CrossMark

Some mathematical properties of the ROC curve and their applications

Camilla Calì[1] · **Maria Longobardi**[1]

Abstract In this paper we present ROC methodology and analyze the ROC curve. We describe first the historical background and its relation with signal detection theory. Some mathematical properties of this curve are given, and in particular the relation with stochastic orders and statistical hypotheses testing are described. We present also a medical application of the Neymann–Pearson lemma.

1 Introduction and historical background

ROC (Receiver Operating Characteristic) methodology belongs to signal detection theory and it was first developed during the Second World War by electrical engineers to analyze radar signals and to study the relation signal/noise, in particular in order to detect correctly enemy objects in battlefields. It was used to quantify the ability of a radar operator to discern between information-bearing patterns (signal) and random

---

✉ Maria Longobardi
  maria.longobardi@unina.it; malongob@unina.it

  Camilla Calì
  camilla.cal@libero.it

1 Dipartimento di Matematica e Applicazioni, Università di Napoli Federico II,
   via Cintia, Naples, Italy

🌍 Springer

patterns that distract from the information (noise). Since the second half of the last century, ROC analysis has been used fully in radar detection theory (Marcum [11]) and in psychophysical researches (Green and Swets [3]). About twenty years later, both interest and application have increased in the field of medicine (Lusted [10]), first in radiology (Goodenough [4]; Hanley and McNeil [5,6]) and then in preventive medicine and for evaluation of diagnostic tests (Erdreich [2])! and its relevance is increasing also in machine learning and data mining research.

ROC analysis is useful to evaluate the performance of any kind of diagnostic tests, and more generally the accuracy of statistical models that divide subjects into two classes. In statistics, one of the first and most important contribution to ROC analysis is due to Donald Bamber [1], who clarified the relationship between the ROC curve and the ordinal dominance (OD) graph and explained the relation with the Mann-Whitney U-statistics.

Let X and Y be two random variables with F and G distribution functions, f and g density functions, and $\bar{F} = 1 - F$ and $\bar{G} = 1 - G$ survival functions. We recall here some of the main stochastic orders (for further informations see Shaked and Shanthikumar [14]).

We say that X is smaller than Y in the *usual stochastic order*, and we write $X \leq_{st} Y$, if

$$\bar{F}(x) \leq \bar{G}(x) \quad \text{for all} \quad x \in (-\infty, +\infty),$$

or equivalently,

$$F(x) \geq G(x) \quad \text{for all} \quad x \in (-\infty, +\infty).$$

Moreover, let $\lambda_F(x) = f(x)/\bar{F}(x)$ and $\lambda_G(x) = g(x)/\bar{G}(x)$ be the hazard rate functions of X and Y, respectively. X is said to be smaller than Y in the *hazard rate order*, $X \leq_{hr} Y$, if

$$\lambda_F(x) \geq \lambda_G(x), \ x \geq 0.$$

Consider now $\phi_F(x) = f(x)/F(x)$ and $\phi_G(x) = g(x)/G(x)$ the reversed hazard rate functions of X and Y, respectively; X is said to be smaller than Y in the *reversed hazard rate order*, and we write $X \leq_{rh} Y$, if

$$\phi_F(x) \leq \phi_G(x), \ x \geq 0,$$

or equivalently

$$X_{(t)} \leq_{st} Y_{(t)} \quad \text{for all} \quad t \geq 0,$$

where $X_{(t)}$ and $Y_{(t)}$ are the residual random variables at age t, i.e.

$$X_{(t)} = \left(X - t | X > t\right)$$

and

$$Y_{(t)} = \left(Y - t \,|\, Y > t\right).$$

Finally, X is said to be smaller than Y in the *likelihood order*, $X \leq_{lr} Y$, if

$$\frac{g(x)}{f(x)} \quad \text{increases when } x \text{ ranges in the union of the supports of } X \text{ and } Y$$

or equivalently if

$$f(x)g(y) \geq f(y)g(x) \quad \text{for all} \quad x \leq y.$$

The following result is well known.

