

METHODOLOGY

Open Access



An information-oriented paradigm in evaluating accuracy and agreement in radiology

Alberto Casagrande¹, Francesco Fabris^{1*}  and Rossano Girometti²

Abstract

The goal of any radiological diagnostic process is to gain information about the patient's status. However, the mathematical notion of information is usually not adopted to measure the performance of a diagnostic test or the agreement among readers in providing a certain diagnosis. Indeed, commonly used metrics for assessing diagnostic accuracy (e.g., sensitivity and specificity) or inter-reader agreement (Cohen κ statistics) use confusion matrices containing the number of true- and false positives/negatives results of a test, or the number of concordant/discordant categorizations, respectively, thus lacking proper information content. We present a methodological paradigm, based on Shannon's information theory, aiming to measure both accuracy and agreement in diagnostic radiology. This approach models the information flow as a "diagnostic channel" connecting the state of the patient's disease and the radiologist or, in the case of agreement analysis, as an "agreement channel" linking two or more radiologists evaluating the same set of images. For both cases, we proposed some measures, derived from Shannon's mutual information, which can represent an alternative way to express diagnostic accuracy and agreement in radiology.

Key points

- Diagnostic processes can be modeled with information theory (IT).
- IT metrics of diagnostic accuracy are independent from disease prevalence.
- IT metrics of inter-reader agreements can overcome Cohen κ pitfalls.

Keywords Diagnostic imaging, Information theory, Observer variation, Research design, Sensitivity and specificity

Background

One can hypothesize that any examination aims to extract as much information as possible about a disease (its presence and/or its gravity or stage), which can be assumed a "hidden," objective status of a patient according to a standard of reference. Analogously, gauging the agreement in

radiology can be seen as measuring how much information is shared between different readers assessing a certain condition, e.g., by using the prostate imaging reporting and data system (PI-RADS) categories for assessing prostate cancer [1]. Of note, the notion of information, rather than being vague, can be expressed quantitatively and rigorously in accordance with the mathematical apparatus underlying the so-called information theory (IT), which is the base of current telecommunication systems technology [2]. Details on the mathematical definition of information and derived measures can be found in the milestone paper with which Claude Shannon founded IT in 1948 [3]. Based on the above assumptions, some IT-derived measures of both diagnostic accuracy and agreement have previously been built [4–7],

*Correspondence:

Francesco Fabris
ffabris@units.it

¹ Dipartimento di Matematica e Geoscienze, Università degli Studi di Trieste, Trieste, Italy

² Istituto di Radiologia, Dipartimento di Area Medica, Università degli Studi di Udine, Udine, Italy

proving their formal consistency. The purpose of this article is to present them to the radiological community, describing their conceptual fundamentals and potential advantages compared to conventional statistics such as receiver operating characteristic (ROC) analysis and Cohen κ .

The information-oriented paradigm

The mutual information (MI) [3] is a measure of the (average) quantity of information exchanged between a sender and a receiver communicating on a channel. It is defined by the formula as follows:

$$MI(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x, y) * \log \frac{p_{XY}(x, y)}{p_X(x) * p_Y(y)}$$

where X and Y are two random variables modeling the input and the output of the channel, \mathcal{X} and \mathcal{Y} are the sets of the possible input and output messages, and $p_X(x)$, $p_Y(y)$, and $p_{XY}(x, y)$ denote the probability for X to equal x , the probability for Y to equal y , and the joint probability for X and Y to equal x and y at the same time, respectively.

MI measures the dependency between X and Y : the more they are related, the higher the mutual information. Whenever the two variables are independent, *i.e.*, their values are non-related, MI drops to 0. The mutual information is symmetric and the role of the channel input and output is not relevant in gauging the information exchanged on the channel itself, *i.e.*, $MI(X, Y) = MI(Y, X)$.

Because of its features, MI can also be used as a correlation index between X and Y , but differently from the Pearson correlation coefficient that exclusively handles linear correlations, MI is a nonlinear index.

Modeling diagnostic information with MI

The diagnostic channel $D \iff X$ is a virtual channel modeling the information flow between the unknown patient condition D (*i.e.*, the channel input) and the outcome X of a diagnostic test (*i.e.*, the channel output) [4, 6–8]. The diagnostic test relates the two random variables D and X as channels connect their input and output, and its goal is to carry the maximum amount of information of the former to the latter.

