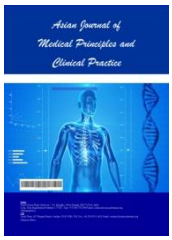


Asian Journal of Medical Principles and Clinical Practice

4(3): 51-65, 2021; Article no.AJMPCP.68338

Measures, Metrics and Indicators Derived from the Ubiquitous Two-by-two Contingency Table, Part I: Background

Muzainah Ali Rushdi[1] and Ali Muhammad Rushdi[2*]

[1]Kasr Al-Ainy Faculty of Medicine, Cairo University, Cairo, 11562, Arab Republic of Egypt.
[2]Department of Electrical and Computer Engineering, King Abdulaziz University, P.O.Box 80200, Jeddah 21589, Saudi Arabia.

Authors' contributions

This work was carried out in collaboration between both authors. Author MAR envisioned and designed the study, performed the analyses, stressed the medical context for the entities considered, and contributed to the literature search. Author AMR managed the literature search and wrote the preliminary manuscript. Both authors read and approved the final manuscript.

Article Information

Editor(s):
(1) Prof. Suprakash Chaudhury, Dr D. Y. Patil Medical College, Hospital and Research Center, India.
Reviewers:
(1) Elias Dritsas, University of Patras, Greece.
(2) Mudide Ramprasad, Jawaharlal Nehru Technological University Hyderabad (JNTUH), India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/68338

Original Research Article

Received 20 March 2021
Accepted 29 May 2021
Published 04 June 20211

ABSTRACT

This paper (the first part of two sibling parts) provides a tutorial exposition of indicators derived of the ubiquitous two-by-two contingency table (confusion matrix) that has widespread applications in many fields, including, in particular, the fields of binary classification and clinical or epidemiological testing. These indicators include the eight most prominent indicators used in diagnostic testing, namely the Sensitivity or True Positive Rate (TPR), the Specificity or True Negative Rate (TNR), the Positive and Negative Predictive Values (PPV and NPV), together with their respective complements, namely the False Negative Rate (FNR), False Positive Rate (FPR), False Discovery rate (FDR) and False Omission Rate (FOR). We consider also some other indicators, such as the total error and accuracy, pre-test prevalence, the diagnostic odds ratio (DOR), the inverse DOR, the F-scores, Youden's Index (Informedness), Markedness and the Index of Association (Matthews Correlation Coefficient (MCC)). We review recent studies asserting that the MCC is the most reliable single metric derivable from the contingency matrix. We suggest that any mean (signed

**Corresponding author: E-mail: arushdi@kau.edu.sa, arushdi@yahoo.com;*

geometric mean, arithmetic mean, or harmonic mean) of Informedness and Markedness might be as effective as the MCC in summarizing the contingency matrix into a single value. We set criteria in terms of basic and composite indicators for identifying the quality of binary classification, going down from the perfect type to the completely-contradictory type, where random-guessing-like classification marks the middle point of transition between good and bad classification. In a sequel paper, we present a potpourri of example or test cases to reveal and unravel many of the properties and inter-relationships among binary and composite indicators.

Keywords: Diagnostic testing; binary classification; sensitivity; specificity; predictive values; F scores; Matthews correlation coefficient; means of Informedness and Markedness.

1. INTRODUCTION

A contingency table is a powerful tool in data analysis employing matrix format for comparing two categorical variables [1-9]. This table (known also by a variety of other names such as the confusion table, the frequency matrix or the agreement table) is a ubiquitous tool of scientific analysis and research. It originates in a variety of applications such as clinical testing, criminal investigations, judicial trials, lie detection, null-hypothesis acceptance/rejection, quality control, industrial management, satellite mapping, text classification, communications, DNA identification, forensic reasoning and machine classification. Our current treatment will be mainly in the context of clinical testing or binary classification. Generally, the table is a means of classification of data describing a discrete sample of an arbitrary kind in which each individual case of the sample either possesses or does not possess a certain attribute, trait, or condition to be detected. This attribute is possibly a categorical binary variable (sick/healthy, guilty/innocent, false/true, black/white, …, *etc*.) or a continuous variable to be dichotomized using a specific threshold. A certain test, operator or metric j (called a reference or a standard) partitions the sample into two groups, one with the attribute and another without it. A second test, operator or metric i (called the assessed metric or predictor) introduces its own partitioning of the sample, again into two groups. Therefore, each individual case among the population sample must fall into one of four categories. The total numbers of cases within these categories are entered into the four cells of the contingency table or matrix (See Fig. 1).

This paper is a tutorial exposition of the most important indicators used in diagnostic testing. The most prominent among these are the Sensitivity or True Positive Rate (TPR), the Specificity or True Negative Rate (TNR), the Positive and Negative Predictive Values (PPV and NPV), together with their respective complements (to 1.0), namely the False Negative Rate (FNR), False Positive Rate (FPR), False Discovery rate (FDR) and False Omission Rate (FOR) [10-20]. Other important indicators included herein are the total error and accuracy, the pre-test prevalence, the diagnostic odds ratio (DOR), the inverse DOR, the F-scores, Youden's Index (Informedness), Markedness and the Index of Association (Matthews Correlation Coefficient) [13]. Other important metrics not discussed herein include the Brier score [21], Cohen's Kappa [22], the K measure [23], the Fowlkes-Mallows index [24] and the H-index [25].

The organization of the rest of this paper is as follows. Section 2 is a brief primer about diagnostic testing and its basic measures. Section 3 introduces the eight most prominent basic indicators of diagnostic testing. It reports and points out the existence of formulas of inter-dependence among the four direct measures among them (the two predictive values, sensitivity, and specificity). These formulas express any one of these four indicators in terms of the other three, under the assumption that each of the four exists, and no division by zero is encountered [16-19]. Section 3 also reports formulas of inter-dependence among the four complementary measures. Section 4 is a brief presentation of F-scores, while Section 5 introduces and comments on Matthews Correlation Coefficient. Moreover, Section 5 suggests that any mean (signed geometric mean, arithmetic mean, or harmonic mean) of Informedness and Markedness might be as effective as the MCC in summarizing the contingency matrix into a single value. Section 5 also sets criteria (in terms of basic and composite indicators) for identifying the quality of binary classification, going down from the perfect type to the completely-contradictory type, where random-guessing-like classification marks the middle point of transition between good and bad

classification. Section 6 presents a brief outline of a novel technique for solving ternary problems of conditional probability, which is instrumental for full characterization of the contingency table. Section 7 concludes the paper.

