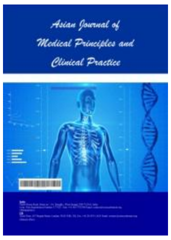

Asian Journal of Medical Principles and Clinical Practice

4(3): 26-50, 2021; Article no.AJMPCP.68420

Measures, Metrics, and Indicators Derived from the Ubiquitous Two-by-Two Contingency Table, Part B: Examples

Ali Muhammad Ali Rushdi[1*] and Sultan Mohammed Alghamdi[1]

[1]*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 the two authors. Author AMAR designed the study, compiled the necessary formulas, provided the test cases and commented on them, managed the literature search, and wrote the manuscript. Author SMA wrote the computer program, and obtained the numerical results. Both authors approved the final manuscript.

Article Information

Editor(s):
(1) Dr. Suprakash Chaudhury, Dr D.Y. Patil Medical College, Hospital & Research Center, India.
Reviewers:
(1) G. Donald Allen, Texas A&M University, USA.
(2) K. Sreelatha, Sreenidhi Institute of Science and Technology, India.
Complete Peer review History: http://www.sdiarticle4.com/review-history/68420

Review Article

Received 08 March 2021
Accepted 13 May 2021
Published 14 May 2021

ABSTRACT

This paper (the second of two sibling papers) continues the tutorial exposition presented in the first part of indicators derived from the ubiquitous two-by-two contingency table (confusion matrix). The indicators considered herein are those given in the context of clinical testing or binary classification. We present a pedagogical program that computes all important indicators 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 $\{Prev,\ Sens_{ij}, Spec_{ij}\}$. The paper presents a potpourri of test cases to reveal and unravel many of the properties and inter-relationships among the indicators studied. All our test cases confirm the theoretical results and arguments in the sister paper. In particular, these test cases collectively assert that the Matthews correlation coefficient (MCC) is the most reliable single metric derivable from the contingency matrix. A concise classification of types of prediction is given in terms of the set of four basic indicators $\{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}\}$ or in terms of MCC alone.

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

Keywords: Diagnostic testing; binary classification; sensitivity; specificity; positive predictive value; negative predictive value; Matthews correlation coefficient.

1. INTRODUCTION

A contingency table (Also called a confusion matrix) is a powerful tool in data analysis employing matrix format for comparing two categorical variables [1-12]. This table (albeit very simple) can be used, and is still being used [12], to derive an amazingly huge number of metrics or indicators in terms of its four entries, called True Positives, False Positives, False Negatives, and True Negatives and denoted TP_{ij}, FP_{ij}, FN_{ij}, and TN_{ij}, where the subscripts ij are used to assert the notion that a test i is assessed, judged or measured relative to a reference or standard test j. One of the derived indicators, the Index of Association (Matthews Correlation Coefficient (MCC)) is noted to be the most reliable single metric derivable from the contingency matrix [13-16]. The aim of this paper is to extend our earlier work in the sister paper [12] and supplement the existing tutorials on quantities derivable from the contingency table [17-27]. The exposition used herein is a novel one as it presents a good number of carefully-selected test cases, and then provides pedagogical comments on the results obtained for each test case. These comments are ultimately summarized in a single table.

The organization of the rest of this paper is as follows. Section 2 is a brief primer about the metrics and indicators considered. Section 3 presents a pedagogical program that computes all important indicators based on knowledge of either (a) the set of the 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 $\{Prev,\ Sens_{ij}, Spec_{ij}\}$. Section 3 also offers a potpourri of test cases to reveal and unravel many of the properties and inter-relationships among the aforementioned metrics and indicators. These test cases collectively assert the claim that the Matthews correlation coefficient (MCC) is the most reliable single metric derivable from the contingency matrix. The results obtained are summarized in a concise classification of the types of prediction in terms of the set of four basic indicators $\{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}\}$ or in terms of MCC alone. Section 4 concludes the paper.

2. METRICS AND INDICATORS CONSIDERED HEREIN

Table 1 (borrowed from the sibling paper [12] and originally adapted from [4]) lists some of the measures or indicators commonly used in diagnostic testing or binary classification. The table expresses each of these quantities in terms of the four elements of the contingency matrix, states its range of values, and identify the value for perfect testing or classification. Many quantities have ranges $[0.0, 1.0]$, but a few belong to $[0.0, \infty)$ or $[-1.0,\ +1.0]$. 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 since they are test-independent.

3. DISPLAY AND COMMENT ON TEST CASES

We implemented all the equations of Table 1 in a program 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 [1-12] were incorporated to attain the needed computations. Table 1 shows fifteen sets of input values used to test our program, which were carefully selected to reveal certain theoretical aspects stressed in [12]. Figs 1-15 display snapshots of computer outputs obtained for the various test cases. Each of these figures was included for a reason, and every figure (except one) has two versions supplied by the two sets of inputs to yield the same output. Many useful comments are included within the captions of these figures. The results obtained for the four basic indicators are checked for consistency in Table 3 according to the novel tests introduced in [8-10].

All our test cases confirm the theoretical results and arguments in the sister paper [12]. In particular, they 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. This is in line with the fact that the MCC has attracted the attention of the diagnostic testing and the machine learning communities as a method that summarizes the contingency matrix into a single value. Table 4 summarizes the results of this paper in a concise classification of the types of prediction in terms of the set of four basic indicators

$\{Sens_{ij}, Spec_{ij}, PPV_{ij}, NPV_{ij}\}$ or in terms of MCC alone. This summary attests once more to the powerfulness of MCC. The sister paper [12] ponders whether novel composite indicators might share this powerfulness, and proposes three novel indicators for this purpose, namely, the arithmetic mean, the harmonic mean, and the signed geometric mean of informedness and markedness.

Table 1. Commonly used quantities pertaining to diagnostic testing (borrowed from the sister paper [12] and originally adapted from [4]). 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

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

Table 2. Various sets of contingency-matrix entries to test the program

TP	TN	FP	FN	Comments	Displayed in Figure
10	990	0	0	Perfect prediction	1
500	500	0	0	Perfect prediction	2
9	900	90	1	Low MCC	3
90	800	100	10	Medium MCC	4
250	250	250	250	Equal entries, zero MCC	5
90	9009	891	10	Gigerenzer et al. [16], Rushdi & Rushdi [3, 6]	6
9	900	1	90	Low MCC	7
90	800	10	100	Medium MCC	8
400	400	100	100	Medium MCC	9
0	1000	0	0	Many NaN results	10
35	35	15	15	Mirror image of case 15	11
4	76	19	1	Chicco et al. [16]	12
95	0	5	0	Chicco [13]	13
90	1	5	4	Chicco [13]	14
15	15	35	35	Negative MCC	15
63	72	28	37	https://en.wikipedia.org/wiki/Receiver_operating_characteris	16
77	23	77	23	tic	17
24	12	88	76		18
76	88	12	24		19

Table 3. Checking consistency among our sets of the four prominent diagnostic indicators. According to the scheme in [8-10], when the sets are consistent they are depicted as uncolored entries, and when they are somewhat problematic they are highlighted in yellow. If the sets are obviously inconsistent they are labelled as orange, and finally if they are dramatically inconsistent, they are highlighted in red. The results of all our cases are wonderfully uncolored (i.e., consistent). There is only some lack of information when original values are missing or an undefined 0/0 is encountered. The diagnostic checking difference (DCD) is admirably equal to 0.0000 in all cases, while the diagnostic checking ratio (DCR) deviates from 1.0000 by no more than 0.0004

#	Original Values				Checking Values		Computed Values			
	$Sens_{ij}$	$Spec_{ij}$	PPV_{ij}	NPV_{ij}	DCD_{ij}	DCR_{ij}	$Sens_{ij}$	$Spec_{ij}$	PPV_{ij}	NPV_{ij}
1a	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
3a	0.9000	0.9091	0.0909	0.9989	0.0000	1.0001	0.9008	0.9098	0.0902	0.9989
4a	0.9000	0.8889	0.4737	0.9877	0.0000	1.0001	0.9003	0.8893	0.4728	0.9877
5a	0.5000	0.5000	0.5000	0.5000	0.0000	#DIV/0!	0.5000	0.5000	0.5000	0.5000
6a	0.9000	0.9100	0.0917	0.9989	0.0000	1.0001	0.9007	0.9106	0.0911	0.9989
7a	0.0909	0.9989	0.9000	0.9091	0.0000	0.9999	0.0902	0.9989	0.9008	0.9098
7b	0.0909	0.9989	0.9008	0.9091	0.0000	1.0000	0.0909	0.9989	0.9008	0.9091
8a	0.4737	0.9877	0.9000	0.8889	0.0000	0.9999	0.4728	0.9877	0.9003	0.8893
8b	0.4737	0.9877	0.9003	0.8889	0.0000	1.0000	0.4736	0.9877	0.9003	0.8889
9a	0.8000	0.8000	0.8000	0.8000	0.0000	1.0000	0.8000	0.8000	0.8000	0.8000
10		1.0000		1.0000	0.0000	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!
11a	0.7000	0.7000	0.7000	0.7000	0.0000	1.0000	0.7000	0.7000	0.7000	0.7000
12a	0.8000	0.8000	0.1739	0.9870	0.0000	0.9999	0.7998	0.7998	0.1741	0.9870
12b	0.8000	0.8000	0.1739	0.9870	0.0000	0.9999	0.7998	0.7998	0.1741	0.9870
13a	1.0000	0.0000	0.9500		0.0000	#DIV/0!	#DIV/0!	0.0000	#DIV/0!	#DIV/0!
14a	0.9574	0.1667	0.9474	0.2000	0.0000	1.0004	0.9575	0.1669	0.9473	0.1998

#	Original Values				Checking Values		Computed Values			
	$Sens_{ij}$	$Spec_{ij}$	PPV_{ij}	NPV_{ij}	DCD_{ij}	DCR_{ij}	$Sens_{ij}$	$Spec_{ij}$	PPV_{ij}	NPV_{ij}
15a	0.3000	0.3000	0.3000	0.3000	0.0000	1.0000	0.3000	0.3000	0.3000	0.3000
16a	0.6300	0.7200	0.6923	0.6606	0.0000	1.0001	0.6300	0.7200	0.6923	0.6606
17a	0.7700	0.2300	0.5000	0.5000	0.0000	#DIV/0!	0.7700	0.2300	0.5000	0.5000
18a	0.2400	0.1200	0.2143	0.1364	0.0000	0.9996	0.2401	0.1200	0.2142	0.1364
19a	0.7600	0.8800	0.8636	0.7857	0.0000	1.0000	0.7599	0.8800	0.8636	0.7858

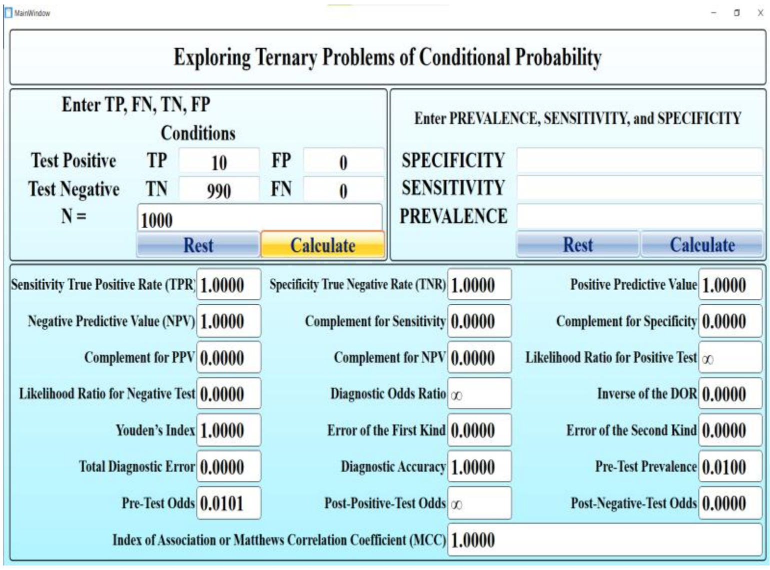

Fig. 1a. First test case with input of contingency matrix entries. This is a case of perfect prediction with negatives more than positives (low prevalence). Except for pre-test prevalence and pre-test odds (which are test-independent), all outcomes are as anticipated in Table 1 for perfect prediction

