ROC Curves in medical decision

Ana Cristina Braga, Lino Costa and Pedro Oliveira

Abstract The accurate medical diagnostic of a disease condition is fundamental for a correct medical decision. Disease screening programs are based, in general, in diagnostic tests which provide a binary response: a subject is classified as positive, if the test result is above a given threshold, and negative, otherwise. Therefore, false positive and false negative classifications can be generated. The performance of test can be evaluated by *ROC* curves which defined, for a given threshold, the compromise between Sensitivity and Specificity, i.e., the True and False Positive fractions. In this work, we address the problem of comparing two diagnostic systems where the corresponding *ROC* curves cross each other. A methodology is developed providing a graphical display that identifies the regions where one curve is superior to the other, with the corresponding Sensitivity and Specificity regions.

1 Introduction

The primary step for the treatment of a disease is its detection. Thus, accurate medical diagnostic of a disease condition is fundamental, specially in the case where the detection of the disease in its early stages can improve the success of possible treatments. Therefore, the disease has to be correctly identified, in general, through the available information obtained from diagnostic tests. In particular, screening populations in order to detect in its early stages diseases as breast cancer, colo-rectal cancer, PKU (phenilketonuria), AIDS, Pap smear and so many more, have become a widespread practice in most health systems. These tests are applied in a large

Ana Cristina Braga
Universidade do Minho, Braga, Portugal, e-mail: acb@dps.uminho.pt

Lino Costa
Universidade do Minho, Braga, Portugal, e-mail: lac@dps.uminho.pt

Pedro Oliveira
Universidade do Porto, Porto, Portugal, e-mail: pnoliveira@icbas.up.pt

scale, with a positive cost benefit relation and, if possible, should be non invasive. Thus, through these widespread screening tests, positive results signal patients to be followed with medical diagnostic. Thus, screening can be viewed as diagnostic test and, in this sense, approached by statistical modeling. There are diagnostic tests that produce a binary result, positive or negative, and so, the subject has or does not have the disease. For instance, the diagnostic of genetic diseases is done by the presence or absence of a specific gene, thus producing a simple Yes or No response. In this work, a continuous or ordinal response is considered and, in general, a threshold is used to classify cases as positive, if above the given threshold, and negative, otherwise. The question to be answered is how to set this threshold so that misclassification is minimized, i.e., positive cases classified as negative and negative cases classified as positive and, in the presence of two alternative diagnostic tests, how to determine the one that performs better.

2 Classification

2.1 Classification for binary tests

In the case of a binary response, the variable D, represents the subject status, and X the result of the diagnostic test,

$$D = \begin{cases} 1 & \text{disease} \\ 0 & \text{non}-\text{disease} \end{cases} \qquad X = \begin{cases} 1 & \text{positive for disease} \\ 0 & \text{negative for disease} \end{cases}$$

Therefore, there are four classification possibilities for a given test result, which are presented in Table 1.

Table 1 Classification results

	$D=0$	$D=1$
$X=0$	True negative	False negative
$X=1$	False positive	True positive

Thus, according to the test results, the False Positive FPF, the True Positive TPF, True Negative TNF and False Negative FNF can be defined,

$$FPF = P[X=1|D=0] \quad TNF = P[X=0|D=0]$$

$$TPF = P[X=1|D=1] \quad FNF = P[X=0|D=1]$$

where the pair (FPF, TPF) provides the probabilities of the errors,

$$TPF + FNF = 1 \qquad TNF + FPF = 1$$

These fractions are, in general, presented as Sensitivity (TPF) and Specificity $(1-FPF)$. Sensitivity can be perceived as the capacity of the diagnostic test to detect the disease in a given subject, whereas Specificity can be interpreted as the capacity of the test to exclude individuals without the disease. Therefore, for any diagnostic test the aim is to have a high Sensitivity and a high Specificity.

2.2 Classification for non-binary tests

The previous section presented the classification of test results in the case of a binary response. However, there are many tests where the response is not binary, and can be presented on a continuous or ordinal scale. Examples of such tests measured on a continuous scale are the biomarkers for cancer such as PSA and CA 125, the diagnostic of kidney disease through the level of creatine, the presence of a heart condition based on the total cholesterol level or the readings of blood pressure; on the other hand, a ordinal scale can be used in the classification of radiologic images (present, possible, absent), or in the classification of clinical symptoms (severe, moderate, mild, not present).