2. ON DIAGNOSTIC TESTING AND ITS BASIC INDICATORS

This section is intended for a brief primer about diagnostic testing and its most basic indicators [10-20]. Fig. 1 demonstrates a two-by-two contingency matrix for test or classification i with respect to test or classification j. Each of the two variables i and j is a dichotomous variable that belongs to the set $\{+1, -1\}$ of indices. The test i reports 'positive' cases (arbitrarily assigned the value $+1$), in which a certain disease, attribute, trait, or condition is present, or reports 'negative' cases (arbitrarily assigned the value -1), in which this condition is absent. This test is assessed or evaluated by a reference or gold standard test j, which has its own labeling of cases, again as positive or negative. The reference test j designates various cases of the assessed test i as "true" or "false," depending on whether it agrees or disagrees with test i, respectively. As a result, the matrix four entries are called True Positives, False Positives, False Negatives, and True Negatives. These entries are usually assigned the standard abbreviations TP, FP, FN, and TN. In the sequel, we will use the subscripted abbreviations TP_{ij}, FP_{ij}, FN_{ij}, and TN_{ij}, where we use the subscripts ij for all measures (and later for indicators derived from them) to assert the notion that i is assessed, judged or measured relative to j. The sum of these four entries is the size of the reported population or the total number of reported cases N. If the tests i and j interchange their roles (so that test j is now assessed relative to test i) then the four measures are relabeled as $TP_{ji}, FP_{ji}, FN_{ji}$, and TN_{ji} such that $TP_{ji} = TP_{ij}$, and $TN_{ji} = TN_{ij}$ but with $FP_{ji} = FN_{ij}$, and $FN_{ji} = FP_{ij}$. This is the reason why omission of the subscripts is not desirable, as it leads to an inadvertent ambiguity as to which assesses which. We will see later that a few indicators (that we call self-transposed ones) might dispense with these subscripts due to their inherent symmetry. These include the F_1 score, the MCC, and the product (or the geometric, arithmetic, or harmonic mean) of the Informedness and Markedness indicators.

We use the symbols $A = \{j = +1\}$ and $B = \{i = +1\}$ to denote the events of positive cases (presence of the considered condition) according to the tests j and i, respectively. Hence, the complementary events $\overline{A} = \{j = -1\}$ and $\overline{B} = \{i = -1\}$ denote the events of negative cases (absence of the considered condition) according to the tests j and i, respectively. There are eight conditional probabilities concerning these two events and their complements, as shown in Fig. 2. These can be identified as the eight most prominent indicators used in diagnostic testing. These are the Sensitivity ($Sens_{ij}$) or True Positive Rate (TPR_{ij}), the Specificity ($Spec_{ij}$) or True Negative Rate (TNR_{ij}), and the Positive and Negative Predictive Values (PPV_{ij} and NPV_{ij}), together with their respective complements (to 1.0), namely the False Negative Rate (FNR_{ij}), False Positive Rate (FPR_{ij}), False Discovery rate (FDR_{ij}) and False Omission Rate (FOR_{ij}) [10-20]. The former four indicators are considered more popular, more basic or more prominent, and they act as direct or agreement measures while the latter four serve as discrepancy or disagreement measures between the two tests i and j. Due to the four complementation relations within pairs of these eight measures, the number of independent quantities among them is at most four. It seems that there is a widespread (and at least implicit) belief that this number is exactly four (usually obtained by counting the four direct indicators $Sens_{ij}, Spec_{ij}, PPV_{ij}$ and NPV_{ij}). We show in Section 3 that this number is, in fact, three, by simply being able to express any of the four direct indicators in terms of the other three [13,16-19].

Table 1 (adapted from [13]) lists some of the measures or indicators commonly used in diagnostic testing or binary classification, including the afore-mentioned eight indicators. The table expresses each of these quantities in terms of the elements of the contingency matrix, states its range of values, and identifies its value for perfect testing or classification. Many quantities have ranges $[0.0, 1.0]$ (and hence might be viewed as probabilities or have probability interpretations), but a few belong to $[0.0, \infty)$ or to $[-1.0, +1.0]$. Direct measures and indicators are highlighted in a greenish color, while inverse or opposite ones are shown with a reddish color. Pre-test quantities are designated neither way since they are test-independent.

i \ j	$+1$	-1
$+1$	TP_{ij} **(True Positives)**	FP_{ij} **(False Positives) (When normalized: Type I Error)**
-1	FN_{ij} **(False Negatives) (When normalized: Type II Error)**	TN_{ij} **(True Negatives)**

Fig. 1. The two-by-two contingency matrix of test or classification i with respect to test or classification j. This matrix has integer entries that add to the total number of cases N. The symbols $A = \{j = +1\}$ and $B = \{i = +1\}$ denote the events of positive cases according to tests j and i, respectively

			B conditioned	
	$P(\bar{A}\|\bar{B}) = P(j = -1\|i = -1) = NPV_{ij}$	$P(A\|\bar{B}) = P(j = +1\|i = -1) = FOR_{ij}$	$P(B\|\bar{A}) = P(i = +1\|j = -1) = FPR_{ij}$	$P(\bar{B}\|\bar{A}) = P(i = -1\|j = -1) = Spec_{ij} = TNR_{ij}$
Conditioning uncomplemented	$P(\bar{A}\|B) = P(j = -1\|i = +1) = FDR_{ij}$	$P(A\|B) = P(j = +1\|i = +1) = PPV_{ij}$	$P(B\|A) = P(i = +1\|j = +1) = Sens_{ij} = TPR_{ij}$	$P(\bar{B}\|A) = P(i = -1\|j = +1) = FNR_{ij}$
		Conditioned uncomplemented		

Fig. 2. Definition of the eight conditional probabilities concerning events $A = \{j = +1\}$ and $B = \{i = +1\}$, which constitute the eight most prominent indicators of diagnostic testing. The four shaded entries are direct indicators, usually taken for the most basic ones. The four unshaded entries are complementary to the horizontally-adjacent shaded ones. Each conditional probability has a 'dual' one obtained by complementing both the conditioned and conditioning events, and also has an inverse or transposed one, obtained by swapping the conditioned and conditioning events. Relations among the ordered set

$$\{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}\} = \{P(B|A), P(\bar{B}|\bar{A}), P(A|B), P(\bar{A}|\bar{B})\}$$

might be replaced by relations for the ordered set

$$\{FNR_{ij}, FPR_{ij}, FOR_{ij}, FDR_{ij}\} = \{P(\bar{B}|A), P(B|\bar{A}), P(A|\bar{B}), P(\bar{A}|B)\}$$

Table 1. Commonly used quantities pertaining to diagnostic testing (adapted from [13]). Direct measures and indicators are highlighted in a greenish color, while inverse ones are shown with a reddish color. Pre-test quantities are designated neither way [13]

