STATISTICS FOR RESEARCHERS

Statistical Evaluation of Diagnostic Tests (Part 1): Sensitivity, Specificity, Positive and Negative Predictive Values

NJ Gogtay, UM Thatte

Introduction to Screening and Diagnostic Tests

In clinical practice, two broad types of tests are used- screening and diagnostic tests. Screening tests are those that are used on a large population [usually healthy individuals or patients who are yet asymptomatic] to identify those likely to need intervention or identify disease early. Examples of screening tests would include routine blood pressure monitoring for diagnosing hypertension, a Pap smear for early diagnosis of cervical cancer, Prostate Specific Antigen [PSA] estimation for detection of prostate cancer or a mammogram for early detection of breast cancer. In real life, these are usually done, for example, for purposes of obtaining an insurance or a routine health checkup or as part of evidence based health policy recommendations in a given population. Screening tests should be easy to use, relatively inexpensive and ensure that they do not miss patients with disease, nor misclassify those without.

The second type of test is the diagnostic test. This is an aid to clinical decision-making, done on patients who are symptomatic and is usually more expensive and can carry more risks than screening tests [a trans-rectal biopsy for confirmation of prostate cancer for instance carries greater risk than a screening blood test for the PSA].[1] Diagnostic tests are also done after a positive screening test to establish a definitive diagnosis and are often called confirmatory tests.

The Challenge of Interpreting Screening and Diagnostic Tests

The interpretation of screening and diagnostic tests can be challenging. For instance, for the diagnosis of malaria, the peripheral smear still remains the "gold standard" or the "reference standard" test for identifying the malarial parasite. For a patient who presents with fever, a physician thinks of the probability of malaria and orders the peripheral smear. The test result is binary - either positive or negative. The easiest approach for a clinician would be to simply classify the patient into one of two groups- "the test is positive and hence the patient has the disease and thus I should treat him for malaria" OR that "the test is negative and hence the patient does not have malaria and so I need to look for alternate diagnoses". But does this really happen in the clinical setting? The answer is maybe not! A physician may still prescribe anti-malarials to a patient who is smear-negative, because the signs and symptoms are classically that of malaria or because the patient has had a past episode of malaria. Similarly, a clinician may ask for a complete blood count along with ESR in a patient with evening rise of temperature and cough for over 3 weeks [wherein he suspects tuberculosis], but would never treat the patient for tuberculosis simply based on the results of an elevated ESR.

Thus, the question that the clinician is really trying to answer is "Given a test result [positive or negative], what is the probability that the patient has [or does not] have the disease?' From the point of view of the patient, his/her thoughts are likely to be a) the test is negative so should I be reassured or continue to worry? b) the test is positive- should I worry or simply ignore it?

Diagnostic and screening tests, thus, should be used *correctly and interpreted appropriately* to make a diagnosis or aid in one. This is dependent upon the *discriminative ability* of the test, i.e., the ability to make a distinction between the two conditions of interest- health and disease.

The Discriminative Ability of a Test – Metrics of Diagnostic Accuracy

The following metrics are used to assess the diagnostic accuracy of a new test (regardless of whether it is a screening or a new diagnostic test, as defined above) [2]

- Sensitivity
- Specificity

Department of Clinical Pharmacology, Seth GS Medical College & KEM Hospital, Mumbai, Maharashtra
Received: 06.05.2017; Accepted: 10.05.2017

Table 1: A 2 x 2 table of depicting the results of a new test *vis* à *vis* a gold standard

	Disease Status as confirmed by the gold standard		
	Present	**Absent**	
Test Result Positive	True Positive [TP] a	False positive [FP] b	a+b
Test Result Negative	False Negative [FN] c	True Negative [TN] d	c+d
	a+c	b+d	

- Positive and Negative predictive values
- Likelihood ratio
- Area under the Receiver Operating Characteristic [ROC] Curve and Youden's index
- Diagnostic odds ratio

In this first article in the Diagnostic tests series, we will understand the concepts of sensitivity, specificity and positive and negative predictive values along with the mathematical formulae to compute them as also their limitations and clinical applications.

