Medical & Biological Engineering & Computing manuscript No.
(will be inserted by the editor)

Informational Analysis: a Shannon Theoretic Approach to Measure the Performance of a Diagnostic Test

Rossano Girometti · Francesco Fabris

Received: date / Accepted: date

Abstract Diagnostic test accuracy, based on sensitivity, specificity, positive/negative predictive values (dichotomous case), and on ROC analysis (continuous case), should be expressed with a single, coherent index. We propose to modelize the diagnostic test as a flow of information between the disease, that is a hidden state of the patient, and the physicians. We assume that: i) sensitivity, specificity, false positive/negative rates are the probabilities of a *Binary Asymmetric Channel*; ii) the diagnostic channel information is measured by *Mutual Information*. We introduce two summary measures of accuracy, namely the *Information Ratio* (IR) for the dichotomous case, and the *Global Information Ratio* (GIR) for the continuous case. We apply our model to a study by Pisano et al. [19], who compared digital versus film mammography, in diagnosing breast cancer in a screening population of 42,760 women. In film mammography, the maximum IR (0.178) corresponds to the standard cut-off of sensitivity and specificity provided by the ROC analysis (GIR 0.200). Maximum IR and GIR for digital mammography are higher (0.201 and 0.229, respectively), but IR corresponds to a cut-off with higher sensitivity but lower specificity, thus suggesting that larger information provided by digital mammography carries the risk of more false positive cases.

Part of this work has been presented at 2014 Congress of the European Society of Abdominal and Gastrointestinal Radiology (ESGAR), and at the Radiological Society of North America (RSNA) 2014 Annual Meeting

R. Girometti
Dipartimento di Scienze Mediche e Biologiche - Istituto di Radiologia Diagnostica, Università degli Studi di Udine, Italy
Tel.: +39-0432-559433
Fax: +39-0432-494301
E-mail: rossano.girometti@uniud.it

F. Fabris
Dipartimento di Matematica e Geoscienze, Università degli Studi di Trieste, Italy
Tel.: +39-040-5582625
Fax: +39-040-5582636
E-mail: ffabris@units.it

Keywords Measure of diagnostic test performance · Multivalued diagnostic test · Binary classifiers · Readers expertise evaluation · ROC analysis

– Total number of words (title page, abstract, text, references, tables, and figure legends): 6331

– Abstract: 199 words

– Number of tables: 1

– Number of figures: 6

1 Introduction

Standard analysis of accuracy, for a dichotomous diagnostic test, is currently based on the calculation of *sensitivity* (*SE*), *specificity* (*SP*), *positive predictive value* (*PPV*) and *negative predictive value* (*NPV*), as well as *SE* and *SP*-derived measures, such as *likelihood ratios* [10], [24]. Even though there may be preference for either specificity or sensitivity (in which case a single measure would hide the required detailed information about different aspects of accuracy), nevertheless a single statistical measure that can summarize the global quality of a dichotomous diagnostic test is of interest in many clinical situations, and is still lacking [18]. In the case of continuous test results, standard analysis includes *Receiver Operating Characteristic* (ROC) curves analysis as the most popular method to assess a test or to compare it with a different one [4], [8], [15], [16]. Different summary measures have been proposed to describe the accuracy with a single, objective index, e.g. the *Area Under the Curve* (AUC) provided by ROC analysis, and each of them shows well known advantages and drawbacks [3], [24]. Another important problem is which decision cut-off should be used to classify continuous test results, and how will the choice of a decision threshold affect comparisons between two diagnostic tests or between two raters [15]. Even these are critical questions when computing sensitivity and specificity, yet the choice for the decision threshold is often arbitrary [15].
Regardless of the method chosen for the summary measure or for the cut-off, standard analysis is based on the comparison between a *Standard of Reference* (SR) and test results, that is on probabilities that a test result really represents the presence or absence of a certain medical condition. Even assuming that SR always provides a correct representation of the patient status, which is not the case [24], standard analysis does not rigorously objectify which is the maximum information on the disease we can achieve depending on that specific test. One might argue this is a potential limitation, since it might be difficult to understand: *i*) whether a disappointing test performance has been conditioned by definite characteristics of a clinical setting (low prevalence, selection bias and so on) or rather by intrinsic limits on the amount of diagnostic information the test can "physically" vehicle; *ii*) whether, for a certain clinical scenario, it could be convenient to improve further an already good test performance, or not. Furthermore, one might expect that quantifying information underlying test results might complement the standard dichotomous or ROC analysis.
To our knowledge, no definite methods to measure test accuracy in informational terms are currently available. We hypothesize that *Information Theory* [21] is the proper framework to define such a method, since it is the mathematical apparatus underlying current telecommunication systems, based on a rigorous, quantifiable notion of information and information-derived measures [23]. As detailed below, we show that a diagnostic test can be modeled using Information Theory, and consequently that diagnostic test accuracy can be expressed coherently, with a single index, in the form of a summary information measure; it can be used also to select the best cut-off in the continuous case.
On this basis, the aim of this paper is manyfold: *i*) to present the above model of Shannon informational analysis of diagnostic accuracy; *ii*) to illustrate how informational analysis can be considered as a useful complement to standard ROC analysis, *iii*) to show that informational analysis is able to overtake some weakness

of the ROC/AUC approach, and *iv*) to illustrate the applicability of the informational model using a dataset from medical literature.

2 Methods

2.1 Basic definitions

In the following, we shall use the $2 \cdot 2$ table with the four possible outcomes deriving from the application of a standard of reference to the diagnostic test set; they are, respectively, the number of *true positives* (*TP*), *false positives* (*FP*), *false negatives* (*FN*) and *true negatives* (*TN*) reports. If we call *reader* the clinician who formulates the diagnostic *report* R, we assume that there are only two mutually exclusive states D of *disease* (or pathologic state) for a patient: it can be present ($D=1$) or absent ($D=0$) (see [25]). Similarly, a report indicating the presence of the disease is called positive ($R = 1$); when indicating its absence is called negative ($R=0$); this corresponds to a dichotomous or *binary* classifier. Once we have *TP, FN, FP* and *TN*, we can specify the following quantities:

$$\textbf{Sensitivity} \qquad SE = p(R = 1/D = 1) = \frac{TP}{TP + FN} \tag{1}$$

$$\textbf{False negative rate} \qquad FNR = p(R = 0/D = 1) = \frac{FN}{TP + FN} \tag{2}$$

$$\textbf{Specificity} \qquad SP = p(R = 0/D = 0) = \frac{TN}{FP + TN} \tag{3}$$

$$\textbf{False positive rate} \qquad FPR = p(R = 1/D = 0) = \frac{FP}{FP + TN} \tag{4}$$

where $p(R=x/D=y)$ is the conditional probability that the report R is x, given that the disease is y. From equations (1) and (2) we note that $p(R = 1/D = 1) + p(R = 0/D = 1) = 1$, since once the disease is present, we necessarily have $R = 1$ or $R = 0$ as a possible diagnosis. The same happens for equations (3) and (4), that is $p(R = 1/D = 0) + p(R = 0/D = 0) = 1$.
So, if we set $FNR = \alpha$ and $FPR = \beta$, we have $SE = 1 - \alpha$ and $SP = 1 - \beta$.
We assume that the disease is a hidden, objective status of the patient, which can be present ($D = 1$) or absent ($D = 0$). The physician makes assumptions on the disease by interpreting the result of a diagnostic test, that can be thought of as the outcome of the *diagnostic channel* of figure 1; it might represent, e.g., a mammography examination interpreted by a radiologist, or a *Prostate Specific Antigen* (PSA) level, telling the urologist whether the cut-off of 10 ng/ml has been exceeded. Since test results give the reader information on the disease, then the accurate diagnostic test (or the accurate reader) will be the one able to extract as much information as possible from the diagnostic channel: the more information on the patient status flows from the disease to the reader on the diagnostic channel, the more accurate the diagnostic test. This means that a coherent measure of the diagnostic test quality is the (maximum) amount of information that can be extracted from the diagnostic channel. We shall describe in details how it is possible to measure precisely such a quantity.