Measure or indicator	Formula in terms of entries of the contingency matrix	Range	Perfect value
Sensitivity (True Positive Rate (TPR), Recall, Probability of Detection)	$Sens_{ij} = TP_{ij}/(TP_{ij} + FN_{ij})$	**[0.0, 1.0]**	**1.0**
Specificity, Inverse recall (True Negative Rate (TNR))	$Spec_{ij} = TN_{ij}/(TN_{ij} + FP_{ij})$	**[0.0, 1.0]**	**1.0**
Precision (Positive Predictive Value (PPV))	$PPV_{ij} = TP_{ij}/(TP_{ij} + FP_{ij})$	**[0.0, 1.0]**	**1.0**
Inverse precision (Negative Predictive Value (NPV))	$NPV_{ij} = TN_{ij}/(TN_{ij} + FN_{ij})$	**[0.0, 1.0]**	**1.0**
False Negative Rate (FNR)	$FNR_{ij} = 1 - Sens_{ij} = FN_{ij}/(TP_{ij} + FN_{ij})$	**[0.0, 1.0]**	**0.0**
False Positive Rate (FPR) (Fall-Out, False Alarm)	$FPR_{ij} = 1 - Spec_{ij} = FP_{ij}/(TN_{ij} + FP_{ij})$	**[0.0, 1.0]**	**0.0**
False Discovery Rate (FDR)	$FDR_{ij} = 1 - PPV_{ij} = FP_{ij}/(TP_{ij} + FP_{ij})$	**[0.0, 1.0]**	**0.0**
False Omission Rate (FOR)	$FOR_{ij} = 1 - NPV_{ij} = FN_{ij}/(TN_{ij} + FN_{ij})$	**[0.0, 1.0]**	**0.0**
Likelihood Ratio for Positive Test	$(LR+)_{ij} = Sens_{ij}/(1 - Spec_{ij})$	**[0.0, ∞)**	**∞**
Likelihood Ratio for Negative Test	$(LR-)_{ij} = (1 - Sens_{ij})/Spec_{ij}$	**[0.0, ∞)**	**0.0**
Diagnostic Odds Ratio	$DOR_{ij} = (TP_{ij} * TN_{ij})/(FP_{ij} * FN_{ij})$	**[0.0, ∞)**	**∞**
Inverse of the DOR	$DOR_{ij}^{-1} = (FP_{ij} * FN_{ij})/(TP_{ij} * TN_{ij})$	**[0.0, ∞)**	**0.0**
Youden's Index (Informedness)	$YI_{ij} = Sens_{ij} + Spec_{ij} - 1$	$[-\mathbf{1.0}, \mathbf{1.0}]$	**1.0**
Markedness	$M_{ij} = PPV_{ij} + NPV_{ij} - 1$	$[-\mathbf{1.0}, \mathbf{1.0}]$	**1.0**
Error of the First Kind	$E1_{ij} = FP_{ij}/N$	**[0.0, 1.0]**	**0.0**
Error of the Second Kind	$E2_{ij} = FN_{ij}/N$	**[0.0, 1.0]**	**0.0**
Total Diagnostic Error	$E_{ij} = (FP_{ij} + FN_{ij})/N$	**[0.0, 1.0]**	**0.0**
Diagnostic Accuracy	$A_{ij} = (TP_{ij} + TN_{ij})/N$	**[0.0, 1.0]**	**1.0**
Pre-Test Prevalence	$PTP_{ij} = (TP_{ij} + FN_{ij})/N$	**[0.0, 1.0]**	–
Pre-Test Odds	$PTO_{ij} = (TP_{ij} + FN_{ij})/(FP_{ij} + TN_{ij})$	**[0.0, ∞)**	–
Post-Positive-Test Odds	$PPTO_{ij} = PTO_{ij}(LR+)_{ij} = TP_{ij}/FP_{ij}$	**[0.0, ∞)**	**∞**
Post-Negative-Test Odds	$PNTO_{ij} = PTO_{ij}(LR-)_{ij} = FN_{ij}/TN_{ij}$	**[0.0, ∞)**	**0.0**
F_1 score	$F_1 = 2\,TP_{ij} / \left(2\,TP_{ij} + FP_{ij} + FN_{ij}\right).$	**[0.0, 1.0]**	**1.0**
Index of Association or Matthews Correlation Coefficient (MCC) $\emptyset_{ij} = \emptyset_{ji}$	$\emptyset_{ij} = \emptyset_{ji} = (TP_{ij} * TN_{ij} - FP_{ij} * FN_{ij})/\,SQRT((TP_{ij} + FN_{ij})(TP_{ij} + FP_{ij})(TN_{ij} + FP_{ij})(TN_{ij} + FN_{ij}))$	$[-\mathbf{1.0}, \mathbf{1.0}]$	**1.0**

3. FORMULAS RELATING THE MOST PROMINENT BASIC INDICATORS

We now express each of the four most prominent basic indicators of diagnostic testing (Specificity, Negative Predictive Value, Sensitivity, and Positive Predictive Value) solely in terms of the other three (provided each of the four indicators exists, and no division by zero is encountered), namely [16-19].

$$Sens_{ij} = \frac{PPV_{ij} * \ NPV_{ij}\,[1 - Spec_{ij}]}{PPV_{ij}\ NPV_{ij} + Spec_{ij}\left[1 - PPV_{ij} - NPV_{ij}\right]}, \tag{1}$$

$$Spec_{ij} = \frac{PPV_{ij} * NPV_{ij}\left[1 - \ Sens_{ij}\right]}{PPV_{ij} * NPV_{ij} + Sens_{ij}\left[1 - \ PPV_{ij} - NPV_{ij}\right]}, \tag{2}$$

$$PPV_{ij} = \frac{Sens_{ij} * Spec_{ij}\,[1 - NPV_{ij}]}{Sens_{ij} * Spec_{ij} + NPV_{ij}\,[1 - \ Sens_{ij} - Spec_{ij}]}, \tag{3}$$

$$NPV_{ij} = \frac{Sens_{ij} * Spec_{ij}\,[1 - PPV_{ij}]}{Sens_{ij} * Spec_{ij} + PPV_{ij}[1 - \ Sens_{ij} - Spec_{ij}]}. \tag{4}$$

We note that relations (1-4) among the ordered set

$$\{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}\} = \{P(B|A), P(\bar{B}|\bar{A}), P(A|B), P(\bar{A}|\bar{B})\}, \tag{5}$$

might be replaced by relations for the following ordered set (arbitrarily obtained by complementing the event B)

$$\{FNR_{ij}, FPR_{ij}, FOR_{ij}, FDR_{ij}\} = \{P(\bar{B}|A), P(B|\bar{A}), P(A|\bar{B}), P(\bar{A}|B)\}. \tag{6}$$

Equations (1-4) might be rewritten as relations among the four complementary indicators (namely the False Negative Rate ($FNR_{ij} = 1 - Sens_{ij}$), False Positive Rate ($FPR_{ij} = 1 - \ Spec_{ij}$), False Discovery rate ($FDR_{ij} = 1 - PPV_{ij}$) and False Omission Rate ($FOR_{ij} = 1 - NPV_{ij}$). The resulting relations are:

$$FNR_{ij} = \frac{FOR_{ij} * \ FDR_{ij}\,[1 - FPR_{ij}]}{FOR_{ij}\ FDR_{ij} + FPR_{ij}\left[1 - FOR_{ij} - FDR_{ij}\right]}, \tag{7}$$

$$FPR_{ij} = \frac{FOR_{ij} * FDR_{ij}\left[1 - \ FNR_{ij}\right]}{FOR_{ij} * FDR_{ij} + FNR_{ij}\left[1 - \ FOR_{ij} - FDR_{ij}\right]}, \tag{8}$$

$$FDR_{ij} = \frac{FNR_{ij} * FPR_{ij}\,[1 - FOR_{ij}]}{FNR_{ij} * FPR_{ij} + FOR_{ij}[1 - \ FNR_{ij} - FPR_{ij}]}. \tag{9}$$

$$FOR_{ij} = \frac{FNR_{ij} * FPR_{ij}\,[1 - FDR_{ij}]}{FNR_{ij} * FPR_{ij} + FDR_{ij}\,[1 - \ FNR_{ij} - FPR_{ij}]}, \tag{10}$$

