A Review of the Basics Statistical Concepts Used in Clinical Tests Interpretation and Decision Support

John Wang

Philips Healthcare, Andover, MA, USA

Abstract

Numerous diagnostic tests are routinely performed clinically 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.

The basic performance statistics required in the test results interpretation, including sensitivity (Se), specificity (Sp), overall accuracy (ACC), pretest likelihood (prevalence), and posttest likelihood including positive predictive value (PPV) and negative predictive value (NPV) are reviewed. The definitions and formulas of these performance measures and their relationships are summarized. Unlike Se and Sp, which are independent of the prevalence of the condition being tested, the other three performance measures PPV, NPV, and ACC depend on the disease prevalence. It is very important to understand the impact of disease prevalence when using these performance measures in reporting and interpreting the test results.

The problem (also known as "accuracy paradox") of using the single performance measure ACC to characterize the test performance is described and explained. Despite its simplicity and frequent use in the published literatures for performance reporting, it is shown that the overall accuracy is not a reliable performance measure and its use should be avoided.

The most important relationship of PPV as a function of the pretest likelihood and the accuracy of the test method (specified by Se and Sp) is clearly described and explained. With a clear understanding of the relationship between the pretest likelihood and PPV, discussion of several topics are presented to show how these basic statistical concepts can be applied in a variety of situations for effective decision support and patient management.

1. Introduction

Numerous diagnostic tests are routinely performed clinically 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.

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 true positive and true negative) to the total test cases.

For a complete description of the performance of a test, both sensitivity and specificity need to be reported. In addition, for most applications, both positive predictive value and negative predictive value are often included as part of the test performance reporting.

ISSN: 2325-887X DOI:10.22489/CinC.2017.256-064

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 **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**	

Because once a test is positive, one is interested to know the predictive value of the test, namely, the likelihood (or probability) that the positive test is indeed a positive case. In addition, overall accuracy is often used to report the performance results because it is a single-valued performance measure that incorporates both sensitivity and specificity. In addition, overall accuracy is often used to report the performance results because it is a single-valued performance measure that incorporates both sensitivity and specificity.

3. 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 disease prevalence. Because of this prevalence-dependent nature of these measures, it is very important to understand the impact of disease prevalence when using these performance measures in reporting and interpreting the test results. In the following, both ACC and PPV will be discussed in detail.

3.1. Overall Accuracy

Overall accuracy (ACC), defined as the ratio of all correct classification (true positive and true negative) to the total test cases, can be expressed in terms of sensitivity, specificity, and prevalence from the following equation:

$$ACC = \frac{TP+TN}{N} = \frac{TP}{N} + \frac{TN}{N}$$

$$= \frac{TP}{TP+FN} \times \frac{TP+FN}{N} + \frac{TN}{FP+TN} \times \frac{FP+TN}{N}$$

$$= \; Se \times Prevalence + Sp \times (1 - Prevalence)$$

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

This equation shows that ACC is a weighted sum of the sensitivity and specificity of the test. The weighting factor for the sensitivity is the prevalence and the weighting factor for the specificity is the complement of the prevalence [1]. Thus, for an algorithm with performance specified by sensitivity and specificity, ACC will vary depending on the disease prevalence of the test population.

Table 2. Overall accuracy (ACC) as a function of the prevalence level.

Prevalence	Overall Accuracy		

	Algorithm A
	Se = 60%
	Sp = 40%
		Algorithm B
		Se = 40%
		Sp = 60%
			Algorithm C
			Se = 60%
			Sp = 60%

90%	58%	42%	60%
50%	50%	50%	60%
10%	42%	60%	60%

To illustrate this relationship, Table 2 shows the ACC values for three different algorithms tested at three different levels of prevalence. When the prevalence is greater than 50%, ACC will be higher for an algorithm with higher sensitivity than specificity. On the other hand, when the prevalence is less than 50%, ACC will be higher for an algorithm with higher specificity than sensitivity. When sensitivity and specificity are the same, regardless of the prevalence level, ACC will be the same as the sensitivity and specificity. When the prevalence level is 50%, ACC will equal to the mathematical average of the sensitivity and specificity.

Accuracy paradox

Because of this prevalence-dependent nature of the overall accuracy, it has created the so-called "accuracy paradox", where a test with a lower positive predictive value may actually have a higher overall accuracy and thus distorted the test validity. In addition, a useless test may have a higher overall accuracy than a more useful test with a lower value of overall accuracy. The accuracy paradox is illustrated in two examples as shown in Tables 3 and 4.

Table 3. Summary of test performance results for two different algorithms at different levels of prevalence.

Algorithm Test #1		D+ (50)	D– (950)	Prevalence (5%)
A	Test +	20	38	PPV = 35%
	Test –	30	912	
	Results	Se = 40%	Sp = 96%	ACC = 93.2%
Algorithm Test #2		D+ (250)	D– (750)	Prevalence (25%)
B	Test +	150	15	PPV = 91%
	Test –	100	735	
	Results	Se = 60%	Sp = 98%	ACC = 88.5%

In Table 3, algorithms A and B are tested using datasets with prevalence of 5% and 25%, respectively. From the test results shown in Table 3, it is clear that although algorithm B has better Se, SP, and PPV than algorithm A, yet algorithm B shows a higher ACC value (93.2% vs. 88.5%). This example shows that the ACC performance measure does not accurately represent the validity of the tests.

