Understanding and using sensitivity, specificity and predictive values

***Rajul Parikh**, MS; **Annie Mathai**, MS; **Shefali Parikh**, MD; **G Chandra Sekhar**, MD; **Ravi Thomas**, MD*

In this article, we have discussed the basic knowledge to calculate sensitivity, specificity, positive predictive value and negative predictive value. We have discussed the advantage and limitations of these measures and have provided how we should use these measures in our day-to-day clinical practice. We also have illustrated how to calculate sensitivity and specificity while combining two tests and how to use these results for our patients in day-to-day practice.

Key words: Predictive values, sensitivity, specificity

Indian J Ophthalmol 2008;56:45-50

Modern ophthalmology has experienced a dramatic increase in knowledge and an exponential increase in technology. A lot of this 'hi-tech' explosion involves diagnostic tests. Regrettably, there is sometimes a tendency to use tests just because they are available; or because they are hi-tech. The basic idea of performing a diagnostic test is to increase (or decrease) our suspicion that a patient has a particular disease, to the extent that we can make management decisions. In this article, we have tried to explain the rationale behind tests and their 'scientific' application in the practical management of a patient.

Diagnostic Tests

For this article, the term 'diagnostic tests' will include everything physicians do to diagnose disease. This includes assessing symptoms and signs, as well as what we conventionally refer to as tests: such as laboratory investigations, gonioscopy, Optical Coherence Tomography (OCT), etc.

Gold Standard

The gold standard is the best single test (or a combination of tests) that is considered the current preferred method of diagnosing a particular disease (X). All other methods of diagnosing X, including any new test, need to be compared against this 'gold' standard. The gold standard is different for different diseases. If we are considering peripheral anterior chamber depth (van Herick test [2]) for the diagnosis of primary angle closure (PAC), the current gold standard is gonioscopy. The gold standard for demonstrating the functional defect in glaucoma is automated perimetry. The gold standard for X may be considered outdated or inadequate, but any new test designed to replace the gold standard *has* to be initially validated against the gold standard. If the new test is indeed better, there are ways to prove that; following which the new test may become the gold standard.

Validity

It is the extent to which a test measures what it is supposed to measure; in other words, it is the accuracy of the test. Validity is measured by sensitivity and specificity. These terms, as well as other jargon, are best illustrated using a conventional two-by-two (2 × 2) table.

The information obtained by comparing a new diagnostic test with the gold standard is conventionally summarized in a two-by-two table [Table 1].

In cell 'a,' we enter those in whom the test in question correctly diagnosed the disease (as determined by the gold standard). In other words, the test is positive, as is the gold standard. These are the true positives (TP).

In cell 'b,' we enter those who have positive results for the test in question but do not have disease according to the 'gold standard test.' The newer test has wrongly diagnosed the disease: These are false positives (FP).

In cell 'c,' we enter those who have disease on the 'gold standard test' but have negative results with the test in question. The test has wrongly labeled a diseased person as 'normal.' These are false negatives (FN).

In cell 'd,' we enter those who have no disease as determined by the 'gold standard test' and are also negative with the newer test. These are true negatives (TN).

Sensitivity (positive in disease)

Sensitivity is the ability of a test to correctly classify an individual as 'diseased' [Table 2].

Sensitivity = a / a+c
= a (true positive) / a+c (true positive + false negative)
= Probability of being test positive when disease present.

Example: One hundred persons with primary angle closure glaucoma (PACG, diagnosed by 'gold standard': gonioscopy) are examined by van Herick test. Seventy-five of them had narrow peripheral anterior chamber depth [Table 3]. The sensitivity of the peripheral anterior chamber depth

L. V. Prasad Eye Institute, Banjara Hills, Hyderabad, India

Correspondence to Dr. Rajul Parikh, Victor villa, 5, Babulnath Road, Mumbai - 400007, India. E-mail: drparikhs@gmail.com

