A New Graphical Method for Reporting Performance Results of a Diagnostic Test

John Wang

Philips Healthcare, Andover, MA, USA

Abstract

Reporting diagnostic performance results using standard performance measures, such as: sensitivity, specificity, and predictive values, have been a standard practice for decades. Issues with reporting using only numerical values include: 1) often only a subset of performance measures are reported, thus results could be misinterpreted and misused, 2) difficult to visualize the complex relationships of the reported performance measures. To overcome these shortcomings, a graphical presentation has been developed to further improve the test results reporting.

The 2-dimentional performance graph uses line segment, area, ratio of line segments, ratio of areas, and sum of areas to represent most of the commonly used performance measures in this single graph.

Advantages of the new graphic presentation are: 1) large number of performance measures can be presented and visualized simultaneously in a single graph, 2) allow the complex relationships of all performance measures to be understood more easily, 3) reduce the need to memorize some of the complex formulas for computing the performance measures, and 4) a great teaching tool in explaining the relationships of the commonly used performance measures.

1. Introduction

Diagnostic tests are routinely performed to: 1) screen for disease, 2) establish or rule out a diagnosis, and 3) track and monitor disease progression and effectiveness of treatment. Thus, interpretation of diagnostic test results is critical in supporting clinical decisions for the most effective patient management. Reporting diagnostic test performance results using standard performance measures, such as: sensitivity, specificity, and predictive values, have been a standard practice for decades [1-7]. Issues with reporting using only numerical values include: only a subset of performance measures are reported and it is difficult to visualize the complex relationships of the reported performance measures. Numerous graphical presentations have been developed and used in many other applications to assist in visualizing the results of numerical data. Examples of commonly used graphical presentations include Venn diagram, pie chart, histogram, scatter plot, Bland-Altman plot, and ROC curve. To further improve the performance reporting of the results of a diagnostic test, a graphical presentation has been developed.

The 2-dimentional performance summary graph allows easy visualizations of most of the commonly used performance measures using line segment, area, ratio of line segments, ratio of areas, and sum of areas.

.

2. Performance measures

The standard statistical performance measures used in reporting test results and their definitions are summarized in Table 1. As shown in the table, the efficacy of a test is entirely captured by the following four basic measurements: true positive (TP), false negative (FN), false positive (FP), and true negative (TN) (presented in a 2x2 contingency sub-table). From these four basic measurements, all the other relevant statistical measures can then be derived.

Sensitivity (Se) indicates the ability of a test to identify positive cases; a test with high sensitivity has few false negatives (incorrectly identify a patients as not having a disease). Specificity (Sp) indicates the ability of a test to identify negative cases; a test with high specificity has few false positives (incorrectly identify a patient as having a disease). Positive predictive value (PPV) provides the probability of being true positive when the test is positive. Negative predictive value (NPV) provides the probability of being true negative when the test is negative.

Positive likelihood ratio (LR+) and negative likelihood ratio (LR–), which combine both the sensitivity and specificity of the test, provide estimates of how much the result of a test will change the odds of being positive and negative, respectively. Finally, overall accuracy (ACC) is a single-valued performance measure calculated as the ratio of all the correct classification (both TP and TN) to the total test cases.

 ISSN: 2325-887X DOI: 10.22489/CinC.2019.409

Table 1. Summary of statistical performance measures and their definitions used in reporting test results.

Test Classification	Reference Classification		Total	Performance Measures
	Positive	Negative		
Positive	True Positive **(TP)**	False Positive **(FP)**	All Positive Test Cases **(TP + FP)**	Positive Predictive Value (PPV) **TP / (TP + FP)**
Negative	False Negative **(FN)**	True Negative **(TN)**	All Negative Test Cases **(FN + TN)**	Negative Predictive Value (NPV) **TN / (FN + TN)**
Total	All Positive Cases **(TP + FN)**	All Negative Cases **(FP + TN)**	All Test Cases **(N = TP+FN+FP+TN)**	Overall Accuracy (ACC) **(TP+TN) / (TP+FN+FP+TN)**
			Prevalence **(TP+FN) / (TP+FN+FP+TN)**	
Performance Measures	Sensitivity (Se) **TP / (TP + FN)**	False Positive Rate = 1 – Specificity **FP / (FP + TN)**	Positive Likelihood Ratio (LR+) **Sensitivity / (1 – Specificity)**	
	False Negative Rate = 1 – Sensitivity **FN / (TP + FN)**	Specificity (Sp) **TN / (FP + TN)**	Negative Likelihood Ratio (LR–) **(1 – Sensitivity) / Specificity**	

Prevalence-dependent performance measures

Unlike sensitivity and specificity, which are independent of the prevalence of the condition being tested, the other three performance measures PPV, NPV, and ACC depend on the prevalence. Because of this prevalence-dependent nature of these measures, it is very important to understand the impact of prevalence when using these performance measures in reporting and interpreting the test results.

