The inclusion of Receiver Operating Characteristic (ROC) analysis as an integral component of Disease Management (DM)

Elbiely M. Mona
Assistant professor
E-mail: elbielym@yahoo.com

Abstract

The Receiver-Operating Characteristic (ROC) analysis has been long used in Signal Detection Theory to depict the tradeoff between hit rates and false alarm rates of classifiers. In the last years, ROC analysis has become largely used in the medical community for visualizing and analyzing the performance of diagnostic tests.

The receiver operating characteristic (ROC) curve is a popular method for characterizing the accuracy of diagnostic tests when test results are not binary. Diagnostic or predictive accuracy concerns are common in all phases of disease management (DM) program.. This article introduces the concepts of a receiver operating characteristic (ROC) Curve, the construction of an ROC graph, how the ROC curve can be used to select optimal cut-off points for a test result. Furthermore, this article introduces the receiver operator characteristic (ROC) analysis as a more appropriate and useful technique for assessing a diagnostic and predictive accuracy in DM. Its advantages include: testing accuracy across the entire range of scores, easily examined visual and statistical comparisons across tests or scores, and independence from outcome prevalence. Therefore the implementation of ROC as an evaluation tool should be strongly considered in the various phases of a DM programme. We illustrate our approach with a prospective study was undertaken with cervicovaginal samples collected from 700 pregnant women at a sample of cervicovaginal secretions obtained during pelvic examinations performed every 2 weeks between 24-36 weeks' gestation. Level of Beta- human Chorionic Gonadotropin ß-hCG in Cervicovaginal were measured by radioimmunoassay.

Keywords: receiver operating characteristics (ROC), ROC analysis, area under the curve (AUC), diagnostic performance, sensitivity, specificity, disease management (DM) program.

Introduction

Disease management (DM) is a system of coordinated interventions aimed to improve patient self management as well as increase doctors' adherence to evidence-based practice guidelines. The assumption is that by augmenting the traditional episodic medical care system with services and support between doctor visits, the overall cost of health care can be reduced [1].

Diagnostic or predictive accuracy concerns are common in all phases of a DM programme, and ultimately play an influential role in the assessment of programme effectiveness. For example,(1) accurate identification of diseased patients is essential for programme inclusion. Most programmes rely on medical and pharmacy claims data for this purpose. (2) Predictive models are typically used as an initial stratification tool to forecast patients' use of services in future periods. These tools also typically rely on past claims experience and are thereby limited in both the accuracy of those data as well as the statistical model used for the prediction. (3) During the initial patients contact, the DM nurse typically performs an assessment of the patient's disease severity level to determine the intensity of DM services that will be required. Misclassification may result in the patient receiving either too much or too little on-going attention. (4) Throughout the programme intervention, accuracy is needed in assessing a patient's level of self-management and their progression across the stages of behavioral. [2]

According to the World Health Organization, preterm delivery is defined as that occurring at less than 37 completed weeks of gestation or less than 259 days from the first day of the last menstrual period. A number of clinical and biochemical markers have been proposed for prediction of preterm delivery. Many investigators have tried to recognize tools that might lead to early and correct identification of preterm labor, or risk of preterm delivery.

Human chorionic Gonadotropin (hCG), in both maternal serum and amniotic fluid are probably the result of direct hCG diffusion from the placenta throughout pregnancy. It has been suggested that the hCG level in vaginal fluid was a useful marker of premature rupture of membranes (PROM) hCG in cervical secretions was found in high concentration until 20 weeks' gestation; but after 20 weeks of gestation, it remained at a stable level, where the median levels of hCG were between 5.6 and 7. mIU/ml. These findings suggested that hCG level in cervical secretion had been concordant with the levels in maternal sera and amniotic fluid. [3]

THE ROC CURVE

The ROC curve is a graphical technique for assessing the ability of a test to discriminate between those with disease and those without disease [4]. ROC curves allow visual analyses of the trade-offs between the sensitivity and the specificity of a test with regard to the various cut-offs that may be used. The curve is obtained by calculating the sensitivity and specificity of the test at every possible cut-off point, and plotting sensitivity against 1-specificity.