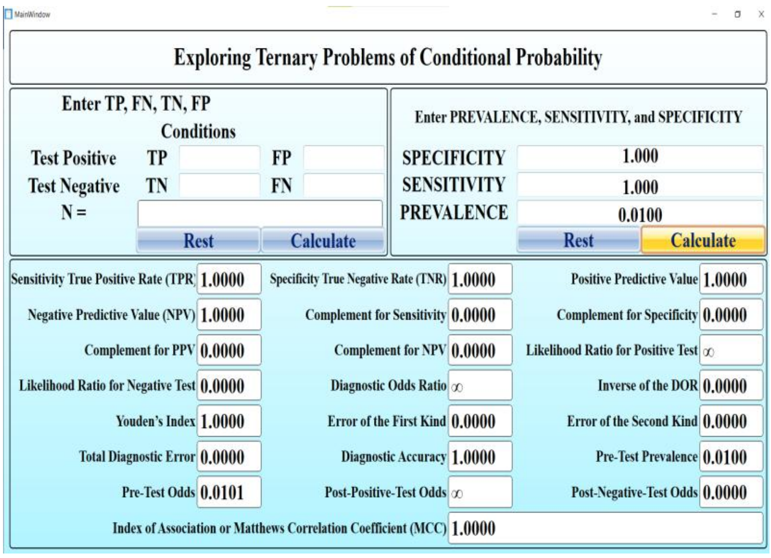

Fig. 1b. First test case with input of pre-test prevalence, sensitivity, and specificity. This is a case of perfect prediction with negatives more than positives (low prevalence). Except for pre-test prevalence and pre-test odds (which are test independent), all outcomes are as anticipated in Table 1 for perfect prediction

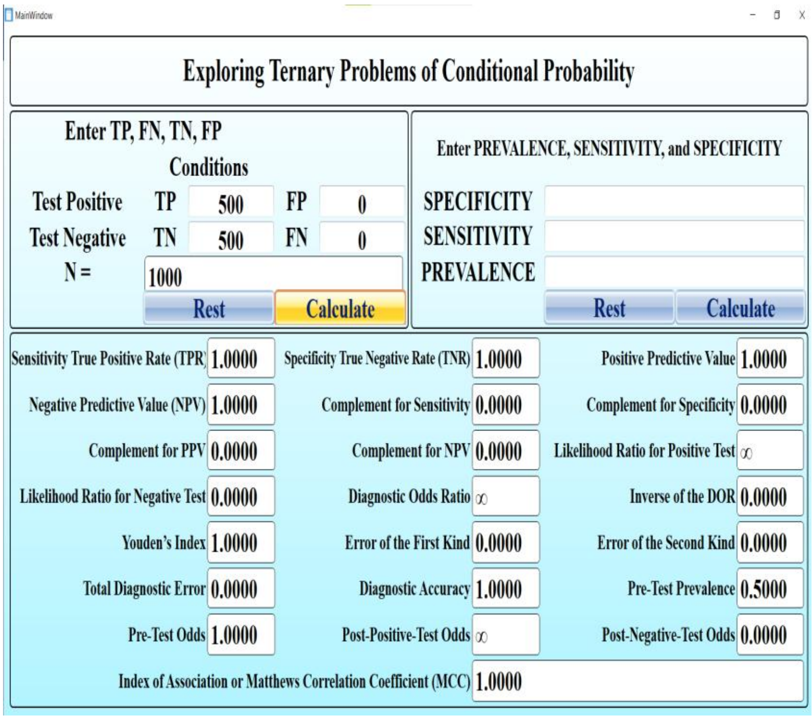

Fig. 2a. Second test case with input of contingency matrix entries. This is a case of perfect prediction with negatives equal to positives (prevalence equal to one half). Again, all outcomes are as anticipated in Table 1 for perfect prediction (except for pre-test prevalence and pre-test odds, which are test-independent)

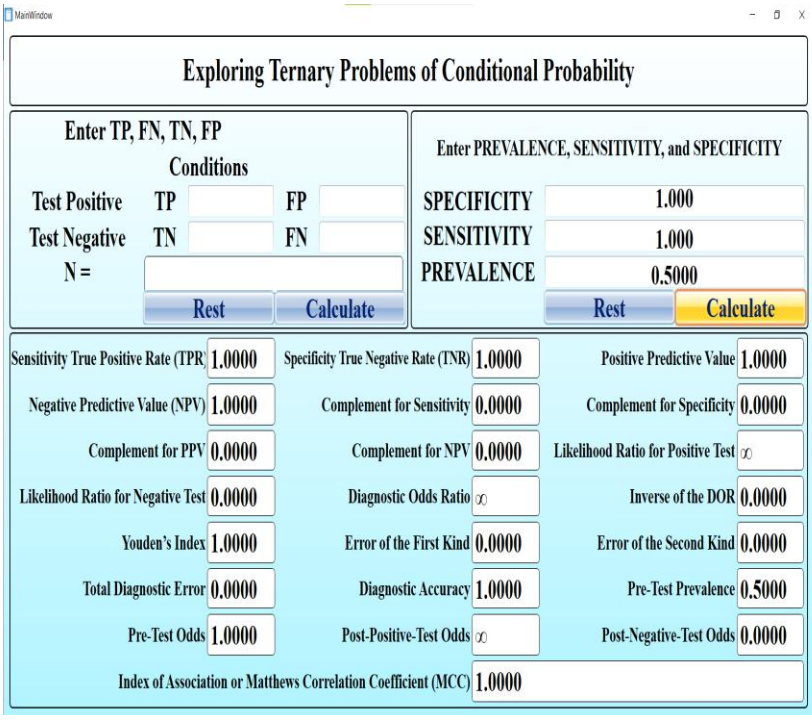

Fig. 2b. Second test case with input of pre-test prevalence, sensitivity, and specificity. This is a case of perfect prediction with negatives equal to positives (prevalence equal to one half). Again, all outcomes are as anticipated in Table 1 for perfect prediction (except for pre-test prevalence and pre-test odds, which are test independent)

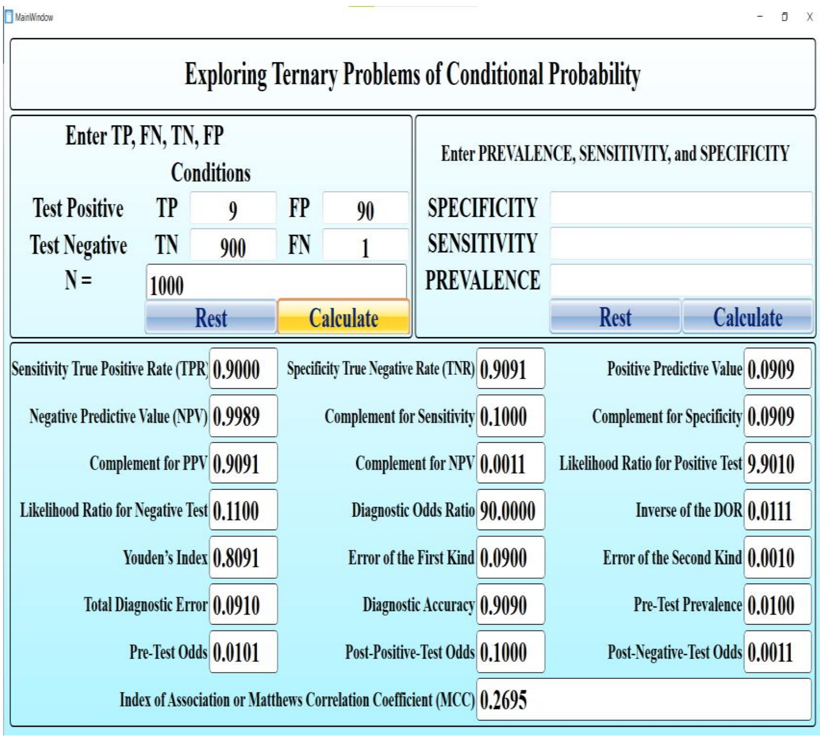

Fig. 3a. Third test case with input of contingency matrix entries. This is a case of a very low PPV, a very high NPV, moderately high sensitivity and specificity and a relatively low MCC

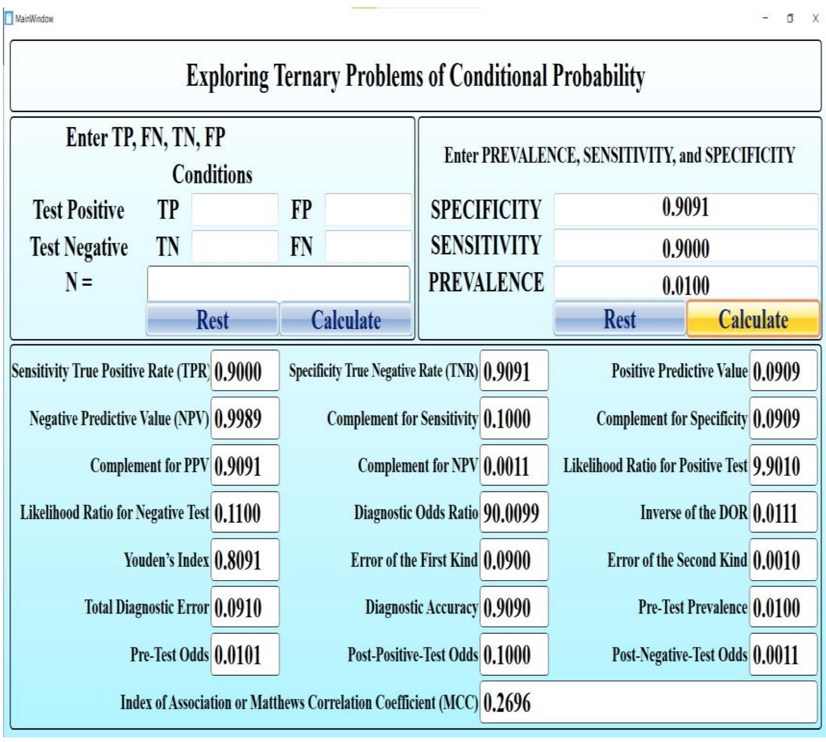

Fig. 3b. Third test case with input of pre-test prevalence, sensitivity, and specificity. This is a case of a very low PPV, a very high NPV, moderately high sensitivity and specificity, and a relatively low MCC. Each numerical value in this figure is the same as the corresponding one in Fig. 3a (to within permissible round-off errors)