3. New graphical presentation of performance test results

The new performance summary graph is shown in Fig. 1 with a test case of Se = 60%, Sp = 80%, and prevalence (PV) = 50%. The axes of the 2-dimensional square graph are: 1) Sensitivity - left vertical axis with a scale of 0 to 1 from top to bottom, 2) Specificity - right vertical axis with a scale of 1 to 0 from top to bottom, and 3) Prevalence - horizontal axis with a scale of 0 to 100% from left to right. The prevalence (50%) is plotted as a vertical line from top to bottom; the sensitivity (0.6) and specificity (0.8) are plotted as horizontal lines from the corresponding vertical axes to the vertical prevalence line.

The four areas, UL (upper-left), UR (upper-right), LL (lower-left), and LR (lower-right), formed by the prevalence, Se, and Sp lines are the normalized true positive (TPn), false positive (FPn), false negative (FNn), and true negative (TNn) respectively as shown below:

$$UL = Se \times Prevalence \times 100 = \frac{TP}{TP+FN} \times \frac{TP+FN}{N} \times 100 = \frac{TP}{N} \times 100 = TP_n \quad (1)$$

$$UR = (1-Sp) \times (1-Prevalence) \times 100 = \frac{FP}{FP+TN} \times \frac{FP+TN}{N} \times 100 = \frac{FP}{N} \times 100 = FP_n \quad (2)$$

$$LL = (1-Se) \times Prevalence \times 100 = \frac{FN}{TP+FN} \times \frac{TP+FN}{N} \times 100 = \frac{FN}{N} \times 100 = FN_n \quad (3)$$

$$LR = Sp \times (1-Prevalence) \times 100 = \frac{TN}{FP+TN} \times \frac{FP+TN}{N} \times 100 = \frac{TN}{N} \times 100 = TN_n \quad (4)$$

$$Where:\ TP_n + FP_n + FN_n + TN_n = 100$$

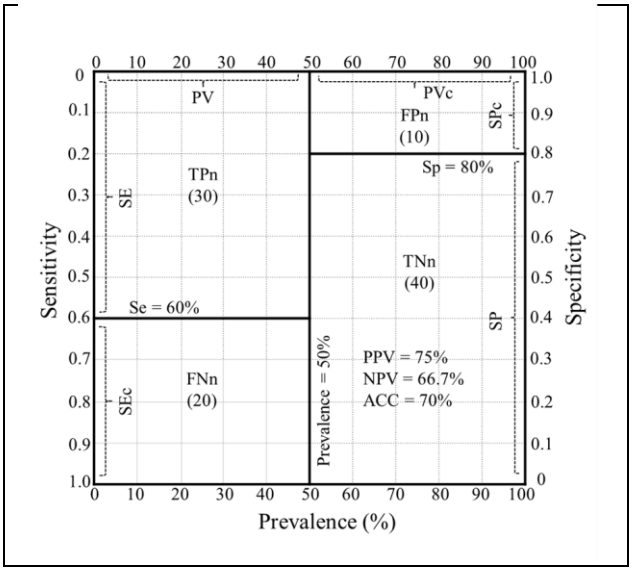

Figure 1. New summary graph for reporting diagnostic test performance results.

From Eqs. (1) – (4) above, performance measures PPV, NPV, and ACC can then be expressed in terms of the four areas as shown in the performance graph Fig. 1:

$$PPV = \frac{TP}{TP+FP} = \frac{TP_n}{TP_n+FP_n} = \frac{UL}{UL+UR} \quad (5)$$

$$NPV = \frac{TN}{FN+TN} = \frac{TN_n}{FN_n+TN_n} = \frac{LR}{LL+LR} \quad (6)$$

$$ACC = \frac{TP+TN}{N} = \frac{TP_n+TN_n}{100} = (UL+LR)/100 \quad (7)$$

$$Where:\ N = TP + FN + FP + TN$$

Two examples are shown below to illustrate the usefulness of the graphical performance plot as described in Fig. 1.

Example 1 - Relationship between PPV and prevalence (See Fig. 2). In this example both Se and Sp remain unchanged at 50%. Graphically, PPV is the ratio of the upper-left area to the sum of the upper-left and upper-right areas. Thus PPV equals to 50% when prevalence is at 50% level because the upper-left and upper-right areas are the same. However, PPV will increase (from 50% to 70%, Fig. 2a) when prevalence is increased because now the upper-left area becomes larger than the upper-right area. On the other hand, PPV will decrease (from 50% to 30%, Fig. 2b) when the prevalence is decreased because now the upper-left area becomes smaller than the upper-right area. These relationships remain the same whether Se > Sp or Sp > Se, because moving the vertical prevalence line only changes the width but not the height of the two areas used in determining the PPV result. Thus, using the graphical performance graph, it is very easy to show that increase prevalence will always improve PPV regardless of the values of Se and Sp.