Theorem 1 *If X and Y are two continuous or discrete random variables such that $X \leq_{lr} Y$, then $X \leq_{hr} Y$ and $X \leq_{rh} Y$, and so also $X \leq_{st} Y$.*

In this paper we discuss some mathematical properties of the ROC curve and an application to medical studies.

In Sect. 2 we describe the so-called "yes-no" tests and introduce the definitions of sensitivity and specificity. The ROC curve is considered as plot of sensitivity (TPR) versus 1-specificity (FPR) considering all possible values of the cut-off c. In Sect. 3 some properties of the ROC curve are studied, especially in relation with stochastic orders. This relation is used in order to obtain a result about the concavity of the ROC curve. In Sect. 4 we analyze statistical hypotheses testing, focusing on the definitions of False Positive Rate and False Negative Rate as error of I and II type, respectively.

The last two sections are devoted to an estimation of the ROC curve and to a medical application.

2 The ROC curve

Consider a signal detection experiment. On each trial of the experiment, we must discriminate two events: a signal event (such as radar signals, stimuli in psycology, etc.) or a noise event (that appears even though there is no signal and so gets the observer confused in the detection of the "real" signal). On each trial, we obtain a sensory impression of the event and infer whether signal or noise has been presented to us by analyzing the strength of this impression. We must say "yes" if we think that it is a signal and "no" otherwise. The result of this analysis can be summarized in a 2×2 contingency table (see Table 1).

This table can be used for all type of dichotomous tests, in particular for the so called "yes-no" tests. The accuracy of these tests can be defined in different ways, according to Table 1. We define:

Table 1 The 2 × 2 table summarizes the results of a dichotomous (yes-no) test

	Event		Total
	Signal ($E+$)	Noise ($E-$)	
Yes ($T+$)	TP	FP	m_1
No ($T-$)	FN	TN	m_0
Total	n_1	n_0	N

TP represents the number of the true positive (the signal is present and the observer answers "yes"), FP the false positive (the signal is absent and the observer says "yes"), TN the true negative (the signal is absent and the observer says "no") and FN the false negative (the signal is present and the observer answers "no")

- **Sensitivity or True Positive Rate, TPR**, that is the conditional probability of correctly classifying the presence of signals:

$$\text{Sensitivity (SE)} = \mathcal{P}(T+|E+) = \frac{TP}{TP+FN} = \frac{TP}{n_1}$$

- **Specificity or True Negative Rate, TNR**, that is the conditional probability of correctly classifying the absence of signals:

$$\text{Specificity (SP)} = \mathcal{P}(T-|E-) = \frac{TN}{TN+FP} = \frac{TN}{n_0}$$

- **False Positive Rate, FPR**, that is the probability of type I error:

$$\text{FPR} = \mathcal{P}(T+|E-) = \frac{FP}{FP+TN} = \frac{FP}{n_0} = 1 - \text{specificity}$$

- **False Negative Rate, FNR**, that is the probability of type II error:

$$\text{FNR} = \mathcal{P}(T-|E+) = \frac{FN}{FN+TP} = \frac{FP}{n_1} = 1 - \text{sensitivity}$$

The true-positive rate is also known as "sensitivity" in particular in biomedical information theory, and "recall" in machine learning. The false-positive rate is also known as the "fall-out".

Many tests give a quantitative result T (for instance, in the above situation, the strength of sensory impression arising from signal events and noise events). In this case it is necessary to define a cut-off value c to separate events into two groups: we classify the signals as present if $T \geq c$ and as absent if $T < c$. Here the previous measures of accuracy are not suitable to define uniquely a "good test", because different cut-off values have different 2 × 2 tables. In particular, as c decreases, sensitivity increases but specificity decreases and viceversa when c increases (see Fig. 1).