Manuscript received: 23.02.07; Revision accepted: 31.07.07

Table 1: Shows 2 × 2 (two-by-two) table

	Gold standard disease present	Gold standard disease absent	
Test positive	True positives (TP)	False positives (FP)	Total test positives:
	a	b	a+b
Test negative	False negative (FN)	True negatives (TN)	Total test negatives:
	c	d	c+d
	Total diseased:	Total normal:	Total population:
	a+c	b+d	a+b+c+d

Table 2: Calculation of sensitivity and specificity

	Disease present	Disease absent
Test positive	a (TP)	b (FP)
Test negative	c (FN)	d (TN)
	Sensitivity:	Specificity:
	a/ (a+c)	d/ (b+d)

TP: True positive, FP: False positive, FN: False negative, TN: True negative

Table 3: Shows example for the calculation of sensitivity and specificity

New test	Gold standard	
	Positive	Negative
Test +ve	75	15
Test –ve	25	85
Total	100	100
	Sensitivity: 75/100	Specificity: 85/100

examination to PACG is therefore –

75 / 100 = 75%.

Specificity (negative in health)

The ability of a test to correctly classify an individual as *disease-free* is called the test's specificity**.** [Table 2]

Specificity = d / b+d
= d (true negative) / b+d (true negative + false positive)
= Probability of being test negative when disease absent.

Example: One hundred persons with normal angles (diagnosed by 'gold standard': gonioscopy) are examined by peripheral angle chamber depth examination. Eighty-five persons had normal peripheral angle chamber depth [Table 3]. The specificity of the peripheral angle chamber depth examination to PACG is therefore –

85 / 100 = 85%.

Sensitivity and specificity are inversely proportional, meaning that as the sensitivity increases, the specificity decreases and vice versa. What do we mean by this? Let us say that an intraocular pressure (IOP) of ≥25 mmHg is test positive and <25 mmHg is test negative. Very few normal subjects would have IOP more than 25 mmHg, and hence the specificity (NIH – negative in health) would be very high. But as a significant number of glaucoma subjects would have an IOP <25 mmHg (remember that close to 50% of glaucomas detected in population are normal-tension glaucomas), the sensitivity (PID – positive in disease) of IOP >25 mmHg in the detection of glaucoma would be low. Suppose we take the IOP cutoff for test positive to be 35 mmHg. Almost no normal subject would have this high an IOP, and the specificity would be very high (>99%); and a highly specific test if positive (for example an IOP >35 mmHg), rules in the disease. Remember this as SpPIN: a highly Specific test if Positive, rules IN disease. Similarly, if we take a cutoff of 12 mmHg, almost no glaucoma subject would have an IOP <12 mmHg (high sensitivity). An eye with an IOP <12 mmHg is extremely unlikely to have glaucoma. A highly sensitive test if negative, rules out the disease. Remember this as SnNOUT: a highly Sensitive test if Negative, rules OUT disease. (Almost all normals would have an IOP >12 mmHg, a very low specificity; but that is a different issue). Another example of SnNOUT would be the absence of venous pulsation in papilledema. The sensitivity of the sign 'absence of venous pulsation' in the diagnosis of papilledema is 99%, and specificity is 90%. So if venous pulsation is present, then we can apply SnNOUT and rule out papilledema. At that point in time, papilledema may be evolving and may still develop a few days or a week later; or patients may have papilledema, but the intracranial pressure at the time of examination is normal.

Positive Predictive Value (PPV)

It is the percentage of patients with a positive test who actually have the disease. In a 2 × 2 table [Table 1], cell 'a' is 'true positives' and cell 'b' is 'false positives.' In real life situation, we do the new test first and we do not have results of 'gold standard' available. We want to know how this new test is doing. PPV tells us about this – how many of test positives are true positives; and if this number is higher (as close to 100 as possible), then it suggests that this new test is doing as good as 'gold standard.'

PPV: = a / a+b
= a (true positive) / a+b (true positive + false positive)
= Probability (patient having disease when test is positive)

Example: We will use sensitivity and specificity provided in Table 3 to calculate positive predictive value.

