J. Fetal Med. (December 2016) 3:151–157
DOI 10.1007/s40556-016-0100-4

REVIEW ARTICLE

Basic Biostatistical Concepts for the Fetal Physician—I: The 2 × 2 Table and Its Derivatives

K. Manikandan[1] · **Sudarshan Suresh**[1] · **Suresh Seshadri**[1]

Received: 15 June 2016 / Accepted: 4 August 2016 / Published online: 30 August 2016
© Society of Fetal Medicine 2016

Abstract In the field of fetal medicine, correct interpretation and optimal utilization of diagnostic tests and signs have a major impact on the pregnancy outcome. The attributes of a diagnostic test, such as sensitivity and specificity, and the attributes of a test result such as positive and negative predictive values, and positive and negative likelihood ratios are important yet poorly utilized concepts in clinical practice. This paper explains these concepts using simple language and examples from the fetal medicine literature.

Keywords Diagnostic test · Sensitivity · Specificity · Likelihood ratio · Pretest probability · Post-test probability · Positive predictive value · Negative predictive value

Introduction

The efficient practice of modern medicine requires the physician to be well versed with the current literature in the areas of his specialty. The enormous rate of increase in the published literature can be overwhelming even to the most ardent academic. Added to this burden is the complexity of statistical concepts that are employed in current scientific papers. Much of the undergraduate medical curricula cover the basic concepts in an abstract way and almost none of the clinical specialty courses formally teach or train residents in these concepts, leaving large lacunae in the specialty training programs. In addition, the advances in the science of biostatistics invariably tend to introduce an increasingly complex range of methods, concepts, and tests that the busy practitioner finds hard to cope up with.

However, it is not impossible to revive the basic concepts back to the clinic as we have seen from our experiences in conducting such workshops. We, therefore, aim to explain in simple clinician's language some of the basic concepts relevant to our day-to-day practice using real world explanations and examples in this series.

Concepts of Pretest and Post-test Probability

Commonly in clinical practice, when the clinician encounters a patient presenting with a symptom, a diagnostic algorithm is initiated. Often this is an informal 'mental' calculation by the physician based on his training, knowledge, and experience. The physician initially shortlists the possible etiologies that would explain the patient's presentation, creates a set of differential diagnoses, and then narrows down to one or two based on further inputs. These further inputs can be in the form of signs elicited, special physical tests, laboratory investigations, imaging, tissue examination, etc.

In a statistical sense, the physician firstly 'calculates' the probability of any of the differential diagnoses being true, and then using the 'further inputs' changes the probability of each of these such that the one with the highest probability is entertained as the diagnosis (Fig. 1). Therefore, the concepts of pretest and post-test probabilities of a disease being present in an individual encountered in the clinic are in fact familiar grounds in medical training. We only objectivize and make it explicit when we talk about a test and its attributes in scientific papers. This makes it easy

✉ K. Manikandan
manimmc@yahoo.com; drmanikandan@mediscan.org.in

[1] Mediscan Systems, 197, Dr Natesan Road, Mylapore, Chennai 600004, India

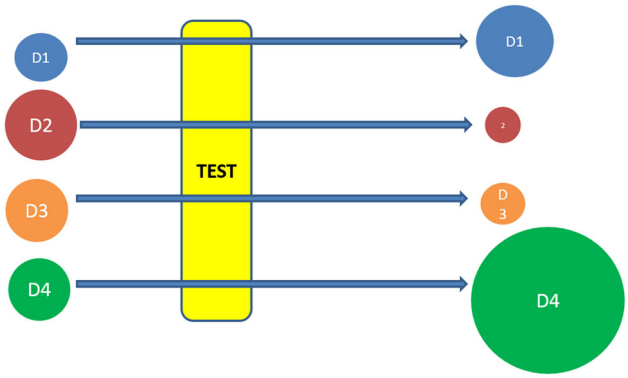


Fig. 1 Diagram showing the interplay of the pretest probability and the diagnostic test in the final post-test probability. *Colors* indicate varied possible diagnoses in a clinical situation and the size of the circles indicates relative probability of these differential diagnoses

and convenient to compare between different tests, different testing scenarios, and different clinical decision-making models.