Note that each conditional probability in Fig. 2 has a 'dual' one obtained by complementing both the conditioned and conditioning events [7,18], and also has an inverse or transposed one, obtained by swapping or interchanging the conditioned and conditioning events [12,15]. Our definitions of duality and transposition mean that each conditional probability P has a dual P^d, a transpose or inverse T, and a dual of its transpose or inverse (a transpose of its dual) T^d. Note that both the duality and transposition operators are involutary or self-inverse operators, i.e., each of them satisfies 'the law of involution' (applying any of them twice to a specific conditional probability leaves it intact). Table 2 defines the four possible sets $\{P, P^d, T, T^d\}$ pertaining to the set of four direct indicators of diagnostic testing [19], and also the four possible sets pertaining to the set of four complementary indicators of diagnostic testing. Equations (1-4, 7-10) might be rewritten in a unified generic form (See Table 2) as [19]

$$P = \frac{T * T^d\,[1 - P^d]}{T * T^d + P^d[1 - T - T^d]} \tag{11}$$

Equation (11) suggests the existence of a universal diagnostic identity

$$P * P^d\,[T + T^d - 1] - \ T * T^d\,[P + P^d - 1] = 0. \tag{12}$$

Rushdi and Serag [19] noted that the quantity $f(P) = (P * P^d\,[T + T^d - 1])$, which is naturally invariant to the replacement of every term by its dual, is also (thanks to (12)) invariant to the replacement of every term by its transpose. With a little abuse of notation, we wrote this function as $f(P)$ rather than $f(P, P^d, T, T^d)$, since P^d, T, and T^d are uniquely determined by P. We suggest that it might serve as a composite diagnostic indicator, which satisfies

$$f(P) = f(P^d) = f(T) = f(T^d) = P * P^d\,[T + T^d - 1] = T * T^d\,[P + P^d - 1]. \quad (13)$$

Table 2 asserts that there are four pairs of dual conditional probabilities $\{Sens_{ij}, Spec_{ij}\}$, $\{PPV_{ij}, NPV_{ij}\}$, $\{FNR_{ij}, FPR_{ij}\}$, and $\{FDR_{ij}, FOR_{ij}\}$. Therefore, Eq. (13) might be rewritten as

$$f(Sens_{ij}) = f(Spec_{ij}) = f(PPV_{ij}) = f(NPV_{ij}) = Sens_{ij} * Spec_{ij}\,[PPV_{ij} + NPV_{ij} - 1] = PPV_{ij} * NPV_{ij}\,[Sens_{ij} + Spec_{ij} - 1], \quad (14)$$

or it might be rewritten as

$$f(FNR_{ij}) = f(FPR_{ij}) = f(FDR_{ij}) = f(FOR_{ij}) = FNR_{ij} * FPR_{ij}\,[FDR_{ij} + FOR_{ij} - 1] = FDR_{ij} * FOR_{ij}\,[FNR_{ij} + FPR_{ij} - 1]. \quad (15)$$

Equations (14) and (15) involve the pair of unbiased indicators called Youden's Index (YI_{ij}) (Informedness (I_{ij})), and Markedness (M_{ij}) defined by [13,19]

$$Informedness_{ij} = YI_{ij} = I_{ij} = Sens_{ij} + Spec_{ij} - 1 = -[FNR_{ij} + FPR_{ij} - 1] = Markedness_{ji}, \quad (16)$$

$$Markedness_{ij} = M_{ij} = PPV_{ij} + NPV_{ij} - 1 = -[FDR_{ij} + FOR_{ij} - 1] = Informedness_{ji}. \quad (17)$$

Each of these two quantities belongs to $[-1.0, +1.0]$. We note that the $Informedness_{ij}$ and $Markedness_{ij}$ indicators is the transpose of each other. They share the same sign, a result that is obvious from the universal identity (12). We also suggest that the sum of the two composite indicators in (14) and (15) is yet a stronger non-negative composite indicator, of values in $[0.0, 1.0]$, given by

$$f(Sens_{ij}) + f(FNR_{ij}) = Sens_{ij} * Spec_{ij} * Markedness_{ij} + (1 - Sens_{ij}) * (1 - Spec_{ij}) * (-Markedness_{ij}) = Informedness_{ij} * Markedness_{ij} = Informedness_{ij} * Informedness_{ji} = Markedness_{ji} * Markedness_{ij}. \quad (18)$$

We now add formulas that include also one or two marginal probabilities representing the pre-test prevalence or true prevalence ($Prev = P(A)$) and the apparent prevalence ($Prev' = P(B)$). We have the following formulas [13,16], in each of which a certain probability is expressed in terms of three others.

$$P(A) = Prev = PPV_{ij}\,Prev' + (1 - NPV_{ij})(1 - Prev'), \quad (19)$$

$$P(B) = Prev' = Sens_{ij}\,Prev + (1 - Spec_{ij})(1 - Prev), \quad (20)$$

$$PPV_{ij} = P(A|B) = \frac{Sens_{ij}\,Prev}{Sens_{ij}\,Prev + (1 - Spec_{ij})(1 - Prev)}. \quad (21)$$

$$NPV_{ij} = P(\bar{A}|\bar{B}) = \frac{Spec_{ij}\,(1 - Prev)}{(1 - Sens_{ij})\,Prev + Spec_{ij}(1 - Prev)}, \quad (22)$$

$$Sens_{ij} = P(B|A) = \frac{PPV_{ij}\,(NPV_{ij} + Prev - 1)}{(PPV_{ij} + NPV_{ij} - 1)Prev}. \quad (23)$$

$$Spec_{ij} = P(\bar{B}|\bar{A}) = \frac{NPV_{ij}\,(PPV_{ij} - Prev)}{(PPV_{ij} + NPV_{ij} - 1)(1 - Prev)}, \quad (24)$$

$$Prev = \frac{(Spec_{ij} + Prev' - 1)}{(Spec_{ij} + Sens_{ij} - 1)}. \quad (25)$$

Table 2. Possible definitions of a probability P, its dual P^d, its transpose or inverse T, and the dual of its transpose or inverse (transpose of its dual) T^d. These definitions pertain to the set of four direct indicators of diagnostic testing $\{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}\}$, and also to the set of four complementary indicators of diagnostic testing $\{FNR_{ij}, FPR_{ij}, FOR_{ij}, FDR_{ij}\}$. Note that there are four pairs of dual conditional probabilities $\{Sens_{ij}, Spec_{ij}\}$, $\{PPV_{ij}, NPV_{ij}\}$, $\{FNR_{ij}, FPR_{ij}\}$, and $\{FDR_{ij}, FOR_{ij}\}$

P	P^d	T	T^d
$Sens_{ij}$	$Spec_{ij}$	PPV_{ij}	NPV_{ij}
$Spec_{ij}$	$Sens_{ij}$	NPV_{ij}	PPV_{ij}
PPV_{ij}	NPV_{ij}	$Sens_{ij}$	$Spec_{ij}$
NPV_{ij}	PPV_{ij}	$Spec_{ij}$	$Sens_{ij}$
FNR_{ij}	FPR_{ij}	FOR_{ij}	FDR_{ij}
FPR_{ij}	FNR_{ij}	FDR_{ij}	FOR_{ij}
FOR_{ij}	FDR_{ij}	FNR_{ij}	FPR_{ij}
FDR_{ij}	FOR_{ij}	FPR_{ij}	FNR_{ij}

4. THE F SCORES

A plausible measure of a test's accuracy would be some mean (arithmetic, geometric or harmonic) of two dual or transposed probabilities, such as the transposed quantities of (a) precision (positive predictive value)

$$PPV_{ij} = TP_{ij}/(TP_{ij} + FP_{ij}) = Sens_{ji}, \tag{26}$$

and (b) recall (sensitivity)

$$Sens_{ij} = TP_{ij}/(TP_{ij} + FN_{ij}) = PPV_{ji} \tag{27}$$

where we make use of the fact that $\{TP_{ji} = TP_{ij}, TN_{ji} = TN_{ij}\ .\ FP_{ji} = FN_{ij}\ ,\ FN_{ji} = FP_{ij}\}$ to assert that precision and recall are swapped if the roles of reference and assessed tests are interchanged. The geometric mean is already in use under the name of the Fowlkes–Mallows index [24]. Of a much wider use is the traditional F-measure or balanced F-score (F_1 score), which is the harmonic mean (the reciprocal of the arithmetic mean of the reciprocals). This score derives its mathematical simplicity from the fact that precision and recall have a common numerator (as can be seen from (26) and (27)), and hence their reciprocals have a common denominator, namely.