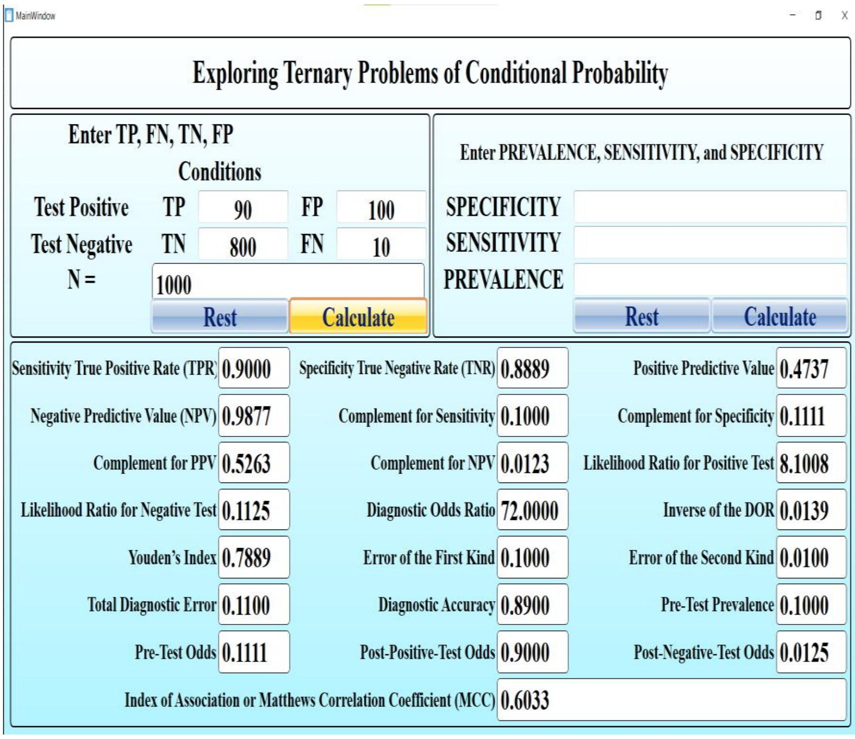

Fig. 4a. Fourth test case with input of contingency matrix entries. This is a case of a low (but not very low) PPV, a very high NPV, moderately high sensitivity and specificity, and a relatively medium MCC

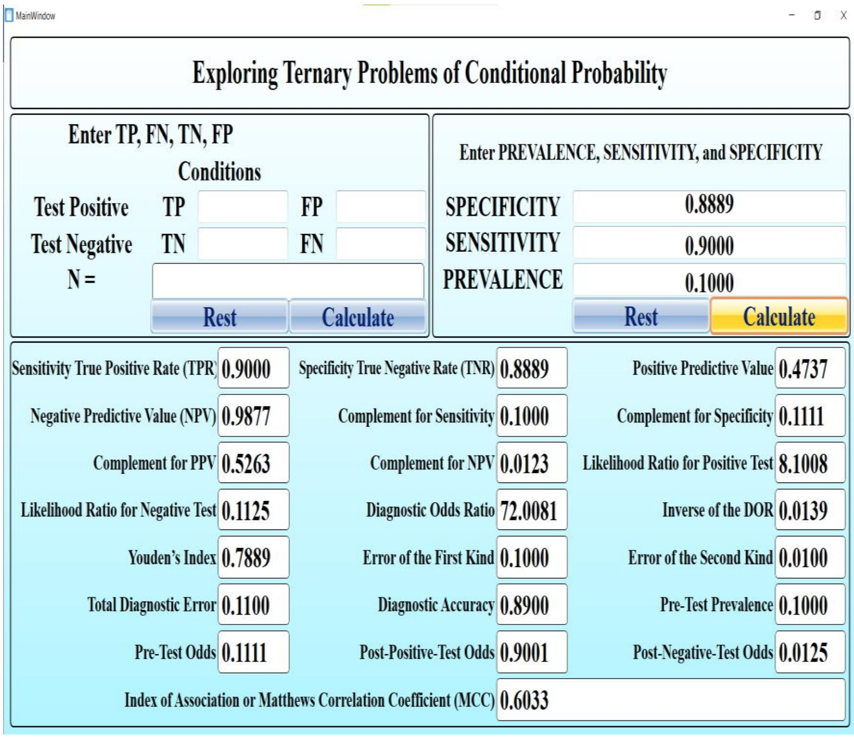

Fig. 4b. Fourth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case of a low (albeit not very low) PPV, a very high NPV, moderately high sensitivity and specificity, and a relatively medium MCC. Each numerical value in this figure is the same as the corresponding one in Fig. 4a (to within permissible round-off errors)

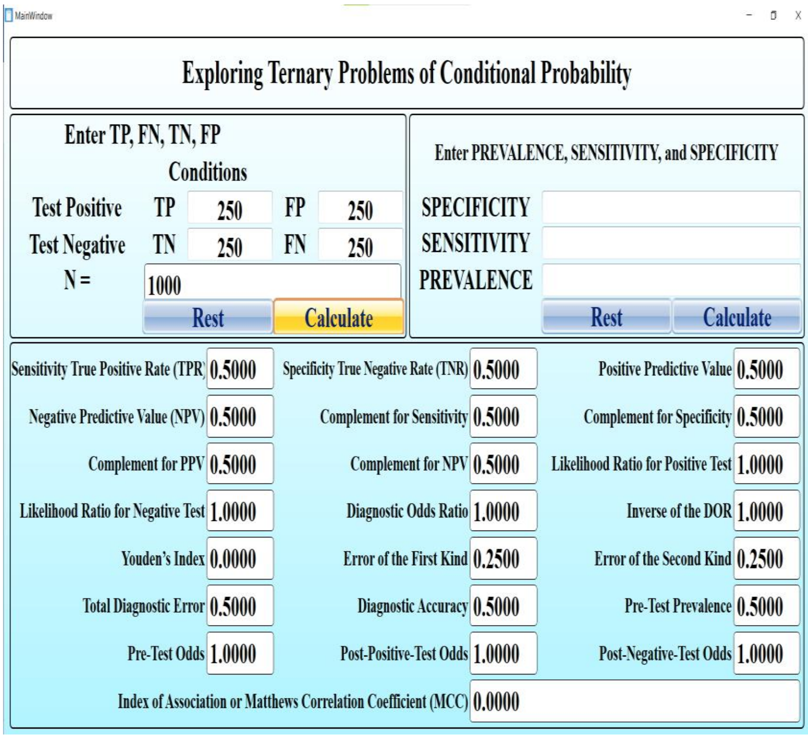

Fig. 5a. Fifth test case with input of contingency matrix entries. This is the equal-entry case with a zero MCC

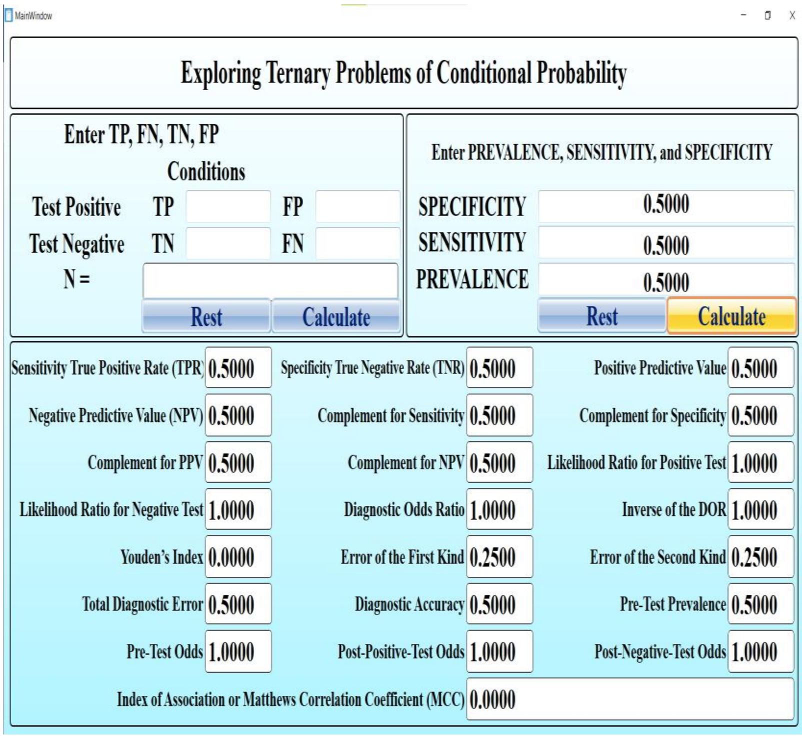

Fig. 5b. Fifth test case with input of pre-test prevalence, sensitivity, and specificity. This is the equal-entry case with a zero MCC. Each numerical value in this figure is the same as the corresponding one in Fig. 5a (to within permissible round-off errors)

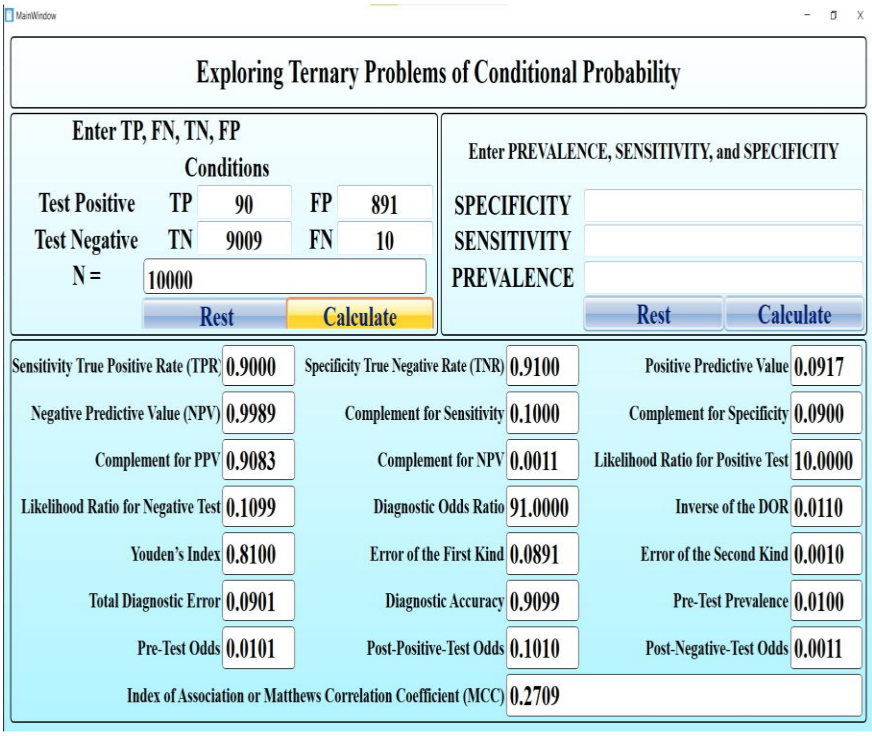

Fig. 6a. Sixth test case with input of contingency matrix entries. This is the celebrated example of Gigerenzer et al. [17] and Rushdi & Rushdi [3, 6], with a poor PPV (apparently despite, but actually because of, high sensitivity and specificity, as well as very high NPV) and a low MCC

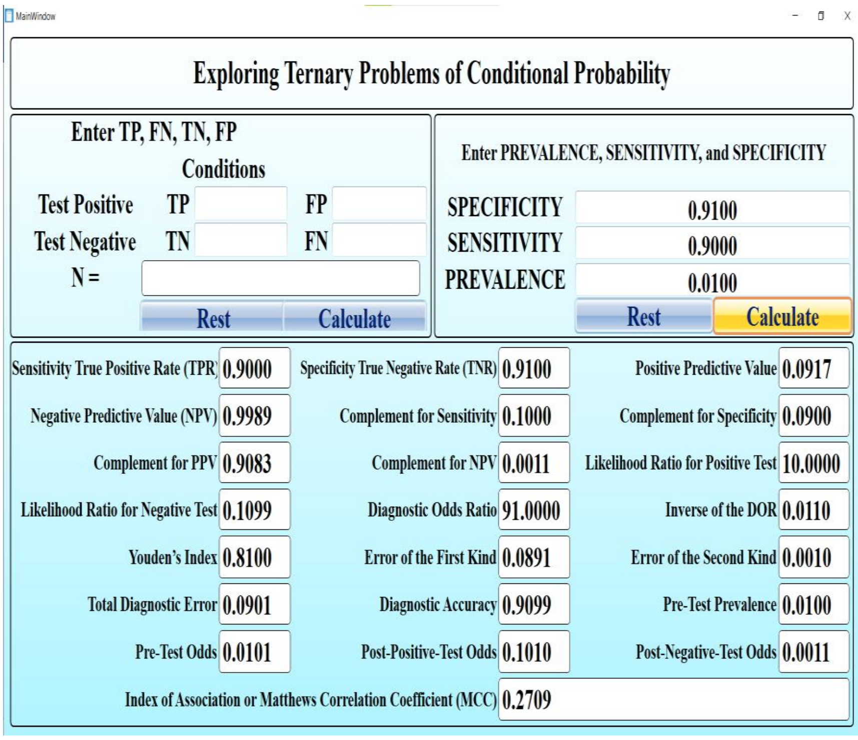

Fig. 6b. Sixth test case with input of pre-test prevalence, sensitivity, and specificity. This is the celebrated example of Gigerenzer et al. [17] and Rushdi & Rushdi [3, 6], with a poor PPV and a low MCC. Each numerical value in this figure is the same as the corresponding one in Fig. 6a (to within permissible round-off errors)