Description of the ROC curve

A typical ROC curve is shown in Figure 1. By convention, sensitivity (the proportion of true positive results) is shown on the y axis, going from 0 to 1 (0–100%) and 1-specificity (the proportion of false positive results) is shown on the x axis, going from 0 to 1 (0–100%) [5].

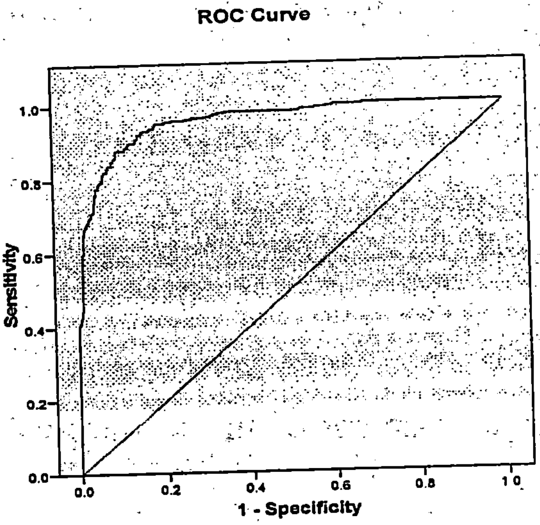

Fig. 1. A typical ROC curve

As shown in Figure 1, the diagonal line on the graph going from the lower left-hand corner (0, 0) to the upper right hand corner (1, 1) serves as a reference line, and represents the characteristics of a test which is completely useless at differentiating between those with disease and those without disease. Points along this line indicate that the test detects an equal number of true and false positives, that is it does not discriminate between those with disease and those without disease[6]. A test that perfectly discriminates between diseased and non-diseased patients would yield a 'curve' that coincided with the left and top sides of the plot [7]. In practice, however, it is unusual to have such a curve, and ROC curves usually lie between these extremes.

Uses of the ROC curve

The ROC curve may be used for three purposes:

1. It allows the determination of the cut-off point at which optimal sensitivity and specificity are achieved.
2. It allows an assessment of the diagnostic accuracy of a test and
3. It allows the comparison of the usefulness of two or more diagnostic tests.

Measures related to ROC curves

The dichotomous decision process is based on a threshold value "V" (cutoff point) which classifies the scores of a continuous variable “Y” (also called “classifier”) into two categories: positive vs. negative. If $Y \geq V$, the subject will be classified as positive; if $Y < V$, the subject will be classified as negative[8].

let us assume that we have a valid procedure of discriminating between the presence and the absence of a disorder (also called “valid diagnosis”), and we differentiate two groups of individuals: with and without a certain disorder. By also administering the test that assesses the value of the Y variable in each subject, we obtain two distributions of Y scores, one for each group. The distribution of test results will overlap, as shown in Figure1.

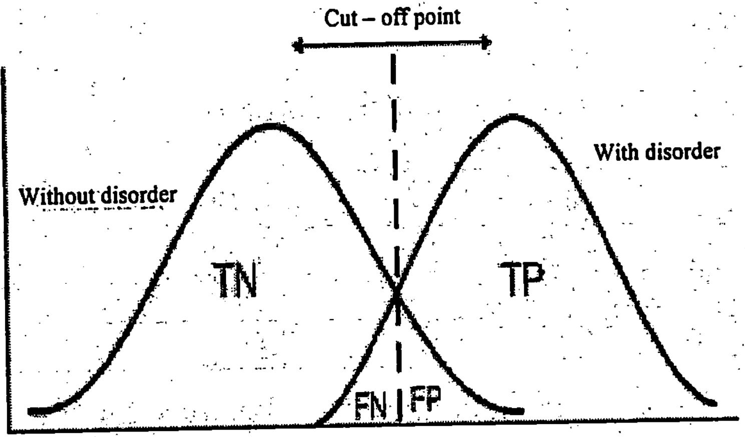

Fig. 2. Four possible outcomes when intersecting a „valid diagnosis” with a „classifier”

