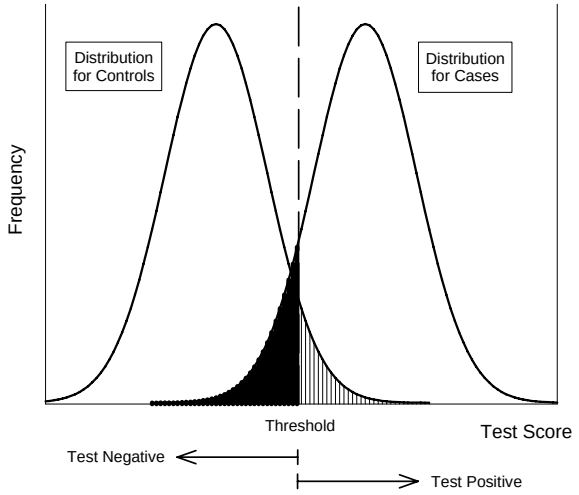

Figure 1. Gaussian curves describing the distribution of a test score for diseased plants ('cases') and healthy plants ('controls'). The unshaded are to the left of the threshold are the true negatives (TN), the unshaded area to the right of the threshold are the true positives (TP). The blackened area shows the false negatives (FN) cases and the hatched area shows the false positive (FP) cases.

Now, say N individuals (test subjects) were classified under the rules of the diagnostic test, X of the N subjects were classified as 'cases', and Y of these as 'controls'. Then the true positive proportion (TPP) is the number of true positive decisions divided by the total number of cases. TPP is referred to as the '*sensitivity*' of the test. The true negative proportion (TNP) is the number of true negatives divided by the number of controls. This is referred to as the '*specificity*' of the test. The false positive proportion (FPP) is the number of false positives divided by the number of controls. The false negative proportion (FNP) is the number of false negatives divided by the number of cases. In probabilistic terms, TPP is an estimate of Prob($T>T_{thresh} \mid D+$) (read as 'the probability of a test score above the threshold, given the presence of disease). Similarly, FNP is an estimate of

Prob($T\#T_{thresh}$ | $D+$), FPP is an estimate of Prob($T>T_{thresh}$ | $D!$), and TNP is an estimate of Prob($T\#T_{thresh}$ | $D!$).

The ROC curve is a plot of TPP (sensitivity) versus FPP (1-specificity) over the full range of possible threshold scores (T)(Figure 2)(4,6,7). The curve passes through the point (0,0), corresponding to the highest threshold value; a value that classifies all individuals as disease free. Under this scenario, no subject would be treated for the disease, i.e., no true positives are identified but, also, no false positives would be declared. The plot also passes through the point (1,1) which corresponds to a threshold of zero and the diagnostic test classifying every individual as diseased. Under this scenario, every subject is treated for the disease, thus, all true positives will be identified, however, at the expense of numerous false positives.

An ROC curve that passes through the point (0,1) shows that the test has both desirable sensitivity and specificity characteristics. Thus, the threshold value corresponding to the point closest to the point (0,1), would be considered the best threshold. The straight line joining the points (0,0) and (1,1) is the 'no discrimination' line (Figure 2). If an ROC curve falls along this line, the test does not discriminate between cases and controls. An intuitive measure of the performance or accuracy of a forecaster would be to simply calculate the proportion of correct decisions, i.e., the ([TP +TN]/N). However, sensitivity and specificity represent two kinds of accuracy, respectively, for cases and controls (4). Unlike the calculations of accuracy from the proportions of correct decisions, and ROC curve analysis does not depend on the proportions of cases and controls in a data set, because sensitivity and specificity are calculated independent of these proportions (6)."

In addition to the specificity and sensitivity of a test, one can calculate the 'the positive predictive value' of a test (10). This statistic represents the probability that an the subject is diseased given that the test predicted disease [=Prob($D+$ | $T>T_{thresh}$)]. It is possible for a test to have high specificity and sensitivity yet have a low positive predictive value when disease prevalence is low. Similarly, 'the negative predictive value' [=Prob($D!$ | $T\#T_{thresh}$)], the probability that a tests negative result is an indicator of disease, decreases when disease prevalence is high, thus offsetting the gains in positive predictive value. Bayes theorem allows one to calculate the positive predictive value from the sensitivity and specificity, however, requires an estimate of that the 'true' level of the disease in the test population (3). The positive predictive value is calculated using:

$$\Pr(T > T_{thresh}) = \frac{\Pr(T > T_{thresh} \mid D+) \cdot \Pr(D+)}{\Pr(T > T_{thresh} \mid D+) \cdot \Pr(D+) + \Pr(T > T_{thresh} \mid D-) \cdot \Pr(D-)} \qquad (3)$$

Yuen et al. (12) has provided a simplified version of this expression based on the use of likelihoods and odds ratios.