Even though sensitivity and specificity do not depend on frequency of presence of signals or noises, they strongly depend on selected cut-off value.
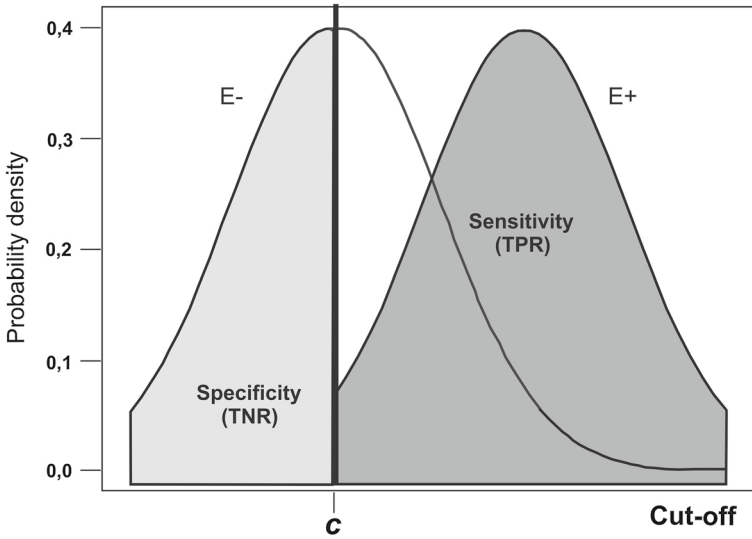
Fig. 1 Sensitivity and specificity in relation with the cut-off point c

A ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. By considering all possible values of the cut-off c, the ROC curve can be constructed as a plot of sensitivity (TPR) versus $1-$specificity (FPR). For any cut-off c, we can define:

$$\text{TPR}(c) = \mathcal{P}(T \geq c|E+)$$
$$\text{FPR}(c) = \mathcal{P}(T \geq c|E-).$$

Thus the ROC curve is

$$\text{ROC}(\cdot) = \{\text{FPR}(c), \text{TPR}(c), \quad c \in (-\infty, +\infty)\}.$$

We also write the ROC curve as:

$$\text{ROC}(\cdot) = \{(t, \text{ROC}(t)), \quad t \in (0, 1)\},$$

where the ROC function maps t to TPR (c), anc c is the cut-off corresponding to $FPR(c) = t$.

3 Properties of the ROC curve

The ROC curve is a monotone increasing function mapping $(0, 1)$ to $(0, 1)$. An uninformative test is one such that $TPR(c) = FPR(c)$ for every threshold c and this situation is represented by ROC curve $ROC(t) = t$, which is a line with unit slope. A perfect test completely separates the signal and noise events, i.e. $TPR(c) = 1$ and
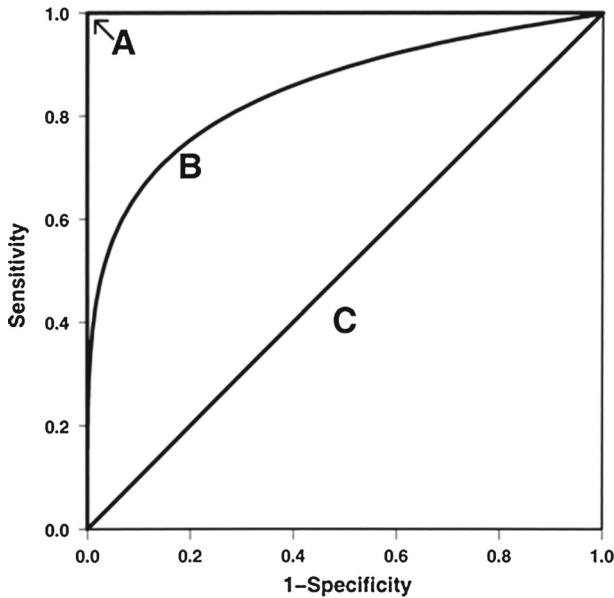
Fig. 2 Three hypothetical ROC curves representing the accuracy of an ideal test (*line A*) on the upper and left axes in the unit square, a typical ROC curve (*curve B*), and a diagonal line corresponding to an uninformative test (*line C*). As test accuracy improves, the ROC curve moves toward line A

$FPR(c) = 1$ for some threshold c. This ROC curve is along the left and upper borders of the positive unit quadrant (see Fig. 2).

