STATISTICS FOR RESEARCHERS

Statistical Evaluation of Diagnostic Tests – Part 2 [Pre-test and post-test probability and odds, Likelihood ratios, Receiver Operating Characteristic Curve, Youden's Index and Diagnostic test biases]

NJ Gogtay, UM Thatte

Introduction

In the previous article on the statistical evaluation of diagnostic tests –Part 1, we understood the measures of sensitivity, specificity, positive and negative predictive values. The use of these metrices stems from the fact that no diagnostic test is ever perfect and every time we carry out a test, it will yield one of four possible outcomes– true positive, false positive, true negative or false negative. The 2 x 2 table [Table 1] gives each of these four possibilities along with their mathematical calculations when a new test is compared with a gold standard test.[1]

In this article, the second in the diagnostic test series, we will discuss single summary statistics that help us understand and use these tests appropriately both in the clinical context and when these summary statistics appear in literature. Before we discuss these, we need to recapitulate a few concepts presented in earlier articles [odds and probability] and also some novel concepts [Bayesian statistics, pre-test and post-test probabilities and odds].

Understanding Probability and Odds and the Relationship between the Two

Let us understand probability and odds with the example of a drug producing bleeding in 10/100 patients treated with it. The probability of bleeding will be 10/100 [10%], while the odds of bleeding will be 10/90 [11%]. This is because odds is defined as the probability of the event occurring divided by the probability of the event not occurring.[2] Thus, every odds can be expressed as probability and every probability as odds as these are two ways of explaining the same concept.

From the example, it follows that

Odds = p/1-p, where p is the probability of the event occurring.

Probability, on the other hand, is given by the formula

p = Odds/1+Odds

Bayesian Statistics, Pre-Test Probability and Pre-Test Odds

A clinician often suspects that a patient has the disease even before he orders a test [screening or diagnostic] on the patient. For example, when a patient who is a chronic smoker and presents with cough and weight loss of a six-month duration, the suspicion of lung cancer has already entered

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

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

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

the physician's mind. Thus, the clinician has already, mentally, identified some *"pre-test" probability* of the patient having the disease; lung cancer in this case.

Clinical decision-making, by and large, requires a combination of clinical acumen along with a correctly performed and interpreted screening or diagnostic test. When the physician allocates a "pre-test probability", what he is applying is a field of statistics called Bayesian statistics. Herein, the knowledge of *prior beliefs* is used and quantified as a numerical value ranging from 0 -100%.[3] This value is then used for subsequent calculations. Bayesian statistics allows us to interpret screening and diagnostic tests in their clinical context.

Logically, the next question would be - what are the ways in which these pre-test probabilities can be allocated? These are listed below

- Subjectively based on informed opinion, consensus guidelines or experience in treating the disease in question
- An understanding of the evolution of the disease and matching it with how the disease has actually evolved in the patient
- Objectively based on available evidence [prevalence data for example]

In the example presented, the treating physician may assign a pretest probability of 60% or even higher based on his clinical acumen and what he sees in practice. How is this calculated? Let us say that the clinician is a lung cancer specialist and he sees 100 patients in three months who are chronic smokers with persistent cough and weight loss. Sixty of them eventually return a diagnosis of lung cancer based on one more tests. The pretest probability for a new patient with a similar history and complaints who presents to him in the fourth month would thus be 60%.

Mathematically, this is calculated as

Pre-test probability =

Number of patients with complaints actually diagnosed to have the disease

Total number of patients who present with the same complaints

[In this case, it would be 60/100 or 60%].

Pretest odds, however, would be 0.6/0.4 or 1.5 (the probability of the event occurring divided by the probability of the event not occurring).

The clinician next orders a test, which he hopes, will confirm [or refute] his diagnosis. The test result and the pre-test probability together will now be used to calculate the post-test probability as described below.