Attributes of a Test: Choosing an Appropriate Test

One of the basic limitations in clinical research is the biological variation among subjects as well as disease manifestation. Consequently, *no test can be described as perfect*. There will always be correct and incorrect results. These correct and incorrect instances of a test can be classified into four categories:

1. True positive results: When the test *correctly detects* disease in a diseased individual.
2. True negative results: When the test *correctly denies* disease in a nondiseased individual.
3. False positive results: When the test *incorrectly detects* disease in a nondiseased individual.
4. False negative results: When the test *incorrectly denies* disease in a diseased individual.

These categories can be presented in a tabular form and this is referred to as the '2 × 2 Table'.

It is quite obvious that a test that has high true positive results and true negative results, and consequently, low false positive and false negative results is most accurate and useful for the clinician. However, this does not always happen and hence, the clinician has to choose the tests based on the understanding of the attributes of the test and the clinical requirement. Before we explain this concept of choosing the test, let us understand the statistical terminology used in the description of the *test attributes*.

1. *Sensitivity* This is the proportion of true positive results among all diseased individuals. It is otherwise called *true positive rate* or *detection rate.*
2. *Specificity* This is the proportion of true negative results among all nondiseased individuals. It is otherwise called true negative rate.

Both sensitivity and specificity can be expressed either as percentage or fraction. The complement of sensitivity is false negative rate (i.e., false negative rate = 100 – sensitivity %) and that of specificity is false positive rate (i.e., false positive rate = 100 – specificity %).

An important concept to understand here is that sensitivity and specificity (and therefore, its complements) are not affected by the disease prevalence since they are attributes of the test itself.

Let us discuss these concepts in a real paper published in the fetal medicine literature. In their paper describing the incidence of hypoplastic or absent nasal bone among midtrimester fetuses that underwent amniocentesis for a variety of indications, Cicero et al. [1] reported these numbers that can be arranged as shown in Table 1. We can see that the sensitivity of the test (nasal bone status) is 21/34 = 61.7 % and the specificity is 970/982 = 98.7 %.

It is important to note that both sensitivity and specificity (and their complements, false negativity, and false positivity) are attributes of the test per se. These attributes should be considered when choosing between different tests in different clinical situations. In general, a high sensitivity test is to be used as a screening test and a high specificity test as a diagnostic test.

Interpreting the Test Result: Obtaining Post-test Probability

As we had seen earlier, the application of a test result to a set of possible diagnoses changes their probability of being present in an individual patient, i.e., the post-test probability changes. In clinical practice, clinician often act upon this posttest probability rather than the test result itself. In this regard, when a test returns a result, the clinician has to be aware of two probability related attributes of the *test result*: positive predictive value (PPV) and negative predictive value (NPV).

In the above example, we see that of 33 fetuses that test positive for abnormal nasal bone, 21 have the disease (Down syndrome). Therefore, the proportion of positive test result fetuses being truly diseased is 21/33 = 63.6 %. This is the PPV of the test in *this population*.

Also, of the 983 fetuses that tested negative, 970 were truly nondiseased. Therefore, the proportion of negative test result fetuses being truly nondiseased is 970/983 = 98.6 %, or the negative predictive value of the test *in this* population.

Table 1 Key results of the Cicero et al. paper, in 2 × 2 table presentation [1]

Test result	Down syndrome	No Down syndrome	Total
Test positive (nasal bone absent/hypoplastic)	21	12	33
Test negative (nasal bone present)	13	970	983
	34	982	1016

The test being nasal bone status, a positive test here is absent or hypoplastic nasal bone

Sensitivity of the test = 21/34 = 61.7 %; Specificity = 970/982 = 98.7 %

Positive predictive value (of a test result) = 21/33 = 63.6 %; Negative predictive value = 970/983 = 98.6 %

Alternatively, we can say, that among this group of fetuses, when the nasal bone test is positive (i.e., nasal bone is hypoplastic) the post-test probability of Down syndrome is 63.6 % and when the test is negative (i.e., nasal bone is not hypoplastic) the post-test probability of not being Down syndrome is 98.6 %.