Example 2 - How to improve PPV? Fig. 3 shows that three lines can be moved to increase TPn (upper-left area) and/or decrease FPn (upper-right area) in order to improve the value of PPV. Which strategy is the best given a possible change of 20% in total? The performance evaluation is done for three strategies as shown in Fig. 3. The three strategies are: 1) increase the upper-left area and decrease the upper-right area by increasing prevalence by 20% (Fig. 3a), 2) increase the upper-left area by increasing the sensitivity by 20% (Fig. 3b), 3) decrease the upper-right area by increasing the specificity by 20% (Fig. 3c).

Instead of improving the performance of the diagnostic algorithm, strategy #1 improves PPV performance by selecting test subjects with higher pretest likelihood. Very often, in many applications this is the most effective strategy in improving PPV since the width of the upper-left area is increased and the width of the upper-right area is decreased simultaneously by increasing the prevalence. Note also that in this case LR+ = 1 (Se = 1-Sp), as such the diagnostic test itself does not improve PPV, rather the improved PPV is a direct result of the higher pretest likelihood (prevalence). For the other two strategies, as expected, a higher LR+ value will result in a higher PPV. Thus, the better strategy is to increase Sp by 20% since it has a higher LR+ value (Fig. 2c). Note also that the ACC values are the same for the two strategies because the sum of the upper-left and lower-right areas remains the same regardless how the 20% performance gain is allocated.

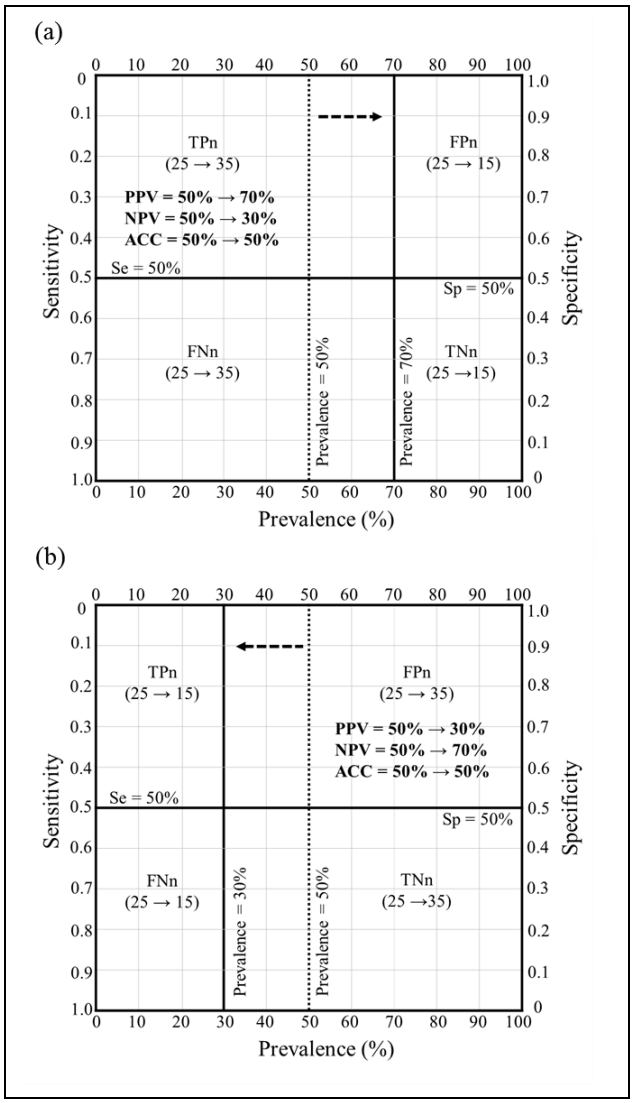

Figure 2. Relationship of PPV and Prevalence: a) PPV increases with higher prevalence value, b) PPV decreases with lower prevalence value.