Post-test Probability and Post-Test Odds

Since the result of a diagnostic test can be either positive or negative, post-test probabilities are either positive or negative. Mathematically,

- Post-test probability = Pre-test probability x Likelihood ratio (see below for explanation), while
- Post-test odds = Post-test probability/1 – post-test probability

The Likelihood Ratio [A Summary Statistic]

Likelihood ratios [LR] combine both sensitivity and specificity into a single measure and are an alternate way of evaluating and interpreting diagnostic tests.[4] They help in making a choice of a diagnostic test or sequence of tests. LR essentially tell us how many times more [or less] a test result is to be found in diseased compared to non-diseased people. LRs are of two types – positive and negative. A positive Likelihood ratio is given by

$$\text{Likelihood ratio [positive] LR+} = \frac{\text{Sensitivity [TP]}}{\text{1 -Specificity [FP]}}$$

while a negative Likelihood ratio is given by

$$\text{Likelihood ratio [negative] LR -} = \frac{\text{1- Sensitivity [FN]}}{\text{Specificity [TN]}}$$

Let us understand this with an example. When physical examination is carried out in patients with suspected acute appendicitis, there-is-rebound tenderness at or about the McBurney's point, pain on percussion, rigidity, and guarding. The positive likelihood ratio for the diagnosis of appendicitis would be the ratio of those with appendicitis who have tenderness at McBurney's point [sensitivity] by those without appendicitis who have tenderness at McBurney's point [falsely positive or 1- specificity]

OR

Likelihood ratio [positive] LR+

The number of patients *with appendicitis* who have localized tenderness at the McBurney's point

The number of patients *without appendicitis* who have localized tenderness at the McBurney's point

The negative likelihood ratio LR- would be

The number of patients *with appendicitis* who don't have localized tenderness at the McBurney's point

The number of patients *without appendicitis* who don't have localized tenderness at the McBurney's point

If we were to express both these mathematically, based on the 2 x2 table, these would be as given below

Likelihood ratio positive or LR +

The probability of obtaining a positive test result in patients with disease [TP]

The probability of obtaining a positive test result in patients without the disease [FP]

On the other hand, a negative likelihood ratio or LR- would be

The probability of obtaining a negative test result in patients with disease [FN]

The probability of obtaining a negative test result in patients without the disease [TN]

Since different tests for the same disease have different sensitivities and specificities, each test would yield a different likelihood ratio for the same disease. Let us understand this with an example. The diagnosis of prostate cancer can be made by both digital rectal examination [DRE] and Trans rectal ultrasonography [TRUS]. Manyahi JP and colleagues[5] in their study found the sensitivity of DRE to be 66.7%, and the specificity to be 88.6%. The values for TRUS were 58.3% and 85.7% respectively. The LR + for DRE thus would be 5.8 [.667/1-.886], while that for TRUS would be 4.1[.583/1-.857]. The LR- for the two tests similarly would be 0.38 [1-.667/.886] and 0.49 [1-.583/.857] respectively.

LRs range from 0 to infinity. LRs more than 1 argue for the presence of the disease and numbers further away from 1 strengthen this argument. They, thus *rule in* the disease. LRs between 0 and 1 argue against the diagnosis of interest. Values of 1 or close to 1 indicate that the test may lack diagnostic value. LR- values below 1 indicate that the result is likely to associated with the absence of the disease.[4]

While LRs are good measures of diagnostic accuracy, these are seldom used in clinical practice as they require a knowledge of probabilities and involve calculations. Nomograms such as the Fagan's nomogram [https://mclibrary.duke.edu/sites/mclibrary.duke.edu/files/public/guides/nomogram.pdf] are available[6] for ease of use of LRs, but may not always be available for a quick bedside diagnosis. The logarithm of the likelihood ratio [log likelihood ratio statistic] is used to compute a p value and then compared with the critical p value of 5% that we use routinely use to check for statistical significance of a LR that is calculated.