When choosing between two different tests (e.g., the open and closed symbols in figure 2), the area under the ROC curve can be used a discriminating function. The area under the ROC curve measures the probability that in randomly paired subjects, one from the cases and one from the controls, the test will correctly classify them. Bamber (1) recognized that this probability is equal to the well-known Wilcoxon rank-sum statistic, W or, equivalently, the Mann-Whitney U-statistic. This finding allows one to evaluate the performance of a diagnostic test in statistical framework through testing of the hypothesis H_0: AUROC=0.5 (i.e., the test does not provide discrimination any better than chance (represented by the 'no discrimination line').

The area under the ROC curve (AUROC curve) and its standard error can be calculated for each according to the methods of Hanley and McNeil (3). In short, the calculations are based on the generalization that the AUROC curve is derived by dividing the Mann-Whitney *U*-statistic by the product of the two sample sizes. The Mann-Whitney *U*-statistic is defined as $\sum_{i=1}^{m}\sum_{j=1}^{n} U_{ij}$, where U_{ij} = 1, ½ , or 0 if the j^{th} case is larger than, equal to, or smaller than the i^{th} control, respectively, *m* and *n* are the number of cases and controls, respectively, and *i* and *j* are indexing variables (8). The distribution of the Mann-Whitney *U*-statistic tends towards normality as the sample size increases, permitting the derivation of a *z* statistic and a test of the hypothesis that the AUROC curve is no different than the area under the line of 'no discrimination', i.e., 0.5. The *z* statistic here is calculated with a continuity correction factor: $(|U - E(U)| - 0.5) / \sqrt{\mathrm{Var}(U)}$, where $E(U) = mn/2$ and $\mathrm{Var}(U) = (mn/12)\cdot[(m + n + 1) - \sum_i (t_i^3 - t_i)/(m+n)(m+n-1)]$ is the ties-corrected variance of *U* (t_i is the number of tied values in the i^{th} set of ties). For a two-sided test the *p*-value is calculated as $2\cdot\Pr(|Z| > z)$. A 95% confidence interval about the AUROC curve can be calculated according to method 5 of Newcombe (9). This can be done using the spreadsheet 'generalisedmw1.xls' provided at the author's website (http://www.cardiff.ac.uk/medicine/epidemiology_statistics/research/statistics/newcombe).

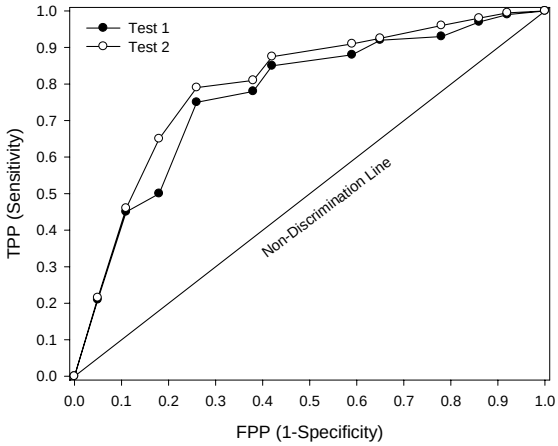

Figure 2. ROC curves for two competing algorithms. Each symbol represents a value for a different threshold

Literature cited:

1. Bamber, D. 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12:387-415.
2. Dewdney, M.M., Biggs, A.R., and Turechek, W.W. 2007. A statistical comparison of the reliability of the blossom blight forecasts of *MARYBLYT* and *Cougarblight* with receiver operating characteristic (ROC) curve analysis. Phytopathology 97:1164-1176.
3. Hanley, J.A., and McNeil, B.J. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29-36.
4. Hughes, G., and Madden, L. V. 2003. Evaluating predictive models with application in regulatory policy for invasive weeds. Agricultural Systems 76:755-774.
5. Hughes, G., McRoberts, N., and Burnett, F.J. 1999. Decision-making and diagnosis in diseases management. Plant Pathology 48:147-153.
6. Metz, C.E. 1978. Basic principles of ROC analysis. Nuclear Medicine 8:283-298.
7. Murtaugh, P.A. 1996. The statistical evaluation of ecological indicators. Ecological Applications 6:132-139.

8. Newcombe, R. G. 2006a. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. Statistics in Medicine 25:543-557.
9. Newcombe, R. G. 2006b. Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation. Statistics in Medicine 25:559-573.
10. Schulzer, M. 1994. Diagnostic tests — a statistical review. Muscle and Nerve 17:815-819.
11. Turechek, W.W., and Wilcox, W.F. 2005. Evaluating predictors of apple scab with receiver operating characteristic curve analysis. Phytopathology 95:679-691.
12. Yuen, J. 2006. Deriving decision rules. *The Plant Health Instructor*. DOI: 10.1094/PHI-A-2006-0517-01