Place figure 1 about here

If the diagnostic channel were perfect (standard of reference), we should have $FNR = FPR = 0$, $SE = SP = 1$, that is $\alpha = 0$ and $\beta = 0$, and the test always

gives us the correct (positive or negative) diagnosis. But some "noise" can affect the results, due to the limits of our diagnostic methods and/or to the inexperience of the reader; this means that α and β are usually strictly greater than zero. The consequence is that we have negative test results in presence of the disease (FN) and positive test results in absence of the disease (FP).

2.2 Modeling dichotomous diagnostic tests

The model above described has already been fully investigated and completely solved by Claude Elwood Shannon in 1948; his milestone paper "A mathematical theory of communication" [21] established the born of a new scientific discipline, the *Information Theory*. All modern digital communication systems that interact over a network (telephones, computers, televisions, mobiles, and so on) are based (also) on this theory and on the work of Shannon.
In the original study the (binary) channel is constituted by a (wired) line of transmission affected by noise, for example a twisted pair, a coaxial cable or other. The effect of noise is that of changing the bit flowing on the channel with a probability that depends on its physical quality. Suppose that D and R describe the (binary) variables at the input and at the output of the channel (see figure 1); if we send, for example, $D = 0$ on the channel, there is a positive probability $p(1/0) = \beta$ to receive $R = 1$ on the output. The same happens if we send $D = 1$, since we have a positive probability $p(0/1) = \alpha$ to receive $R = 0$. On the contrary, if the channel works well we send $D = 1$ and receive $R = 1$, but this occurs only with a probability $p(1/1) = 1 - \alpha$; the same happens if we send $D = 0$ and receive $R = 0$ (with probability $p(0/0) = 1 - \beta$). So, the behavior of this *Binary Asymmetric Channel* (BAC) is described by the (stochastic) *transition matrix* Γ [1]

$$\Gamma = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \tag{5}$$

In our setting, we have nothing to do than using Shannon Information Theory to evaluate the (maximum) quantity of information flowing through the diagnostic channel of figure 1; this will correspond to the accuracy of the diagnostic test. Shannon precisely defined, in mathematical terms, this flow of information; he introduced the concept of *Mutual Information* [21], and we are going to use it to measure the quality of the diagnostic test [13].
Now we have clear the terms of the problem; we have assigned:

- $\mathcal{D} = \mathcal{R} = \{0, 1\}$, that are the two binary sets of the outcomes of disease and test results;
- a (binary) probability distribution $P_D = \{p_D(0), p_D(1)\}$ for the disease D, where

$p_D(0) = \Pr\{D = 0\} = \Pr\{\text{disease is absent}\}$
$p_D(1) = \Pr\{D = 1\} = \Pr\{\text{disease is present}\}$, or *pre-test probability of disease* or *prevalence*;

[1] In the special case of communication channels the matrix is usually symmetric ($\alpha = \beta$), since the system has a symmetric behavior with respect to 0 or 1.

– a (binary) probability distribution $P_R = \{p_R(0), p_R(1)\}$ for the test result R, where

$p_R(0) = \Pr\{R = 0\} = \Pr\{\text{test result is negative}\}$
$p_R(1) = \Pr\{R = 1\} = \Pr\{\text{test result is positive}\}$

The bond between D and R is fixed by the transition matrix (5) of the diagnostic channel of figure 1, and the flow of information between D and R is measured by the *Mutual Information* (*MI*)(see [21], [5]); it is defined as

$$I(D,R) = \sum_{\substack{d\in\mathcal{D}\\ r\in\mathcal{R}}} p(d,r)\log\frac{p(d,r)}{p(d)p(r)} \tag{6}$$

where $p(d,r) = \Pr\{D = d, R = r\}$ $(d \in \mathcal{D},\ r \in \mathcal{R})$ are the *joint* probabilities, $p(d) = \Pr\{D = d\}$ and $p(r) = \Pr\{R = r\}$ are the *marginals*, and the logarithm is taken to the base 2. By taking into account of the Bayes rule $p(d,r) = p(d)p(r/d)$, and that $p(r) = \sum_{d\in\mathcal{D}} p(d,r) = \sum_{d\in\mathcal{D}} p(d)p(r/d)$ as its consequence, we have

$$I(D,R) = \sum_{\substack{d\in\mathcal{D}\\ r\in\mathcal{R}}} p(d)p(r/d)\log\frac{p(r/d)}{p(r)} = \sum_{\substack{d\in\mathcal{D}\\ r\in\mathcal{R}}} p(d)p(r/d)\log\frac{p(r/d)}{\sum_{d\in\mathcal{D}} p(d)p(r/d)} \tag{7}$$

When decoded in terms of SE , FNR , FPR and SP we have

$$\begin{aligned} I(D,R) = {} & PREV\cdot SE\cdot\log\frac{SE}{PR} + PREV\cdot FNR\cdot\log\frac{FNR}{NR} + \\ & + (1-PREV)\cdot FPR\cdot\log\frac{FPR}{PR} + (1-PREV)\cdot SP\cdot\log\frac{SP}{NR} \end{aligned} \tag{8}$$

where $PREV$ is the prevalence, while PR and NR are, respectively, the probability of a positive and of a negative test result. This means that the quantity of information the reader can extract from the diagnostic channel depends on two factors:

– the terms $p(r/d)$ (SE , FNR , FPR and SP), that are the components of the transition matrix Γ (5); they are related to the quality of the diagnostic channel;
– the pre-test probability distribution P_D of the disease ($PREV$).

This implies that we can fix a certain P_D and compare different diagnostic tests, or we can search for the P_D which maximizes $I(D,R)$ for a single diagnostic test. In the Shannon model $I(D,R)$ is a measure of the amount of information exchanged between the two random variables D and R [5]; it is greater than or equal to 0, and it is a measure of the pseudo-distance[2] between two probability distributions, that are the joint probability distribution $P(D,R)$, and the product of the marginals $P(D)\cdot P(R)$. When $P(D,R) = P(D)\cdot P(R)$, then $p(d,r) = p(d)p(r)$ for

[2] From the mathematical point of view, $I(D,R)$ is not a *distance*, because the symmetry and the triangle inequality are not satisfied.