The major pitfall in the utility of PPV and NPV is that these values change with the prevalence of the disease in question. Thus for the same test, with the same test attributes (sensitivity and specificity), the test result attributes (positive and negative predictive values) change with the background prevalence. In this example, we know that the background prevalence of Down syndrome changes with maternal age [2] and hence, the exact PPV and NPV from this particular study cannot be applied to the test results of women of significantly different age group.

Likelihood Ratios (LR)

Another attribute of the test result that is particularly useful in fetal medicine is likelihood ratios. The concept is intuitively simple to understand. It simply refers to the likelihood of the disease being present when the test result is positive (or disease not present when the test result is negative). A deeper understanding of the concept enables one to compare the utility of different tests and to understand the implications of this attribute.

There are different ways of arriving at the likelihood ratio. The simplest way of obtaining is to find the ratio of the prevalence of a positive test among diseased individuals to that among nondiseased individuals. In the above example,

$$\begin{aligned}&\text{Likelihood ratio of a positive nasal bone test (LR+)}\\&= [21/34]/[12/982]\\&= [\text{sensitivity}]/[100-\text{specificity}\,\%]\\&= 61.7/1.2 = 51.4\end{aligned}$$

We can clearly see that likelihood ratio calculated here is a derivative of sensitivity and specificity which are prevalence independent. Therefore, the LR+ calculated is also *prevalence independent* and hence applicable to different populations with different background risks (or background prevalence).

With this LR+ value, we understand that the disease is about 51 times more likely if the nasal bone test was positive (i.e., hypoplastic nasal bone). However, there is a small statistical adjustment to be made before the post-test *probability* can be arrived at.

Let us take the case of a 35-year-old mother presenting at midtrimester with hypoplastic nasal bone in the fetus. We know the background risk for this age is about 1 in 300, or a pretest probability of 0.33 %. Now the LR+ for this test result is 51.4. In order to calculate the post-test probability, we cannot directly multiply the LR+ with the prevalence; we need to convert the pretest probability to pretest *odds.*

$$\begin{aligned}\text{Pretest odds} &= \text{Probability}/(1-\text{Probability})\\&= 0.0033/0.9967\\&= 0.0033(\text{approx.})\end{aligned}$$

Now, LR+ can be used directly.

$$\begin{aligned}\text{Post-test odds} &= \text{Pretest odds} \times \text{LR}\\&= 0.0033 \times 51.4\\&= 0.17\end{aligned}$$

From this post-test odds, we can recalculate the posttest *probability*

$$\begin{aligned}\text{Posttest probability} &= \text{Posttest odds}/(1+\text{Posttest odds})\\&= 0.17/(1.17)\\&= 0.15\end{aligned}$$

Or a risk of 15 in 100 (1 in 7).

On the same note, we can calculate what happens to the disease post-test probability when the test is negative (i.e., nasal bone is not hypoplastic). We see from the above example,

Negative LR− = Prevalence of negative finding in diseased population/prevalence of negative finding in nondiseased population

$$= [13/34]/[970/982]$$
$$= [100 - \text{Sensitivity}\,\%]/[\text{Specificity}]$$
$$= 38/98.7$$
$$= 0.39$$

Intuitively, we can understand that when nasal bone is not hypoplastic, the disease is 0.39 times less likely to be present than the background prevalence. When we apply this into the above example,

$$\text{Post-test odds} = \text{Pretest odds} \times \text{LR}-$$
$$= 0.0033 \times 0.39 = 0.0013$$

$$\text{Post-test probability} = 0.0013/(1 + 0.0013) = 0.0013$$
$$= 13/10000 \text{ or } 1 \text{ in } 770$$

Having understood the direct application of the likelihood ratio of a test and its effects on the post-test probability of the disease, we will now see the other utility of likelihood ratio.