Clinical Application – putting Together Probability, Odds and the Likelihood ratio

Having understood the concepts of probability and odds, pre-test and post-test probabilities and the likelihood ratios we need to put all of them together to see how they actually help in clinical decision making; the sequence for which is given below

- Calculate Pre – test probability (p)
- Derive Pre- test odds as p/1-p
- Conduct the test [screening or diagnostic] with an appreciation of its sensitivity and specificity
- See the result – positive or negative
- Calculate Post-test odds = Pre-test odds x Likelihood ratio [a positive LR is used for a positive test and *vice versa*]
- Calculate Post-test probability = Post-test odds/(1+ post-test odds)
- Make a decision regarding the diagnosis

Let us understand this with the same hypothetical example. Let us say that a 60-year old male patient with 20 pack years of smoking presents with cough and weight loss of 6 months' duration. The treating physician knows from literature that the pre-test probability of lung cancer is 60% in those with 20 pack years or more in the 50-75 age group.

- Thus, pre-test probability = 60% or 0.6

We now convert pre-test probability into pre-test odds

- Pre-test odds = 0.6/ 1-0.6 or 0.6/0.4 or 1.5

We now conduct a CT scan [low dose] which returns a diagnosis of lung cancer. *In other words, the test is "positive"*. Literature tells us[7] that low dose CT has an approximate sensitivity of 80% and a specificity of 90%. Thus, the positive likelihood ratio would be

- LR + = Sensitivity [.8]/ 1-specificity [1-0.9] = 8 [this LR + indicates that the test result is more likely in someone with lung cancer than someone without]

We now calculate the post-test odds as pre-test odds x likelihood ratio

- Thus, post-test odds = 1.5 x 8 = 12

Finally, we want to convert the post-test odds into post-test probability

- i.e., 12/1 + 12 = 12/13 0r 0.92 or 92% [indicating a high probability that the patient has lung cancer]

What if the CT scan results had been negative?

Here, the pre-test probability of 0.6 and the pre-test odds of 1.5 would have remained unaltered. However, we would now need to calculate the negative LR or LR-

Negative Likelihood ratio [LR-] = 1- sensitivity/specificity

- Or 1-0.8/0.9 = 0.22

Now, the post-test odds would be pretest odds x LR-

- Or 1.5 x 0.22 = 0.33

Post-test probability would be

0.33/1 + 0.33 = 0.25 or 25% [a much lower probability of the patient having lung cancer]

Based on these single summary statistics [92% or 25%], the physician will take the next steps towards management. However, as stated earlier, because LRs involve tedious calculations that include conversion of odds to probabilities and thus are rarely used in clinical practice.

Table 2: Area under the ROC curve and interpretation of the diagnostic accuracy of the test[9]

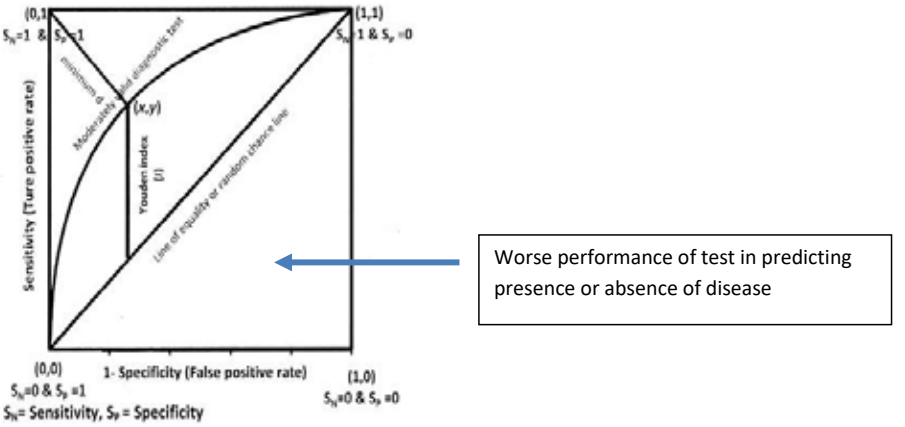
Area under the ROC curve	Interpretation of the test accuracy
1	Perfect
0.9-1	Excellent
0.8-0.9	Good
0.7-0.8	Fair
0.5-0.7	Poor

Receiver Operating Characteristic [ROC] Curve and its Interpretation

The ROC curve is a plot of the sensitivity or true positive rate on the y-axis and *1 minus Specificity* or the false positive rate on the x-axis. Figure 1 depicts the various components of the ROC curve and these are described below.