In Table 4, algorithms A and B are tested using a dataset with 10,000 cases and 5% prevalence. Algorithm A has Se, Sp, and PPV equal to 90%, 90%, and 32% respectively. Algorithm B, on the other hand, is a totally useless algorithm. The algorithm call all cases positive (Sp = 100%) and fails to detect any positive cases (Se = 0%). However, algorithm B has a higher ACC (95%) than algorithm A (90%).

Table 4. Summary of test performance results for two different classification algorithms A and B.

Algorithms Tested		D+ (500)	D– (9,500)	Prevalence (5%)
A	Test +	450	950	PPV = 32%
	Test –	50	8,550	
	Results	Se = 90%	Sp = 90%	Acc = 90%
B	Test +	0	0	PPV = N/A
	Test –	500	9,500	
	Results	Se = 0%	Sp = 100%	Acc = 95%

Although, it is intuitively reasonable to assume that the overall accuracy should be a very useful single-valued performance measure, these two examples clearly show that the overall accuracy is not a reliable performance measure, its use should be avoided.

3.2. Positive predictive value

Given the performance of a test as specified by the sensitivity and specificity, the PPV as a function of the prevalence can be calculated using the following equation:

$$PPV = \frac{[Se/(1-Sp)] \times [Prevalence/(1-Prevalence)]}{1 + [Se/(1-Sp)] \times [Prevalence/(1-Prevalence)]}$$

As an example to show the relationship of prevalence and PPV, four separate tests, each with 10,000 test cases, are conducted. The TP, FN, FP and TN numbers for all four tests are shown in Table 5. From these numbers, the sensitivity, specificity, PPV, and prevalence are calculated and summarized in Table 6. The results show that while all tests have the same sensitivity (95%) and specificity (95%), the PPVs for these four tests are very different. The PPVs are 95%, 68%, 16%, and 2% for prevalence levels of 50%, 10%, 1%, and 0.1%, respectively. The PPV depends on the pretest prevalence. A low prevalence yields a low PPV. The first test with a prevalence level of 50% produces a high PPV of 95%. On the other hand the fourth test with a prevalence of 0.1% produces a very low PPV value of only 2%, which is not a very useful test.

Since the prevalence can significantly impact the test performance results it is very important to know approximately the prevalence in order to interpreting the test results [2]. While for many applications, one can potentially achieve a higher PPV by selecting only cases to be tested with high pretest likelihood (prevalence). However, for some applications, for example real-time

Table 5. Summary of test results for an example of understanding the relationship between positive predictive value and prevalence. D+ = Disease positive, D– = Disease negative

Test Results	Test #1 (100,000 Cases; Prevalence = 50%)		Test #2 (100,000 Cases; Prevalence = 10%)		Test #3 (100,000 Cases; Prevalence = 1%)		Test #4 (100,000 Cases; Prevalence = 0.1%)	
	D+ (50,000)	D– (50,000)	D+ (10,000)	D– (90,000)	D+ (1,000)	D– (99,000)	D+ (100)	D– (99,900)
Test positive	47,500	2,500	9,500	4,500	950	4,950	95	4,995
Test negative	2,500	47,500	500	85,500	50	94,050	5	94,905

arrhythmia monitoring this may not be an option, since very often all patients are monitored regardless whether they are likely to have arrhythmia or not [3].

It is also important to know that while database used in the testing of a diagnostic algorithm must contain sufficient number of positive cases (thus, high prevalence) in order to accurately measure the sensitivity of the algorithm being evaluated, the performance results thus obtained may not be clinically relevant since the prevalence of the targeted patient population for the test may have a very different prevalence level (usually much lower). Thus, it is very important that the post-test likelihood needs to be reported using the actual prevalence level of the patient population the test is targeted for.

Table 6. Summary of test performance calculated from the results provided in Table 5.

Performance Results	Test #1	Test #2	Test #3	Test #4
Sensitivity	95%	95%	95%	95%
Specificity	95%	95%	95%	95%
Positive Predictive Value	95%	68%	16%	2%
Prevalence	50%	10%	1%	0.1%

PPV is also a very important performance measure when trying to compare an algorithm with higher sensitivity but lower specificity to an algorithm with lower sensitivity but higher specificity. An example is shown in Table 7. It is shown that even with a large increase of 30% on sensitivity (from 40% to 70%) PPV will drop by 11% (from 60% to 49%) with a small 5% drop on specificity (from 97% to 92%) at low prevalence level of 10%. The drop of PPV is less significant at 3% when the prevalence level is higher at 50%. Thus, a risk and benefit analysis is needed in order to make the proper trade-off in determining whether such an improvement is indeed clinically beneficial.

Table 7. Using PPV for performance trade-off evaluation

Se	Sp	PPV	
		Prevalence 10%	Prevalence 50%
40%	97%	60%	93%
70%	92%	49%	90%

4. Conclusion

It is important to understand the dependency of the overall accuracy and positive predictive value on the disease prevalence of the test patient population. It is recommended that diagnostic performance be specified by sensitivity and specificity but not by the single-valued overall accuracy performance measure. It is also recommended to report clinical-relevant positive predictive value based on the disease prevalence of the targeted test patient population instead of the value obtained from the test dataset with high disease prevalence.

References

[1] 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.

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

[3] Wang, J. Proposed new Requirements for Testing and Reporting Performance Results of Arrhythmia Detection Algorithms. *Computing in Cardiology* 2013;40:967-970.

Address for correspondence:

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