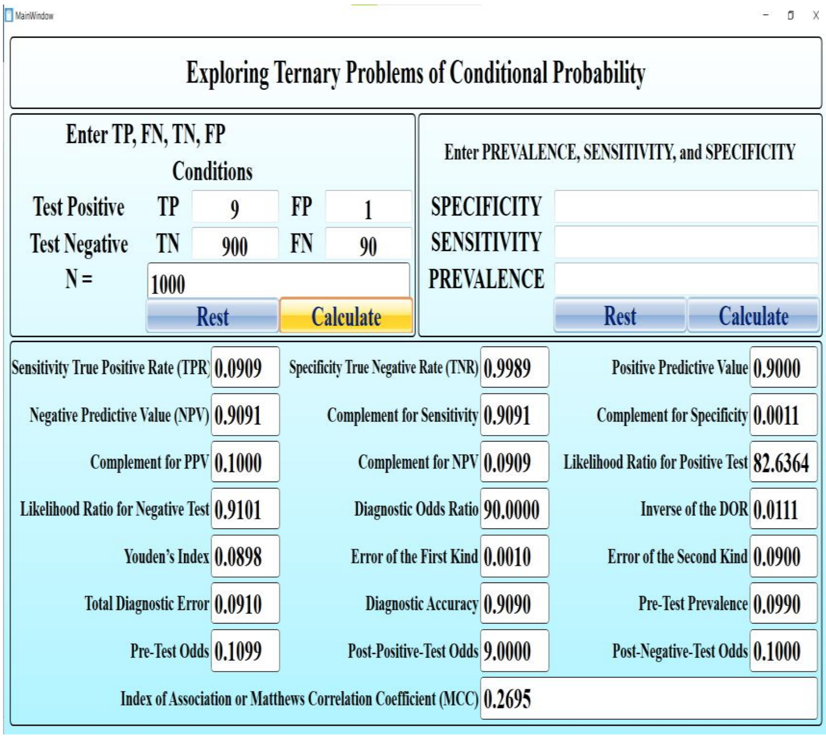

Fig. 7a. Seventh test case with input of contingency matrix entries. This is a case with a poor sensitivity and a low MCC. The poor sensitivity does not contradict (but actually results from) a combination of high predictive values with a very high specificity

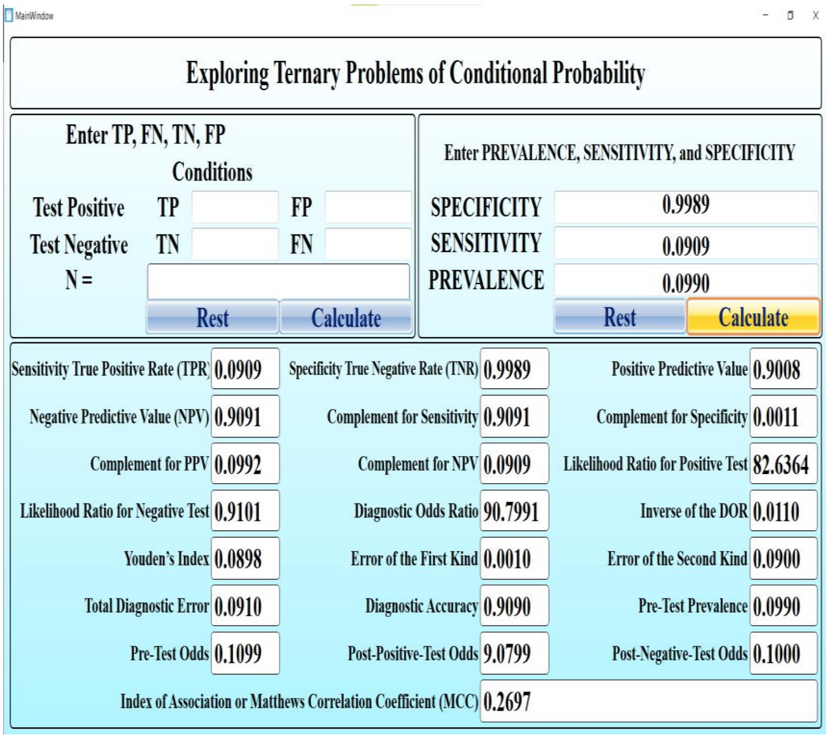

Fig. 7b. Seventh test case with input of pre-test prevalence, sensitivity, and specificity. This is a case with a poor sensitivity and a low MCC. Each numerical value in this figure is the same as the corresponding one in Fig. 7a (to within permissible round-off errors)

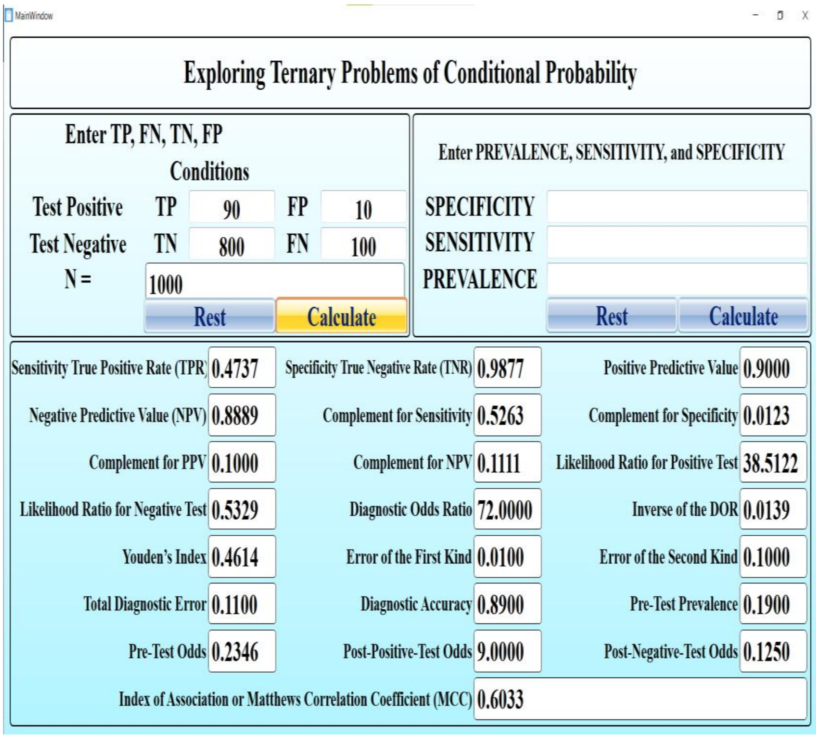

Fig. 8a. Eighth test case with input of contingency matrix entries. This is a case with intermediate sensitivity and MCC

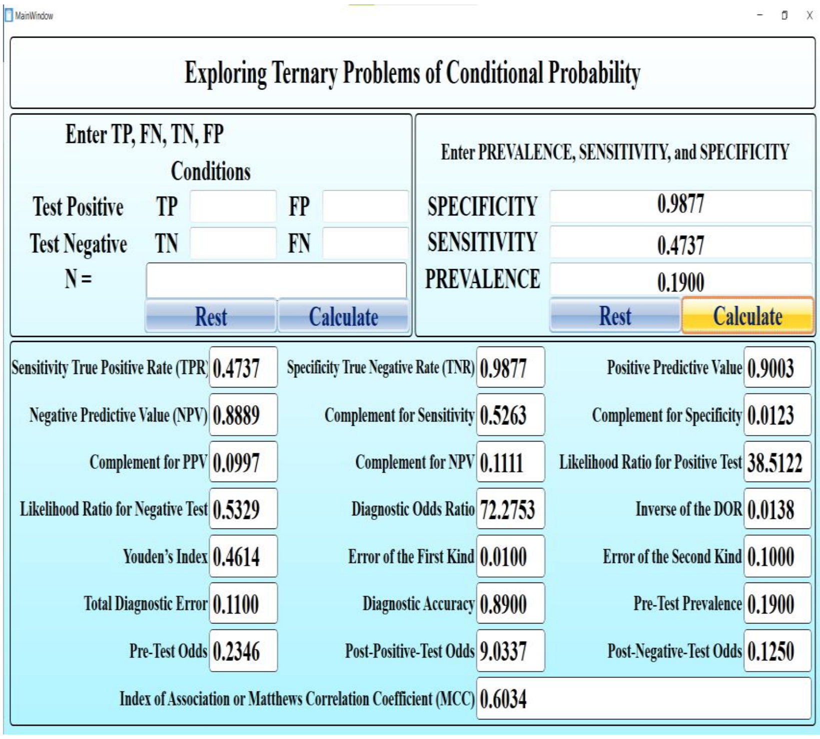

Fig. 8b. Eighth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case with intermediate sensitivity and MCC

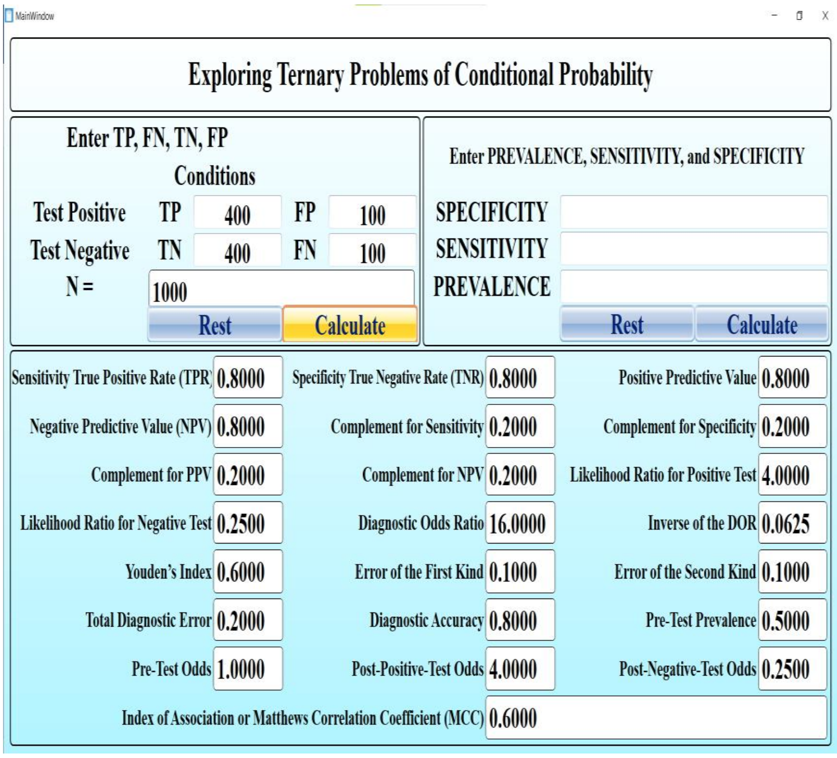

Fig. 9a. Ninth test case with input of contingency matrix entries. This is a case with 'reasonable' direct metrics and an intermediate MCC