The intersection of the “valid diagnosis” with the “classifier” generates four possible outcomes. If the valid diagnosis is positive and it is correctly classified as positive, the outcome is counted as a true positive (TP); if the same outcome is incorrectly classified as negative, it is counted as a false negative (FN). If the valid diagnosis is negative and it is correctly classified as negative, the outcome is counted as a true negative (TN); if the same outcome is incorrectly classified as positive, it is counted as a false positive (FP) [9].

The outcomes identified in Figure 1 can be conceptualized in a contingency table that we will call "decision matrix" [10]. Such an approach is presented in table 1.

Table 1. The decision matrix diagnosis

Test	Disease	No Disease	Total
Positive test	TP	FP	T^+
Negative test	FN	TN	T^-
Total	D^+	D^-	N

In the “decision matrix”, besides the four outcomes described above, we have included the following values: the total number of subjects who have the disorder (D+), the total number of subjects who do not have the disorder (D-), the total number of subjects with a positive test result (T+), the total number of subjects with a negative test result (T-) and the total number of subjects analyzed (n) [11, 12]. It is important to note that in research situations we have a decision matrix for each possible cut-off point. From the outcomes described in the "decision matrix", we can calculate the following measures (metrics):

(1) Sensitivity = TP/D+

(2) Specificity = TN/D-

(3) Positive likelihood ratio = Sensitivity / (1-Specificity)

(4) Negative likelihood ratio = (1-Sensitivity) /Specificity

(5) Positive predictive value = TP/T+

(6) Negative predictive value = TN/T-

(7) Accuracy = (TP+TN)/n

Sensitivity, also called the true positive rate (when expressed as a percentage) is defined as the probability that a test result will be positive when the disorder is present. Specificity, also called the true negative rate (when expressed as a percentage), represents the probability that a test result will be negative when the disorder is not present. These two indicators are essential for ROC curves analysis.

The positive likelihood ratio, is the ratio between the probability of a positive test result given the presence of the disorder and the probability of a positive test result given the absence of the disorder. Similarly, the negative likelihood ratio is defined as the ratio between the probability of a negative test result given the presence of the disorder and the probability of a negative test result given the absence of the disorder. Positive predictive value, also called precision, is defined as the probability that the disorder is present when the result of the test is positive, while the negative predictive value is defined as the probability that the disorder is not present when the result of the test is negative.

The last indicator presented here is the diagnostic accuracy of a test, or the clinical performance of a test. It can be described in terms of diagnostic accuracy, or the ability to correctly classify subjects into clinically relevant subgroups. Diagnostic accuracy refers to the quality of the information provided by the classification device

Other measures that are worth mentioning here, even though they are not actually used in the following analysis, are prevalence and pre-test and post-test odds. The prevalence (D+/n) refers to the proportion of cases exhibiting the disorder; the pre-test odds (prevalence/1-prevalence) refers to the odds that the patient suffers from the target disorder before the test is carried out, while the post-test odds (pre-test odds* Positive likelihood ratio) reflects the odds that the patient suffers from the target disorder after the test is carried out [13].

Determining the optimal cut-off point

A perfect medical test would have 100% sensitivity and 100% specificity, and such a test will identify all people with disease and all those without disease. The point on the ROC curve which corresponds to this perfect scenario (100% sensitivity and 100% specificity) would be at the upper left-hand corner (0, 1). In practice, however, few tests are perfect, and one has to strike a balance between sensitivity and specificity. Generally speaking, the closer the ROC curve gets to the upper left-hand corner (0, 1), the better the test is at discriminating between cases and non-cases[14].