$$\begin{aligned} 1/F_1 &= ((1/PPV_{ij}) + (1/Sens_{ij}))/2 \\ 2/F_1 &= (TP_{ij} + FP_{ij})/TP_{ij} + (TP_{ij} + FN_{ij})/TP_{ij}. \\ F_1 &= 2\,TP_{ij}\ /(2\,TP_{ij} + FP_{ij} + FN_{ij}). \end{aligned} \tag{28}$$

The value of F_1 varies from 0.0 ($TP_{ij} = 0$) to 1.0 ($FP_{ij} = FN_{ij} = 0$, i.e., no error). Compared to the arithmetic mean, the harmonic mean punishes a low value of precision or recall more. In other words, for the F_1 score to be high, *both* precision and recall should be high. The F_1 score assigns an equal weight to precision and recall. It might be generalized to a more generic F score, which assigns different weights to them, thereby valuing one of them more than the other.

Since the F score does not take true negatives into account, it is deemed less informative than measures such as the Matthews correlation coefficient (MCC). In Section 5, we will see that the F_1 score can be misleading, since it does not fully consider the size of the four classes of the confusion matrix (contingency table) in the final score computation.

5. THE MATTHEWS CORRELATION COEFFICIENT (MCC)

The Matthews correlation coefficient (MCC) [26, 27] (known also as the Index of Association [13], Yule phi coefficient [28] or Pearson phi coefficient [29,30]) is commonly used in machine learning as a measure of the quality of binary (two-class) classifications. The MCC enjoys a striking balance in its dependence on the four entries of the contingency table (TP_{ij}, FP_{ij}, FN_{ij}, and TN_{ij}), a balance that renders it one of the most beautiful or elegant mathematical formulas [13,31].

$$\emptyset_{ij} = \emptyset_{ji} = (TP_{ij} * TN_{ij} - FP_{ij} * FN_{ij})/ SQRT((TP_{ij} + FN_{ij})(TP_{ij} + FP_{ij})(TN_{ij} + FP_{ij})(TN_{ij} + FN_{ij})) \quad (29)$$

The MCC is a partially-symmetric function in TP_{ij} and TN_{ij}, as well as in FP_{ij} and FN_{ij}. Unlike other metrics of diagnostic testing in Table 1, the MCC does not depend on which metric or test is assessed relative to which ($\emptyset_{ij} = \emptyset_{ji}$). The MCC value belongs to $[-1.0, +1.0]$, with a value of $+1.0$ representing a total agreement or perfect prediction, a value of 0 being no better than random prediction and a value of -1.0 indicating total disagreement or completely contradictory prediction (See Table 3). The MCC can also be calculated by a formula comprising the eight most prominent indicators of diagnostic testing depicted in Fig. 2, namely

$$\emptyset_{ij} = \emptyset_{ji} = SQRT(Sens_{ij} * Spec_{ij} * PPV_{ij} * NPV_{ij}) - SQRT(FDR_{ij} * FNR_{ij} * FPR_{ij} * FOR_{ij}) \quad (30)$$

Equation (30) is also strikingly beautiful. By contrast, the F_1 score utilizes only two of these eight metrics in (28). Chicco and Jurman [32] assert that the Matthews correlation coefficient (MCC), produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives), proportionally both to the size of positive elements and the size of negative elements in the dataset. If we assume P is one of the four direct basic indicators, then (30) might be rewritten in the more generic form

$$\emptyset_{ij} = \emptyset_{ji} = SQRT(P * P^d * T * T^d) - SQRT((1 - P) * (1 - P^d) * (1 - T) * (1 - T^d)). \quad (30a)$$

We can use the universal diagnostic identity (12) to prove that the condition $\{P + P^d = 1.0\}$ is equivalent to the condition $\{T + T^d = 1.0\}$, and then use (30a) to show that any of these two conditions is equivalent to the condition that the prediction cannot be distinguished from random guessing ($\emptyset_{ij} = \emptyset_{ji} = 0$).

Chicco [31] considers a faulty algorithm which always predicts positive, and by applying this only-positive predictor to an imbalanced validation set (of 95 positives and 5 negatives), the four contingency matrix categories obtained are $TP_{ij} = 95, TN_{ij} = 0, FP_{ij} = 5$, and $FN_{ij} = 0$. These values translate into assuring performance values of accuracy $= 0.95$ and F_1 score $= 180/185 = 0.9744$. However, MCC yields an undefined value of $0/0$, thereby alerting the user that some problem exists. Alternatively, a set of matrix entries, $TP_{ij} = 90, TN_{ij} = 1, FP_{ij} = 5$, and $FN_{ij} = 4$ (obtained by another algorithm on the same validation set) yields accuracy $= 0.91$, F_1 score $= 180/189 = 0.95$, and MCC $= (90 - 20)/SQRT(94 * 95 * 5 * 6) = 0.1352$. Here, both the accuracy and F_1 score are too high for a classifier unable to recognize negative data elements, albeit doing well with positive ones. Chicco [31] also notes that the F_1 score behaviour depends on which class is defined as the positive class. If one applies an only-negative predictor to the earlier imbalanced validation set (of 95 positives and 5 negatives), the four contingency matrix categories obtained are $TP_{ij} = 0, TN_{ij} = 5, FP_{ij} = 0$, and $FN_{ij} = 95$. In this case, the F_1 score catches the faulty behaviour of the algorithm by reporting an extremely cautionary value of 0. By contrast, the MCC does not depend on which class is the positive one and which is the negative one.

Several scientific studies (mostly by Chicco and his co-workers [31-38]) show why the Matthews correlation coefficient (MCC) is more informative and trustworthy than confusion-entropy error, accuracy, F1 score, bookmaker informedness, markedness, balanced accuracy, and the diagnostic odds ratio (DOR). A high Matthews correlation coefficient (close to $+1$) means always high values for all the four basic indicators $Sens_{ij}, Spec_{ij}, PPV_{ij}$ and NPV_{ij}. Chicco et al. [36] identify it as the only metric that possesses this property. For comparison, a high F_1 score means high values for just two of these basic indicators $Sens_{ij}$, and PPV_{ij}. Chicco et al. [36] still plan to compare the Matthews correlation coefficient with other metrics, such as Brier score [21], Cohen's Kappa [22], K measure [23], Fowlkes-Mallows index [24] and H-index [25]. The results of Chicco et al. [36] are in line with the fact that the MCC has attracted the attention of the machine learning and the diagnostic testing communities as a method that summarizes the contingency matrix into a single value [37,38].