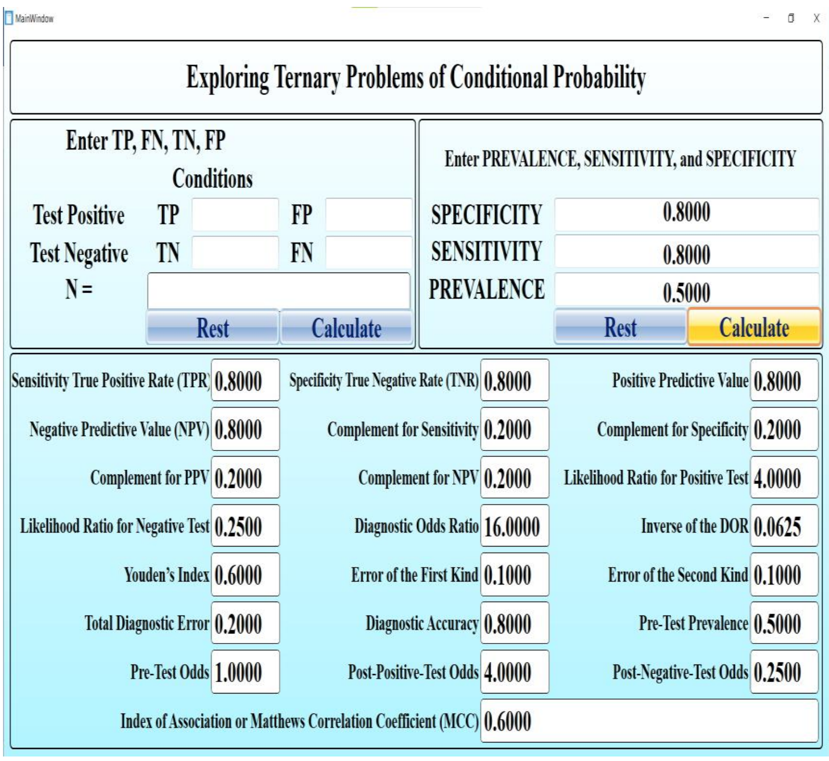

Fig. 9b. Ninth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case with 'reasonable' direct metrics and an intermediate MCC

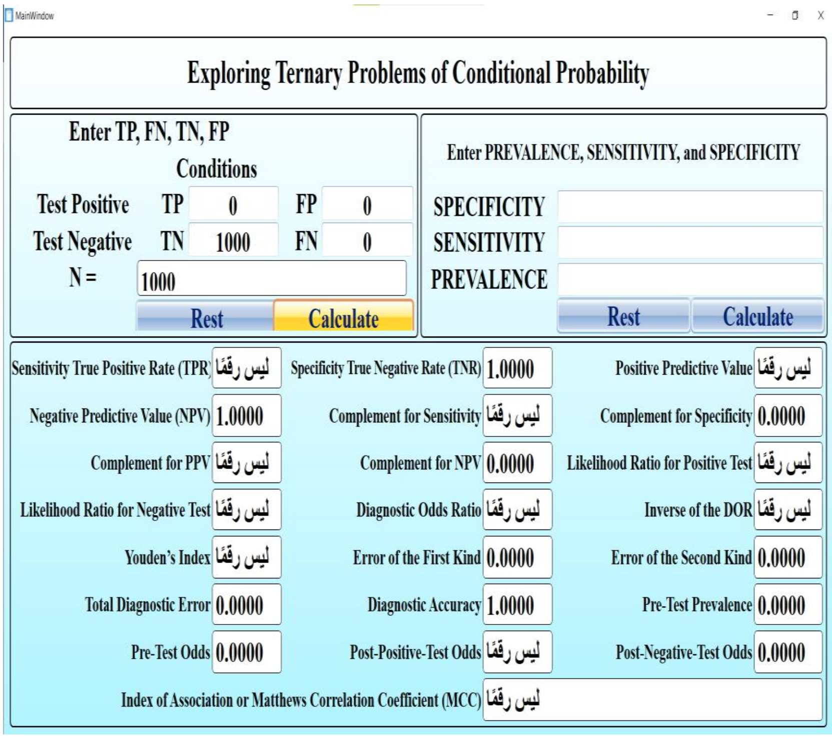

Fig. 10. Tenth test case with input of contingency matrix entries. This is an extreme case in which most computed values are undefined, and designated as NaN or Not a Number (ليس رقما). We deliberately used the Arabic script for NaN to alert the reader that computations are incomplete. Since sensitivity is undefined, this figure is not duplicated

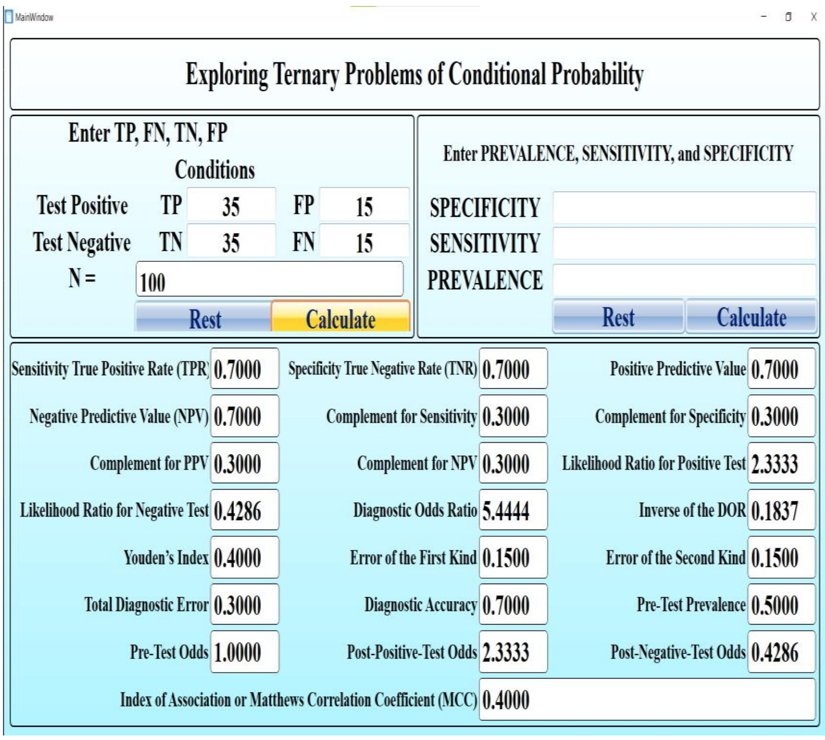

Fig. 11a. Eleventh test case with input of contingency matrix entries. This is a case with somewhat 'reasonable' direct metrics and a low MCC

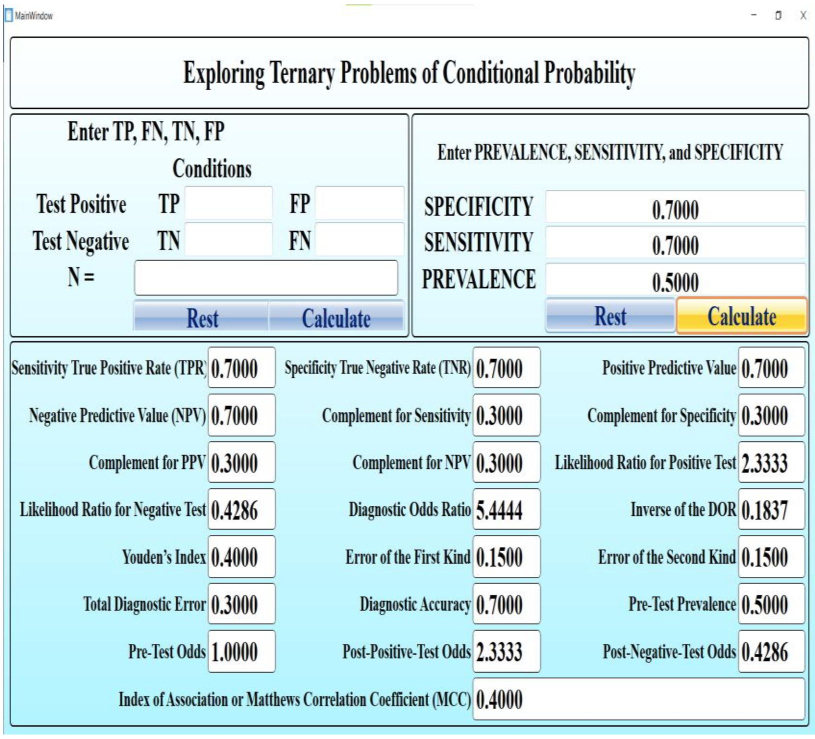

Fig. 11b. Eleventh test case with input of pre-test prevalence, sensitivity, and specificity. This is a case with somewhat 'reasonable' direct metrics and a low MCC

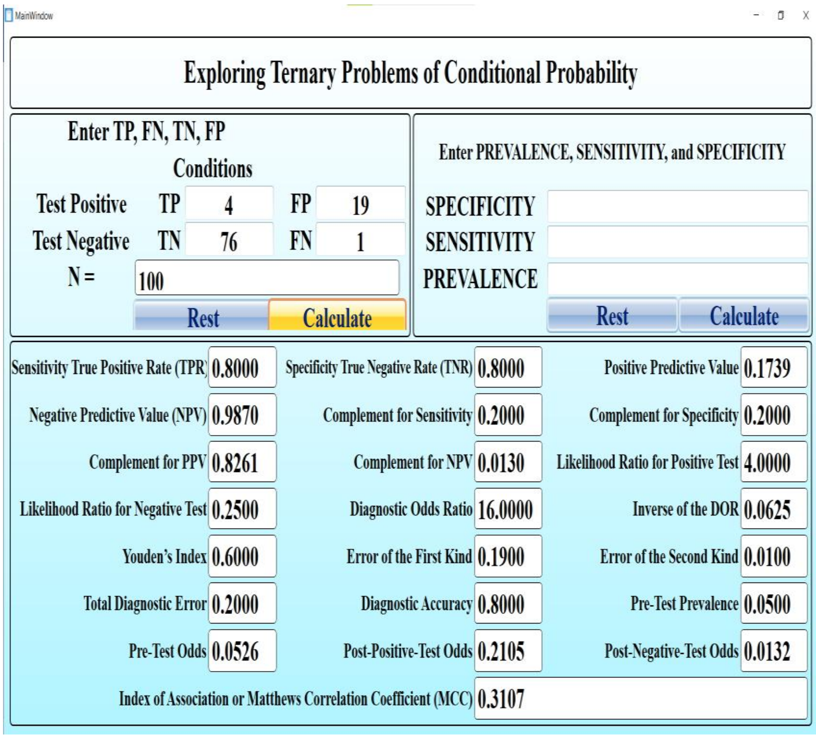

Fig. 12a. Twelfth test case with input of contingency matrix entries. This is a case with a poor PPV and a low MCC. The poor PPV does not contradict (but actually results from) a combination of high sensitivity and specificity with a very high NPV

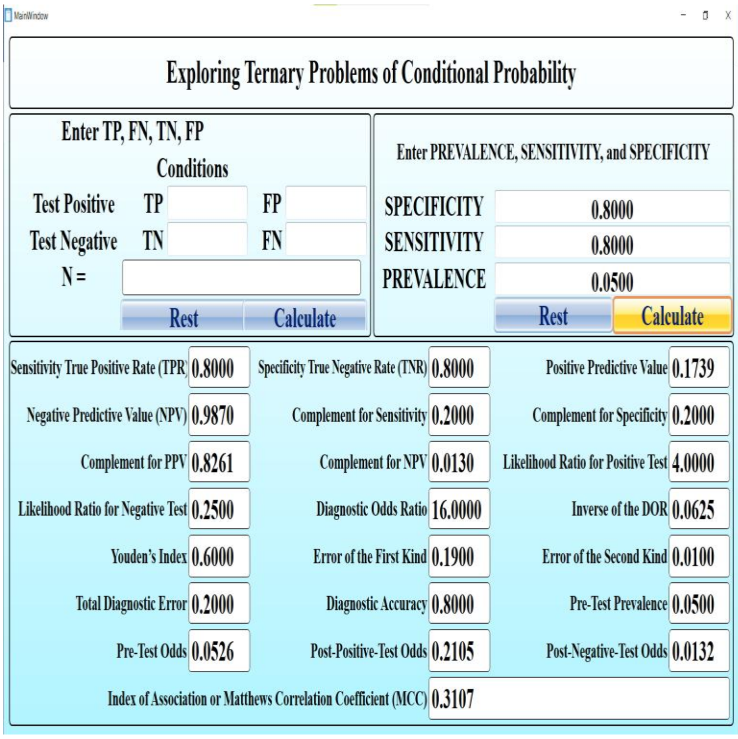

Fig. 12b. Twelfth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case with a poor PPV and a low MCC. The poor PPV does not contradict (but actually results from) a combination of high sensitivity and specificity with a very high NPV