Two methods for identifying optimal cut-off points using sensitivity, specificity and the ROC curve are commonly used [15]. The first method assumes that the best cut-off point for balancing the sensitivity and specificity of a test is the point on the curve closest to the (0, 1) point. In this method, optimal sensitivity and specificity are defined as those yielding the minimal value for

$$(1 - \text{sensitivity})^2 + (1 - \text{specificity})^2. \quad (1)$$

The cut-off point corresponding to these sensitivity and specificity values is the one closest to the (0, 1) point and is taken to be the cut-off point that best differentiates between people with disease and those without disease[16].

The second method that may be used to determine the optimal cut-off point for a test is the Youden index (*J*) [17]. *J* is defined as the maximum vertical distance between the ROC curve and the diagonal or chance line and is calculated as

$$J = \text{maximum } \{\text{sensitivity} + \text{specificity} - 1\}. \quad (2)$$

Using this measure, the cut-off point on the ROC curve which corresponds to J, that is, at which (sensitivity + specificity − 1) is maximized, is taken to be the optimal cut-off point. An intuitive interpretation of J is that it corresponds to the point on the curve farthest from chance [18].

Ideally, one would want to have a test which is both highly sensitive and highly specific but this is not always possible. As earlier discussed, when the cut-off point between normal and abnormal is changed to increase either sensitivity or specificity, there is usually a concomitant decrease in the other.

In general, when it is very important not to miss a diagnosis (for instance, when there is a disease with high mortality but for which a cure is available), a test with high sensitivity is needed. On the other hand, when the consequences of having a false positive test is very serious (e.g. the psychological problems associated with falsely diagnosing a person to have HIV), a test with a high specificity is important.

Assessing diagnostic accuracy

The ROC curve is also important because the area under the curve (AUC) is a reflection of how good the test is at distinguishing between patients with disease and those without disease. The AUC serves as a single measure, independent of prevalence, that summarizes the discriminative ability of a test across the full range of cut-offs[19]. The greater the AUC, the better the test.

A perfect test (as described earlier) will have an AUC of 1.0, while a completely useless test (one whose curve falls on the diagonal line) has an AUC of 0.5. The AUC of many tests used in clinical practice fall between these two values. In general the closer the AUC is closer to 1, the better the overall diagnostic performance of the test, and the closer it is to 0.5, the poorer the test.

The interpretation of the AUC of a test is the following: the AUC is the probability that the test yielded a higher value for a randomly chosen individual suffering from the disorder than for a randomly chosen individual not suffering from the disorder [20].

Going back to AUC utility in determining the ability of a test to discriminate between groups, in interpreting AUC, Streiner and Cairney [21] show that the accuracy of tests with AUC between 0.50 and 0.70 is low; an accuracy between 0.70 and 0.90 is moderate, while an AUC over 0.90 indicates high accuracy.

CONFIDENCE INTERVALS

It should be noted that all measures of diagnostic accuracy including the AUC are statistical estimates and should be reported with confidence intervals [22]. In this study the AUC of the ROC curve for BHCG is reported to be 0.956 with a 95% confidence interval (95% CI) of 0.942–0.970 (Fig. 1). The 95% CI inform the reader about the interval in which 95% of all estimates of AUC will fall if the study was repeated over and over again. In other words, one can be 95% certain that the true value of the AUC of the ROC curve for Beta- human chorionic Gonadotropin lies between 0.942 and 0.970 [23].

Results

In analyzing our results, we first checked if the AUC is significantly different from the area under the diagonal determined of a random test. Table 2 presents the results of this analysis for the β-HCG.

Table 2. The AUC for β-HCG scale

Area under the ROC curve (AUC)	0.956
Standard error	0.007
95% Confidence interval	0.942 to 0.970
z statistic	23.338
Significance level P (Area=0.5)	0.000

As Table 2 shows, the Z test performed with MedCalc indicates a significant difference from the random area, with a probability of error smaller than 1% (Z=23.33, p<.01). The same table shows an AUC for the β-HCG. Scale of .956, which, according to Streiner and Cairney [24] indicates, a high discriminative capacity of the β-hCG scale.