Without loss of generality, it will be considered that larger values of X are indicative of the presence of the disease. In order to classify the results, a dichotomous decision rule will be used based on a given threshold. This rule underlies the medical decision, which supports the decision to treat or not to treat a particular patient. For instance, in the case of the biomarker for prostate cancer the threshold is 4.0 ng/mL of serum PSA and for total cholesterol is 190 mg/dL. Fixing a giving threshold has to take into consideration the consequences of the decision. In the case of AIDS, wrongly classifying an individual as diseased or failing to detect that the individual has the disease, has serious implications, for the individual in the first place, but also for the society at large. Furthermore, other factors must be taken into consideration such as the costs of the test and the treatment of the disease in a more advanced status, the severity of the disease and the characteristics of treatment (for instance, surgery required or not).

The evaluation of the performance of diagnostic tests, the definition of a threshold, has been studied quite a long time and firstly in other areas such as signal detection theory and psychology. It should be mentioned the pioneering works of Fechner (1801–1887), Thurstone (1887–1955) Blackwell (1940) and later the contributions of [10, 11, 7, 6, 2, 9].

2.3 The ROC curve

For a continuous decision variable X, as the threshold varies, different pairs of TPF and FPF are defined. The projection of these pairs on a plane defines a curve, denoted as Receiver Operating Characteristic (*ROC*) Curve. Let us consider as c the threshold, or cut-off value, which classifies the patient as Positive if $X \geq c$ and Negative if $X < c$. Therefore, the True (TPF) and False (FPF) Positive Fractions depend on the threshold c,

$$TPF(c) = P(X \geq c|D = 1) \quad FPF(c) = P(X \geq c|D = 0)$$

For a given c, each pair (FPF, TPF) defines a point on the *ROC* curve which can be defined as

$$ROC(.) = \{(FPF(c), TPF(c)), c \in (-\infty, +\infty)\}.$$

The most usual performance measure is the area under the curve AUC,

$$AUC = \int_{-\infty}^{+\infty} TPF(c)\, dFPF(c).$$

Figure 2.3 presents the relation between the threshold value c and the true (TPF) and false positive (FPF) fractions for the distribution of diseased and non-diseased cases and the corresponding *ROC* curve. Dashed lines correspond to diagnostic systems with different discrimination between diseased and non-diseased subjects.

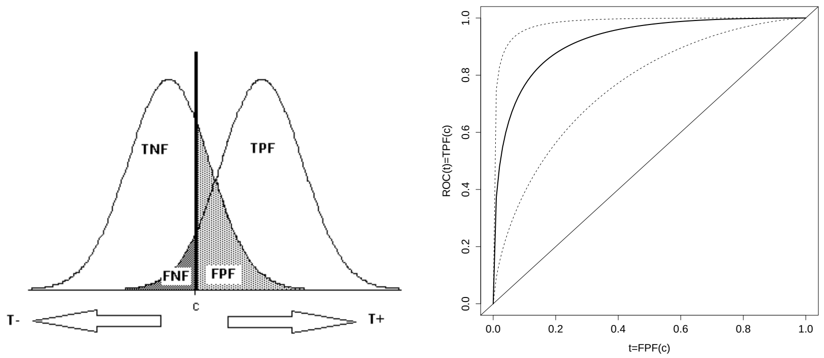

Fig. 1 TPF and FPF and *ROC* curves.

For a perfect test the curve will pass on the upper left corner since the distribution of diseased and non-diseased cases will be completely disjoint, defining an area equal to unity; for a non informative test, the distributions of diseased and non-diseased cases will be identical, and the curve will coincide with the diagonal. Thus, it is possible to compare two diagnostic systems, even using different measurement scales, since the *ROC* curve transforms both tests to a common reference scale. Moreover, the threshold can be determined in such a way to minimize both errors, False Positive and False Negative fractions, or some function of these fractions that takes into consideration decision costs.

The *ROC* can be defined as

$$ROC(t) = TPF(FPF^{-1}(t))$$

where, for a given threshold c, $t = FPF(c)$, with $t \in (0,1)$. For two diagnostic systems, in which test A is uniformly better than test B (please refer to Figure), then

$$ROC_A(t) \geq ROC_B(t) \quad \forall t \in (0,1),$$

as well as

$$AUC_A \geq AUC_B.$$

3 Global and partial comparison

The usual performance measure for *ROC* curves is the area under the curve (*AUC*). When comparing two diagnostic systems, based on the *AUC* index, the best system, the one which better discriminates positive from negative results, corresponds to the system that exhibits greater *AUC*. This is true, provided that the two curves do not cross each other. If the curves cross each other, this may not be the case.