Beginning the Assessment of a New Test - The 2 X 2 Table

The assessment of any new test [also called as the index test] begins with testing in two groups of individuals - those who have the disease and those who do not. It is good to remember here that any test can return results as binary [positive or negative as seen with peripheral smear for malaria] or on a continuous scale as seen with blood sugar, serum cholesterol or plasma phenytoin levels.

The next step would be the construction of the two by two [2x2] table [Table 1]. The disease status (as assessed with the Gold Standard {see below}) is conventionally put in the top row and the test result in the first column.

Table 1 represents the four distinct possibilities that follow after a test is conducted on an individual. The test is

- Positive and the individual has the disease [TP] - a
- Positive, but the individual does not have the disease [FP] - b
- Negative and the individual does not have the disease [TN]- d
- Negative, but the individual has the disease [FN] -c

For the purpose of calculating the various metrices described above, the data is summated conventionally as follows:

- a+c gives the total number of individuals WITH the disease
- b+d is the total number of individuals WITHOUT the disease
- a+b gives the total number of positive test results
- c + d gives the total of negative test results

Evaluating and Understanding the Metrics

Sensitivity

This is defined as the probability that an individual with disease will have a positive test and represents the "true positivity rate" or TPR of the test. Simply put, it is the ability of a test to correctly classify an individual as 'diseased'.[2]

Mathematically, this would be

Those WITH the disease who test positive [a]

ALL those WITH the disease [a+c]

OR

$$\frac{TP}{TP + FN}$$

When we say that a test has a sensitivity of 90%, it means that of the 100 individuals who have the disease and are tested, the test will pick up 90/100 [90%] with the disease. The corollary would be that 10/100 [10%] would be missed [false negative].

Specificity

This is defined as the probability that a disease-free individual will have a negative test and represents the "true negativity rate" or TNR of the test. Simply put, this would be the ability of a test to correctly classify an individual as being *disease- free.*[2]

Mathematically, this would be

Those WITHOUT the disease who test negative [d]

ALL those WITHOUT the disease [b+d]

OR expressed mathematically as

$$\frac{TN}{TN + FP}$$

Thus, when we say that a test has a specificity of 90%, it means that of the 100 individuals who do not have the disease and are tested, the test would show 90/100 [90%] individuals as not having the disease. The corollary would be that 10/100 [10%] of them would be wrongly picked up as having the disease [false positive].

Sensitivity and specificity primarily address the question "How accurately does the test being evaluated discriminate between individuals with disease and without?" Both are test characteristics or test properties and are independent of the disease prevalence of the population where they are tested as we will see a little later.

Positive and Negative Predictive values

As stated earlier, what the clinician wants to know is "Given a certain test result, what is the probability of the disease?" which brings us to understanding the "predictive value" concept.

Positive predictive value (PPV) is the probability that an individual with a positive test truly has the disease. In other words, an individual has a positive test; how worried should he be? Mathematically, this would be a ratio and expressed as the proportion of all those tested who have the disease AND a positive test [a] to all those screened who return a positive test [a+b]. Thus

Table 2: Calculation of sensitivity, specificity, positive and negative predictive values of a test using the 2x2 table[3]

Positive predictive value = a/a +b

	Disease		
Test	Present	Absent	
Positive	True Positive [TP] a	False positive [FP] b	a+b
Negative	False Negative [FN] c	True Negative [TN] d	c+d

Sensitivity = a/a +c

Specificity = d/b+d

Negative predictive value = d/d +c

Table 3: Diagnosis of microfilaraemia in a village with a prevalence of 5% using a test with 90% sensitivity and 90% specificity

Test	Disease		
	Present	Absent	
Positive	True Positive [45]	False positive [95]	a+b [140]
Negative	False Negative [5]	True Negative [855]	c+d [860]
	50	950	1000

Table 4: Diagnosis of microfilaraemia in a village with a prevalence of 20% using a test with 90% sensitivity and 90% specificity

Test	Disease		
	Present	Absent	
Positive	True Positive [180]	False positive [80]	a+b [260]
Negative	False Negative [20]	True Negative [720]	c+d [740]
	200	800	1000

mathematically, *Positive predictive value* is given by a/a+b.