The point where the x and y axis begin [0,1] depicts 0% sensitivity and 100% specificity. Both sensitivity and specificity are 0 [0,0] where the x axis ends. The upper end of the y axis would be the ideal test with 100% sensitivity and 100% specificity [1,1]. If we were to draw yet another x-axis at the top parallel to the one below, its outer end would depict 100% sensitivity and 0% specificity [0,1] [Figure 1]. The line that connects the beginning of the lower x-axis to the end of the upper x-axis is called the line of equality or random chance line where x [false positive] = y [true positive]. Thus, any ROC curve that appears *below* this line indicates that the test performs worse than random guessing.

Each point on the ROC curve represents a sensitivity-specificity pair corresponding to a certain decision threshold. An ideal test would be one that has 100% sensitivity and 100% specificity and thus the curve will pass through the upper left corner [Figure 1]. Since no test is really ideal and we tradeoff between sensitivity and specificity, the closer the curve is to the upper left corner, the better is its accuracy. The area under



[Reproduced with permission from Indian Pediatrics][10]

Fig. 1: A typical receiver operating characteristic curve and its components

the ROC curve, is taken as 1 and is a useful metric for evaluating the performance of a test. The closer the value of the AUC is to 1, the better is the discriminatory ability of the test [Table 2 and Figure 1]. Since the curve is based on the metrics of sensitivity and specificity alone, the ROC curve is independent of disease prevalence.[8]

Applications of the ROC curve- Any ROC curve helps serve the following four purposes[10]

a. Finding the cut off that least misclassifies diseased and non-diseased individuals
b. Assessing the discriminatory ability of the test
c. Comparing the discriminatory ability of two or more diagnostic tests for assessing the same disease
d. Comparing two or more observers performing the same test [inter- observer variability]

The Youden's index [a Summary Statistic]

It is useful to summarize the information from a ROC curve into a single statistic or index. One of the commonly used indices in the Youden's index "J". This index gives the maximum vertical distance from the line of equality to point [x, y] [Figure 1]. In other words, the Youden index J is that point on the ROC curve that is furthest away from the line of equality [the diagonal line] and maximizes the difference between the sensitivity [true positivity rate] and the false positivity rate [1-specificity].[10,11] It is calculated by deducting 1 from the sum of the test's sensitivity and specificity expressed not as percentage but as a part of a whole number. In other words, it is (sensitivity + specificity) – 1. For a test with poor diagnostic accuracy, Youden's index equals 0, and a perfect test will have a Youden's index of 1.

Diagnostic Odds Ratio [A Summary Statistic]

The Diagnostic odds ratio [DOR] is yet another summary statistic for diagnostic accuracy, that is used for the evaluation of the discriminative abilities of diagnostic procedures as also for the comparison of diagnostic accuracies between two or more diagnostic tests. DOR of a test is defined as the ratio of the odds of positivity in individuals with disease relative to the odds of positivity in individuals without disease. It is calculated similar to the odds ratio as seen in an earlier article[12] as a cross product from the 2 x 2 [Table 1] and given by the formula

DOR = TP x TN ÷ FP x FN

DOR as seen with its calculation depends significantly on the

Table 3: A 2 x 2 table depicting the calculation of the diagnostic odds ratio as a cross product ratio

	Disease present	Disease absent
Test positive	TP	FP
Test negative	FN	TN

sensitivity and specificity of a test. A test with a high specificity and sensitivity [i.e., low rates of false positives and false negatives] will have a high DOR. It is also important to remember here that the same DOR may be achieved with different combinations of sensitivity and specificity. As an illustration, the DOR of 4 can have four combinations of sensitivity and specificity [Table 4].[13]