PPV = a (true positive) / a+b (true positive + false positive)
= 75 / 75 + 15 = 75 / 90 = 83.3%

Negative Predictive Value (NPV)

It is the percentage of patients with a negative test who do not have the disease. In 2 × 2 table [Table 1], cell 'd' is 'true negatives' and cell 'c' is 'false negatives.' NPV tells us how many of test negatives are true negatives; and if this number is higher (should be close to 100), then it suggests that this new test is doing as good as 'gold standard.'

NPV:= d / c+d
= d (true negative) / c+d (false negative + true negative)
= Probability (patient not having disease when test is negative)

Example: We will use sensitivity and specificity provided in Table 3 to calculate negative predictive value.

NPV = a (true negatives) / c+d (false negative + true negative)
= 85 / 85 + 25 = 85 / 110 = 77.3%

Positive and negative predictive values are directly related to the prevalence of the disease in the population [Fig. 1]. Assuming all other factors remain constant, the PPV will increase with increasing prevalence; and NPV decreases with increase in prevalence. This is illustrated by the following example.

A new test has been developed to diagnose primary angle closure glaucoma (PACG). To clarify the terminology used in the example, we will repeat definitions of primary angle closure (PAC) and PACG. PAC is defined as a person with an occludable angle (>180° of posterior trabecular meshwork not visible) with peripheral anterior synechiae with or without raised intraocular pressure (IOP). Optic disc and visual field do not show glaucomatous damage. PACG is defined as PAC with optic disc and visual field changes. PAC affects approximately 3 to 4% of population, while PACG affects approximately 1% of population.

This new test has been performed in 1,000 patients that had documented PACG (disease positive) on gonioscopy (gold standard) and 1,000 normal persons as controls. The authors found that 900 were correctly classified as PACG by the 'new test,' and 950 were correctly labeled as open angle [Table 4a]. The authors would report the sensitivity and specificity of a test as 90 and 95% respectively. With a sensitivity of 90% and a specificity of 95%, the new test appears to be an excellent test.

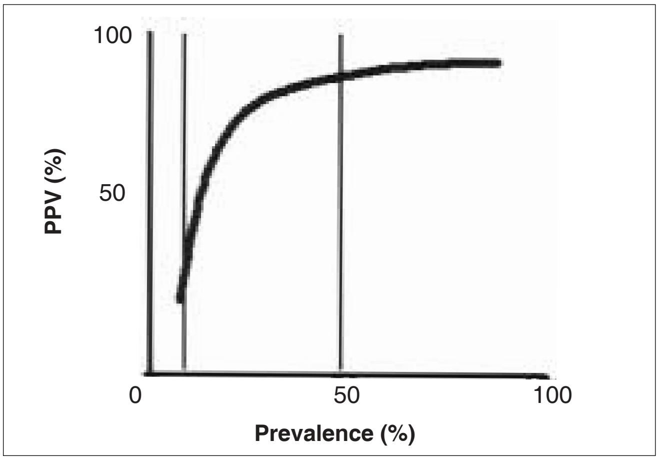

Figure 1: As the disease prevalence increases, the positive predictive value also increases

Let's apply this test to a million people where only 1% is affected with PACG. Of the million people, 10,000 would be affected with PACG. Since our new test is 90% sensitive, the test will detect 9,000 (TP) people who are actually affected with PACG and miss 1,000 (FN). Looking at those numbers, we would think that our test is very good because we have detected 9,000 out of 10,000 PACG-affected people. However, of the original 1 million, 990,000 are not affected. If we look at the test results on the normal population (remember, the specificity of the test is 95%), we find that while 940,500 are found to be not affected by the new test (TN), we have 49,500 individuals who are found to be positive by the new test (FP).