Since MI is symmetric, the diagnostic channel can be represented as symmetric too: the information flows from the patient condition to the test outcome exactly as it flows from the test outcome to the patient condition. Thus, if the diagnostic channels $D \iff X$ and $D \iff Y$ model two different diagnostic tests or readers, we can relate the outcomes X and Y of the two tests/readers by joining $D \iff X$ and $D \iff Y$ in the single

channel $X \iff D \iff Y$, or, in short, $X \iff Y$. This last channel is the agreement channel [5–7].

The mutual information can be used to both evaluate the accuracy of the test modeled by a diagnostic channel $D \iff X$ and measure the agreement of two tests/readers linked by an agreement channel $X \iff Y$.

An information measure of diagnostic accuracy

Image analysis can be seen as a way to extract information from the patient to correctly diagnose the disease. The most accurate diagnosis will be, in turn, the one extracting as much information as possible. Consequently, the more information flows on the diagnostic channel from the disease to the radiologist, the more accurate the diagnosis is.

The dichotomous case

As shown in [4], whenever the evaluation has only two possible outcomes (dichotomous case), the amount of information flowing in a diagnostic channel, *i.e.*, MI, can be expressed in terms of sensitivity (SE), specificity (SP), and prevalence of the disease PREV as follows:

$$MI(SE, SP, PREV) = h(SP - (SE + SP - 1) * PREV) + (h(SP) - h(SE)) * PREV - h(SP)$$

where $h(x) = -x * \log_2 x - (1 - x) * \log_2(1 - x)$ is the (binary) *Shannon entropy* [3].

The IT-based approach avoids the dependency from the prevalence of the disease by assessing diagnostic accuracy in terms of the area under the curve (AUC) subtended by the *MI-curve* obtained by varying the prevalence for all possible values in the interval [0, 1], *i.e.*, a prevalence of disease ranging from [0] to 100%.

The information ratio (IR) [4] is an information measure that normalizes MI-curve AUC with respect to the best possible performance. In the dichotomic case, IR can be computed by the formula

$$IR(SE, SP) = (\ln 4) * \int_0^1 MI(SE, SP, PREV) dPREV = 1 + \frac{(1 - SP) * h(1 - SE) + (1 - SE) * h(1 - SP)}{SE + SP - 1} * \ln 2 + \frac{(1 - SE) * \ln(1 - SE) + (1 - SP) * \ln(1 - SP)}{SE + SP - 1}$$

where $h(x)$ is the binary Shannon entropy as above. Figure 1 depicts two exemplificative MI-curves: MI for SE 0.9 and SP 0.8 as the prevalence changes and MI for the reference standard, *i.e.*, SE and SP equal to 1. The ratio between the AUC of the former and that of the AUC of the latter is the IR(0.9, 0.8).

IR represents the normalized amount of information carried by a diagnostic process as a value in the interval [0, 1]: the higher diagnostic accuracy, the nearer IR to 1.