Table 3. Types of prediction in terms of indicators

	Direct Basic Indicators $\{\boldsymbol{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}}\}$	**Complementary Basic Indicators** $\{\boldsymbol{FNR_{ij}, FPR_{ij}, FOR_{ij}, FDR_{ij}}\}$	**Good Composite indicators** $\boldsymbol{M \in \{I_{ij}, M_{ij}, MCC, SGM, AM, HM\}}$
Perfect Prediction	$Sens_{ij} + Spec_{ij} = 2.0,$ $PPV_{ij} + NPV_{ij} = 2.0,$ $Sens_{ij} = Spec_{ij} = PPV_{ij} = NPV_{ij} = 1.0$	$FNR_{ij} + FPR_{ij} = 0.0,$ $FOR_{ij} + FDR_{ij} = 0.0,$ $FNR_{ij} = FPR_{ij} = FOR_{ij} = FDR_{ij} = 0.0$	$M = +1.0$
Good Prediction	$1 < Sens_{ij} + Spec_{ij} \leq 2,$ $1 < PPV_{ij} + NPV_{ij} \leq 2,$	$0 \leq FNR_{ij} + FPR_{ij} < 1,$ $0 \leq FOR_{ij} + FDR_{ij} < 1,$	$0.0 < M \leq 1.0$
Random-Guessing-Like Prediction	$Sens_{ij} + Spec_{ij} = 1.0,$ $PPV_{ij} + NPV_{ij} = 1.0,$	$FNR_{ij} + FPR_{ij} = 1.0,$ $FOR_{ij} + FDR_{ij} = 1.0,$	$M = 0.0$
Bad Prediction	$0 \leq Sens_{ij} + Spec_{ij} < 1,$ $0 \leq PPV_{ij} + NPV_{ij} < 1,$	$1 < FNR_{ij} + FPR_{ij} \leq 2,$ $1 < FOR_{ij} + FDR_{ij} \leq 2,$	$-1.0 \leq M < 0.0$
Completely-contradictory Prediction	$Sens_{ij} + Spec_{ij} = 0.0,$ $PPV_{ij} + NPV_{ij} = 0.0,$ $Sens_{ij} = Spec_{ij} = PPV_{ij} = NPV_{ij} = 0.0$	$FNR_{ij} + FPR_{ij} = 2.0,$ $FOR_{ij} + NPV_{ij} = 2.0,$ $FNR_{ij} = FPR_{ij} = FOR_{ij} = FDR_{ij} = 1.0$	$M = -1.0$

In passing, we claim that the new composite indicator introduced in (18)

$$Informedness_{ij} * Markedness_{ij} = (Sens_{ij} + Spec_{ij} - 1) * (PPV_{ij} + NPV_{ij} - 1) = (FNR_{ij} + FPR_{ij} - 1,) * (FDR_{ij} + FOR_{ij} - 1), \quad (31)$$

has two elegant formulas in terms of all the basic indicators, and would have proven to be a strong competitor to the MCC indicator had it not lost the sign information of its constituent elements ($Informedness_{ij}$ and $Markedness_{ij}$). Further, we note that

- A high MCC (close to $+1$) means high values for $Sens_{ij}, Spec_{ij}, PPV_{ij}$ and NPV_{ij} (each close to $+1$) [36]. Likewise, a low MCC (close to -1) means low values for $Sens_{ij}, Spec_{ij}, PPV_{ij}$ and NPV_{ij} (each close to 0).
- A high $Informedness_{ij}$ (close to $+1$) means high values for $Sens_{ij}$ and $Spec_{ij}$ and at least one of PPV_{ij} and NPV_{ij} [36]. Likewise, a low $Informedness_{ij}$ (close to -1) means low values for $Sens_{ij}$ and $Spec_{ij}$ and at least one of PPV_{ij} and NPV_{ij}.
- A high $Markedness_{ij}$ (close to $+1$) means high values for PPV_{ij} and NPV_{ij} and at least one of $Sens_{ij}$ and $Spec_{ij}$ [36]. Likewise, A low $Markedness_{ij}$ (close to -1) means low values for PPV_{ij} and NPV_{ij} and at least one of $Sens_{ij}$ and $Spec_{ij}$.

Therefore, a high value (close to $+1$) of our new product indicator in (18) or (31) means high magnitudes of $Informedness_{ij}$ and $Markedness_{ij}$ (with both close to $+1$ or both close to -1), which, in turn, means high values (close to $+1$) for all the four basic direct indicators $Sens_{ij}, Spec_{ij}, PPV_{ij}$ and NPV_{ij}, or low values (close to 0) for all of them. Therefore, a better competitive indicator would be a signed geometric mean of the two indicators $Informedness_{ij}$ and $Markedness_{ij}$, which inherits their common sign as its own sign. This signed geometric mean might be as effective as MCC in summarizing the contingency matrix into a single value. However, its formula lacks the elegance of that of MCC, as it is given by.

$$SGM = SGM_{ij} = SGM_{ji} = SGN\left(Markedness_{ij}\right) * SQRT\left(Informedness_{ij} * Markedness_{ij}\right). \quad (32)$$

In addition to this signed geometric mean, there are two other means, the arithmetic mean and the harmonic mean given by

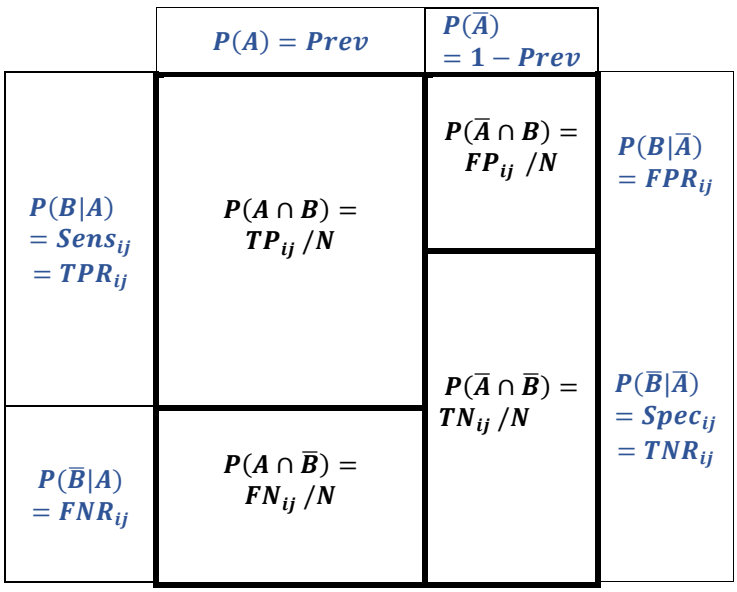
$$AM = AM_{ij} = AM_{ji} = (Informedness_{ij} + Markedness_{ij})/2, \quad (33)$$

$$HM = HM_{ij} = HM_{ji} = 2 * Informedness_{ij} * Markedness_{ij}/(Informedness_{ij} + Markedness). \quad (34)$$

In (34), we assign to HM the value 0 (rather than the undefined 0/0) when it is the harmonic mean of two zeroes. Each of these three means enjoys inherent symmetry and belongs to the interval $[-1.0, +1.0]$ like MCC, and might be a plausible competitor to it. These three novel metrics are included in Table 3 under the umbrella of good composite indicators, used in addition to the eight basic indicators, to identify various types of prediction, going down from the perfect type to the completely-contradictory type.