all d and r, all the ratios in equation (6) are equal to 1, and all the logarithms are equal to zero; when this happens we have $I(D,R)=0$, which means that D and R are *independent* each other; in this case there is no flow of information between the two variables. On the contrary, if $p(d,r) \neq p(d)p(r)$ we have $I(D,R) > 0$, and this value is a measure of the distance *from* the condition of independence. The greater $I(D,R)$, the more tied are the two variables, the more information is exchanged between them (see [5] for the mathematical details).
The maximum of Mutual Information is the *Capacity* of the channel [21]

$$C = \max_{P_D} I(D,R) = \max_{P_D}[H(D) - H(D/R)] = \max_{P_D}[H(R) - H(R/D)] \quad (9)$$

where

$$H(X) = -\sum_{i=1}^{K} p(x_i)\log p(x_i) \quad (10)$$

is the *Shannon Entropy* [21], and

$$H(X/Y) = -\sum_{x,y} p(x,y)\log p(x/y) = \sum_{y} p(y)H(X/Y=y) \quad (11)$$

is the *conditional Entropy*, for two random variables X and Y [21]. The capacity corresponds to the maximum amount of information that can flow on the diagnostic channel, when varying the prevalence. This is the best performance the diagnostic test can achieve. Under this framework, comparing the quality of two (or more) diagnostic tests means comparing the capacities of the corresponding diagnostic channels.
We stress on the fact that the Shannon measure of information based on entropy (10) is not only *one of the possible* approaches to measure information, but it is *the only one*; as a matter of fact there exists a theorem, proved by Khinchin in 1956 [12], which shows that the Shannon entropy based on the logarithm is *the only* measure of information which satisfies some basic and reasonable postulates necessary to coherently define an information measure [1]. In this setting an event with probability p carries an amount of information equal to $-\log p$; note that a sure event, for which $p=1$, carries no information, while a very unlikely event, with p tending to 0, carries a quantity of information that approaches infinity.
To evaluate the best performance of a diagnostic test, we have to maximize the *MI* (6) for the binary asymmetric diagnostic channel of figure 1, over all possible prevalences; this leads to the maximum amount of information that a diagnostic test can provide, that is to the capacity of the channel. After some mathematical manipulation (see [2]), *MI* can be expressed as

$$I(D,R) = h(u) + (h(\alpha) - h(\beta))p_D(0) - h(\alpha) \quad (12)$$

where $h(.)$ is the binary Shannon entropy function defined as

$$h(p) = -p\log_2 p - (1-p)\log_2(1-p) \quad (13)$$

and $u = \alpha + p_D(0)(1-\alpha-\beta)$ $(0 \leq u \leq 1,\ \alpha+\beta<1)$. Note that $I(D,R)$ is in the range $0 \ldots 1$. From [14] we have

$$C_{BAC} = \frac{\beta}{1-\alpha-\beta}\cdot h(\alpha) - \frac{1-\alpha}{1-\alpha-\beta}\cdot h(\beta) + \log_2(1+z) \quad (14)$$

where

$$z = 2^{\frac{h(\beta)-h(\alpha)}{1-\alpha-\beta}} \tag{15}$$

The prevalence $p_D^*(1)$ that achieves this capacity is given by

$$p_D^*(1) = \frac{(1-\beta)(1+z)-1}{(1-\alpha-\beta)(1+z)} \tag{16}$$

In figure 2a we can see the behavior of capacity as a function of $SE = 1-\alpha$ and $\beta = 1-SP$, while in figure 2b we see the corresponding prevalence $p_D^*(1)$ that achieves capacity. Note that the capacity is high also when α and β tend to 1, that is SE and SP tend to 0; this is not surprising, since the complement of a systematically incorrect test result is systematically correct.

Place figure 2 about here

As matter stands, we could compare the quality of different diagnostic tests by comparing the values of the associated capacities (14); note, however, that this approach is insidious, since we are not able to control the prevalence, that is to set it to the value $p_D^*(1)$; as a matter of fact this value is in the range $0,4...0,6$ for the vast majority of α and β (see figure 2b). Note, further, that the concept of "pre-test probability of disease" is ambiguous in this context, since a diagnostic test is usually administered to people who are suspected to have a specific disease, and not to a sample of population. This means that the effective probability that is supplying the diagnostic channel, is something similar to $\Pr\{$*disease/knowing the patient is suspected to have a disease*$\}$, which is practically impossible to evaluate. All these problems can be avoided by assessing the accuracy of the diagnostic test in terms of the area under curve (AUC) subtended by the *MI-Curve*, that is the plot of *MI* variation over $p_D(1)$ changes. Calculating the AUC corresponds to evaluate the diagnostic test globally, over all possible pre-test probabilities of disease. This approach requires to evaluate the definite integral of function (12). If we set $p_D(0) = x = 1-p_D(1)$ and $MI = I(D,R)$ for sake of simplicity, we need to calculate

$$\int_0^1 MI \,\mathrm{d}x = \int_0^1 \left(h(u) + (h(\alpha)-h(\beta))x - h(\alpha)\right) \,\mathrm{d}x \tag{17}$$

where $u = \alpha + x(1-\alpha-\beta)$ $(0 \le u \le 1,\ \alpha+\beta < 1)$. This integral can be evaluated by parts; here is the final value

$$\int_0^1 MI \,\mathrm{d}x = \frac{1}{1-\alpha-\beta}\left\{\frac{(-\beta)^2\log(\beta) + (1-\beta) - (1-\beta)^2\log(1-\beta)}{\log 4} - \frac{(\alpha-1)^2\log(1-\alpha) + \alpha - \alpha^2\log(\alpha)}{\log 4}\right\} - \frac{h(\alpha)+h(\beta)}{2} \tag{18}$$

where the log are now to the natural base e.
Note that the *MI-Curve* area of the standard of reference can be computed by substituting $\alpha = \beta = 0$ in the formula (18); this leads to

$$\int_0^1 MI \,\mathrm{d}x \,\Big|_{\alpha=\beta=0} = \frac{1}{\log 4} = 0.7213 \tag{19}$$

Figure 3 shows the *MI-Curve* of the standard of reference (SR), for which $SE = SP = 1$, and the *MI-Curves* of two hypothetical diagnostic tests with $SE = SP =$

0.99 (A) and $SE = 0.80, SP = 0.90$ (B), respectively. Moreover, we show the *MI-Curve* of the diagnostic test for which $SE = SP = 0.5$ (C '+'), which gives random results; the curve is compressed to zero, and lies on the segment 0...1. Since the AUC subtended by the standard of reference is associated to the best possible performance, we can relate to it the AUC of the *MI-Curve* we want to evaluate, so as to obtain a real number in the interval $0 \dots 1$. We call it the *Information Ratio* (IR) of the diagnostic test.

Place figure 3 about here

$$\mathrm{IR} = \log 4 \int_0^1 MI \, \mathrm{d}x \tag{20}$$

In the examples of figure 3, the *IR*s of the diagnostic tests are respectively 0.91 (A), 0.38 (B) and 0 (C). Note that the *IR* is very sensitive to small variation of α and β; e.g., the transition from $\alpha = \beta = 0.01$ to $\alpha = \beta = 0.02$ leads to a change of IR from 0.907 to 0.841. It is obvious that we can restrict the integral limits in the case we have assigned a specific interval for prevalence.
If we compare *IR* and *standard accuracy*[3] for different values of α $(= \beta)$ in the range $0 \dots 0.5$, we can note that the random diagnostic test with $\alpha = 0.5$ has $IR = 0$, since the situation is equivalent to a random choice of the diagnosis, and no information flows on the diagnostic channel[4]; on the contrary, standard accuracy has a residual value of 50%, that is completely misleading.

2.3 Modeling multi-value diagnostic tests with variable threshold

The above model can be easily applied to the case of diagnostic tests expressing the results on an interval or rank scale, for which ROC analysis is commonly used to handle different thresholds of $\beta = 1 - SP$ [15]; in this case we can obtain a IR value for each β. This leads to three main consequences. First, it is possible to find the optimal threshold that leads to the maximum IR for the specified test. Second, we can draw the iso-informational IR curves on the ROC plane $1 - SP/SE$ (see figure 4), that can be used to directly quantify the IR associated to each point of the ROC curve. Third, we can build an *Information Ratio Curve* (IRC) by plotting these IR values vs β.
The AUC of the IRC can be related with the AUC of the *Limit Information Curve* (LIC), drawn by fixing $SE = 1$ $(\alpha = 0)$ for all values of β, which corresponds to the curve associated with the maximum amount of information we can gain for each value of SP. We call this ratio the *Global Information Ratio* (GIR) of that test.