Negative predictive value (NPV) is the probability that individuals with a negative screening test truly do not have the disease. In other words, the individual has tested negative, so how reassured should he be? Mathematically, this would be expressed as the proportion of all those tested who DO NOT have the disease AND are negative [d] to ALL those who test negative [d/c+d].[2]

In the 2 x 2 table presented earlier, the four concepts of sensitivity, specificity, positive predictive value and negative predictive value can be easily understood with the four arrows and the direction of their movement (Table 2).

The Relationship of Predictive Values with Prevalence

Unlike sensitivity and specificity, PPV and NPV are not fixed characteristics of the test, but depend upon the prevalence of the disease.[3,4] Let us say that we testing for microfilaraemia in a population of 1000 patients in village 1 with a prevalence of 5%. This means that 50 individuals in this village have the disease, and 950 are disease free. A test used to diagnose microfilaraemia has a sensitivity of 90% and specificity of 90%. Thus, of the 50 individuals with the disease, the test would correctly identify 45 as having the disease [and miss 5] and of the 950 without the disease, the test would correctly identify 855 as not having the disease [and falsely label 95 as having the disease]. Let us now fit the 2 x 2 table with the actual values based on this information.

Microfilaraemia prevalence – 5%, Test sensitivity 90% and test specificity 90%

Positive predictive value = 45/140 or 0.32 or 32%

Thus, in this population, this test *may not* be such a good test.

In village 2, the prevalence of microfilaraemia is much higher at 20%. Then, the 2 x 2 table would look as given below for the same sensitivity and specificity

Microfilaraemia prevalence – 20%, test sensitivity 90%, test specificity 90%

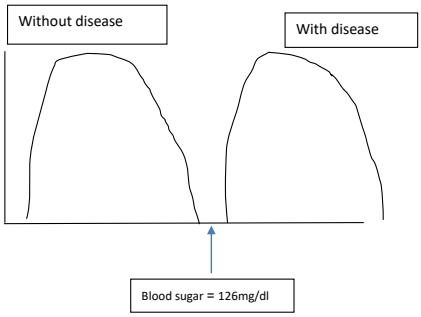

Fig. 1: An ideal test that clearly discriminates between those with disease and those without with a clear cut-off

Positive predictive value = 180/260 or 0.69 or 69%

In village 2, the same test *is a much better test*!

Thus, it is now clear that while sensitivity and specificity remain unaltered as they are properties of the test itself, the positive and negative predictive values are not fixed and vary with variation in the disease prevalence. The corollary to this statement is that because sensitivity and specificity are only test performance features and do not address the problem of prevalence in diverse populations; positive and negative predictive values become important in interpretation of test results.

The Tradeoff between Sensitivity and Specificity

When we finally choose a test, we often have to often accept a trade-off between sensitivity and specificity. Figure 1 depicts an ideal scenario where a fasting plasma glucose of 126mg/dl {based on the guidelines of the American Diabetes Association [ADA]} clearly identifies individuals with diabetes and those without.

This however rarely happens in clinical practice and let us understand this with an example. Measuring fasting blood sugar is one of the screening tests that is used to make the provisional diagnosis of diabetes. Let us say we have a pre-identified sample of

Table 3: Results with a new screening test for diabetes with 85% sensitivity and 30% specificity

Test	Disease		
	Present	Absent	
Positive	True Positive n = 17	False positive n = 14	a+b n = 31
Negative	False Negative n = 3	True Negative n = 6	c+d n = 9
	20	20	

Table 4: Results with a new screening test for diabetes with 25% sensitivity and 90% specificity

Test	Disease		
	Present	Absent	
Positive	True Positive n = 5	False positive n = 02	a+b n = 07
Negative	False Negative n = 15	True Negative n = 18	c+d n = 33
	20	20	

n = 20 patients who have diabetes and n = 20 who do not, based on a gold standard test [the oral glucose tolerance test; OGTT]. We have with us a new screening test that does not need a finger prick but rather measures fasting blood sugar by the principle of infra-red light passing through the skin and is thus noninvasive. When this test is used for measuring blood sugar in this sample of patients we find that it has a sensitivity of 85% and specificity of 30% (Table 3).