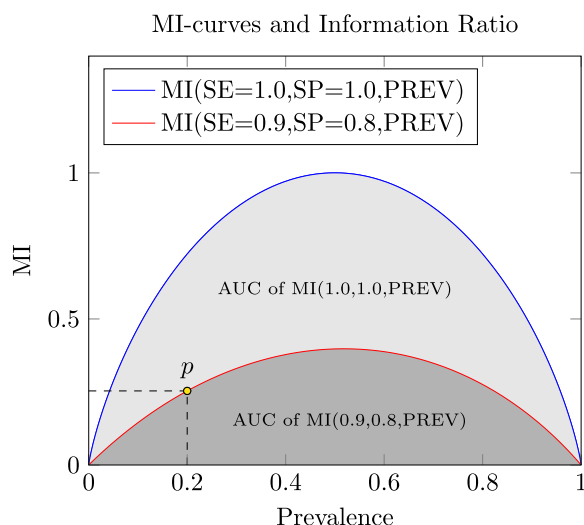


Fig. 1 The red line is the MI-curve for SE 0.9 and SP 0.8 as the prevalence of the disease ranges in the interval [0, 1]. The point p represents the value of MI (value on the MI axis) when the prevalence is 0.2 (value on the prevalence axis). The dark gray region is the area under the MI-curve for SE 0.9 and SP 0.8. It summarizes all the values of MI as the prevalence changes for a diagnostic method whose sensitivity and specificity are 0.9 and 0.8, respectively. This AUC does not range in an interval [0, 1], and it never reach 1. In order to normalize it in the [0, 1] range and get the information ratio, we must divide it by the maximum of the AUC among those obtainable by changing the sensitivity and the specificity. This is the AUC of the diagnostic method that has the best sensitivity and specificity, *i.e.*, 1 and 1. The blue line depicts the MI-curve for SE 1.0 and SP 1.0, *i.e.*, the gold standard, and the gray region (the light gray plus the dark gray regions) represents its AUC. In [4] it is proved that the latter equals $1/\ln 4$; hence, the information ratio of the method whose SE and SP are 0.9 and 0.8, respectively, is the AUC of the red curve multiplied by $\ln 4$. *AUC* Area under the curve, *MI* Mutual information, *SE* Sensitivity, *SP* Specificity, *PREV* Prevalence

Table 1 reports a simulated clinical scenario in which magnetic resonance imaging is used to identify clinically significant prostate cancer in a population of biopsy-naïve men with 40% cancer prevalence [9].

By setting the cutoff for auctioning prostate biopsy to prostate imaging reporting and data system (PI-RADS) category 3, we get Table 2. Since the SE and SP of the data in this table are 0.95 (95%) and 0.5 (50%), respectively, $IR(0.95, 0.5) \approx 0.195$. Hence, due to the imbalance between SE and SP (a typical state-of-the-art clinical scenario in this field), MRI can extract only 19.5% of the possible information about a patient prostate cancer condition on average.

The prevalence issue

While the IR value depends on the SE and SP values, as shown in the formula, IR is calculated by integrating them over all possible prevalence values included between 0 and 1; thus, in fact making the metric independent from

Table 1 Simulated clinical scenario in which magnetic resonance imaging is used to identify clinically significant prostate cancer in a population of 1000 biopsy-naïve men with 40% cancer prevalence

	PI-RADS category	Target biopsy	
		Positive	Negative
Magnetic resonance imaging	5	93	37
	4	201	128
	3	86	135
	2	18	131
	1	2	169
Total number of subjects		400	600

it [4]. Therefore, the results expressed by using it are also valid in cohorts different from the one in which a study has been performed, *e.g.*, cohorts with different prevalences of the disease. Note, however, that many studies have suggested that both SE and SP are related to the prevalence itself [10–13]. This is a problem of spectrum bias, which is a type of sampling bias. In these cases, sampling from a patient population with a higher disease prevalence may include more severely diseased patients making the test performing better. Our analysis does not account for the effect of bias in study design like the abovementioned one. We are currently working on mathematical solutions for trying to overcome this problem.

The multivalue case

Radiologists frequently provide diagnoses by using ranks (*e.g.*, the PI-RADS) or measuring continuous variables (*e.g.*, the apparent diffusion coefficient from diffusion-weighted imaging [14]). In these cases, the receiver operating characteristics (ROC) analysis is commonly used to discover the most appropriate cutoff for the investigated condition identification and assess the diagnostic approach effectiveness. ROC analysis uses each of the values in the considered rank or continuous domain as a possible cutoff for a curated set of diagnoses, and it builds a 2×2 confusion matrix which relates the golden standard and the positive/negative results due to the specific cutoff. SE and SP of each of these matrices depend on the corresponding cutoff value: the greater the cutoff value, the smaller the sensitivity, and the greater the specificity and vice-versa, depending on the clinical scenario. By representing the possible cutoffs as points in a $(1-SP)$ versus SP graph, ROC analysis produces a curve that depicts the effectiveness of the diagnostic method as the cutoff changes (Fig. 2a). The effectiveness is quantified by the AUC of the curve itself which is proven to be the probability for the rank/value assigned to a subject

Table 2 Simulated PI-RADS diagnoses by cut-off selection

		Target biopsy		
		Positive	Negative	
Magnetic resonance imaging	Positive (PI-RADS category ≥ 3)	380	300	Positive predictive value 55.8% (380/680)
	Negative (PI-RADS category ≤ 2)	20	300	Negative predictive value 93.7% (300/320)
		Sensitivity 95% (380/400)	Specificity 50% (300/600)	
		Cohen κ	0.403	
		Information ratio	0.195	

By arbitrary setting the cutoff to PI-RADS category 3 (i.e., categories below 3 are considered a negative response, categories above 2 correspond to a positive response), we obtain Table 2, showing a sensitivity of 0.95 (95%) and a specificity of 0.50 (50%). It follows that the information ratio of the scenario is about 0.195, while for the same data Cohen κ equals 0.403

not having the investigated condition by the diagnostic method to be lower than that of a subject having the condition [15].