Proposition 1 *The ROC curve is invariant with respect to strictly increasing transformations of T.*

Proof Let $(\text{FPR}(c), \text{TPR}(c))$ be a point of the ROC curve for T and let h be a strictly increasing transformation of T:

$$W = h(T) \quad \text{and} \quad d = h(c).$$

Then we have:

$$\mathcal{P}(W \geq d|E-) = \mathcal{P}(T \geq c|E-) \quad \text{and} \quad \mathcal{P}(W \geq d|E+) = \mathcal{P}(T \geq c|E+),$$

and hence the same point lies in the ROC curve for W. □

The test score of a trial of the experiment can be viewed as a random variable. We divide trials of the experiment into two groups: signal events and noise events. The test score of a signal event is defined as a random variable X with cumulative distribution function F and density f. Similarly the test score of a noise event can be represented as a random variable Y with cumulative distribution function G and density g. Some formal properties of the ROC curve can be described as follows.

Proposition 2 *Let $\bar{F}(x) = 1 - F(x)$ and $\bar{G}(y) = 1 - G(y)$ be the survival functions for F and G, respectively. Then the ROC curve has the representation:*

$$ROC(t) = \bar{G}(\bar{F}^{-1}(t)) \quad t \in [0, 1],$$

where

$$\bar{F}(x) = \mathcal{P}(X \geq x | E-) \quad and \quad \bar{G}(y) = \mathcal{P}(Y \geq y | E+).$$

Proof Let $c = \bar{F}^{-1}(t)$, so that c is the threshold corresponding to the false positive rate t, and hence $\mathcal{P}(X \geq c | E-) = t$. The corresponding true positive rate is

$$\mathcal{P}(Y \geq c | E+) = \bar{G}(c).$$

Then the TPR that corresponding to the $FPR = t$ is

$$ROC(t) = \bar{G}(c) = \bar{G}(\bar{F}^{-1}(t)).$$

The statement is proved. □

Note that the ROC curve can be represented also using only a distribution function:

$$ROC(t) = 1 - G(F^{-1}(1 - t)) \quad t \in [0, 1],$$

with a simple change of variables.

The following result can be applied only to continuous tests.

Proposition 3 *The ROC curve is the function on $[0, 1]$ such that $ROC(0) = 0$ and $ROC(1) = 1$ and which has slope:*

$$\frac{\partial ROC(t)}{\partial t} = \frac{g(\bar{F}^{-1}(t))}{f(\bar{F}^{-1}(t))},$$

where f and g denote the probability densities of X and Y, respectively.

Proof We have:

$$\frac{\partial ROC(t)}{\partial t} = \frac{\partial \bar{G}(\bar{F}^{-1}(t))}{\partial t} = \frac{\bar{G}(\bar{F}^{-1}(t))}{\partial \bar{F}^{-1}(t)} \cdot \frac{\bar{F}^{-1}(t)}{\partial t}$$

$$= -g(\bar{F}^{-1}(t)) \cdot \frac{\partial \bar{F}^{-1}(t)}{\partial t}.$$

The statement follows because

$$\frac{\partial \bar{F}^{-1}(t)}{\partial t} = \frac{1}{\frac{\partial}{\partial w}(\bar{F}^{-1}(w))} = \frac{1}{-f(w)},$$

evaluated at $w = \bar{F}^{-1}(t)$. □

The slope can be interpreted as the likelihood ratio

$$\mathcal{L}R(c) = \mathcal{P}(T = c|E+)/\mathcal{P}(T = c|E-),$$

at the threshold c corresponding to the point $(t, \mathrm{ROC}(t))$, that is $c = \bar{F}^{-1}(t)$. As a consequence, we have:

$$\mathcal{L}R(c) = \frac{g(c)}{f(c)} \quad \text{i.e.} \quad \frac{\partial \mathrm{ROC}(c)}{\partial c} = \mathcal{L}R(c),$$

with $c = \bar{F}^{-1}(t)$.