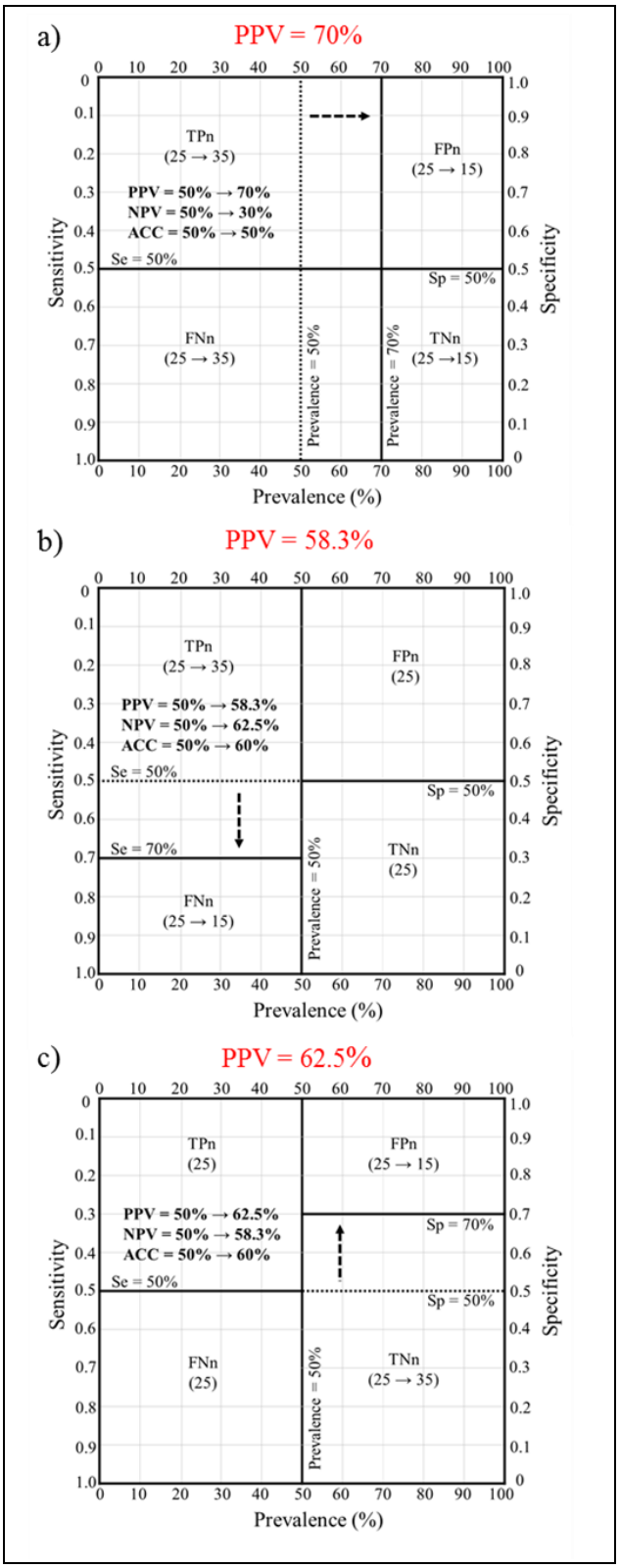

Figure 3. Comparison of strategies for improving PPV.

4. Summary and Conclusions

A new graphical presentation is described for reporting test performance results. The graph provides a complete performance presentation since all the relevant performance measures are included. The ability to visualize the performance measures directly allows the complex relationships of these performance measures to be understood more easily. This ability also makes the graph a highly useful tool in better understanding of the meaning of these performance measures. The graphical presentation also reduces the need to memorize some of the difficult equations for calculating the performance measures. Using the normalized values for TP, FP, FN, and TN also allows easier calculation in impact assessment for different parameter values. Because the graph is easy to use and it provides a complete performance reporting, it should be considered as a standard performance reporting tool so that reporting can be standardized and meaningful performance comparison can be performed.

References

[1] Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BJM Open* 2016;6:e012799. Doi: 10.1136/bmjopen-2016-012799.

[2] Scherokman B. Selecting and interpreting diagnostic tests. *The Permanente Journal;* Fall 1997, Vol.1, No.2, pp 4-7.

[3] Alberg AJ, Park JW, Hager, BW, et al. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J Gen Intern Med* 2004;19:460-5.

[4] Simundic A. Measures of diagnostic accuracy: Basic definitions. *EJIFCC* 2009;19(4):203-211.

[5] Oken UM, Okoro CN. Evaluating measures of indicators of diagnostic test performance: Fundamental meanings and formulars. *J Biomet Biostat* 2012,3:132. Doi: 10.4172/2155-6180.1000132.

[6] Wang, J. Proposed new requirements for testing and reporting performance results of arrhythmia detection algorithms. *Computing in Cardiology* 2013;40:967-970.

[7] Wang J. A review of basic statistical concepts in clinical test interpretation and de ision support. *Computing in Cardiology* 2017;44. Doi: 10.22489/CinC.2017.088-132.

Address for correspondence:

John Wang
Philips Healthcare, MS-4308
3000 Minuteman Road
Andover, MA 01810-1099, USA
E-mail: john.j.wang@philips.co