If we start using this new test without confirmatory testing on the gold standard gonioscopy, we would diagnose 49,500 people, or approximately 5% of the population, as PACG when in reality, they are not. The sensitivity and specificity of the test have not changed. The sensitivity and specificity were however determined with a 50% prevalence of PACG (1,000 PACG and 1,000 normals) with PPV of 95%. We are now applying it to a population with a prevalence of PACG of only 1%. With a 1% prevalence of PACG, the new test has a PPV of 15%. Although the sensitivity and specificity of the test have not changed, the PPV has changed drastically. If the prevalence (also known as the pre-test probability in this situation) of the disease is low, such as with glaucoma or sight-threatening diabetic retinopathy in the general population, the number of false-positive results will be far higher than the number of true-positive results.[3] This leads to a number of problems, including labeling of normal as abnormal resulting in unnecessary treatment.

The NPV of the test also change depending on the prevalence of the disease and usually in reverse direction to PPV. In the above example, in high-prevalence situation (50% prevalence) [Table 4a], the NPV was 90%. In low-prevalence situation [Table 4b], the NPV increased to 99%. So why not use a test for the NPV value? If the prevalence is already so low, the NPV will certainly reduce it further but still not to zero.

The PPV can increase if we repeat the test in certain situations. For example, in HIV, if we repeat ELISA with different kit in

Table 4a: Showing example of calculation of predictive value at 50% prevalence

New test	Gold standard		Predictive values
	Test +ve	Test -ve	
Test +ve	900	50	900/950 = 94.7%
Test -ve	100	950	950/ 1050 = 90.5%
Total	1000	1000	

Table 4b: Showing example of calculation of predictive values at 1% prevalence

New test	Gold standard		Predictive values
	Test +ve	Test -ve	
Test +ve	9000	49500	9000/ 58500 = 15.4%
Test -ve	1000	940500	940500/ 941500 = 99.9%
Total	10,000	9,90,000	

the group that is already ELISA positive, the specificity and PPV will increase. However, if the same test is repeated, then concordance will be a problem.

Everything we have discussed so far has assumed that the sensitivity and specificity do not change as one deals with different groups of people. Sensitivity and specificity, however, can change if the population tested is dramatically different from the population you serve, especially if the spectrum of the disease is different. In more severe disease, we are more likely to be able to make a diagnosis; and thus sensitivity goes up.

What if the new test is actually *better* than the gold standard? There is no shortcut to the process of comparing it to the existing gold standard. The new (presumably better) test will detect more disease than the 'gold standard.' In the 2 × 2 table, the subjects labeled as 'diseased' by the new test (but still 'normal' on the 'gold standard') will go in cell 'b' (false positives). If on follow-up, a significant number of these patients actually develop disease (gold standard positive), then the new test is in fact detecting disease earlier than, and is better than, the gold standard. In some instances, there may be other strategies available to determine straight away whether the new test is in fact better.[4]

Clinical application

So far we have discussed how to calculate sensitivity, specificity, positive and negative predictive values using 2 × 2 table. Now we will discuss the clinical application of these parameters.

The sensitivity, specificity of IOP, torch light test, van Herick test are shown below [Table 5].

Which test should we use to screen the population for angle closure glaucoma? The prevalence and PPV discussed above (and other reasons provided in the reference) should have convinced you that this is a bad idea.[3] So let's take an example in a clinic. Table 5 shows the sensitivity and specificity of various tests we can use for detecting PACG. Gonioscopy is the 'gold standard' for diagnosis of angle closure, and that's why we should do gonioscopy in all patients we see in clinics. All other tests (IOP, torch light test and van Herick test) have poor specificity.[2,5,6] Even with specificity as high as 90%, the PPV will be poor. The prevalence of angle closure (as opposed to angle closure glaucoma) is approximately 3%. With this prevalence, PPV of IOP would be 15%; torch light test, 7.6%; and for van Herick test, 15%. These results mean that if we use IOP or van Herick test to diagnose angle closure, only 15% of suspected angle closure patients will really have disease, and the other 85% would be FP. The sensitivity of these tests is moderate and will miss most of the disease.