The following result shows the ROC curve trend under particular hypotheses involving likelihhod ratio order.

Proposition 4 *Let X and Y be two random variables with distribution functions F and G and densities f and g, respectively. Then $X \leq_{lr} Y$ if and only if $\bar{G}\bar{F}^{-1}$ is concave.*

Proof Using the definition of likelihood ratio order, we have

$$\frac{\partial}{\partial t} \frac{g(t)}{f(t)} \geq 0.$$

The statement follows now from an easy differentiation, proving that

$$\frac{\partial^2 \left(\bar{G}\bar{F}^{-1} \right)}{\partial t^2} \leq 0.$$

$\square$

Considering the results of Propositions 2 and 4, it is easy to conclude that if $X \leq_{lr} Y$ (where X and Y represent the test scores of a signal event and a noise event, respectively) then the corresponding ROC curve is concave.

We can conclude that if $X \leq_{lr} Y$ then the ROC curve lies above the diagonal in the unit square because $\mathrm{ROC}(0) = 0$ and $\mathrm{ROC}(1) = 1$.

Remark 1 As an application of Theorem 1, we can state that under the hypotheses of Proposition 4, the two random variables X and Y are ordered in the usual stochastic order, in the hazard rate order and in the reversed hazard rate order.

4 Analogy with statistical hypotheses testing

In statistical hypoteses tests, we define:

– H_0: the null hypothesis;
– H_1: the alternative hypothesis.

Table 2 The analogy between statistical hypotheses testing and a general diagnostic testing

	Statistical hypotheses testing	Diagnostic testing
Possible states	H_0 vs H_1	$D-$ vs $D+$
Information	Data for n subjects	Test results for a subject
	Denoted by W	Denoted by T
Type I error	$\alpha = \mathcal{P}(\text{reject } H_0 \mid H_0)$	$\text{FPR} = \mathcal{P}(\text{classify } D + \mid D-)$
Type II error	$\beta = \mathcal{P}(\text{accept } H_0 \mid H_1)$	$\text{FNR} = \mathcal{P}(\text{classify } D - \mid D+)$
$\mathcal{L}R$ function	$\mathcal{L}R(W) = \mathcal{P}(W \mid H_1)/\mathcal{P}(W \mid H_0)$	$\mathcal{L}R(W) = \mathcal{P}(T \mid D+)/\mathcal{P}(T \mid D-)$

and the region of rejection, or critical region, $\mathcal{C}$ as the set of values of the test statistic for which H_0 is rejected.

In statistical hypotheses tests two types of errors occur:

- error of the first kind (type I error), when the true null hypothesis is rejected; its occurrence probability is denoted by α;
- error of the second kind (type II error), when the false null hypothesis is accepted; its occurrence probability is denoted by β.

These errors are in strict relation with Table 1; in fact, type I error may be compared with false positive (FP) and type II error with false negative (FN). Power of test $(1 - \beta)$ is the probability for the test of correctly rejecting the null hypothesis (sensitivity). Size of test (α) is the probability for the test of incorrectly rejecting the null hypothesis. The complement of size is specificity (Table 2).