In order to compare two diagnostic systems, it is necessary to determine the regions, on the *ROC* plane, where one system outperforms the other and vice-versa. For that purpose, we propose a methodology [1] based on sampling lines which can estimate the areas under the curves as well as, the regions where one curve is above the other.

Figure 2 shows, for illustration purpose, an empirical *ROC* curve, the line segments that define this curve, as well as three sampling lines that cross the empirical *ROC* curve. The set of $n+1$ points, (x_i, y_i), defining the *ROC* curve are joined by n line segments r_i, given by,

$$r_{i+1}(s) = y_{i+1} + m_{i+1}(s - x_{i+1}) \text{ where } m_{i+1} = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \text{ for } x_i \neq x_{i+1}. \tag{1}$$

where m_{i+1} is the slope of the line segment r_{i+1}. The sampling lines, with origin (s_r, t_r), corresponding to $(1,0)$ on the *ROC* plane, with variable slope, are defined as

$$l_k(s) = t_R - m_k(s - s_R) \text{ where } m_k = \tan\left(\frac{(K+1-k)\pi}{2(K+1)}\right) \text{ for } k = 1, \ldots, K, s \leq s_R. \tag{2}$$

where $K+2$ is the number of sampling lines. The sampling lines define a set of triangles, the sides of which are the distance from the reference point to the intersection point, given by

$$(s_k, t_k) = (\frac{t_R - y_i + m_i x_i + m_k s_R}{m_k + m_i}, t_R - m_k(s_k - s_R)), \tag{3}$$

whose distance from the reference point is given by

$$d_k = \sqrt{(s_k - s_R)^2 + (t_k - t_R)^2} \tag{4}$$

and, thus, the area of each triangle is given by

$$A_k = \frac{1}{2} d_k d_{k-1} \sin\left(\frac{\pi}{2(K+1)}\right) \text{ for } k = 1, \ldots, K+1. \tag{5}$$

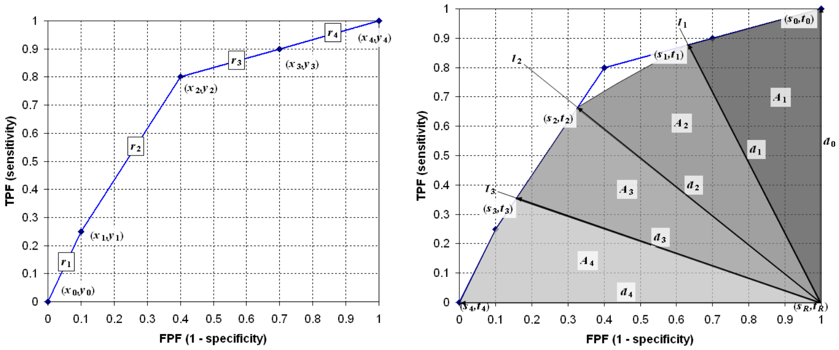

Fig. 2 Empirical *ROC* curve and sampling lines [1].

The number of sampling lines is arbitrarily fixed, but as its number increases, the better will be the estimation of the area [1]. Figure 3 shows an example of two empirical *ROC* curves that cross each other and how the sampling lines can be used for the determination of the area between the two curves. Moreover, this figure shows the region of the *ROC* space where one curve is superior to the other and vice versa. In the adjacent graph, the horizontal axis shows the angles from the sampling lines (for this example, 100 lines have been used) and in the vertical axis, the areas between the two curves are represented. Positive points are the points where curve 1 is better than curve 2. Thus, this approach provides extension and location measures for the comparison of two curves, identifying and quantifying the regions where the differences between the curves occur, and its relation with Sensitivity and Specificity.

In order to evaluate if there are significant differences between the two areas under the curve and to establish, as well, confidence intervals, a permutation test [4] is used. Based on the areas along the *ROC* space, calculated by equation , the differences between the curves, in terms of areas, can be calculated. It should be noted that the differences of the areas between the curves are exchangeable. Considering TS as the summation of the differences between the areas, a large number of samples can be generated by exchanging the areas, through the reassignment of the plus and minus sign to the areas. Thus, it is possible to have the overall distribution of TS