It is uncommon to find tests that can considerably affect the post-test probability to reach diagnostic threshold. This is especially true in fetal medicine, for example in screening for Down syndrome. Since likelihood ratios act on the post-test probability, we can use a series of tests (and their corresponding likelihood ratios) to arrive at a compound post-test probability. The reader is referred to the meta-analysis of second trimester ultrasound markers of trisomy 21 by Agathokleous et al. [3]. The pooled estimates of positive and negative likelihood ratios for the sonographic markers for trisomy 21 are given as follows [presented as marker, positive LR, negative LR]: [Intracardiac echogenic focus, 5.83, 0.80]; [Ventriculomegaly, 27.52, 0.94]; [Increased nuchal fold, 23.30, 0.80]; [Echogenic bowel, 11.44, 0.90]; [Mild hydronephrosis, 7.63, 0.92]; [Short femur, 3.72, 0.80]; [Aberrant right subclavian artery, 21.48,0.71]; and [Hypoplastic nasal bone. 23.2, 0.46]

In the above example, when a comprehensive genetic sonogram is performed and among the soft markers, hypoplastic nasal bone, mild hydronephrosis, and aberrant right subclavian artery are noted, the post-test probability is calculated as follows:

$$\begin{aligned}\text{Post-test odds} &= \text{Pretest odds} \times (\text{LR}_1+) \times (\text{LR}_2+) \\ &\times (\text{LR}_3+) \times (\text{LR}_4-) \times (\text{LR}_5-) \times (\text{LR}_6-) \\ &\times (\text{LR}_7-) \times (\text{LR}_8-) \\ &= 0.0033 \times 23.2 \times 7.63 \times 21.48 \times 0.80 \\ &\times 0.94 \times 0.80 \times 0.90 \times 0.80 \\ &= 5.44\end{aligned}$$

$$\text{Post-test probability} = 5.44/[1 + 5.44] = 5.44/6.44$$
$$= 0.84 \text{ or greater than } 1 \text{ in } 2$$

However, if the only positive finding was an echogenic bowel, then the post-test probability will be calculated as follows:

$$\begin{aligned}\text{Post-test odds} &= 0.0033 \times 11.44 \times 0.8 \times 0.94 \times 0.80 \\ &\times 0.92 \times 0.8 \times 0.71 \times 0.46 \\ &= 0.0055\end{aligned}$$

$$\text{Post-test probability} = 0.0055/1.0055 = 0.0055$$
$$= 55/10000 = 1 \text{ in } 180$$

Another instance commonly encountered can also be used to illustrate the usefulness of LRs. If, for example, the woman in the previous example has undergone the second trimester screening test (quadruple screening test or triple screening test) at, say, 16 weeks and has a final risk of 1 in 270, a detailed genetic sonogram can help in this rather equivocal situation. The final risk from the biochemical screening becomes the pretest probability for applying the LRs from the soft markers (1 in 270 or 0.0037).

Therefore, in this fetus if two soft markers were to be present such as mild hydronephrosis and increased nuchal fold (and other markers are negative), we can calculate the post-test probability as follows:

$$\text{Pretest odds} = 0.0037/(1 - 0.0037) = 0.0037$$

$$\begin{aligned}\text{Post-test odds} &= \text{Pretest odds} \times \text{LRs} = 0.0037 \times 23.3 \\ &\times 7.63 \times 0.90 \times\ 0.80 \times 0.94 \times 0.80 \\ &\times 0.46 \times 0.71 = 0.12\end{aligned}$$

Post-test probability = 0.12/1.12 = 0.11, which is 11 in 100 or 1 in 9, where an invasive testing is strongly justified.