Place figure 4 about here

To build the *Limit Information Curve* we need to set $\alpha = 0$ in equation (20); this leads to

$$IR(\beta) \Big|_{\alpha=0} = \log 4 \int_0^1 MI \, \mathrm{d}x \Big|_{\alpha=0} = \frac{(-\beta)^2}{1-\beta} \log(\beta) + 1 - (1-\beta) \log(1-\beta) - \log 4 \, \frac{h_2(\beta)}{2} \tag{21}$$

[3] Standard accuracy sums up the fraction of good reports with respect to the total number of reports. It corresponds to the probability of a correct (positive or negative) diagnosis, that is $(TP + TN)/(TP + FP + TN + FN)$.

[4] In the Shannon context this is a special case of the so called *useless channel*, characterized by the relation $\beta = 1 - \alpha$, that leads to a transition matrix with two identical rows.

that represents the LIC as a function of β. In figure 5 we can see the behavior of the LIC, together with the IRC curve derived from a hypothetical simulated test. The GIR is the ratio between the AUC of the IRC and the AUC subtended by the LIC; to obtain the last one we need to integrate the function (21) on β

Place figure 5 about here

$$\begin{aligned} AUC_{LIC} &= \int_0^1 IR(\beta)|_{\alpha=0}\ \mathrm{d}\beta = \\ &= \int_0^1 \frac{(-\beta)^2}{1-\beta}\log(\beta)+1-(1-\beta)\log(1-\beta)-\log 4\ \frac{h_2(\beta)}{2}\mathrm{d}\beta = \\ &= Li_2(1-\beta)+2\beta-\beta\log\beta\ \Big|_0^1 = 2-\frac{\pi^2}{6} = 0.35506 \end{aligned} \tag{22}$$

where $Li_2(x) = \sum_{k=1}^{\infty} x^k/k^2$ is the *Polylogarithm function*. So, we have

$$GIR = \frac{AUC_{IRC}}{0.35506} \tag{23}$$

which represents the summary measure of diagnostic performance of the variable-threshold test in terms of informational analysis.
Even if high values of GIR are usually associated with high values of maximum IR (and *vice-versa*), it is not difficult to show cases where two identical values of maximum IR leads to one GIR significantly greater than the other. So IR and GIR need to be analyzed separately.

2.4 Application to a real scenario

We exemplify our model by using data from the study by Pisano et al. [19], who compared *Digital Mammography* (DM) versus *Film Mammography* (FM) in diagnosing breast cancer in a screening population of 42,760 women recruited over 33 referral centers. The standard of reference was represented by a breast biopsy performed within 15 months after the study entry, or a follow-up mammogram obtained at last 10 months after the study entry. The accuracy of FM and DM was assessed by performing ROC analysis on the entire study cohort and several subgroups (including age, breast density and menopausal status), using a *7-points malignancy scale* (1 = definitely not malignant; 2 = almost definitely not malignant; 3 = probably not malignant; 4 = possibly malignant; 5 = probably malignant; 6 = almost definitely malignant; 7 = definitely malignant); in the analyses, scores of 4, 5, 6, and 7 were defined as positive, while scores of 1, 2, and 3 were defined as negative. After extracting data from the Table 3 of paper [19], we applied our model and recalculated unfitted, empirical operating points of the ROC curve using the same 7-points malignancy scale on the entire study cohort after 455 days of follow-up. We refer to the "main threshold" of $1 - SP/SE$ at ROC analysis as the ones determined with the decision rule explained in the original article, which corresponds to a threshold in column 4. Analysis was performed using a dedicated software, developed in GNU Octave, and freely available on request.

3 Results

The table 1 shows $1 - SP$ and SE of film and digital mammography calculated using *TP, FN, TN* and *FP* extracted from our reference work [19], for each cut-off point. Maximum IR for film mammography is found in the point corresponding to the main threshold used in original ROC analysis (column 4). On the contrary, maximum IR for digital mammography corresponds to a higher cut-off compared to that used with ROC analysis (column 3), with an increase in SE (49 cancers) and in *FP*s (2175 cases). Figures 6a and 6b show the IRC curves and the corresponding LIC for FM and DM, together with the corresponding GIR values. The original ROC analysis showed AUCs of 0.735 and 0.753 for FM and DM, respectively, while the GIR values of IRC analysis are 0.200 and 0.229 for FM and DM, respectively.

Place table 1 about here

Place figure 6 about here

4 Discussion

The American mathematician and engineer C. Shannon (1916-2001) built the foundations of Information Theory in 1948 [21]; it formalizes the mathematical rules underlying telecommunications and defines the general properties that communications systems should have in order to transmit information reliably and affordably [23]. Since the theory provides objective measures of information, it is a pillar in telecommunication technologies and has been used to model a variety of phenomena in different fields, including physics [11], neuroscience [22], molecular biology [7] and others.

We propose an information theoretic model in which SE , SP , FPR and FNR are interpreted as the probabilities of correct and incorrect signal transmission through an asymmetric binary diagnostic channel; here a certain quantity of average information (namely, the Mutual Information) flows from the "hidden" status of the disease to the reader, e.g., a radiologist interpreting mammograms. The higher the MI, the higher the diagnostic information available to the physician interpreting the test. Not surprisingly, this measure has already been hypothesized as a possible tool to compare different classifiers [13]. However, the use of Information Theory-derived indexes is quite limited in medical-statistics literature; an example is given by the Akaike information criterion [20], based on entropy, which is a measure of the relative quality of a statistical model for a given set of data; nevertheless it is not useful to assess the information associated with a diagnostic test [6].

In our model, IRC analysis provides two different informational measures of test accuracy, namely the *Information Ratio* and the *Global Information Ratio*. IR expresses, within the interval 0...1, how much information on the disease a dichotomous test carries on compared to the standard of reference, that is compared to the ideal channel operating without the "noise" associated to incorrect diagnosis. Of note, IR results from a *MI-Curve* representing the MI values at all possible pre-test probabilities of disease. Consequently, this measure is not only independent from this probability in a given clinical experiment, but shows the additional advantage of expressing the maximum information for *all* clinical scenarios *at one time* (e.g., screening or high-pre-test probability populations). The typical *MI-Curve* behavior of figure 3 derives from a mathematical structure of the corresponding formula, which is similar to that of the binary entropy (13) [5]; so the curve tends

to zero when the prevalence approaches 0 or 1, as it should be from the intuitive point of view: a diagnostic test applied to a population where all the members are ill (or healthy) carries no information. On the contrary, the maximum amount of diagnostic information is associated to the situation where the prevalence is close to 1/2.

Even though there may be preference for either specificity or sensitivity, due to the kind of diagnostic test (screening or not) or to the costs for FPs and FNs, nevertheless a single statistical measure that can summarize the global quality of a dichotomous diagnostic test is still lacking [18] (p.340). IR gives such a measure. While the IR is applicable to tests providing a dichotomous result, the GIR applies to tests yielding continuous values (e.g. prostate PSA) or ordered categories (e.g. 7-points scale in breast imaging), corresponding to multiple cut-offs of specificity. Indeed, the GIR is the ratio between the areas under the *Information Ratio Curve* of the test and the *Limit Information Curve*, i.e. the reference curve representing the maximum amount of information we can gain by varying 1-*SP* on the abscissa when *SE* is fixed to 100%. In this setting, the GIR plays the role of an informational summary measure [16], that expresses the accuracy by using a single number, and facilitates the comparison between diagnostic tests.