5 ROC curve estimation

The most important numerical index used to describe the behavior of the ROC curve is the area under the ROC curve (AUC), defined by

$$\text{AUC} = \int_0^1 \text{ROC}(t)dt.$$

Referring to Fig. 2, line A represents a perfect test with $AUC = 1$, curve B represents a typical ROC curve (for example $AUC = 0.85$), and a diagonal line (line C) corresponding to uninformative test with $AUC = 0.5$. As test accuracy improves, the ROC curve moves toward A, and the AUC approaches 1. The AUC has an interesting interpretation. It represents both the average sensitivity over all values of FPR and the probability that test results from a randomly selected pair of signal and noise are correctly ordered, namely $\mathcal{P}(Y > X)$ (Bamber [1], Hanley and McNeil [5]). If two tests are ordered with test A uniformly better than test B (see Fig. 2) in the sense that

$$\text{ROC}_A(t) \geq \text{ROC}_B(t),$$

then their AUC statistics are also ordered:

$$\text{AUC}_A \geq \text{AUC}_B.$$

Suppose that test results $x_1, \ldots, x_{n_0}$ are available for n_0 noises and $y_0, \ldots y_{n_1}$ for n_1 signal trials. Many approaches have been proposed for estimating the ROC curve and so the related AUC: parametric, nonparametric and semiparametric methods ([13,15,17]).

The simplest parametric approach is to assume that F and G follows a parametric family of distributions, and fit these distributions to the observed test results. When we suppose that test result are normally distributed we obtain the so-called *binormal ROC curve* (for details see [13]).

Nonparametric ROC methods do not require any assumption about the test result distributions and generally do not provide a smooth ROC curve. The simplest nonparametric approach involves estimating F and G by the empirical distribution functions (for details see Hsieh and Turnbull [7,8]). An alternative nonparametric approach is to fit a smoothed ROC curve by using the kernel density estimation of F and G (for details see Lloyd [9]).

In recent years a lot of semiparametric ROC methods have been introduced. A popular approach is to assume a so-called *binormal* model, that postulates the existence of some unspecified monotonic transformation H of measurement scale that simultaneously converts F and G distributions to normal ones. Without loss of generality these can be taken, respectively, to be $N(0, 1)$ and $N(\mu, \sigma^2)$ and then we can analyze ROC curve through parametric methods (for details see [7,8]). Another semiparametric approach is represented by a generalized linear model (GLM) (see [12]).

6 A medical application

In medicine a diagnostic test is any kind of medical test performed to aid in the diagnosis or detection of a disease. In particular, a diagnostic test should correctly classify patients into "healthy" and "diseased" categories (Table 2). In many situations, the determination of these two categories is a difficult task. In order to estimate the accuracy of the diagnostic test, it is necessary to compare the result of the test with real status of the patient. Such "true disease state" may be determined in different ways, such as clinical follow-up or biopsy. The results of this analysis can be summarized in a 2×2 contingency table similar to Table 1.

The accuracy of a diagnostic test can be defined in various ways, by using the possibilities described in Table 3. As in Sect. 2, we define:

– **Sensitivity or True Positive Rate, TPR**, the conditional probability of correctly classifying diseased patients;
– **Specificity or True Negative Rate, TNR**, the conditional probability of correctly classifying healty patients;
– **False Positive Rate, FPR**, the probability of type I error;
– **False Negative Rate, FNR**, the probability of type II error.

Table 3 The 2 × 2 table summarizes the results of a dichotomous diagnostic test

	Disease state		Total
	Present ($D+$)	Absent ($D-$)	
Positive test ($T+$)	TP	FP	m_1
Negative test ($T-$)	FN	TN	m_0
Total	n_1	n_0	N

TP represents the number of the true positive (diseased patients with positive test), FP the false positive (healthy patients with positive test), TN the true negative (healthy patients with negative test) and FN the false negative (diseased patients with negative test)

In diagnostic medicine an ideal test is called *golden standard* and it satisfies the conditions:

$$\text{FPR} = 0 \quad \text{and} \quad \text{TPR} = 1.$$

Many diagnostic tests give a quantitative result T, so that it is necessary to define a cut-off value c to separate patients into two groups: we classify the patients as diseased if $T \geq c$ and healthy if $T < c$. It is clear that this is a particular situation in which the ROC analysis can be used in order to evaluate the power of a diagnostic test. ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

All the considerations about the ROC analysis can be traslated in medical context. Most diagnostic tests currently available are not perfect, and so strategies to combine informations obtained by multiple tests may provide better diagnostic tools.