Reporting of Studies using Diagnostic Tests - The STARD and QUADAS Checklists

STARD stands for "Standards for Reporting Diagnostic Accuracy Studies" and is a checklist of n = 30 items developed by the STARD steering group; an independent group of researchers who formulated this checklist in an attempt to ensure both completeness and transparency of reporting by authors and also for editors and peer reviewers to assess adequacy and quality of information. Authors need to use this checklist in manuscripts that report studies that involve screening or diagnostic tests and reporting their accuracy. STARD can be viewed at http://www.stard-statement.org/.[14] The Quality Assessment of Diagnostic Accuracy Studies (QUADAS - 2) tool is a 14-item checklist to help in the evaluation of diagnostic accuracy studies primarily for use in preparing and presenting systematic reviews.[15]

Table 4: Diagnostic odds ratios for varying combinations of sensitivity and specificity[13]

Specificity [%]	Sensitivity [%]						
	50	60	70	80	90	95	99
50	1	2	2	(4)	9	19	99
60	2	2	(4)	6	14	29	149
70	2	(4)	5	9	21	44	231
80	(4)	6	9	16	36	76	396
90	9	14	21	36	81	171	891
95	19	29	44	76	171	361	1881
99	99	149	231	396	891	1881	9801

Statistical Tests to be Used when Diagnostic Tests are Compared

When two screening or diagnostic tests are conducted on the same patient, the results would amount to "paired" data and since the outcomes are either positive or negative, these constitute "binary outcomes. The McNemar's test is used for this type of comparison. When these two tests are conducted on independent populations, then we use the chi-square or Fisher's exact test.[16]

Understanding Biases when Using Diagnostic Tests - Spectrum Bias and the Imperfect Gold Standard Bias

An important and often overlooked aspect of diagnostic tests evaluation is spectrum bias. In general, patients who present later in the course of a disease are easier to diagnose than those who present early, as with the latter, signs maybe subtle and difficult to pick up. Spectrum bias is a form of selection bias that results when a test is used for a disease that has a wide spectrum of severity.[17] Thus, values of sensitivity and specificity obtained for any test are driven by the *population* that is being studied and different populations would yield different values of the two metrices.

Let us understand this with an example. If we are evaluating a test for detecting patients with diabetes, we could have in our "disease" population, patients with very mild diabetes at one end to severe or even uncontrolled diabetes at the other end of the spectrum. Any diagnostic test study that limits the diabetic patients to the "sickest of the sick" will overestimate the sensitivity of a test, while similarly, another study that uses only the "wellest of the well" [those who are truly non-diabetic; for instance, the very young] will overestimate specificity.[18]

Another bias is the "imperfect gold standard" bias.[19] When a new test [also called as the index test] is being tested, it is compared with an existing "gold standard" or reference test. An ideal gold standard test would be one that "rules in" ALL patients with disease and "rules out" ALL those without. Unfortunately, gold standards are rarely perfect and can themselves misclassify those with and without disease leading to what we call an "imperfect gold standard". Let us understand this with the example of malaria diagnosis. The current gold standard is the peripheral smear. In the hands of trained and expert technicians, the test sensitivity is 50 parasites/ml of blood and results are made available within 30 minutes.[20] The use of this "gold standard" will logically result in declaring parasitemias of less than 50parasites/ml as falsely negative. The polymerase chain reaction [PCR], on the other hand, that detects specific nucleic acid sequences of the parasite has a much higher sensitivity at 5 parasites/ml. However, it is time consuming, technically demanding,

expensive and also detects non-viable parasites that may be present even after successful anti-malarial treatment and can confuse the treating physician.[21] Thus, with its inherent limitations of much lower sensitivity [relative to the PCR], the peripheral smear still remains the "gold standard" [albeit imperfect] for the diagnosis of malaria. Some other biases include uninterpretable or indeterminate test bias and inter-observer bias.[10]

Conclusions

Few topics in the medical field are more important than screening and diagnostic tests as these are ordered nearly every day as an important aid to clinical decision making. Diagnoses are made based on a combination of patient history and physical examination. Tests are often ordered to confirm initial impressions or rule out alternatives, and it is estimated that 10% of all diagnoses are not considered final until clinical laboratory testing is complete.[22] The utility of any test must be assessed bearing in mind its discriminatory ability [to distinguish between health and disease], the nature and severity of the disease under question, the ease of availability of the tests and risks associated with their use, understanding the several diverse metrics [with their limitations] that go into interpreting the results, cost considerations and finally impact on patient management based on the results of the test.