Using this test, a majority (17/20) of those who have diabetes have been picked up correctly, but there are also a large number of individuals who have been falsely labelled as being diabetic (14/20, because of the low specificity 30%).

Subsequently, we decide to use another test that is also noninvasive and uses saliva to measure glucose. This test is found to have a sensitivity of 25% and a specificity of 90% (Table 4).

As seen from Table 4, the test has labelled a majority without the disease correctly as not having the disease, but has missed a large number of patients who actually have the disease.

When do you Need a Test to be Highly Sensitive or Highly Specific?

When a test has high sensitivity, the maximum number of patients with the disease are picked up, [meaning the false negatives are very few]. Thus, *a test with high sensitivity actually rules out* the disease as a negative test in fact indicates absence of the disease. [3, 4, 5]

There are three clinical scenarios where high sensitivity is required for a test

a. When there is an important penalty for missing a patient with the disease. For example, in a blood bank, a highly sensitive test like the ELISA is needed as missing an HIV positive donor can have serious consequences for the recipient.
b. When the probability of the disease is low and the sole purpose of the test is to discover asymptomatic individuals. This is classically seen with screening for diabetes in diabetes detection "camps" where apparently normal individuals are screened *en masse*.
c. In early stages for the work up of a disease. Here a "negative" test tells the clinician that a particular disease is highly unlikely in the patient and that he should be looking at differential diagnoses.

Examples of tests with high sensitivity include a positive D-Dimer test for deep vein thrombosis [sensitivity 89%] or the positive corneal reflex [sensitivity 92%] for favorable prognosis following non-traumatic coma.[5]

Likewise, *tests with high specificity actually "rule in" the disease* as a positive test indicates that the patient has the disease in all likelihood. There are two clinical scenarios where a highly specific test is useful

a. When a false positive test can harm the patient physically or emotionally [for example declaring a person to be HIV positive or declaring the diagnosis of cancer. Here the clinician has to be absolutely sure that the patient does indeed have the disease]
b. To *rule in* a diagnosis suggested by other tests [for example a biopsy that will rule in the final diagnosis of breast or prostate cancer that has been suggested by a mammogram or PSA test]

The serum ferritin test [90% specificity] for iron deficiency anemia is an example of a test with high specificity.[5]

How Varying the Cut-Off Points can Impact Sensitivity and Specificity

Sensitivity and specificity are inversely proportional, meaning that as the sensitivity increases, the specificity decreases and *vice versa*. Let us understand this with an example of Prostate specific antigen [PSA] in the diagnosis of prostate cancer. Worldwide, most studies have used a PSA cut off of 4ng/ml for the diagnosis of the disease [i.e., those below are likely not have the disease, and those above likely to].[6] At this cut off, the sensitivity of the test is approximately 20% and specificity 90%. Table 5 depicts this information for 50 patients with proven prostate cancer and 50 individuals who are disease free.

When the PSA cut off is lowered from 4ng/ml to 3 ng/ml, logically, more patients with the disease will be picked, but more individuals without disease will now be labelled as having the disease. Table 6 depicts this information.