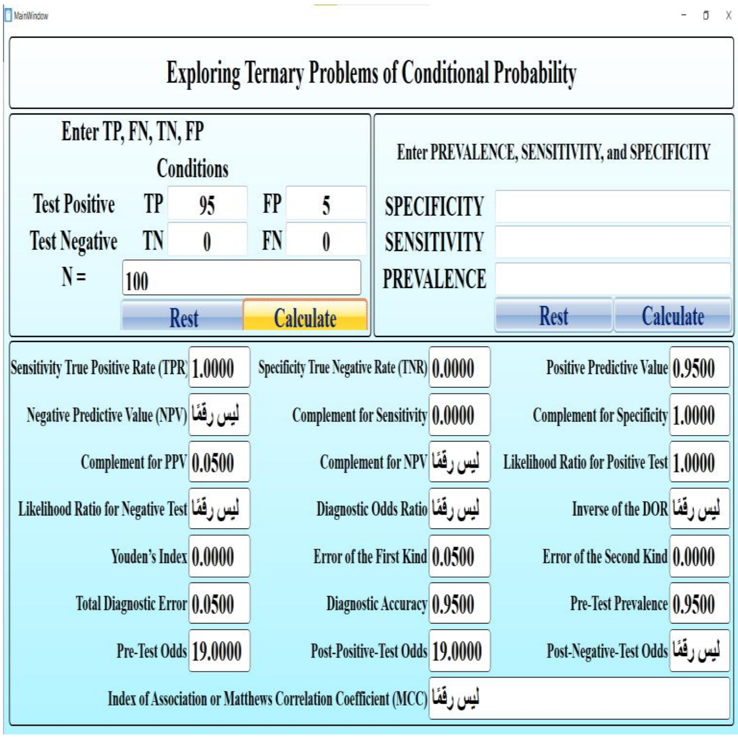

Fig. 13a. Thirteenth test case with input of contingency matrix entries. This is an extreme case in which many computed values are undefined, and designated as NaN or Not a Number (ليس رقما). Unlike Fig. 10, this figure can be duplicated since none of sensitivity, specificity and pre-test prevalence is undefined

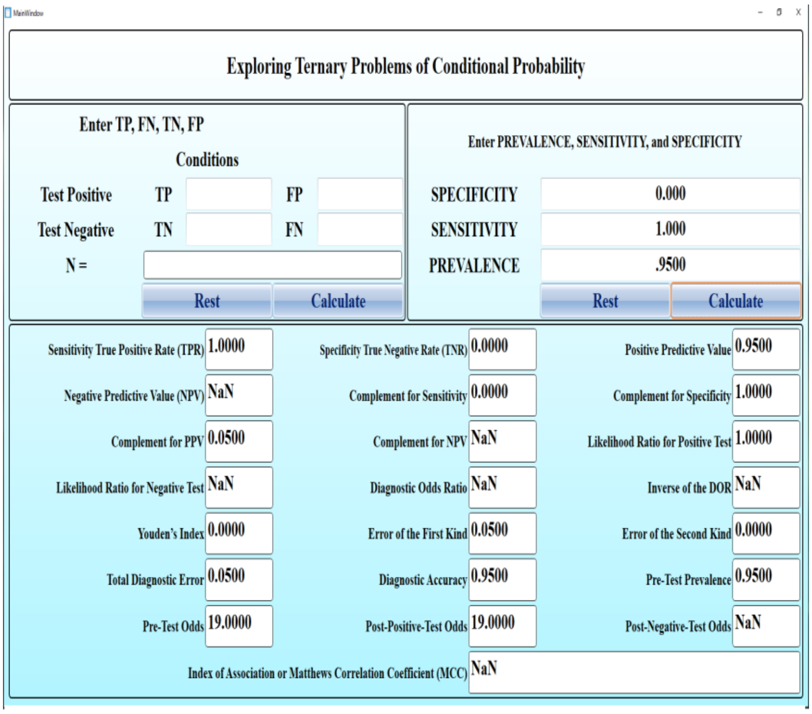

Fig. 13b. Thirteenth test case with input of pre-test prevalence, sensitivity, and specificity. This is an extreme case in which many computed values are undefined, and designated as NaN or Not a Number. Unlike Figs. 10 and 13a, this figure has a true English screen in which the standard notation (NaN) replaces its Arabic equivalent (ليس رقما).

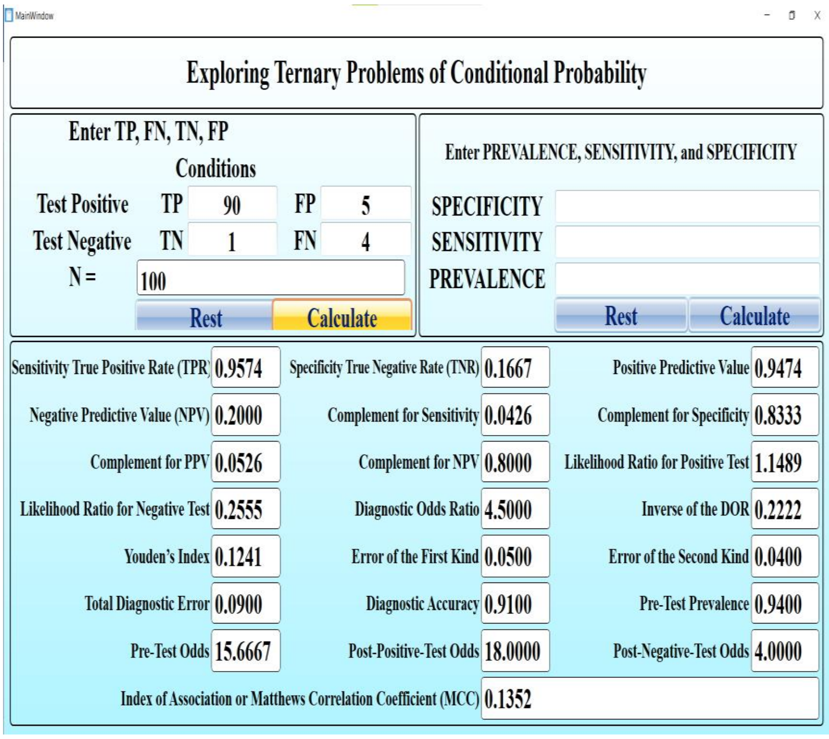

Fig. 14a. Fourteenth test case with input of contingency matrix entries. This is a case with a poor specificity, a poor NPV and a low MCC

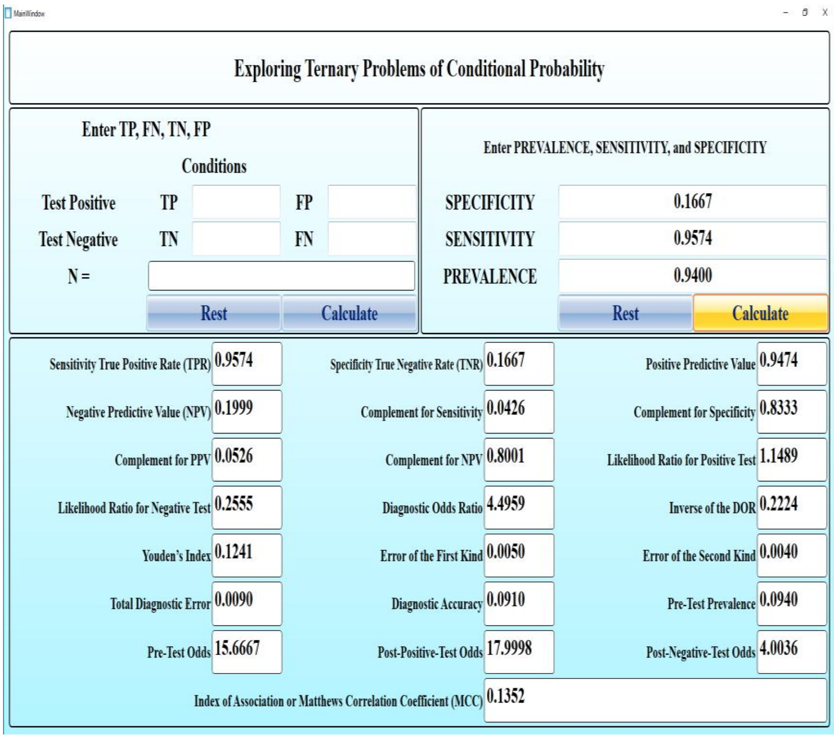

Fig. 14b. Fourteenth test case with input of pre-test prevalence, sensitivity and specificity. This is a case with a poor specificity, a poor NPV and a low MCC

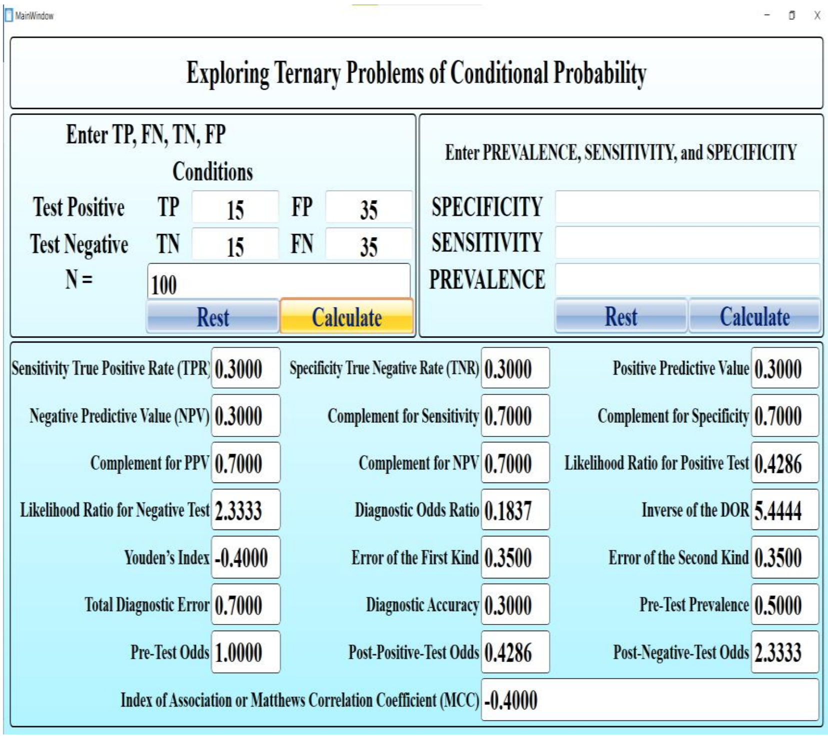

Fig. 15a. Fifteenth test case with input of contingency matrix entries. This is a case with a negative MCC and a negative informedness (Youden's index)