In day-to-day clinical practice, we can however combine results of two independent tests to be more confident of the diagnosis – for example, combining IOP and optic disc changes for primary open angle glaucoma (POAG), IOP and peripheral angle chamber depth for diagnosis of PACG, history of diabetes and frequency doubling technology (FDT) defect for diabetic retinopathy.[3]

Table 5: Shows sensitivity, specificity of intraocular pressure, torch light test and van Herick test

Test	Sensitivity (%)	Specificity (%)
Intraocular pressure	47	92
Torch light test	80	70
van Herick test	61.9	89.3

Case 1

A 54-year-old male patient was diagnosed to have POAG. He did not have any ocular or systemic complaints. The vision was 20/20, N6 in each eye. The IOPs were 25 mmHg in both eyes on several occasions. Corneal pachymetry was normal and the angle was reported to be open. The optic discs showed changes suggestive of glaucoma, and there were corresponding early visual defects. The patient was started on a unilateral trial of timolol 0.5% twice daily.

The van Herick test when the patient was examined 2 weeks later is shown in Fig. 2. The peripheral anterior chamber depth was less than one-fourth the peripheral corneal thickness in both eyes.

With this IOP and van Herick test, a diagnosis of POAG becomes unlikely. Let us examine the rationale behind this statement. The specificity of IOP for glaucoma is 90%. That in itself is not enough for a SpPIN, or 'rule in,' and doesn't help too much. The specificity of the van Herick test for angle closure is 85%, which again, on its own is not of much help either. However, the two tests can be combined to increase the specificity and perhaps apply SpPIN and 'rule in' diagnosis. The specificity of the two tests can be combined in the following manner[6]:

Specificity of combined test = 1 − (1 − specificity of test 1) × (1 − specificity of test 2)

Plugging in the values for our patient,

$1 - (1 - 0.9) \times (1 - 0.85) = 1 - 0.1 \times 0.15 = 1 - 0.015$

$= 0.985$, or 98.5%

This combined specificity of 98.5% definitely allows us to invoke SpPIN and rule in a diagnosis: until proved otherwise, this patient has angle closure. (We assume that the IOP specificity of 90% holds for angle closure glaucoma too.)

The 'open angle' described earlier is shown in Fig. 3. The angles on repeat gonioscopy (indentation) are shown in Fig. 4.

One valid objection to combining tests in this manner is that the resultant sensitivity becomes the product of the sensitivities of the two tests – that is, the product of the sensitivity of an IOP >21 mmHg (50%) and the sensitivity of the van Herick test (69%) = 0.50 × 0.69 = 34.5%. While 35 is a low sensitivity as far as tests in general are concerned, it doesn't really matter here as we are utilizing the 'rule in' specifically to make the diagnosis in an individual patient.

Let's take another example: a patient has repeatable IOP measurements of 24 mmHg with normal pachymetry, and the angles this time are *really* open. The specificity of the IOP measurement is 90%. And, while not too useful a measure, the cup disc ratio is 0.7 (specificity of CDR >0.55 is 73%). The combined specificity of IOP and disc now becomes 1 − (1 − specificity of IOP) × (1 − specificity of Disc) = 1 − (1 − 0.90)×(1 − 0.73) = 1− (0.1)×(0.27) = 1 − 0.027 = 97.3%.

This specificity is high enough to "rule in" the diagnosis of POAG, without further testing. Any further testing is probably

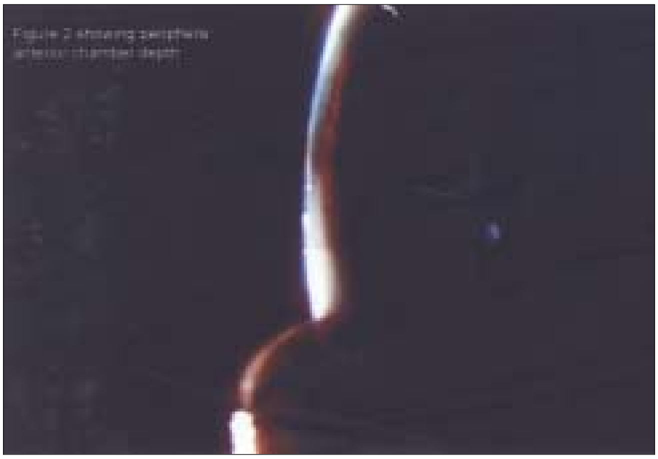

Figure 2: Van Herick test showing shallow peripheral anterior chamber depth (< one-fourth the peripheral corneal thickness)