In a similar way, the IT-based approach evaluates the IRs of the cutoff-specific matrices, and it plots the information ratio curve (IRC), i.e., the curve of IR as the cutoff is raised or, equivalently, as the specificity decreases.

The IRC AUC does not range in the interval [0, 1]. In order to normalize it, it must be divided by the AUC of the limit information curve (LIC) that is the curve produced by the best theoretical diagnostic method: the one that always has the maximum sensitivity (i.e., 1) for any admissible value of the specificity (Fig. 2b). This ratio is the global information ratio (GIR). Since LIC AUC

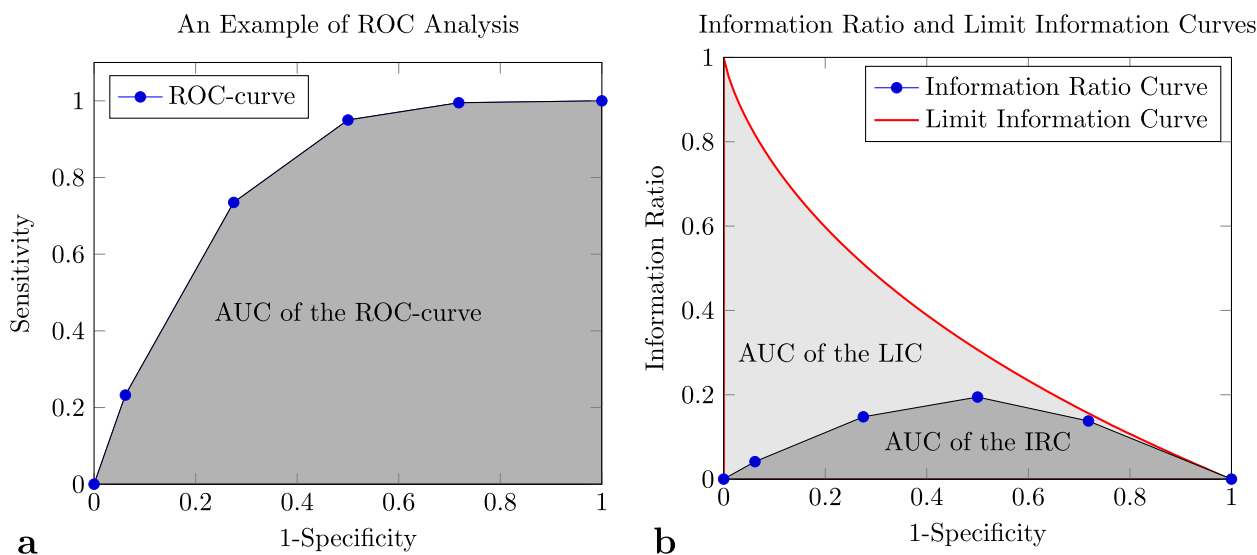


Fig. 2 Receiver operating characteristic (ROC) and global information ratio (GIR) analyses of data reported in Table 1. The blue points in the figures denotes the 6 possible cutoffs, i.e., all the category below the i th one, where $i \in [0, 5]$ are considered negative diagnoses. **a** The ROC analysis plots the cutoff points in the $(1 - SP) \times SE$ space, and it connects them by using the ROC curve. The area under this curve is a cutoff-independent measure of the effectiveness of the diagnostic approach: the higher the area, the better the approach. The ROC area under the curve (AUC) ranges in the $[0, 1]$ interval. The figure represents the ROC curve and its AUC as a black line and a dark gray region, respectively. In the depicted scenario, the AUC is about 0.793. **b** The GIR analysis of the same data depicts the cutoff points in the $(1 - SP) \times IR$ space. The point themselves are connected by the information ratio curve (IRC) which is represented as a black line. As the ROC AUC, the IRC AUC (the dark gray region in the figure) is a measure of the effectiveness of the diagnostic approach, but, since it is computed by using IR, it is prevalence-independent. Unfortunately, it does not range in the interval $[0, 1]$, and to normalize it, it must be divided by the IRC AUC of the best theoretical diagnostic approach, i.e., those whose sensitivity is always 1: the limit information curve (LIC). The LIC AUC (the light and dark gray regions in the figure) always equals $2 - \pi^2/6$ [4]; thus, the ratio between IRC and LIC AUCs, i.e., the global information ratio (GIR), equals IRC AUC divided by $2 - \pi^2/6$. In the scenario depicted by panel **b**, IRC AUC and GIR are about 0.116 and 0.326, respectively