The lowering of the PSA cut off from 4ng/ml to 3 ng/ml has resulted in the following: an increase in

Table 5: Evaluation of prostate cancer with a PSA cut off of 4ng/ml with test sensitivity of 20% and test specificity of 90%

PSA [test] 4ng/ml	Prostate cancer diagnosed by the current gold standard		
	Present	Absent	
Positive	True Positive n = 10	False positive n = 05	a+b n = 55
Negative	False Negative n = 40	True Negative n = 45	c+d n = 45
	50	50	

Table 6: Evaluation of prostate cancer with a PSA cut-off of 3ng/ml

PSA [test] 3ng/ml	Prostate cancer diagnosed by the current gold standard		
	Present	Absent	
Positive	True Positive n = 15	False positive n = 10	a+b n = 55
Negative	False Negative n = 35	True Negative n = 40	c+d n = 45
	50	50	

detection of true positives [from 10 to 15 individuals]. However, the number of false positives has also gone up [from 5 individuals to 10]. This results in a **test sensitivity of 30%.** Similarly, the reduction of detection of true negatives [from 45 to 40 individuals] and improved false negative rate [which has dropped from 40 to 35 individuals] **giving a test specificity of 80%**.

Thus, when we vary cut offs or boundaries, the following aspects need to be borne in mind

- Different cut off points or boundaries will yield different sensitivities and specificities
- The cutoff point is crucial in that it labels patients as having the disease or otherwise
- A cutoff point that identifies more true negatives, will also yield more false negatives
- A cutoff point that identifies more true positives, will also yield more false positives

What a Gold Standard is

If we go back to the example of peripheral smear for the diagnosis of malaria, the test is available only in select centers, and requires considerable technical expertise, time and skill in identifying the parasite. Hence, several rapid diagnostic tests that use malarial antigens have been introduced where the presence or absence of a "line" along with the control line on the test strip [which requires nothing more than a drop of the patient's blood] give the diagnosis with ease and great rapidity. Now, these "new" tests have to be "compared" for their performance with the existing gold standard. A gold standard is the one that is universally accepted as being the benchmark test for that condition to make a definitive diagnosis or the most accurate test at that point in time. For example, the gold standard test for the diagnosis of prostate cancer as stated earlier would be the trans-rectal ultrasound guided biopsy and that for coronary artery disease would be a coronary angiography. A gold standard test may be a "single" best test or a combination of tests.

Use of Multiple Tests

More often than not, in clinical practice, clinicians tend to use multiple rather than single tests and this needs to be remembered. For example, for primary open angle glaucoma, the most prevalent form of glaucoma, the diagnosis is made on a combination of measuring intra ocular pressure [IOP] and assessing optic disc changes with a slit lamp examination. Similarly, the initial diagnosis of prostate cancer is usually made with a combination of Digital rectal examination [DRE] and the serum Prostate specific antigen estimation [PSA] and confirmed by trans-rectal ultrasound guided biopsy.[7]

Conclusions

In summary, results of both screening and diagnostic tests need to be interpreted in the context of performance of the test [as assessed by the metrics of sensitivity and specificity] and disease prevalence [as assessed by the positive and negative predictive values]. Given that both benefit [identifying individuals with disease and ruling out those without] and harm [falsely labelling an individual as having disease or missing disease in others] can accrue with the use of these tests, their use by clinicians should be judicious and made with a clear understanding and appreciation of the implications for diagnosis and subsequent management, test limitations, financial considerations and finally, impact on the patient's quality of life.

Acknowledgements

The authors are grateful to Dr. Seema Kembhavi, Radiation Oncologist from the Tata Memorial Hospital, Mumbai for her helpful inputs on the manuscript.

References

1. Streiner DL. Diagnosing tests: using and misusing diagnostic and screening tests. *J Pers Assess* 2003; 81:209-19.
2. Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* 2009; 19:203-211.
3. Schulz KF, Grimes DA. Uses and abuses of screening tests. In The Lancet Handbook of essential concepts in clinical research. Elsevier. 2006.
4. Parikh R, Mathai A, Parikh S, et al. Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology* 2008; 56:45-50.
5. http://learn.chm.msu.edu/epi/Coursepack/EPI546_Lecture_5_course_notes.pdf, accessed on 24th April 2017.
6. Ankerst DP, Thompson IM. Sensitivity and specificity of prostate-specific antigen for prostate cancer detection with high rates of biopsy verification. *Arch Ital Urol Androl* 2006; 78:125-9.
7. Mistry K, Cable G. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *J Am Board Fam Pract* 2003; 16:95-101.