ROC analysis, which is the state-of-the-art method for describing the diagnostic accuracy of a test, is universally used and shows well known advantages [4], [8], [15]. Which are the potential, additional benefits of using the informational analysis? In the Appendix, we briefly summarize some technical points and show how it is possible to go over some weakness of the classical ROC/AUC approach.

On this basis we believe that, as usual for a new-born instrument, both theoretical refinements and concrete clinical applications will probably clarify in which scenarios our method can support or even outperform ROC analysis; however, whether this potential is realized will be assessed by clinicians. At present, we suggest the use of IRC analysis as a complement to strengthen ROC analysis results with a formally consistent and independent method. Why? In IRC analysis, accuracy is calculated directly from the amount of information flowing through a diagnostic channel, which in turn depends on the channel properties. In standard analysis, information related to test properties can be inferred only indirectly: we suppose something on diagnostic information based on the count of the expected outcomes, but we do not have a direct access inside the "black box" of the test, in particular to quantify the information made available by the test to obtain the sensitivity and specificity measured by ROC analysis under certain clinical conditions. Informational indexes do provide such a quantification. Three potential consequences may occur in this respect. The first one is exemplified by the coincidence between highest IR and the cut-off of 1-SP/SE of ROC analysis we observed for FM in the paper of Pisano et al. [19]. In this situation, best-coupled sensitivity/specificity corresponds to the maximum information provided by the test, i.e. IRC analysis shows the upper limit of tests capabilities. One might assume that if they correspond to suboptimal test accuracy, no further improvement is permitted unless changing the diagnostic channel (e.g., by incorporating a technological development such as digital mammography) or supplying it with additional diagnostic tools (change in diagnostic workup). The second scenario also emerges from our clinical example [19]: the IR cut-off for DM "moved" to the right of the ROC analysis cut-off, leading to an increase in SE (49 cancers) at expense of a significant increase in FPs (2175 cases). Our explanation for this finding is that higher infor-

mation provided by DM translated in better visibility of subtler anatomical details mimicking malignancy (e.g., spiculated margins or microcalcifications) that might have acted as confounders, thus increasing sensitivity and decreasing specificity. In a third, theoretical scenario, the IR cut-off moves "to the left" of the ROC analysis cut-off, suggesting that information provided by a certain test tends to decrease sensitivity and increase specificity. This might occur if the higher information carried by this test is used to enlighten the signs that tend to exclude the disease. The last two scenarios suggest that, regardless a rigorous quantification of information, the comparison between ROC analysis and IRC analysis has the potential to indicate the "direction" (sensitivity or specificity) towards which information associated to a new technology "move" when interacting with the reader. Again, one can suppose that such a knowledge is of help in refining diagnostic tools (e.g., by guiding the technological upgrade) and/or diagnostic strategies. Whether this is clinically acceptable/relevant or not depends on the specific setting, the type of diagnostic test, and the decision rules used to establish the proper cut-off. We hypothesize that such a capability is of potential relevance at least for all those tests providing the diagnosis with inherently different technologies, i.e. channels with different informational properties.

The IRC analysis suffers, however, some limitations.

First, similarly to ROC analysis [15], our method is based on the weak assumption that, when human readers participate into the diagnostic process (as occurs for radiologists), they always interpret correctly information provided by the test, so that FN and FP cases depends on the test only. In other words IRC analysis, in its present form, probably overestimates diagnostic accuracy and does not account for intra- and inter-reader variability. How to overcome these limitations? We suggest that informational indexes might be obtained for each of the readers of a "multireader factorial" study design, which is currently recommended for conventional analysis of accuracy [17]. In this multireader/multicase model, the same patients undergo all of the diagnostic tests under study and an adequate number of readers interpret the results from all the diagnostic tests. One can assume that the comparison of IR or GIR values measured on the same clinical setting is *per se* indicative of readers-related variability. On the other hand, similarly to conventional analysis, IRC analysis might be complemented by calculation from raw data of classical indexes of intra- and/or inter-rater agreement for different number of raters and types of variable, such as *Cohen's kappa* or *Intraclass Correlation Coefficient* (ICC). Lines of future improvement of IRC analysis include the development of methods: *i*) to adjust α and β errors of the transition matrix for readers-related variability; ii) to adjust for variability in the specific setting of multireader/multicase approach, possibly in accordance with the "random effects" model, that is by treating the variation in readers performance as an independent source of variability [17].

Second, our model does not yet include: *i*) a strategy to fit properly the empiric GIR curve; and *ii*) the statistical rules to compare different GIRs, that is to attest if different values of GIRs are statistically significant. On the other hand both points *i*) and *ii*) constitutes by themselves an entire field of research [15], as it happended at the very beginning for ROC analysis. We believe all the above points can be faced confidently in future works, and their current lack does not influence qualitatively our results.

5 Conclusions

We proposed a conceptual framework based on Information Theory, in which: *i*) the diagnostic test is equivalent to an asymmetric binary transmission channel; *ii*) the diagnostic process can be modeled using Mutual Information, which is a well-established measurement of the information exchanged through the channel between a transmission source (assumed to be represented by the disease) and a receiver (assumed to be represented by the physician). On this basis, the accuracy of a diagnostic test can be expressed, in informational terms, through two summary measures, namely the IR and GIR, which apply to tests providing dichotomous and continuous results, respectively. Potential advantages of using informational indexes are related to the independency from pre-test probability and the capability to refine conventional analysis of accuracy (e.g., ROC analysis) by assessing the type and limits of diagnostic information that a certain test can vehicle, that is by objectifying how much diagnostic information of a certain type (e.g., breast cancer) is transmitted given test properties (e.g., being a film or digital image obtained using X-rays). Using data from a previous study, we showed that informational analysis is applicable to a real clinical scenario and can represent an informational counterpart of standard analysis of accuracy based on ROC curves.

6 Appendix

Here we discuss some technical points regarding the additional benefits of using the IRC analysis instead of (or in conjunction with) the classical ROC/AUC approach.

The cut-off problem - When we are dealing with multi-value diagnostic tests, we have the problem of choosing the best decision threshold. Even though there are several possible approaches to solve this problem, the choice for the optimal cut-off is substantially arbitrary [15], and depends from the context. Here are some examples: (*i*) from the geometrical point of view it could be obvious to select the point $P = (1 - SP, SE)$ on the ROC curve which minimizes the Euclidean distance between the higher-left corner of the ROC space, whose coordinates are (0,1), and P itself. But this choice is acceptable, in clinical practice, only when SE or SP have high values; the informational approach and the figure 4 explain us why. (*ii*) from the classifier point of view, the best reasonable choice is to use the decision threshold which maximize the summation $TP + TN$ [3], or something more elaborated, such as any objective function that is a linear combination of true and false positive rates via the *convex hull* [8]. (*iii*) another possible method is maximizing $SE + SP$, that is equivalent to the use of the so-called Youden index [9]; it uses the maximum vertical distance of ROC curve from the point (x, y) on the diagonal (chance) line. So Youden index maximizes the difference between SE and 1-SP, that is $SE + SP - 1$. (*iv*) the clinical practice of the 7-points scale we used in the Pisano's example, uses a cut-off essentially based on the semantics of the graded scale.