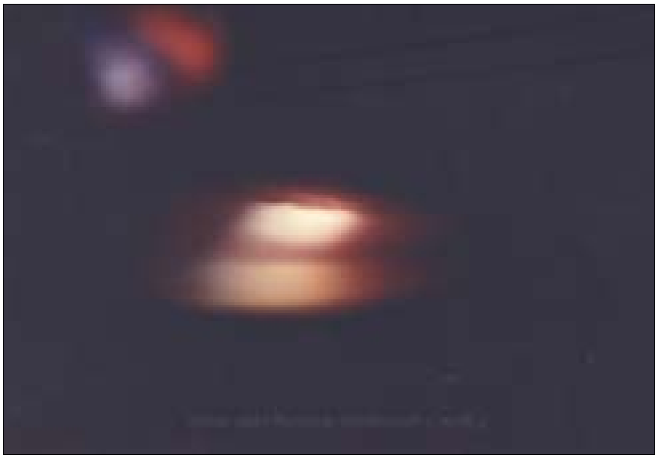

Figure 3: 'Open angle' in an inappropriate testing condition

required for monitoring. Of course, whether we treat or not is a different matter.

Some of us want even more evidence than this. The approach we describe allows incorporation of further testing (including optimal and effective use of modern imaging techniques) too. The GDX 'number (NFI)' in the above patient is more than 32 (specificity of about 85%). If we combine this with just the IOP, can you calculate the combined specificity?

1 − (1 − specificity of IOP)) × (1 − specificity of 'number' >30)

You should get 98.5%.

This should be confirmatory; but if you are still not satisfied and want to take it further, you can use the IOP, Disc and the GDX. 1 − (1 − specificity of IOP) × (1 − specificity of Disc)×(1 − specificity of 'number' >30).

Did you get 99.5%? As a 'rule in,' this is (almost) as good as it gets. Regrettably, there is no absolute certainty. According to our clinical Bible, absolute certainty is limited to theologians and like-minded clinicians.[1] And as the tests are 'independent,' our estimate of specificity should work. If the tests were not independent, there would be some 'convergence,' as it is technically called. When we use three tests, such convergence would have minimal clinical significance.

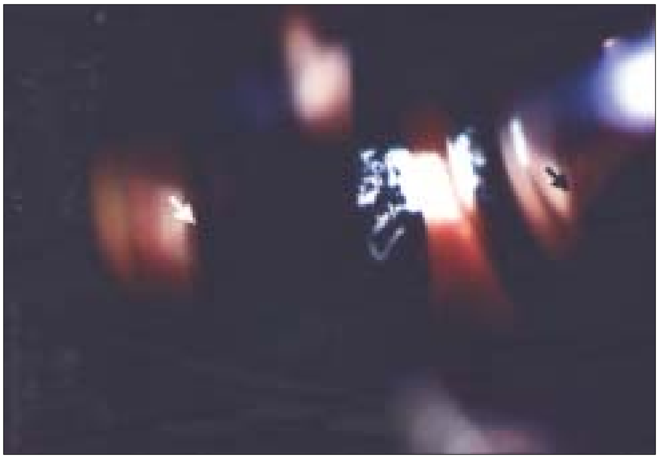

Figure 4: Gonioscopy in an appropriate condition showing closed angle (white arrow) and presence of a peripheral anterior synechia on indentation (black arrow)

Case 2

A 40-year-old male is suspected to have sarcoidosis. It is an idiopathic multi-system granulomatous disease, where the diagnosis is made by a combination of clinical, radiological and laboratory findings. The gold standard is a tissue biopsy showing noncaseating granuloma. Ocular sarcoidosis could present as anterior, intermediate, posterior or panuveitis; but none of these are pathognomonic. Therefore, one has to rely on ancillary testing to confirm the diagnosis.