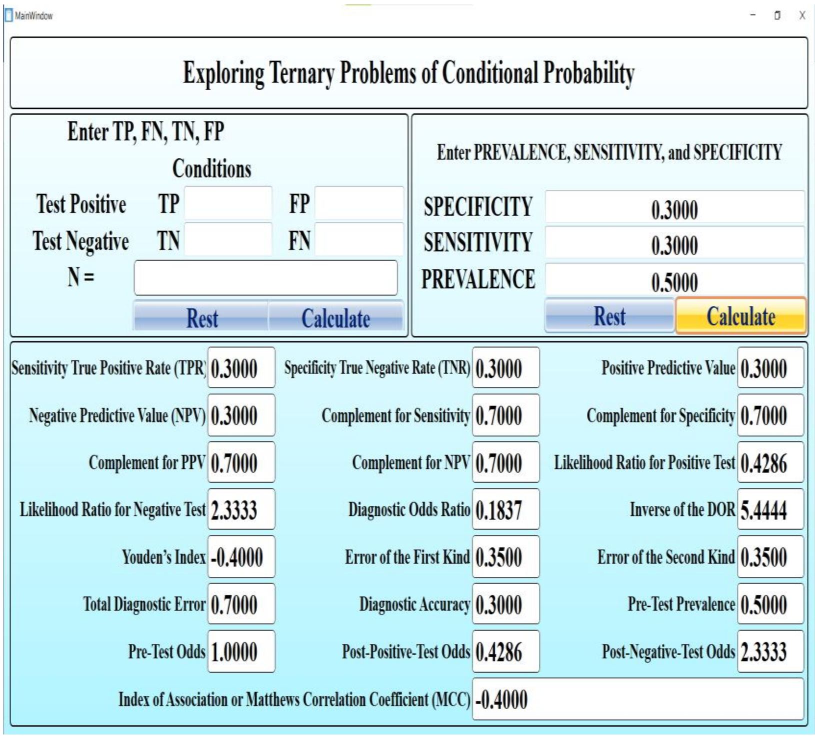

Fig. 15b. Fifteenth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case with a negative MCC and a negative informedness (Youden's index)

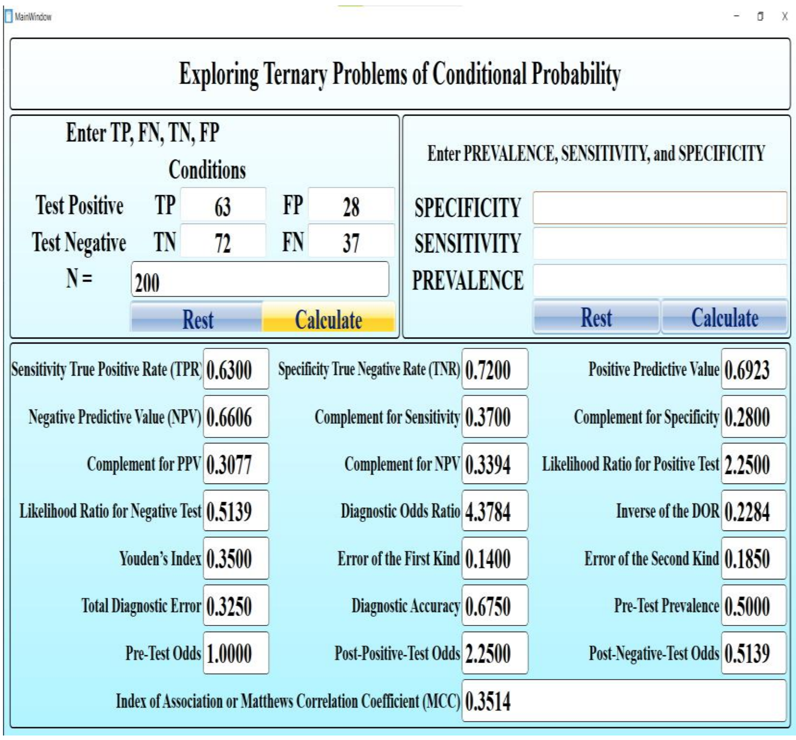

Fig. 16a. Sixteenth test case with input of contingency matrix entries. This is a case of somewhat good prediction, with sensitivity (considerably) greater than the False Positive Rate (1.0 − specificity) and diagnostic accuracy greater than 0.5, but the MCC is below 0.5

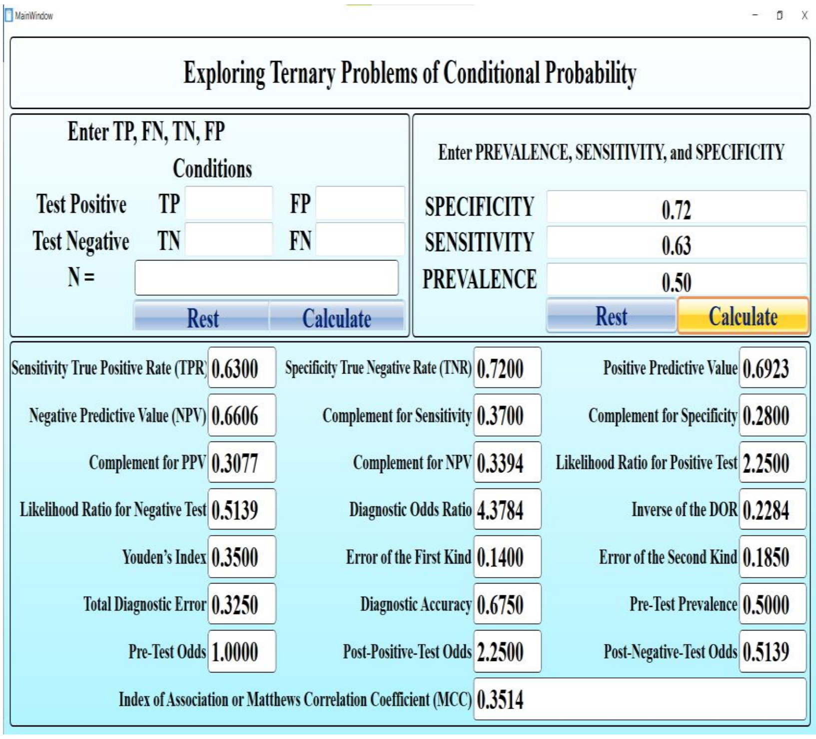

Fig. 16b. Sixteenth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case of somewhat good prediction, with sensitivity (considerably) greater than the False Positive Rate (1.0 − specificity) and diagnostic accuracy greater than 0.5, but the MCC is below 0.5.

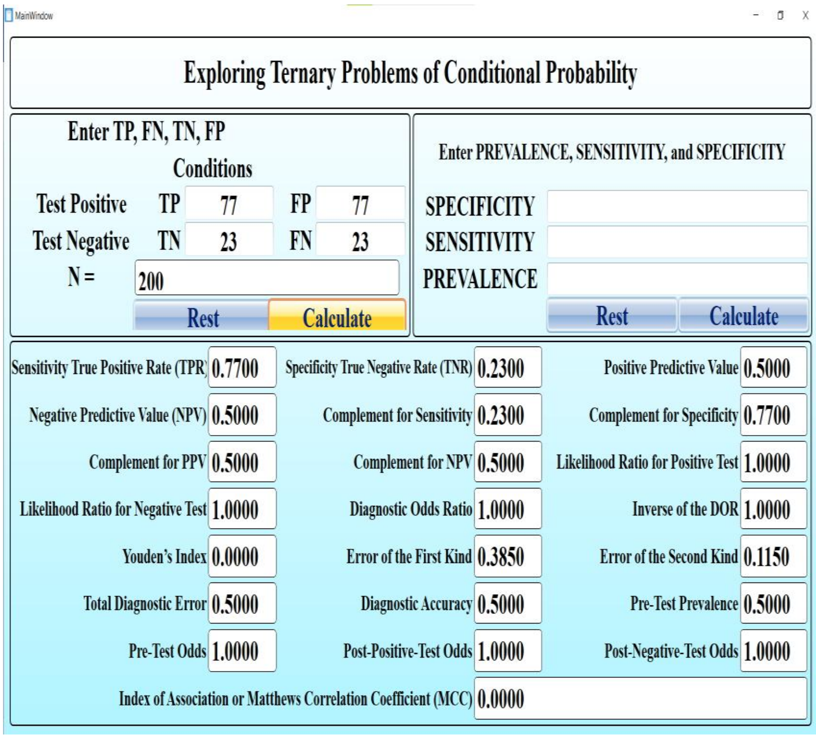

Fig. 17a. Seventeenth test case with input of contingency matrix entries. This is a case on the random guess line, with sensitivity equal to the False Positive Rate (1.0 − specificity) and diagnostic accuracy equal to 0.5, but with a zero MCC and a zero informedness (Youden's index)

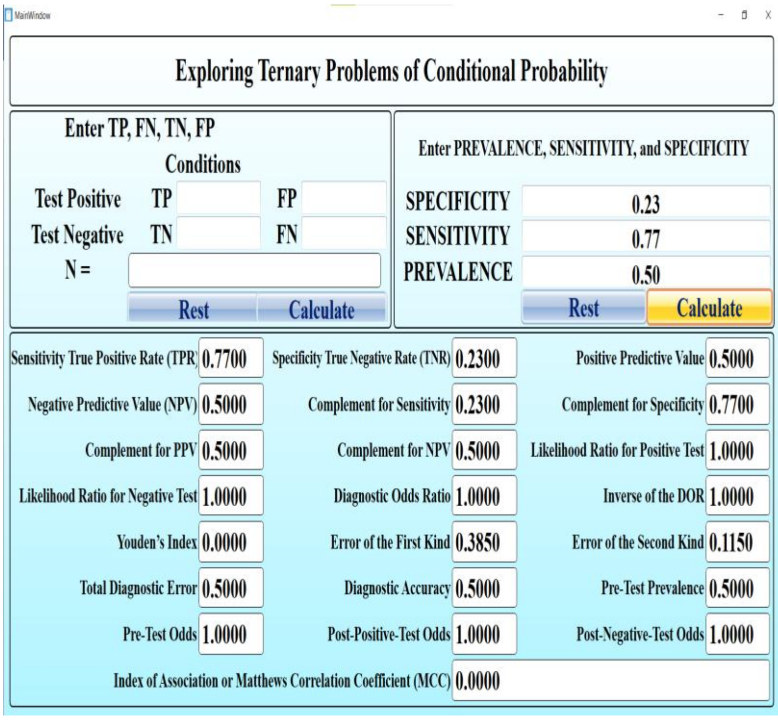

Fig. 17b. Seventeenth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case on the random guess line, with sensitivity equal to the False Positive Rate (1.0 − specificity) and diagnostic accuracy equal to 0.5, but with a zero MCC and a zero informedness (Youden's index)

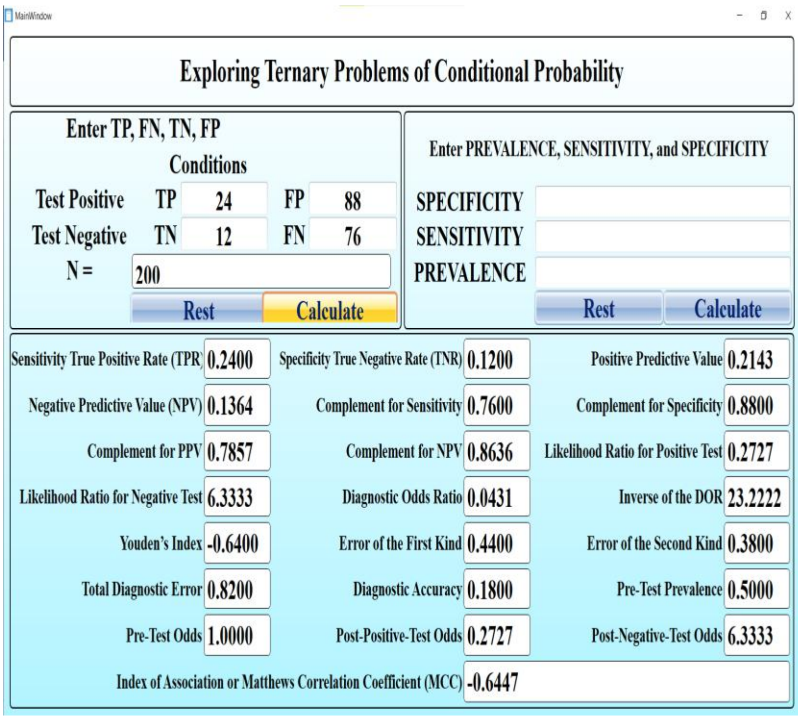

Fig. 18a. Eighteenth test case with input of contingency matrix entries . This is a case below the random guess line, with sensitivity less than the False Positive Rate (1.0 − specificity) and diagnostic accuracy less than 0.5, and with a negative MCC and a negative informedness (Youden's index)