always equals $2 - \pi^2/6 \approx 0.355$ [4], GIR can be computed as follows:

$$GIR = \frac{AUC_{IRC}}{2 - \pi^2/6}$$

Advantages of the IT-based model for diagnostic accuracy

IR and GIR are guaranteed to be global, normalized, and prevalence-independent measures for diagnostic accuracy. The prevalence-independency is a certain advantage with respect to the standard Cohen κ and ROC analysis. Another issue related to the ROC analysis is the arbitrary choice of the optimal cutoff, which, even if reasonable (as in the case of the Youden index [16]), is not supported by any motivation related to the aim of any diagnostic process, which is to extract as much information as possible from the test. The GIR analysis is the only one to offer such a criterion, based on the maximization of the information flow between the patient and the clinician, which is the scope of any good diagnostic system.

An information measure of inter-reader agreement

As already noticed above, the agreement channel $X \iff Y$ relates the outcomes of two diagnostic tests or readers. Hence, the mutual information over this channel, *i.e.*, $MI(X, Y)$, can be used as a correlation index between the outcomes of the two readers, *i.e.*, X and Y . Since $MI(X, Y) \leq \min\{H(X), H(Y)\}$ [3, 5] where $H(X)$ and $H(Y)$ are the entropy associated with the variables X and Y , respectively, we can normalize $MI(X, Y)$ in the interval $[0, 1]$ by dividing it by its maximum value, *i.e.*, $\min\{H(X), H(Y)\}$. This leads to the information agreement (IA):

$$IA(X, Y) = \frac{MI(X, Y)}{\min\{H(X), H(Y)\}}$$

whose value ranges in the interval $[0, 1]$. As pointed out in [5], IA provides an exact and coherent measure of the stochastic distance between P_{XY} and $P_X P_Y$, that is the joint distribution and the product of the marginals. One might argue that such a determination is measured at a less formally rigorous and precise extent when using Cohen κ [17]. However, IA can be thought of as a (normalized) measure of the degree of dependence between two readers when making a diagnosis. The greater the IA value, the higher the inter-reader agreement, *i.e.*, the inter-dependence of the readers. Of note, the measure can be used in both the dichotomous and multi-valued scale ratings [5], while it does not currently apply to continuous variables.

Advantages of the IT-based model for inter-reader agreement

IA overcomes some of the well-known Cohen κ flaws [18–20]. This is partially testified by the three artificial inter-reader agreements presented in Table 3.

Tables 3, 4, and 5 compare three pairs of readers on the bases of 20,000 diagnoses. Since the readers considered in Table 3 agree 73.4% of the time, while 99.4% of the diagnoses coincide in Table 4, the latter scenario seems to deserve an agreement value greater than that presented in Table 3. However, both the tables have the very same Cohen κ , *i.e.*, 0.5. IA better fits common sense in this case because Tables 3 and 4 IAs equal 0.311 and 0.638, respectively.

Tables 4 and 5 are almost identical: they differ for less than 1% of the diagnoses. Because of this similarity, one may expect that their agreement values are almost the same. Still, their Cohen κ is quite different: 0.5 for Table 4 and 0.24 for Table 5. On the contrary, their IAs are 0.638 and 0.580, respectively. Once more, IA offers a more convincing agreement measure with respect to Cohen κ .