Research studies that publish findings using diagnostic tests must be critically appraised using the STARD criteria as also an appreciation of whether the population on whom the test was used is similar or different from the one that a physician actually sees in his practice. Finally, laboratorians who carry out diagnostic testing, clinicians who treat patients and clinician- researchers who interpret evidence need to work in tandem. This enables better linkage of results of the diagnostic testing with the patient. When coupled with continued monitoring of the effectiveness of these tests, we would ensure both optimal outcomes for an individual patient as also decisions that would drive health policy for nations.

Acknowledgements

The authors are grateful to Dr. Seema Kembhavi, from the Department of Radiodiagnosis at the Tata Memorial Hospital for constructive inputs that helped refine the manuscript.

References

1. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology* 2008; 56:45-50.
2. Bland MJ. The odds ratio. *BMJ* 2000; 320:1468.
3. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural Bayesians. *BMJ* 2005; 330:1080–3.
4. McGee S. Simplifying Likelihood Ratios. *Journal of General Internal Medicine* 2002; 17:647-650.
5. Manyahi JP, Musau P, Mteta AK. Diagnostic values of digital rectal examination, prostate specific antigen and trans-rectal ultrasound in men with prostatism. *East Afr Med J* 2009; 86:450-3.
6. Fagan TJ. Nomogram for Bayes theorem. *N Engl J Med* 1975; 293:257.
7. Toyoda Y, Nakayama T, Kusunoki Y, Iso H, Suzuki T. Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography. *British Journal of Cancer* 2008; 98:1602-1607.
8. Linden A. Measuring diagnostic and predictive accuracy in disease management: an introduction to receiver operating characteristic (ROC) analysis. *Journal of Evaluation in Clinical Practice* 2006; 12:132-39.
9. Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC* 2009; 19:203-211.
10. Kumar R, Indrayan A. Receiver Operating Characteristic [ROC] curve for medical researchers. *Ind Peds* 2011; 48:277-287.
11. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection. *Biometrical journal Biometrische Zeitschrift* 2008; 50:419-430.
12. Gogtay NJ, Deshpande S, Thatte UM. Measures of association. *J Assoc Phy Ind* 2016; 64:70-73.
13. http:// methods. cochrane. org/ sites/ methods. cochrane. org. sdt/ files/ public/ uploads/ DTA% 20Handbook% 20Chapter%2011%20201312.pdf, accessed on 3rd June 2017.
14. http://www.stard-statement.org/, accessed on 13th May 2017.
15. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, LeeflangMM, Sterne JA, Bossuyt PM; QUADAS-2 Group.. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 2011; 18;155:529-36.
16. Deshpande SP, Gogtay NJ, Thatte UM. Which test where? *J Assoc Phy Ind* 2016; 64:64-66.
17. Schmidt LR, Factor ER. Understanding sources of bias in Diagnostic Accuracy Studies. *Arch Pathol Lab Med* 2013; 137:558-565.
18. Willis HB. Spectrum bias- why clinicians need to be cautious when applying diagnostic test studies. *Farm Pract* 2008; 25:390-96.
19. Kohn AM, Carpenter RC, Newman BT. Understanding the direction of bias in studies of diagnostic test accuracy. *Academic Emergency Medicine* 2013; 20:1194-1206.
20. Moody A. Rapid Diagnostic Tests for Malaria Parasites. *Clinical Microbiology Reviews* 2002; 15:66-78.
21. Srinavasan SAH, Moody A, Chiodini PL. Comparison of blood-film microscopy, the OptiMAL® dipstick, Rhodamine 123 and PCR for monitoring anti-malarial treatment. *Ann Trop Med Parasitol* 2000; 94:227–232.
22. Wahner – Roedler DL, Chaliki SS, Bauer BA et al. Who makes the diagnosis? The role of clinical skills and diagnostic test results. *J Eval Clin Pract* 2007; 13:321–5.