Recall that the likelihood ratio function for a single test result T is defined as

$$\mathcal{L}R(T) = \mathcal{P}(T|D+)/\mathcal{P}(T|D-).$$

The following result on the function $\mathcal{L}R(T)$ is known.

Proposition 5 *The optimal criterion based on T for classifying subjects as positive for disease is*

$$\mathcal{L}R(T) > c,$$

in the sense that it achieves the highest true positive rate (TPR) among all possible criteria based on T with false positive rate (FPR) equals to

$$\mathcal{P}(\mathcal{L}R(T) > c|D-).$$

The $\mathcal{L}R$ function is a useful tool when we consider multiple tests. If $\mathbf{T} = (T_1, \ldots, T_k)$, we define

$$\mathcal{L}R(\mathbf{T}) = \frac{\mathcal{P}(\mathbf{T}|D+)}{\mathcal{P}(\mathbf{T}|D-)}.$$

Proposition 5 remains true also for multidimensional tests.

This result is essentially an application of Neymann–Pearson lemma for statistical hypotheses testing. Let $D-$ and $D+$ denote the null and the alternative hypothesis, respectively, and let $\mathbf{T}$ be the sample data. The screening test based on $\mathbf{T}$ is the analogue of the rule for rejecting the null hypothesis $D-$ in favour of the alternative $D+$. Then the likelihood ratio function of the k-screening test, $\mathcal{L}R(\mathbf{T})$, and the rules based on its exceeding a threshold achieves the highest TPR among all screening tests based on $\mathbf{T}$ with FPR $= \mathcal{P}(\mathcal{L}R(T) > c|D-)$.

References

1. Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. **12**, 387–415 (1975)
2. Erdreich, L.S.: Use of relative operating characteristic analysis in epidemiology. Am. J. Epidemiol. **114**, 649–662 (1981)
3. Green, D.M., Swets, J.A.: Signal Detection Theory and Psychophysics. Wiley, New York (1966)
4. Goodenough, D.J., Rossmann, K., Lusted, L.B.: Radiographic applications of receiver operating characteristic (ROC) analysis. Radiology **110**, 89–95 (1974)
5. Hanley, J., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology **143**, 29–36 (1982)
6. Hanley, J., McNeil, B.J.: A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases. Radiology **148**, 839–843 (1983)
7. Hsieh, F.S., Turnbull, B.W.: Non- and semi-parametric estimation of the receiver operating characteristic curve. Technical Report 1026, School of Operations Research, Cornell University (1992)
8. Hsieh, F.S., Turnbull, B.W.: Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Ann. Stat. **24**, 25–40 (1996)
9. Lloyd, C.J.: Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. J. Am. Stat. Assoc. **93**, 1356–1364 (1998)
10. Lusted, L.B.: Signal detectability and medical decision-making. Science **171**, 1217–1219 (1971)
11. Marcum, J.I.: A Statistical Theory of Target Detection by Pulsed Radar, The Research Memorandum, RAND report RM-754 (1947)
12. Pepe, M.S.: An interpretation for the ROC curve and inference using GLM procedures. Biometrics **56**, 352–359 (2000)
13. Pepe, M.S.: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, Oxford (2003)
14. Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer, New York (2007)
15. Shapiro, D.E.: The interpretation of diagnostic tests. Stat. Methods Med. Res. **8**, 113–134 (1999)
16. Swets, J.A.: Measuring the accuracy of diagnostic systems. Science **240**, 1285–1293 (1988)
17. Zou, K.H., Liu, A., Bandos, A.I., Ohno Machado, L., Rockette, H.E.: Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis, Chapman and Hall/CRC Biostatistics Series, London (2012)