Conclusions

IT is a rigorous mathematical tool widely used in electronic telecommunication systems. We presented some IT-inspired diagnostic accuracy and inter-reader agreement measures. While the underlying conceptual framework of these measures may be harder to be understood compared to currently used statistical indexes, it brings

Table 3 Agreement matrices leading to paradoxical values for Cohen κ

		Reader 1	
		Positive	Negative
Reader 2	Positive	7,210	5,200
	Negative	120	7,470
		<i>Cohen κ</i>	0.5
		<i>Information agreement</i>	0.311

Table 4 Agreement matrices leading to paradoxical values for Cohen κ

		Reader 1	
		Positive	Negative
Reader 2	Positive	19,818	116
	Negative	5	61
		<i>Cohen κ</i>	0.5
		<i>Information agreement</i>	0.638

Table 5 Agreement matrices leading to paradoxical values for Cohen κ

		Reader 1	
		Positive	Negative
Reader 2	Positive	19,818	156
	Negative	1	25
		<i>Cohen κ</i>	<i>0.240</i>
		<i>Information agreement</i>	<i>0.580</i>

Tables 3 and 4 have the same Cohen κ (*i.e.*, 0.5), but the number of accordant diagnoses in the former (*i.e.*, 73.4%) is significantly lower than those of the latter (*i.e.*, 99.4%). Information agreement properly measures this difference and assigns the values 0.311 and 0.638 to Tables 3 and 4, respectively. On the contrary, Tables 4 and 5 report quite similar scenarios, as they differ for less than 1% of the total diagnoses. However, the Cohen κ of the former (0.5) is much higher than that of the latter (0.240) whereas their information agreement (0.638 and 0.58, respectively) do not significantly deviate. In all these cases, the information agreement exhibits a behavior more consistent to the common sense of agreement than Cohen κ

some appealing advantages in the presentation and interpretation of radiological research, such as (i) providing summary measures of accuracy not depending on the prevalence of disease, and (ii) assessing diagnostic accuracy and inter-reader agreement without potential pitfalls of conventional analysis.

On this basis, we suggest that the above-presented information-based indexes of diagnostic accuracy could complement conventional ones by objectively showing how much information on the patient status is truly captured by a diagnostic tool, given a definite set of SE and SP. This knowledge could be useful to assess whether any refined diagnostic strategy using that tool or new ones truly translates into a gain in information on the disease or compare the amount of information provided by different tools. This is potentially relevant in some settings such as testing artificial intelligence-based tools, which are supposed to extract additional information from medical images. Additional information could be quantified more precisely compared to conventional qualitative images or when comparing various algorithms.

Concerning agreement analysis, the index we proposed could be used as an alternative to Cohen κ as it is not affected by the abovementioned biases related to unbalanced data distribution in source 2×2 tables.

However, some points should be faced before information measures can be fully appreciated and used. First of all, easy-to-use software tools for evaluating them are missing. We are currently working on an online platform designed to obtain the accuracy or agreement information measures by entering data directly from a database and hope this can be readily available. Second, reference values for IR, GIR, and IA are not yet defined, thus making it difficult to establish which values can be qualified as “high or low accuracy”

or rather can express different levels of agreement (*e.g.*, low, moderate, substantial, excellent), and in turn, limiting potential concrete applications. Furthermore, the rules for comparing different GIR values have not been established. Those limitations suggest that some additional mathematical work is needed to refine information indexes. Finally, a direct comparison of those measures with conventional ones in real study cohorts will be indispensable to understand how much informative and impacting the information indexes are compared to conventional ones, in the light of the potential advantages we described and make radiologists familiar with new indexes.

Abbreviations

AUC	Area under the curve
GIR	Global information ratio
IA	Information agreement
IR	Information ratio
IRC	Information ratio curve
IT	Information theory
LIC	Limit information curve
MI	Mutual information
PI-RADS	Prostate imaging reporting and data system
PREV	Prevalence of disease
ROC	Receiver operating characteristic
SE	Sensitivity
SP	Specificity

Acknowledgements

The authors thank anonymous reviewers for useful comments and suggestions.

Authors' contributions

Casagrande and Fabris: mathematical aspects. Girometti: radiological aspects. The authors read and approved the final manuscript.

Funding

This work has not received any funding.

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

R. Girometti is a member of the European Radiology Experimental Editorial Board. He has not taken part in the review or selection process of this article. The remaining authors declare that they have no competing interests.