Angiotensin-converting enzyme (ACE) has a sensitivity of 73% and a specificity of 83% to diagnose sarcoidosis. Abnormal gallium scan has a sensitivity of 91% and a specificity of 84%.[7] Though individually the specificity of either test is not impressive, when we combine both the tests, the specificity becomes –

$$1 - (1 - 0.84) \times (1 - 0.83) = 1 - (0.16 \times 0.17)$$
$$= 1 - 0.03 = 0.97 = 97\%$$

The combination sensitivity becomes = 0.73×0.91 = 0.66 = 66%.

Sensitivities can be used in the same manner to rule out diagnoses. Let us assume that the cup disc ratio (usually useless without a mention of the disc size, but having a sensitivity of 50% for a cutoff of >0.55) is 0.6; and the IOP is 21 mmHg (GHT, sensitivity of only 50%). But you feel the disc is suspicious or the patient has a family history or has been referred or whatever. Based on the above information, could the patient still have glaucoma? The combined sensitivity is calculated as:

1 − (1 − sensitivity of IOP)×(1 − sensitivity of CDR >0.55).

Did you try to calculate that? You should get 75%. That's certainly not good enough to rule out a disease like glaucoma. The visual fields, specifically the glaucoma hemifield test (sensitivity 95%), are normal. The combined specificity now becomes 1 − (0.25)×(1 − 0.95) = 98.75. You should be able to rule out 'functional' glaucoma now. Actually a normal field with a normal GHT with a sensitivity of 95% is on its own a good enough 'rule out,' but we know that the field may be normal with a lot of disc damage. So you can use the GDX to combine information about the nerve fiber layer. The 'number' on GDX is 31, the sensitivity of which is 74%. What is the combined sensitivity now? 98.8%. Can we send the patient home now?

In summary, we have provided the basic knowledge to calculate sensitivity, specificity, PPV and NPV. More importantly, we have discussed the advantage and limitations of these measures and provided how we should use these measures in our day-to-day clinical practice. We also have illustrated how to calculate sensitivity and specificity while combining two tests and how to use the results for our patients in day-to-day practice.

References

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Editors. Clinical Epidemiology: A Basic Science for Clinical Medicine. 2nd Edition. New York, Little, Brown and co., 1991, 163-7.
2. Van Herick W, Shaffer RN, Schwartz A. Estimation of width of angle of anterior chamber: Incidence and significance of the narrow angle. Am J Ophthalmol 1969;68:626-9.
3. Parikh R, Naik M, Mathai A, Kuriakose T, Muliyil J, Thomas R. Role of frequency doubling technology perimetry in screening of diabetic retinopathy. Indian J Ophthalmol 2006;54:17-22.
4. Sackett DL, Haynes RB, Guyatt GH, Tugwell P, editors. Clinical epidemiology: A basic science for clinical medicine. Little, Brown and Co: New York; 1991. p. 51-68.
5. Vargas E, Schulzer M, Drance SM. The use of the oblique illumination test to predict angle closure glaucoma. Can J Ophthalmol 1974;9:104-5.
6. Thomas R, George T, Braganza A, Muliyil J. The Flashlight and van Herick's Test are poor predictors of occludable angles. Aust N Z J Ophthalmol 1996;24:251-6.
7. Power WJ, Neves RA, Rodriguez A, Pedroza-Seres M, Foster CS. The value of combined serum angiotensin-converting enzyme and gallium scan in diagnosing ocular sarcoidosis. Ophthalmology 1995;102:2007-11.

Source of Support: Nil, **Conflict of Interest:** None