If we apply all these different criterions on table 1, we would get completely different cut-offs.

Among all these reasonably, but arbitrary choice of the optimal cut-off, the IRC analysis is *the only* to offer a criterion based on the maximization of the

informational flow between the patient and the clinician, that is the scope of any good diagnostic system. Since we have only one coherent measure of information (remember the unicity theorem of Khinchin [12]), this method is even the only "objective" in the case we assign the same cost to FP and FN.

Fallacy of the undistributed middle - ROC curves, even if widely used, have at least one significant pitfall, that is the so called "fallacy of the undistributed middle" [8]: all random models score an AUC of 0.5, but not every model that scores an AUC of 0.5 is random; in other words $AUC = 0.5$ does not necessarily imply that the classifier is no better than random guessing. An interesting example of this situation is given in figure 4a of [3], where the "two-thresholds classifier" assumes that a certain quantity q, being below a threshold t_1 or exceeding a threshold $t_2 > t_1$, indicates *disease*; while $q \in [t_1, t_2]$ means *normal*. Systolic blood pressure or expression of a gene could be an example of such a situation. If we use ROC analysis, the AUC equals 0.5 (figure 4b of [3]), and so the AUC approach is not able to discriminate this good classifier from a random one. On the contrary, if one use the IRC approach, one can clearly discriminate this classifier, characterized by $GIR = 0.342$, from the random one ($GIR = 0$).

IRC analysis gives us a global evaluation of the diagnostic test performance (GIR) and a choice for the best threshold of specificity. As a consequence it could be considered, at least in principle, as a competitor of ROC curves. Unfortunately, up to this moment, the IRC tool is not yet complete, since it lacks a strategy to fit properly the empiric GIR curve and a statistical rule to compare different GIRs, that is to attest if different values of GIRs are statistically significant.
Nevertheless we can appreciate the fact that IRC analysis is able to strengthen ROC analysis, taking it to a deeper level. We suggest here two issues in this perspective:

Informational chart - The iso-informational curves of figure 4, that are drawn on the ROC plane, could be used as a sort of graph paper (informational chart) over which one can draw a ROC curve; this allow to directly visualize the information associated to each point of the ROC curve, integrating the ROC and the IRC approaches together.

The worst ROC curves are good - Always from figure 4, we can note that each iso-level curve of the upper left side of the diagram has a same-level correspondence in the lower right side of the same diagram. Even if this is not usual in clinical practice, this means that, in an informational sense, a bad diagnostic test, whose ROC curve stands entirely below the diagonal, performs as a good one, whose ROC curve stands entirely above the diagonal. This is not surprising, since the complement of a systematically incorrect test result is systematically correct. This would suggest that a more correct measure of the diagnostic test performance using the ROC approach, with a ROC curve entirely standing above or below the diagonal, should be the AUC* between the ROC curve and the diagonal, that is $AUC^* = |AUC - 0.5|$.

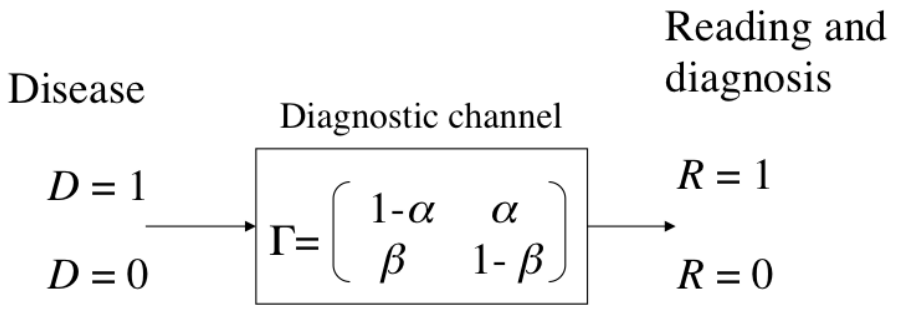
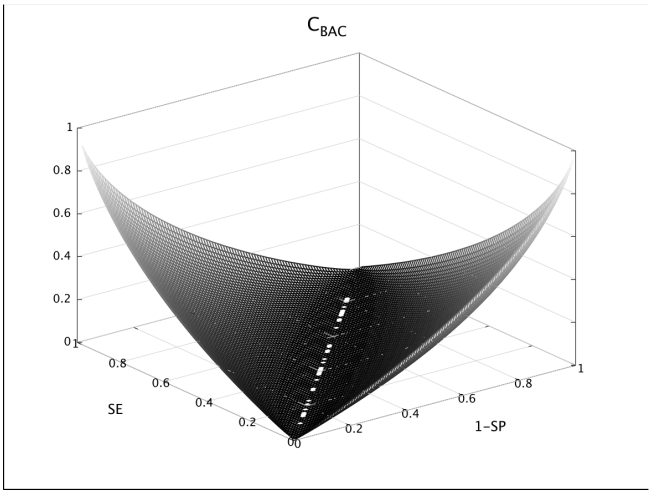
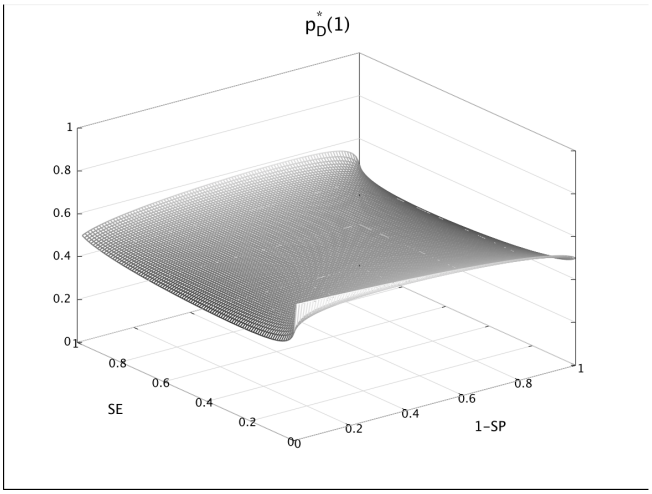


Fig. 1 The disease is a hidden, objective status of the patient that can be revealed to the physician through the "black box" of a diagnostic test. It corresponds to a binary asymmetric diagnostic channel, whose behavior is described by the transition matrix Γ (5).



(a) Capacity



(b) Pre-test probability of disease

Fig. 2 Capacity of the diagnostic channel and pre-test probability of disease $p_D^*(1)$ that achieves this capacity.