Table 3. Some indicators of the efficiency of the medical division rely on β-hCG

Sensitivity	92.03%	88.88% to 94.52%
Specificity	82.96%	78.31% to 86.97%
Positive Likelihood Ratio	5.40	4.22 to 6.91
Negative Likelihood Ratio	0.10	0.07 to 0.14
Disease prevalence	55.57%	51.80% to 59.29%
Positive Predictive Value	87.10%	83.47% to 90.19%
Negative Predictive Value	89.27%	85.12% to 92.59%

Table 4. Criterion values and coordinates of the ROC curve for β-hCG.

Criterion	Sensitivity	1- Specificity	Specificity
2	1.000	1.000	.000
3	1.000	0.997	.003
6	1.000	0.841	.159
17	0.990	0.269	.731
20	0.974	0.151	.849
21	0.935	0.115	.885
22	0.897	0.095	.905
23	0.881	0.072	.928
24	0.835	0.056	.944
25	0.800	0.041	.959
26	0.755	0.028	.972
27	0.719	0.021	.979
28	0.677	0.005	.995
57	0.000	0.000	1.000

The sensitivity and specificity of a test depend on the level that has been chosen as the cut-off point for normal or abnormal, that is they depend on the definition of what constitutes an abnormal test. The sensitivity and specificity at various β-hCG cut-offs is shown in Table 4. You can see from this table that the sensitivity and specificity of the β-hCG change according to what level was taken as the cut-off for normal or abnormal. For example, the sensitivity and specificity at a cut-off of 6 was 100% and 16%, respectively, while at a cut-off of 17, sensitivity and specificity were, respectively, 99% and 73%. It is also clear from Table 4 that as the cut-off level of normal is increased, the sensitivity of the test decreases while specificity increases. This illustrates an important point: sensitivity and specificity are inversely related according to the choice of cut-off value. When increasing values of a test result are associated with disease, higher cut-off values are generally associated with lower sensitivities and higher specificities, while lower cut-offs are associated with higher sensitivities and lower specificities [25]. Thus changing the cut-off point to try and increase the sensitivity or specificity of a test will result in a reduction of the other. This relationship between sensitivity and specificity has important implications. First, for any diagnostic test, we would like to select a cut-off value such that the optimal sensitivity and specificity are achieved, that is the cut-off point

at which the test is most useful in helping to make the diagnosis.

Second, it is obvious that sensitivity and specificity at a single cut-off value do not describe the test's performance at other potential cut-off values. Third, the selected cut-off value should be taken into account when comparing different diagnostic tests. One way of addressing all these issues is to use the ROC curve.

DISCUSSION AND CONCLUSIONS

ROC analysis used in this article illustrates how this procedure works in a real clinical research situation, and consequently, the results must be interpreted in this particular context. However, our findings replicate previous studies on β-hCG.

The sensitivity and specificity of clinical tests whose results are quantitative vary according to what cut-off point is chosen to define normal or abnormal. The ROC curve allows analyses of the trade-offs between sensitivity and specificity at all possible cut-off points. The curve may be used to select optimal cut-off values for a test result, to assess the diagnostic accuracy of a test.

The utility of ROC analysis is demonstrated as a means of assessing accuracy in the programmatic domains of any DM programme. There are several reasons why DM programme evaluators should consider using ROC analysis in lieu of the more conventional methods. First, it thoroughly investigates model or test accuracy across the entire range of scores. Second, it allows for visual examination of scores on one curve or a comparison of two or more curves using a similar metric. Third, the calculations for establishing the sensitivity and 1-specificity coordinates for individual decision thresholds are not especially complicated; an exhaustive iterative process is required to determine all points along the ROC continuum.

REFERENCES

1- **Altman DG, Bland JM.**(1994) Diagnostic tests 3: receiver operating characteristic plots, British Medical Journal, 109- 188.