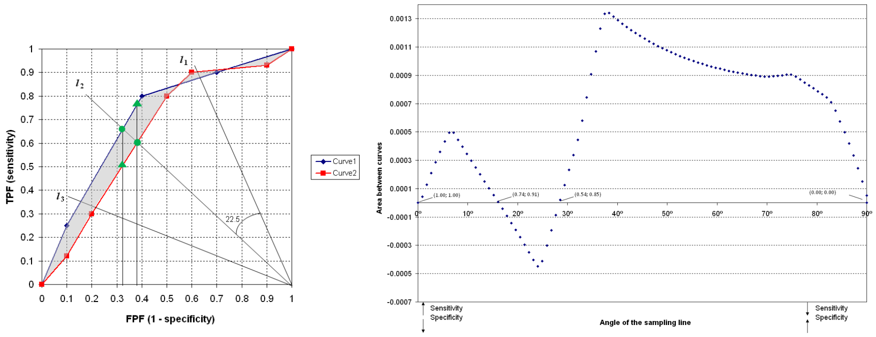

Fig. 3 Area between *ROC* curves, extension and location [1].

and, therefore, to perform a statistical test. The distribution of the distances of the intersection point to the reference point can be generated by bootstrapping, which can be perceived as a *ROC* curve bootstrap distribution, allowing the construction of the percentile bootstrap confidence intervals. Thus, it is possible to calculate confidence intervals for the location and extension measures, which define confidence bands as ilustrated in Figure 4.

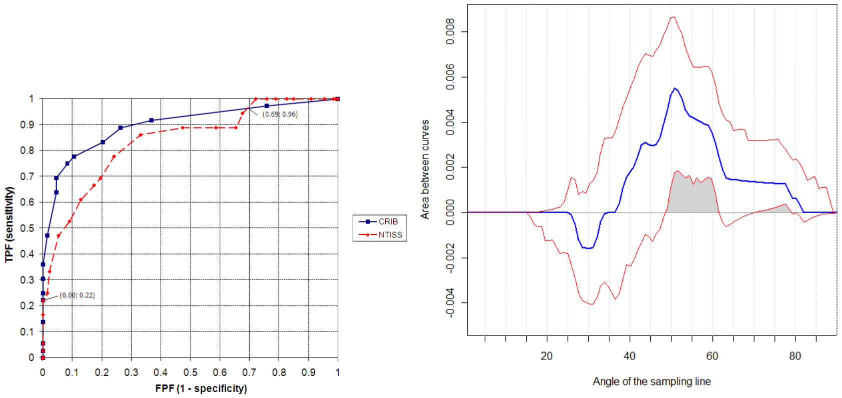

Fig. 4 *ROC* curves comparison and confidence bands for extension and location [1].

4 Conclusions

In this work a methodology to compare *ROC* curves that cross each other has been presented, where a graphical display identifies the regions where one curve is superior to the other, with the corresponding Sensitivity and Specificity regions. Thus, the proposed approach avoids the need of partial comparison of areas [12, 3], providing a global evaluation along the *ROC* space. This approach, based on bootstrap, is a non parametric alternative for the comparison of curves that cross each other, regardless the number of crossings.

References

1. A. C. Braga, L. Costa, and P. Oliveira. An alternative method for global and partial comparison of two diagnostic systems based on ROC curves. *Journal of Statistical Computation and Simulation*, 1–19, online, 2011.
2. E. R. DeLong, D. M. DeLong, and D. I. Clarkepearson. Comparing The Areas Under 2 or More Correlated Receiver Operating Characteristic Curves - A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988.
3. L. E. Dodd and M. S. Pepe. Partial AUC estimation and regression. *Biometrics*, 59(3):614–623, 2003.
4. P. Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag, 2nd edition edition, 2000.
5. J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
6. J. A. Hanley and B. J. McNeil. A Method of Comparing the Areas Under Receiver Operating Characteristic Curves Derived from the Same Cases. *Radiology*, 148(3):839–843, 1983.
7. C. E. Metz. Statistical analysis of roc data in evaluating diagnostic performance. In Myers R Herbert D, editor, *Applications in the Health Sciences*, volume 13, 365–84. American Institute of Physics, 1986.
8. C. E. Metz, P.-L. Wang, and H. Kronman. A new approach for testing the significance of differences between roc curves measured from correlated data. In *Proceedings of the 8th Conference in Information Processing in Medical Imaging*, 432–45, 1983.
9. M. S. Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series. Oxford University Press, 2003.
10. J. A. Swets and R. M. Pickett. *Evaluation of Diagnostic Systems Methods from Signal Detection Theory*. Academic Press, London, 1982.
11. J. A. Swets. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. LEA, New Jersey, 1996.
12. D. D. Zhang, X. H. Zhou, D. H. Freeman, and J. L Freeman. A Non-Parametric Method for The Comparison of Partial Areas Under ROC Curves and Its Application to Large Health Care Data Sets. *Statistics In Medicine*, 21(5):701–715, 2002.