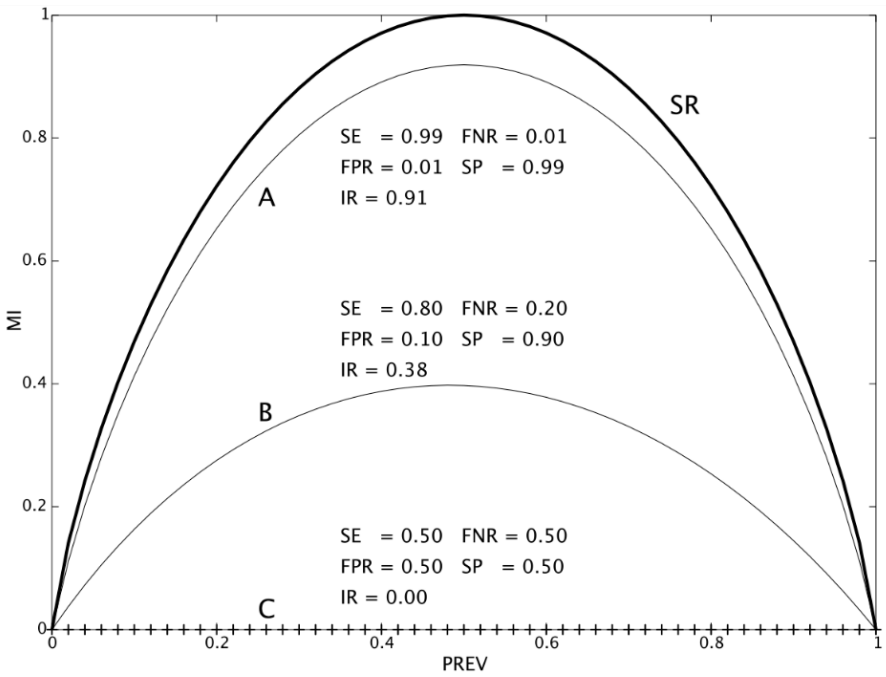


Fig. 3 *Mutual Information Curves* (*MI-Curves*) and *Information Ratios* (IR) of *i)* the standard of reference (SR), for which $SE = SP = 1$ *ii)* a diagnostic test with $SE = SP = 0.99$ (A) *iii)* a diagnostic test with $SE = 0.80, SP = 0.90$ (B) and *iv)* the random test for which $SE = SP = 0.5$ (C, '+')

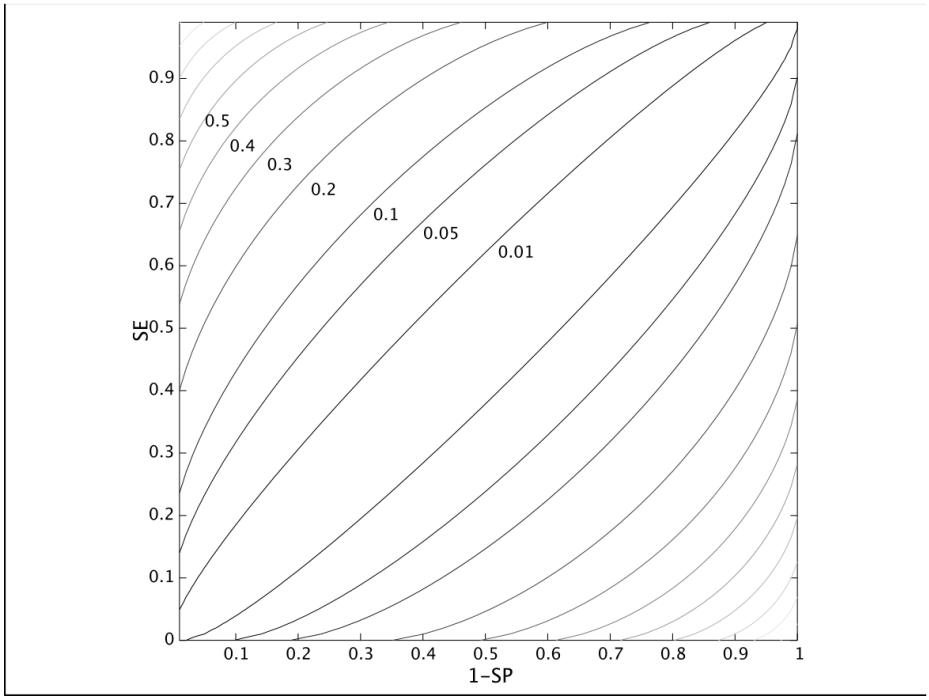


Fig. 4 Iso-informational IR curves as a function of $SE = 1 - \alpha$ and $\beta = 1 - SP$ of the ROC plane.

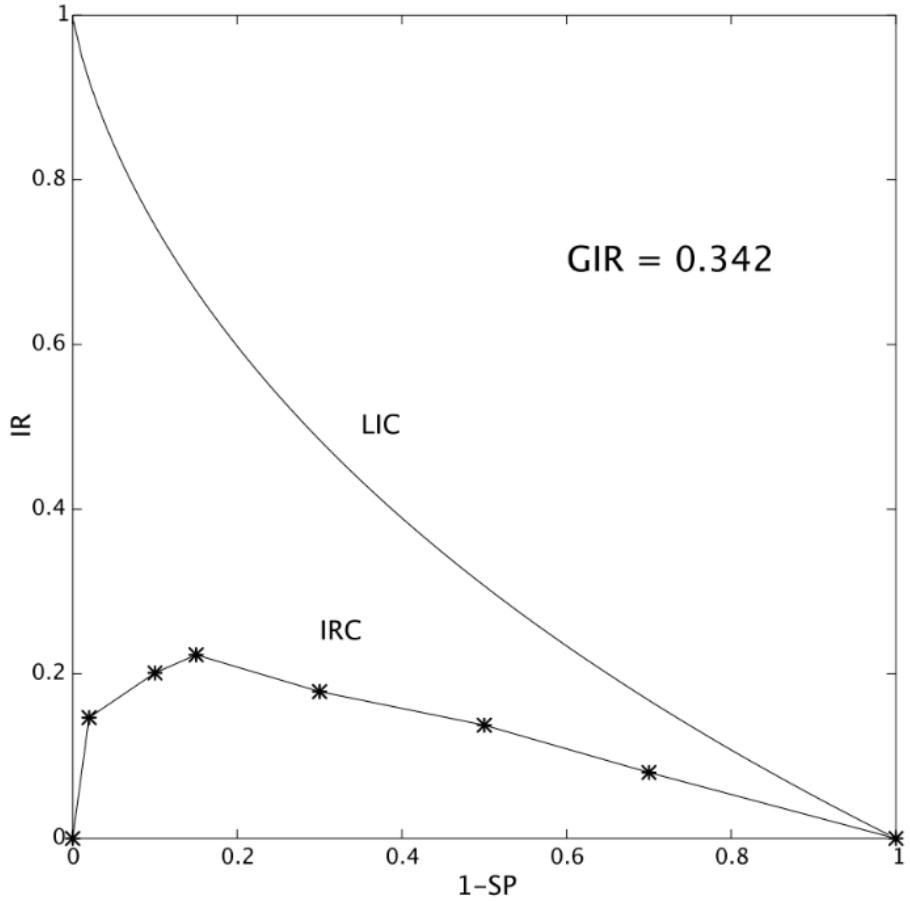
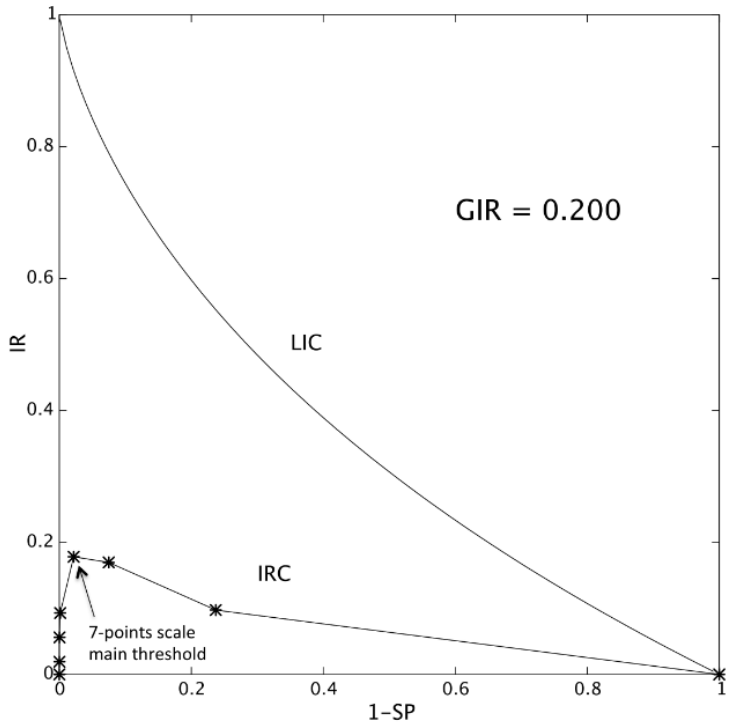