2- **Brown I, Shaw T, Wittlake WA.** (2005) Does leucocytosis identify bacterial infections in febrile neonates presenting to the emergency department? Emerg Med J, 22: 256–9.

3- **Brown, C. D., & Davis, H. T.** (2006) Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Systems, 80: 24-38.

4- **Crichton N.** (2002) Information point: receiver operating characteristic (ROC) curves. J Cin Nurs, 11: 136.

5- **Fischer JE, Bachman LM, Jaeschke R.** (2003) A readers' guide to the interpretation of diagnostic test properties: clinical example of sepsis. Intensive Care Med, 29: 1043–51.

6- **Greiner M, Pfeiffer D, Smith RD.** (2000) Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med, 45: 23–41.

7- **Hanley JA, McNeil BJ.** (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(l):29-36.

8- **Hanley, J.A., and McNeil, B.** (1982) The meaning and use of the area under the receiver operating characteristic (ROC) curve, *Radiology*, **143**: 29-36.

9- **Hardie A.D.** (2009) Identifying abnormal iron deposition in chronic liver disease using a quantitative MRJ technique (T2* decay): a roc study. Journal of gastroenterology. Vol. 8.

10- **Lasko TA, Bhagwat JG, Zou HH, et al.** (2005) The use of receiver operating characteristic curves in biomedical informatics. J Biomed Inform, 38(5):404-415.

11- **Last J. M.** (2001) A dictionary of epidemiology, New York: Oxford University Press.

12- **Linden A., Adams J. & Roberts N.** (2004) The generalizability of disease management program results: getting from here to there. Managed Care Interface 7: 38–45.

13- **Molodianovitch, K., Faraggi, D., and Reiser, B.** (2006). Comparing the areas under two correlated ROC curves: Parametric and non-parametric approaches. Biometrical Journal 48: 745–757.

14- **Obuchowski NA, Lieber ML.** (1998) Confidence intervals for the receiver operating characteristic area in studies with small samples. Academic Radiology, 5:561–71.

15- **Obuchowski, N. A.** (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. Statistics in Medicine 25: 481–493.

16- **Pepe, M. S., Longton, G., Anderson, G. L., Schummer, M.** (2001). Selecting differentially expressed genes from microarray experiments. Biometrics 59:133–142.

17- **Pepe, S. M., and Cai, T.** (2001) Semi-parametric receiver operating characteristic analysis to evaluate biomarkers for disease, *Statistics in Medicine*, **15**: 361-387.

18- **Perkins NJ, Schisterman EF.** (2006) The inconsistency of 'optimal' cutpoints obtained using two criteria based on the receiver operating characteristics curve. Am J Epidemiol, 163: 670–5.

19- **Qing L.U, Yuehua C., Yuehua C.,** (2010) Bagging optimal ROC curve method for predictive genetic tests, with an application for Rheumatoid Arthritis. Journal of Biopharmaceutical statistics, Vol. 20:401-414.

20- **Schisterman EF, Faraggi D, Reiser B, Trevisan M.** (2001) Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. Am J Epidemiol, 154: 174–9.

21- **Streiner, D. L., Cairney, J. (2007).** What's under the ROC? An introduction to Receiver Operating Characteristics Curves. The Canadian Journal of Psychiatry, 52: 121- 128.

22- **Washington DC.** (2004) Definition of Disease Management, Retrieved: June, 23 ,2009, from: http://www.dmaa.org/definition.htm.

23- **Wray. N.R., Michael. E.** (2010) The genetic interpretation of area under the ROC curve in Genomic profiling. PLOS Genetics, Vol. 6: 1-9.

24- **Zhou XH, Obuchowski NA, McClish DK** (2002) Statistical methods in diagnostic medicine. Wiley, New York

25- **Zhou, X. H., Castelluccio, P., and Zhou, C.** (2005). Nonparametric estimation of ROC curves in the absence of a gold standard. Biometrics 61: 600–609.