If the fetus were to show no softmarkers at all after a detailed genetic sonogram (negative for all markers), then

$$\begin{aligned}\text{Post-test odds} &= 0.00370 \times 0.80 \times 0.92 \times 0.90 \times 0.80 \\ &\times\ 0.94 \times 0.80 \times 0.46 \times 0.71 = 0.0005\end{aligned}$$

Post-test probability = 0.0005/1.0005 = 0.0005, which is 5 in 10,000 or 1 in 2000, and therefore, the couple can be counseled based on these figures.

Simplifying the LR Approach

In actual practice, the conversion of probability to odds and the reconversion back to probability may appear daunting to the busy practitioner. Several work-arounds are available. In most fetal medicine units, the LRs of soft markers can be incorporated into their reporting software that can then compound the LRs and give the adjusted posterior risk (Fig. 2).

For tests results that cannot be incorporated into a software, conversion of pretest probability directly to post-

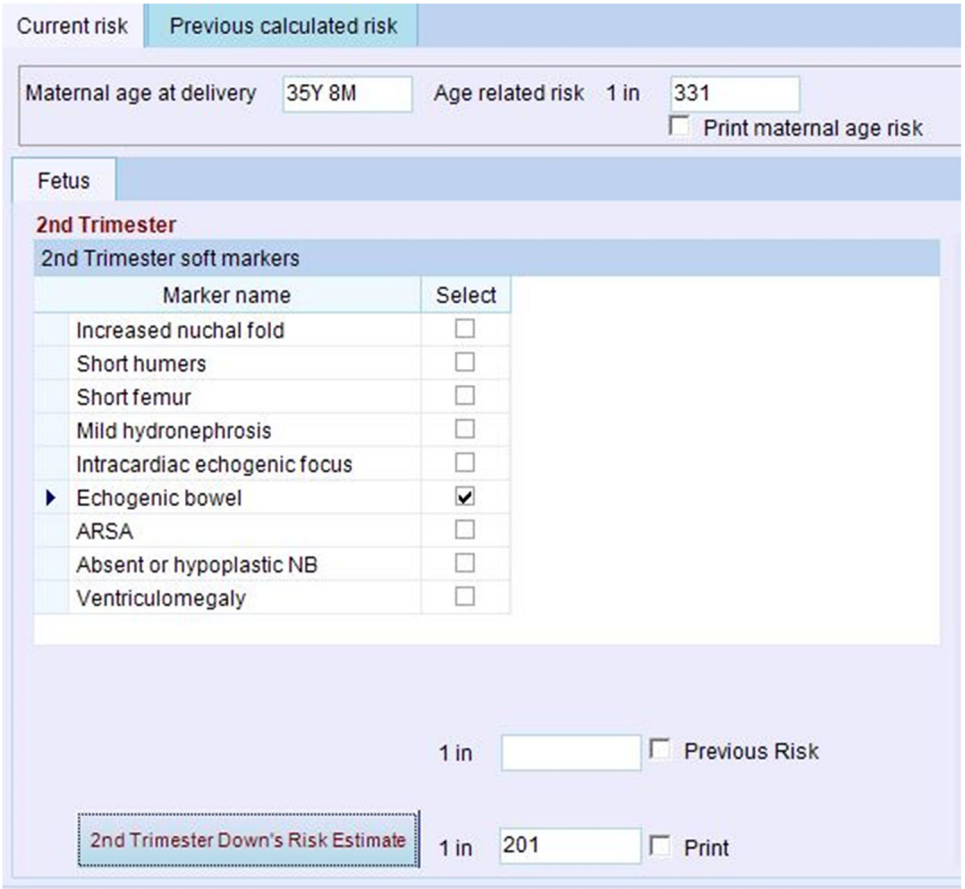


Fig. 2 Use of ultrasound image archiving and reporting software that has integrated likelihood ratio calculators for each soft marker. Physician simply checks the *boxes* corresponding to the soft markers present and final risk is calculated instantly [Sonocare™, with permission from publishers]. The software can be obtained by contacting the publishers: MEDIALOGIC SOLUTIONS (P) LTD., 197, Dr. Natesan Road, Mylapore, Chennai – 600004. http://medialogicindia.com/