Fig. 5 The *Limit Information Curve*, obtained when $\alpha = 0$, and the IRC of a hypothetical simulated test

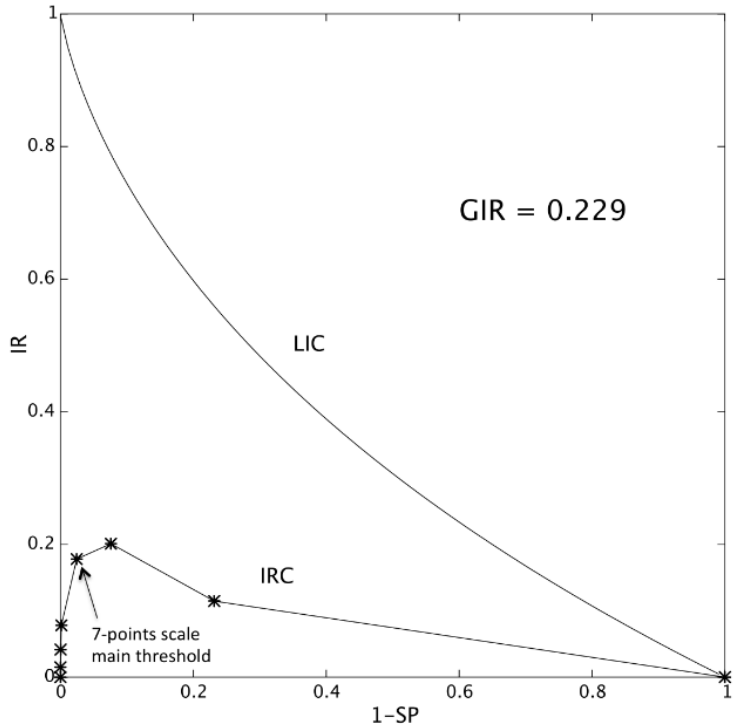
Table 1 *Information Ratio* values calculated on data extracted from Table 3 of the study [19]. Highest IRs are highlighted in bold. Asterisks indicate the values of *SE* /1-*SP* 7-points scale threshold used in the ROC analysis of the original study.

	Film Mammography						
Score	7	6	5	4	3	2	1
IR	0.019	0.056	0.093	**0.178**	0.170	0.098	0.000
TP	13	37	62	136	171	204	335
FN	322	298	273	199	164	131	0
TN	42406	42401	42356	41488	39232	32355	0
FP	4	9	54	922	3178	10055	42410
SE	0.039	0.110	0.185	0.406*	0.510	0.609	1.000
1-SP	0.000	0.000	0.001	0.022*	0.075	0.237	1.000

	Digital Mammography						
Score	7	6	5	4	3	2	1
IR	0.015	0.042	0.078	0.178	**0.201**	0.115	0.000
TP	10	28	53	138	187	212	334
FN	324	306	281	196	147	122	0
TN	42235	42224	42180	41204	39029	32466	0
FP	1	12	56	1032	3207	9770	42236
SE	0.030	0.084	0.159	0.413*	0.560	0.635	1.000
1-SP	0.000	0.000	0.001	0.024*	0.076	0.231	1.000



(a) Film mammography: the threshold corresponds to the point associated with the maximum information



(b) Digital mammography: the threshold does not correspond to the point associated with the maximum information

Fig. 6 IR curves for film and digital mammography, with the data of table 1 derived from [19]

References

1. Aczél J, Daróczy Z. (1975) On Measures of Information and their Characterizations. Mathematics in Science and Engineering, vol. 115. Academic Press, New York-London

2. Amblard P, Michel OJJ, Morfu S (2005) Revisiting the asymmetric binary channel: joint noise-enhanced detection and information transmission through threshold devices. Proc. SPIE 5845, Noise in Complex Systems and Stochastic Dynamics III, Kish LB, Lindenberg K, Zoltan G, Editor(s)

3. Berrar D, Flach P (2012) Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform. Jan;13(1):83-97. Epub 2011 Mar 21.

4. Bewick V, Cheeck L, Ball J (2004) Statistic review 13: Receiver operating characteristic curves. Crit. Care 8(6):508-12.

5. Cover TM, Thomas JA (1991) Elements of Information Theory. John Wiley & Sons, Inc.

6. El Khouli RH, Macura KJ, Kamel IR, Jacobs MA, Bluemke DA (2011) 3-T Dynamic Contrast-Enhanced MRI of the Breast: Pharmacokinetic Parameters Versus Conventional Kinetic Curve Analysis. Am. J. Röentgenol. 197(6):1498-1505.

7. Fabris F (2002) Shannon Information Theory and Molecular Biology. J. of Interdisc. Math. 5:203-20.

8. Flach P (2010) ROC Analysis. Encyclopedia of Machine Learning, In: Sammut C, Webb, GI. Editor(s), Berlin/Heidelberg:Springer:86974.

9. Fluss R, Faraggi D, Reiser B (2005) Estimation of Youden index and its associated cutoff point. Biom J. 47:458-72.

10. Gehlbach SH (1993) Interpretation: sensitivity, specificity, and predictive value. In: Gehlbach SH, ed. Interpreting the medical literature. McGraw-Hill, New York, 129-39

11. Keyl M (2002) Fundamentals of quantum information theory. Phys. Rep. 369:431-548

12. Khinchin AI (1957) Mathematical Foundations of Information Theory. Dover, New York

13. MacKay DJC (2003) Information Theory, Inference, and Learning Algorithms. Cambridge University Press, Cambridge

14. Moser SM (2009) Error probability analysis of binary asymmetric channels. Dept. El. & Comp. Eng., Nat. Chiao Tung Univ.

15. Obuchowsky NA (2005) ROC analysis. Am. J. Röentgenol. 184(2):364-72

16. Obuchowsky NA (2003) Receiver Operating Characteristic Curves and their use in Radiology. Radiology 229:3-8

17. Obuchowski NA, Beiden SV, Berbaum KS, Hillis SL, Ishwaran H, Song HH, Wagner RF (2004) Multireader multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad. Radiol. 11(9):980-95

18. Peacock J, Peacock P (2010) Oxford Handbook of Medical Statistics, Oxford University Press, Oxford

19. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S et al. (2005) Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. N. Engl. J. Med. 353(17):1773-83

20. Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. Syst. Biol. 53(5):793808

21. Shannon CE (1948) A mathematical theory of communication. Bell Syst. Tech. J., 27:379-423, 623-56

22. Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5:42-64

23. Verdu S (1998) Fifty years of Shannon Theory. IEEE T. Inform. Theory 44:2057-78

24. Weinstein S, Obuchowski NA, Lieber ML (2005) Clinical evaluation of diagnostic tests. Am. J. Röentgenol. 184(1):14-9

25. Zhou XH, Obuchowski NA, McClish DK (2002) Statistical Methods in Diagnostic Medicine. Wiley Series in Probability and Statistics. A John Wiley and Sons, Inc., Publication, New York