Received: 27 September 2022 Accepted: 26 January 2023

Published online: 20 March 2023

References

1. Turkbey B, Rosenkrantz AB, Haider MA et al (2019) Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging

- reporting and data system version 2. *Eur Urol* 76:340–351. <https://doi.org/10.1016/j.eururo.2019.02.033>
2. Verdu S (1998) Fifty years of Shannon theory. *IEEE Trans Inf Theory* 44:2057–2078. <https://doi.org/10.1109/18.720531>
 3. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
 4. Girometti R, Fabris F (2015) Informational analysis: a Shannon theoretic approach to measure the performance of a diagnostic test. *Med Biol Eng Compu* 53:899–910. <https://doi.org/10.1007/s11517-015-1294-7>
 5. Casagrande A, Fabris F, Girometti R (2020) Beyond kappa: an informational index for diagnostic agreement in dichotomous and multivalued ordered-categorical ratings. *Med Biol Eng Compu* 58:3089–3099. <https://doi.org/10.1007/s11517-020-02261-2>
 6. Casagrande A, Fabris F, Girometti R (2020b) Extending information agreement by continuity. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp 1432–1439. <https://doi.org/10.1109/BIBM49941.2020.9313173>
 7. Casagrande A, Fabris F, Girometti R (2022) Fifty years of Shannon information theory in assessing the accuracy and agreement of diagnostic tests. *Med Biol Eng Comput* 60(4):941–955. <https://doi.org/10.1007/s11517-021-02494-9>
 8. Benish WA (2015) The channel capacity of a diagnostic test as a function of test sensitivity and test specificity. *Stat Methods Med Res* 24:1044–1052. <https://doi.org/10.1177/0962280212439742>
 9. Dorst FH, Osses D, Nieboer D et al (2020) Prostate magnetic resonance imaging, with or without magnetic resonance imaging-targeted biopsy, and systematic biopsy for detecting prostate cancer: a cochrane systematic review and meta-analysis. *Eur Urol* 77:78–94. <https://doi.org/10.1016/j.eururo.2019.06.023>
 10. Ransohoff DF, Feinstein AR (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 299:926–930. <https://doi.org/10.1056/NEJM197810262991705>
 11. Brenner H, Gefeller O (1997) Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. *Stat Med* 16:981–991. [https://doi.org/10.1002/\(SICI\)1097-0258\(19970515\)16:9%3C981::AID-SIM510%3E3.0.CO;2-N](https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9%3C981::AID-SIM510%3E3.0.CO;2-N)
 12. Mulherin SA, Miller WC (2002) Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med* 137:598–602. <https://doi.org/10.7326/0003-4819-137-7-200210010-00011>
 13. Sardanelli F, Di Leo G (2020) Assessing the value of diagnostic tests in the coronavirus disease 2019 pandemic. *Radiology* 296(3):E193. <https://doi.org/10.1148/radiol.2020201845>
 14. Penn A, Medved M, Abe H et al (2022) Safely reducing unnecessary benign breast biopsies by applying non-mass and DWI directional variance filters to ADC thresholding. *BMC Med Imaging* 22:171. <https://doi.org/10.1186/s12880-022-00897-0>
 15. Flach PA (2016) ROC Analysis. In: Sammut, C., Webb, G. (eds) *Encyclopedia of machine learning and data mining*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7502-7_739-1
 16. Youden WJ (1950) Index for rating diagnostic tests. *Cancer* 3:32–35. [https://doi.org/10.1002/1097-0142\(1950\)3:1%3C32::aid-cnrc2820030106%3e3.0.co;2-3](https://doi.org/10.1002/1097-0142(1950)3:1%3C32::aid-cnrc2820030106%3e3.0.co;2-3)
 17. Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Measur* 20:37–46. <https://doi.org/10.1177/001316446002000104>
 18. Thompson WD, Walter SD (1988) A reappraisal of the kappa coefficient. *J Clin Epidemiol* 41:949–958. [https://doi.org/10.1016/0895-4356\(88\)90031-5](https://doi.org/10.1016/0895-4356(88)90031-5)
 19. Vach W (2005) The dependence of Cohen's kappa on the prevalence does not matter. *J Clin Epidemiol* 58:655–661. <https://doi.org/10.1016/j.jclinepi.2004.02.021>
 20. de Vet HCW, Mokkink LB, Terwee CB et al (2013) Clinicians are right not to like Cohen's Kappa. *BMJ* 346. <https://doi.org/10.1136/bmj.f2125>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen® journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)