6. ON THE SOLUTION OF TERNARY PROBLEMS OF CONDITIONAL PROBABILITIES

A quick glance at our earlier equations (1-4, 7-11, 19-25) suggest that techniques of solving ternary problems of conditional probability [10-20,39-42] are essential for full characterization of the two-by-two contingency table. A normalized version of this table might be enhanced by being interpreted as a probabilistic Universe of Discourse. However, it still suffers from two inter-related shortcomings, arising from lack of length/area proportionality and a potential misconception concerning a false assumption of independence between the two underlying events. Rushdi and Serag [16] proposed the use of Fig. 3 as a remedy of these two shortcomings by modifying the normalized contingency matrix into a new Karnaugh-map-like diagram that resembles an eikosogram [43,44]. Furthermore, they suggested the use of the pair of functionally complementary versions of this diagram (shown in Fig. 3) to handle any ternary problem of conditional probability. The two diagrams split the unknowns and equations between themselves in a fashion that allows the use of a divide and-conquer strategy to handle such a problem. This methodology is particularly useful in various areas of diagnostic testing such as clinical or epidemiological testing, though it is still conveniently applicable in other types of problems of general nature involving conditional probabilities. Rushdi and Serag [16,17] explained why and how a conditional-probability problem (with exactly three appropriate quantities being given or pre-specified) can be solved. They also identified the case when an arithmetic solution is possible and differentiated this case from the case when an algebraic solution is warranted. The methodology proposed in [16] can be used to recover all known relations involving quantities pertinent to or derivable from the two-by-two contingency table. As a particularly significant offshoot, this methodology shows that the four most prominent indicators of diagnostic testing (Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value) constitute three rather than four independent quantities. This observation is virtually unheard of, though it is implicit in earlier solutions of the ternary problem of conditional probability [12,13,15].



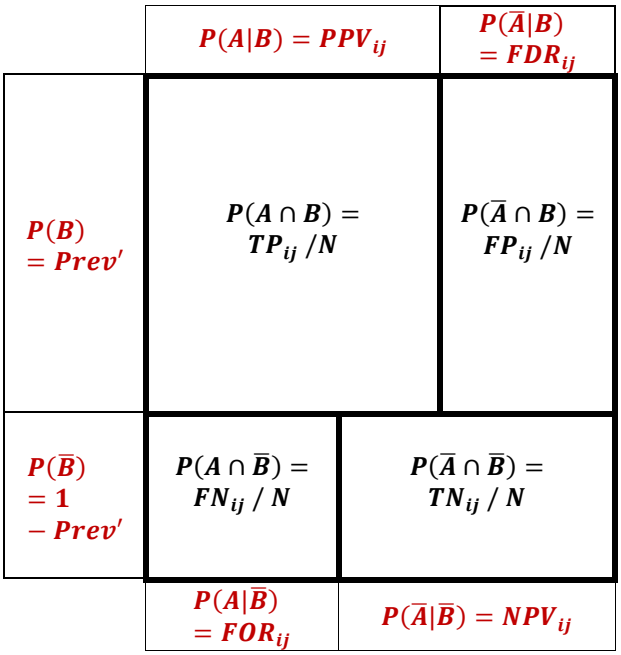


Fig. 3. The probability Universe of Discourse being replaced by two different length/area-proportional Karnaugh-Map-like Diagrams (Eikosogram Diagrams). Each diagram is a liaison among the four conjunctive probabilities, two specific marginal probabilities and four specific conditional probabilities. Each map supplies four independent equations, each of which expresses a conjunctive probability (as a product of a conditional probability and a marginal one), as well as two additive relations for conditional probabilities. These twelve basic equations are supplemented by an independent equation (an additive relation for two marginal probabilities or four conjunctive probabilities)

In passing, we note that all the quantities with values in the unit interval [0.0, 1.0] look like probabilities, and, in fact, have probability interpretations [10-16,45-49]. These interpretations are quite obvious for the basic indicators, and have been shown explicitly in Fig. 2. However, certain considerations and deliberations might be needed for other indicators, such as the F-scores [50,51]. Probabilistic interpretation is also possible for metrics located in the interval [– 1.0, 1.0] (such as the MCC [52]), by mapping this interval to the unit interval [0.0, 1.0].

7. CONCLUSIONS

This paper dealt with indicators derived of the ubiquitous two-by-two contingency table (confusion matrix) that has widespread applications in many fields, including, in particular, the fields of binary classification and clinical or epidemiological testing. The paper presented a variety of these indicators, and stressed the fact that among these the Index of Association (Matthews Correlation Coefficient) has particular advantages. Other, novel and hopefully advantageous, metrics derived from the celebrated Informedness and Markedness metrics have been proposed herein.

We implemented all the equations in Table 1 to compute all the metrics and indicators therein based on knowledge of either (a) the set of four entries of the contingency table $\{TP_{ij}, FP_{ij}, FN_{ij}, TN_{ij}\}$, or (b) the set of true (pre-test) prevalence, sensitivity, and specificity $\{Prev, Sens_{ij}, Spec_{ij}\}$. Techniques of solving ternary problems of conditional probability [10-20,37-40] were incorporated to attain the needed computations. We used a potpourri of test cases to reveal and unravel many of the properties and inter-relationships among these indicators. Our results, reported in a sequel of this paper [53], assert that the MCC is the most reliable single metric that can be derived from the contingency table, and that all the four basic indicators $Sens_{ij}, Spec_{ij}, PPV_{ij}$ and NPV_{ij} must be high for the MCC to be high. Further work is warranted to assess the new indicators proposed herein in comparison with the MCC.

ACKNOWLEDGEMENT

The authors are grateful to Dr. Rufaidah Ali Rushdi (Kasr Al-Ainy Faculty of Medicine, Cairo University, Cairo, Arab Republic of Egypt) for fruitful discussions on diagnostic testing in medical context. Material in this paper is heavily based on her paper [13], and our present Table 1 is adapted from [13].

NOTE ADDED IN PROOF

Equation (29) can be used to prove exact equivalence between the MCC and the signed geometric mean SGM of Informedness and Markedness. This equivalence has already been known in the open literature (see, e.g., [36,54-56]). Although we regret oversight on our part leading us to overlook this equivalence, we are still investigating the utility and comparative merits of the two other means we introduced.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

REFERENCES

1. Anderson TW, Finn JD. Summarizing multivariate data: association between categorical variables, Chapter 6 in the new statistical analysis of data. Springer Science & Business Media; 1996.
2. Johnson KM. The two by two diagram: A graphical truth table. Journal of Clinical Epidemiology. 1999;52(11):1073-1082.
3. Flach PA. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In Proceedings of the 20th international conference on machine learning (ICML-03). 2003;194-201.
4. Azzimonti Renzo JC. Failures of common measures of agreement in medicine and the need for a better tool: Feinstein's paradoxes and the dual vision method. Scandinavian Journal of Clinical and Laboratory Investigation. 2003;63(3):207-216.
5. Freeman JV, Julious SA. The analysis of categorical data. Scope. 2007;16(1):18-21.
6. Texel PP. Measure, metric, and indicator: An object-oriented approach for consistent terminology. In 2013 Proceedings of IEEE Southeastcon 2013 Apr 4. IEEE. 2013;1-5.
7. Canbek G, Sagiroglu S, Temizel TT, Baykal N. Binary classification performance measures / metrics: A comprehensive visualized roadmap to gain new insights. In 2017 International Conference on Computer Science and Engineering (UBMK) 2017 Oct 5. IEEE. 2017;821-826.
8. Brzezinski D, Stefanowski J, Susmaga R, Szczęch I. Visual-based analysis of classification measures and their properties for class imbalanced problems. Information Sciences. 2018;462:242-261.
9. Neth H, Gradwohl N, Streeb D, Keim DA, Gaissmaier W. Perspectives on the 2× 2 Matrix: Solving Semantically Distinct Problems Based on a Shared Structure of Binary Contingencies. Frontiers in Psychology. 2020;11(567817):1-31.
10. Rushdi RA, Rushdi AM, Talmees FA. Novel pedagogical methods for conditional-probability computations in medical disciplines. Journal of Advances in Medicine and Medical Research. 2018; 25(10):1-15.
11. Rushdi AMA, Talmees FA. An exposition of the eight basic measures in diagnostic testing using several pedagogical tools, Journal of Advances in Mathematics and Computer Science. 2018;26(3):1-17.
12. Rushdi RA, Rushdi AM. Common fallacies of probability in medical context: A simple mathematical exposition. Journal of Advances in Medicine and Medical Research. 2018;26(1):1-21.
13. Rushdi RA, Rushdi AM. Karnaugh-map utility in medical studies: The case of Fetal Malnutrition. International Journal of Mathematical, Engineering and Management Sciences. 2018;3(3):220-244.
14. Rushdi AMA, Talmees FA. Computations of the Eight Basic Measures in Diagnostic Testing. Chapter 6 in Advances in Mathematics and Computer Science, Vol. 2, Book Publishers International, Hooghly, West Bengal, India. 2019;66-87.
15. Rushdi RAM, Rushdi AMA. Mathematics and Examples for Avoiding Common Probability Fallacies in Medical Disciplines. Chapter 11 in Current Trends in Medicine and Medical Research, Book Publishers International, Hooghly, West Bengal, India. 2019;1: 106-132.
16. Rushdi AMA, Serag HA. Solutions of ternary problems of conditional probability with applications to mathematical epidemiology and the COVID-19 pandemic. International Journal of Mathematical,