test probability can be achieved by the use of Fagan nomogram [4] as shown in Fig. 3. Firstly, the physician has to determine the pretest probability of the diagnosis in question and join this probability with the LR of the test result. Extending this line up to the post-test probability line will give the final probability of the diagnosis. There are online resources that can compute the sensitivity, specificity, likelihood ratios, and therefore, post-test probabilities of diagnosis that can be of help to the busy practitioner. One such resource is from the department of medical education of the University of Illinois, Chicago at: http://araw.mede.uic.edu/cgi-bin/testcalc.pl.

Another way of simplifying the LR is to make use of the linear relationship between probability and logarithm of odds, when the probability falls between 10 and 90 %. This means that the post-test probability can be obtained by adding a constant value. McGee [5] presented the simplified, rule of thumb approach for the utilization of LR by remembering just three values of LR: 2, 5, and 10.

When the LR value is 1, this adds nothing to the post-test probability. For LR of value 2, 5, and 10, respectively, the *increment* in post-test probability is the first three multiples of 15 (i.e., 15, 30 and 45 %, respectively)—this needs to be added to the pretest probability. When the LR is less than 1, the decrement in post-test probability can be obtained by simply inverting the same three critical values i.e., 1/2, 1/5, and 1/10 (or LRs of 0.5, 0.2, and 0.1): a decrease of 15, 30 and 45 %, respectively from the pretest probability. These are approximate values and are useful for quick bedside application.

Tests with Continuous Results

In many instances, a test result will be a continuous output rather than just positive or negative. Common examples include blood pressure, blood glucose levels, etc. In such tests, a cutoff value is used to define positive and negative test result. There is a certain tool in biostatistics that can be employed to find out the right cut-off value that will optimize the correct identification of cases (i.e., true positive cases) while minimizing the false identification (false positive cases)—receiver operating characteristic curve (ROC curve).

In order to construct this curve, firstly the ROC space is defined. This space is a unit square, with length of any side equal to 1. Recall that the sensitivity or specificity of any test cannot exceed 100 % (or in fractions, cannot exceed

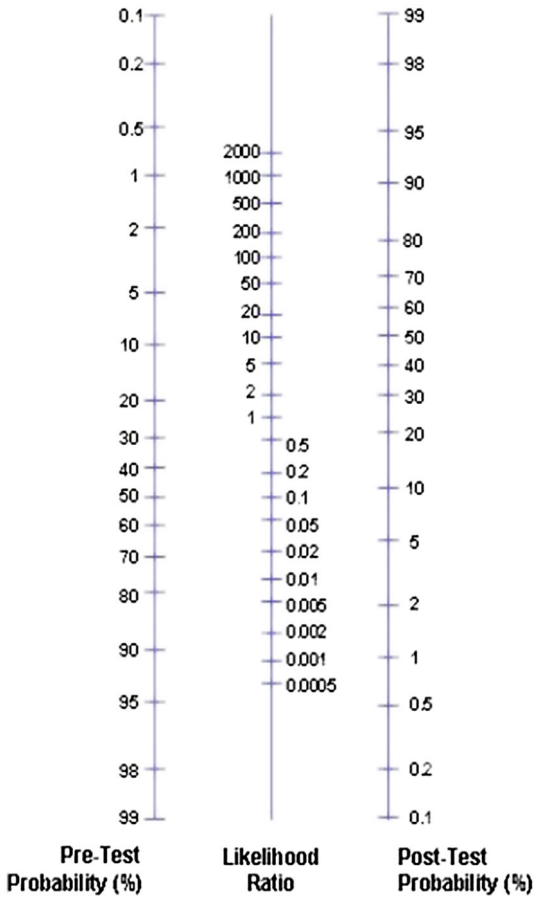


Fig. 3 The Fagan nomogram. A *straight-line* is drawn connecting the pretest probability of the disease in question with the likelihood ratio of the applied test result and extended further rightwards to meet the post-test probability line