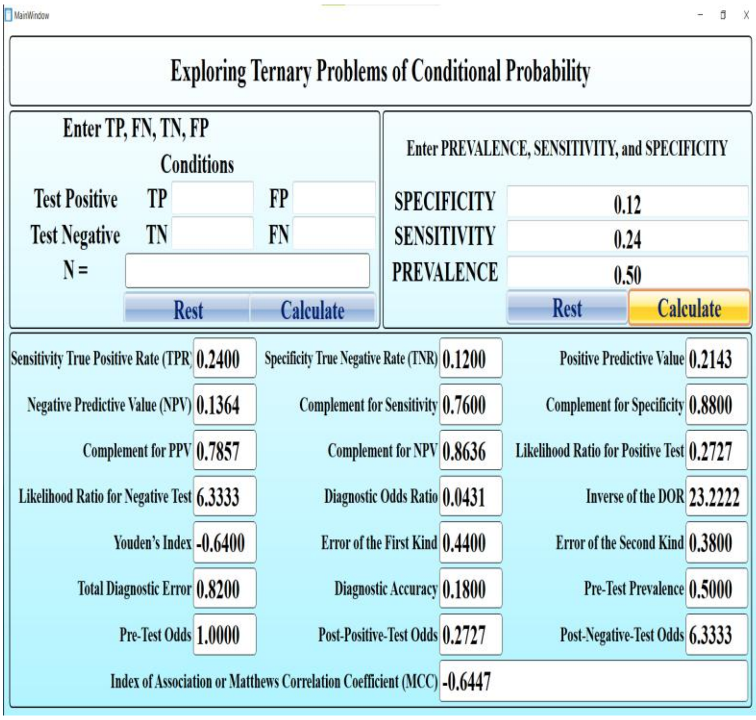

Fig. 18b. Eighteenth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case below the random guess line, with sensitivity less than the False Positive Rate (1.0 - specificity) and diagnostic accuracy less than 0.5, and with a negative MCC and a negative informedness (Youden's index)

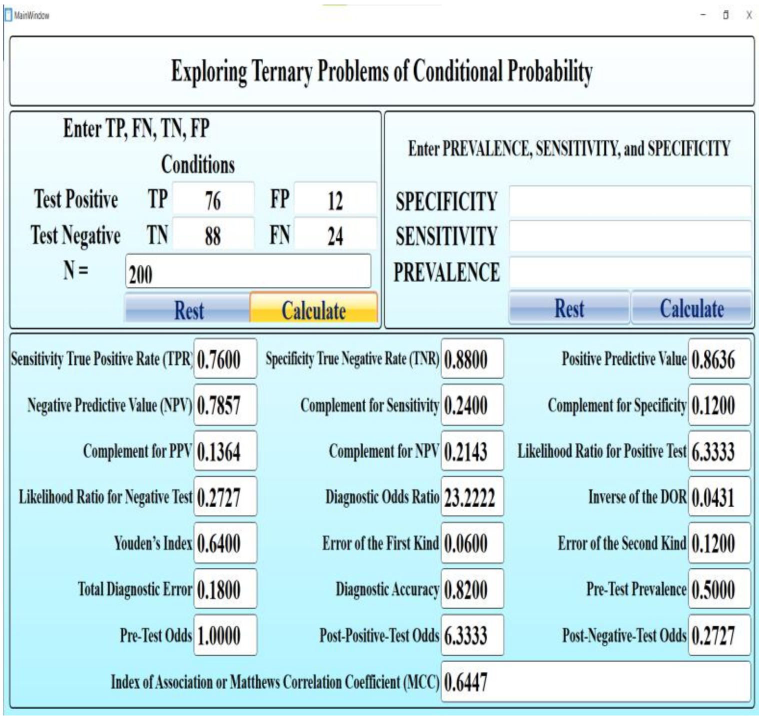

Fig. 19a. Nineteenth test case with input of contingency matrix entries. This is a case in which prediction decisions in Fig. 18 are reversed. It is a mirror image of the case in Fig. 18 with the values of sensitivity, specificity and accuracy replaced by their complements to 1.0, while MCC switched sign. This proves that the output of a consistently bad predictor could simply be inverted to obtain a good predictor

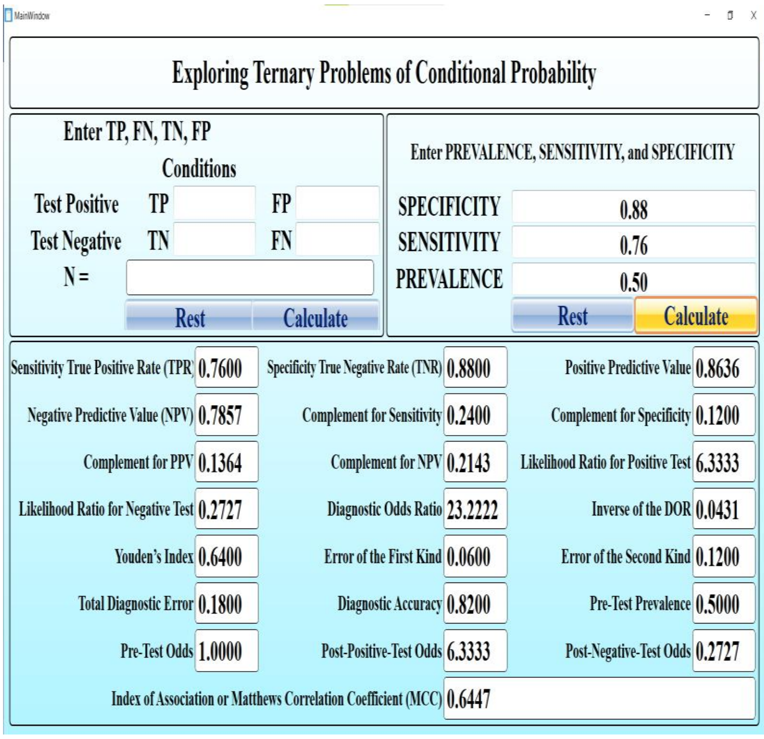

Fig. 19b. Nineteenth test case with input of pre-test prevalence, sensitivity, and specificity. This is a case in which prediction decisions in Fig. 18 are reversed. It is a mirror image of the case in Fig. 18 with the values of sensitivity, specificity and accuracy replaced by their complements to 1.0, while MCC switched sign. This proves that the output of a consistently bad predictor could simply be inverted to obtain a good predictor

Table 4. Types of prediction in terms of the four basic indicators and in terms of Mathew Correlation Coefficient, borrowed from the sister paper [12]

	Direct Basic Indicators $\{\boldsymbol{Sens_{ij}}, \boldsymbol{Spec_{ij}}, \boldsymbol{PPV_{ij}}, \boldsymbol{NPV_{ij}}\}$	**Mathew Correlation Coefficient** $\boldsymbol{M}$
Perfect Prediction	$Sens_{ij} + Spec_{ij} = 2.0$, $PPV_{ij} + NPV_{ij} = 2.0$, $Sens_{ij} = Spec_{ij} = PPV_{ij} = NPV_{ij} = 1.0$	$M = +1.0$
Good Prediction	$1.0 < Sens_{ij} + Spec_{ij} \leq 2.0$, $1.0 < PPV_{ij} + NPV_{ij} \leq 2.0$,	$0.0 < M \leq 1.0$
Random-Guessing-Like Prediction	$Sens_{ij} + Spec_{ij} = 1.0$, $PPV_{ij} + NPV_{ij} = 1.0$,	$M = 0.0$
Bad Prediction	$0.0 \leq Sens_{ij} + Spec_{ij} < 1.0$, $0.0 \leq PPV_{ij} + NPV_{ij} < 1.0$,	$-1.0 \leq M < 0.0$
Completely-contradictory Prediction	$Sens_{ij} + Spec_{ij} = 0.0$, $PPV_{ij} + NPV_{ij} = 0.0$, $Sens_{ij} = Spec_{ij} = PPV_{ij} = NPV_{ij} = 0.0$	$M = -1.0$

4. 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. The paper presented a potpourri of test cases to reveal and unravel many of the properties and inter-relationships among these indicators. The tests serve as further verification of the utility of the Matthews Correlation Coefficient as the most informative single metric that can be derived from the contingency table.

ACKNOWLEDGEMENT

The first-named author (AMAR) is grateful to Dr. Muzainah Ali Rushdi (Kasr Al-Ainy Faculty of Medicine, Cairo University, Egypt) for fruitful discussions on diagnostic testing in medical context. Material in this paper is based on her paper [12], and our present Table 1 and Table 4 are borrowed from [12]. The authors are also sincerely grateful and greatly indebted to Engineer Hamza Serag (Ph. D. candidate at KAU, Jeddah, Saudi Arabia) for computing the results of Table 3 using techniques in [8-10].

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Rushdi AMA, Talmees FA. Computations of the eight basic measures in diagnostic testing. Chapter 6 in Advances in Mathematics and Computer Science. Book Publishers International, Hooghly, West Bengal, India: 2019;2:66-87.
6. 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, Vol. 1, Book Publishers International, Hooghly, West Bengal, India. 2019;1:106-132.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Rushdi MA, Rushdi AM. Measures, metrics and indicators derived from the ubiquitous two-by-two contingency table, Part A: Background. Asian Journal of Medical Principles and Clinical Practice; 2021.
13. Chicco D. Ten quick tips for machine learning in computational biology. BioData Mining. 2017;10(1):1-7.
14. 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.
15. 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.
16. 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.
17. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping doctors and patients make sense of health statistics. Psychological Science in the Public Interest. 2007;8(2):53-96.
18. Glaros AG, Kline RB. Understanding the accuracy of tests with cutting scores: The sensitivity, specificity, and predictive value model. Journal of Clinical Psychology. 1988;44(6):1013-1023.
19. Gaddis GM, Gaddis ML. Introduction to biostatistics: part 3, sensitivity, specificity, predictive value, and hypothesis testing. Annals of Emergency Medicine. 1990; 19(5):591-597.
20. Buckland M, Gey F. The relationship between recall and precision. Journal of the American Society for Information Science. 1994;45(1):12-19.
21. Brenner H, Gefeller OL. Variation of sensitivity, specificity, likelihood ratios and predictive values with disease prevalence. Statistics in Medicine. 1997;16(9):981-891.
22. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8(1):12-17.
23. Chu K. An introduction to sensitivity, specificity, predictive values and likelihood ratios. Emergency Medicine. 1999; 11(3):175-181.
24. Grimes DA, Schulz KF. Uses and abuses of screening tests. The Lancet. 2002; 359(9309):881-884.
25. 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 (pp. 345-359). Springer, Berlin, Heidelberg.
26. Akobeng AK. Understanding diagnostic tests 1: sensitivity, specificity and predictive values. Acta Paediatrica. 2007; 96(3):338-341.
27. Parikh R, Mathai A, Parikh S, Sekhar GC, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian Journal of Ophthalmology. 2008; 56(1):45-50.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/68420