Engineering and Management Sciences. 2020;5(5):787-811.
17. Serag HA, Rushdi AMA. Checking consistency among the four basic indicators of diagnostic testing in Saudi medical journals, Asian Journal of Medical Principles and Clinical Practice. 2021;4(1): 14-27.
18. Rushdi AMA, Serag HA. Inter-relationships among the four basic measures of diagnostic testing: A signal-flow-graph approach. Journal of King Abdulaziz University: Computing and Information Technology Sciences. 2021;10(1):49-72.
19. Rushdi AM, Serag HA. Has the pandemic triggered a 'paperdemic'? Towards an assessment of diagnostic indicators for COVID-19. International Journal of Pathogen Research. 2021;6(2):28-49.
20. Rushdi, RA, Rushdi, AM, Talmees, FA. Review of Methods for Conditional-Probability Computations in Medical Disciplines, a Chapter in Highlights on Medicine and Medical Research, Book Publishers International, Hooghly, West Bengal, India. 2021: 76-94.
21. Rufibach K. Use of Brier score to assess binary predictions. Journal of Clinical Epidemiology. 2010;63(8):938-939.
22. Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 1960;20(1): 37-46.
23. Sebastiani F. An axiomatically derived measure for the evaluation of classification algorithms. In Proceedings of the 2015 International Conference on the theory of Information Retrieval. Sep 27, 2015; 11-20.
24. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association. 1983;78(383):553-569.
25. Campagner A, Sconfienza L, Cabitza F. H-accuracy, an alternative metric to assess classification models in medicine. In Pape-Haugaard LB, et al. (Editors), Digital Personalized Health and Medicine; Studies in Health Technology and Informatics; IOS Press: Amsterdam, The Netherlands. 2020; 242-246.
26. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure. 1975;405(2):442-451.
27. Setiawan AW. Image Segmentation Metrics in Skin Lesion: Accuracy, Sensitivity, Specificity, Dice Coefficient, Jaccard Index, and Matthews Correlation Coefficient. In 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM). IEEE. 2020;97-102.
28. Yule GU. On the methods of measuring association between two attributes. Journal of the Royal Statistical Society. 1912;75(6): 579-652.
29. Guilford JP, Perry NC. Estimation of other coefficients of correlation from the phi coefficient. Psychometrika. 1951;16(3): 335-46.
30. Cox DR, Wermuth N. A comment on the coefficient of determination for binary responses. The American Statistician. 1992;46(1):1-4.
31. Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining. 2017;10(1):1-7.
32. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):1-3.
33. Yao J, Shepperd M. Assessing software defection prediction performance: Why using the Matthews correlation coefficient matters. In Proceedings of the Evaluation and Assessment in Software Engineering. 2020;120-129.
34. Zhu Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognition Letters. 2020;136:71-80.
35. Chicco D, Starovoitov V, Jurman G. The Benefits of the Matthews Correlation Coefficient (MCC) Over the Diagnostic Odds Ratio (DOR) in Binary Classification Assessment. IEEE Access. 2021;9:47112-471124.
36. Chicco D, Tötsch N, Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Mining. 2021;14(1): 1-22.
37. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16(5): 412-24.
38. Jurman G, Riccadonna S, Furlanello C. A

comparison of MCC and CEN error measures in multi-class prediction. PloS One. 2012;7(8,e41882):1-8.
39. Carles M, Huerta MP. Conditional probability problems and contexts. The diagnostic test context. In Proceedings of the Fifth Congress of the European Society for Research in Mathematics Education, CERME. 2007;5(2):702-710.
40. Huerta MP. On conditional probability problem solving research–structures and contexts. International Electronic Journal of Mathematics Education. 2009;4(3):163-94.
41. Huerta MP. Researching conditional probability problem solving. In Probabilistic Thinking. Springer, Dordrecht. 2014;613-639.
42. Huerta MP, Cerdán F, Lonjedo MA, Edo P. Assessing difficulties of conditional probability problems. In Proceedings of the Seventh Congress of the European Society for Research in Mathematics Education. 2011;807-817.
43. Oldford RW, Cherry WH. Picturing probability: The poverty of Venn diagrams, the richness of eikosograms. Retrieved from 2006. Available:http://www.stats.uwaterloo.ca/~rwoldfor/papers/venn/eikosograms/paperpdf.pdf
44. Pfannkuch M, Budgett S. Reasoning from an eikosogram: An exploratory study. International Journal of Research in Undergraduate Mathematics Education. 2017;3(2):283-310.
45. Diamond GA, Forrester JS. Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease. New England Journal of Medicine. 1979;300(24): 1350-1358.
46. Diamond GA, Hirsch MI, Forrester JS, Staniloff HM, Vas R, Halpern SW, Swan HJ. Application of information theory to clinical diagnostic testing. The electrocardiographic stress test. Circulation. 1981;63(4):915-921.
47. SOX Jr HC. Diagnostic decision: Probability theory in the use of diagnostic tests: An introduction to critical study of the literature. Annals of Internal Medicine. 1986;104(1):60-66.
48. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009;45(4):427-437.
49. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS ONE. 2015;10(3),e0118432:1-23.
50. Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval 2005 Mar 21. Springer, Berlin, Heidelberg. 2005;345-359.
51. Lipton ZC, Elkan C, Naryanaswamy B. Thresholding classifiers to maximize F1 score, 2014. arXiv stat.ML. 1402.1892v2.
52. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS ONE. 2017;12(6),e0177678: 1-17.
53. Rushdi AMA, Alghamdi SM. Measures, metrics, and indicators derived from the ubiquitous two-by-two contingency table, Part B: Examples. Asian Journal of Medical Principles and Clinical Practice. 2021;4(3):26-50.
54. Powers, DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies. 2011;2(1):37-63.
55. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition. 2019;91:216-231.
56. Ge W, Fazal Z, Jakobsson E. Using optimal f-measure and random resampling in gene ontology enrichment calculations. Frontiers in Applied Mathematics and Statistics. 2019;5:20-33.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/68338