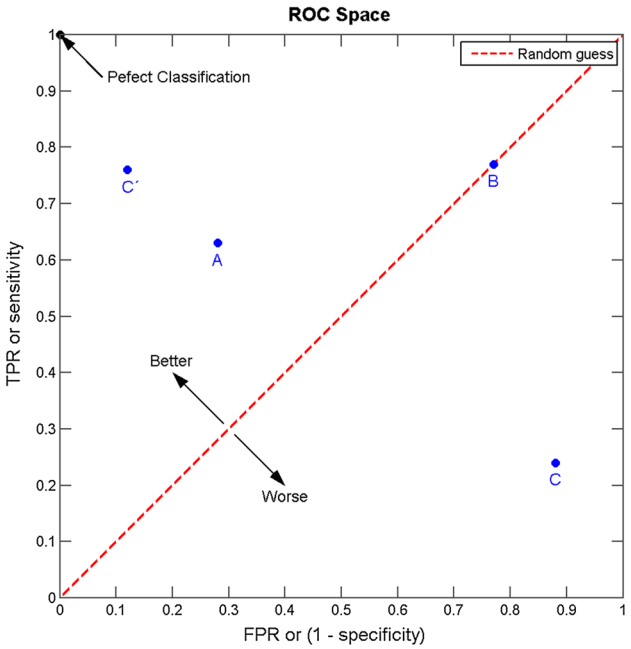


Fig. 4 The ROC space—a square of area 1 (or 100, if the parameters are expressed in percentages). The false positive rate of the test forms the X-axis and the true positive rate forms the Y-axis adapted from https://commons.wikimedia.org/wiki/File:ROC_space-2.png

1). Now the X-axis of this unit square is calibrated with false positive rate (or 1 − specificity, in fraction) and the Y-axis of this unit square is calibrated with true positive rate (or sensitivity, in fraction) as shown in the Fig. 4.

Let us now take the example of the results from the paper published by Arya et al. [6]. In their retrospective analysis of fetuses suspected to be at risk of coarctation of aorta, Arya et al. attempted to study if the distance between the left common carotid and left subclavian artery (LSCA) would be a good test to differentiate fetuses that developed neonatal coarctation of aorta.

Since distance is a continuous variable, we use the ROC curve to find out the optimal cut off for this test. In all, 40 fetuses had LSCA distance measured. Of these fetuses, 20 had coarctation and 20 did not have coarctation, in post-natal follow-up. For each measured value, if kept as test positive, we can generate sensitivity and specificity (and therefore, 1 − specificity) in diagnosing (predicting) coarctation. In the ROC space, each of these measured values of distance can be plotted against its sensitivity and 1 − specificity (or 100 − specificity, if expressed in percentage; Fig. 5). Note the diagonal line starting from [0, 0] and reaching [100, 100] coordinates. This line represents the line of no-utility of the test. A test that plots to the left of this line has a better diagnostic value than mere guessing. The farther from the line, the more useful is the test. The perfect test would lie on the top left corner (0, 100). Hence for a given test, we choose the test result that lies closest to the top left corner as the best cut off point to declare the test as positive or negative (discriminating point). In this example, a distance of 4.5 mm between the left subclavian and left common carotid has the best combination of sensitivity and false positivity (100 − specificity). Thus, ROC curve is useful to identify an optimal cutoff value of a continuous test result to declare the test as positive or negative.

Another important use of ROC curve is to compare the diagnostic values of two tests that have continuous results. We know that the area of the ROC space is 1 (since it is a square of side 1). The curve drawn using the test results covers an area under it. This is referred to as the "area under the curve" or AUC. In a perfect test, this AUC is 1. In practice, this is usually less than 1. Intuitively, we can understand that the area under the line of no- discrimination is 0.5.

Now, to compare the diagnostic utility of two tests, we simply compare the AUC of the two tests. In the same

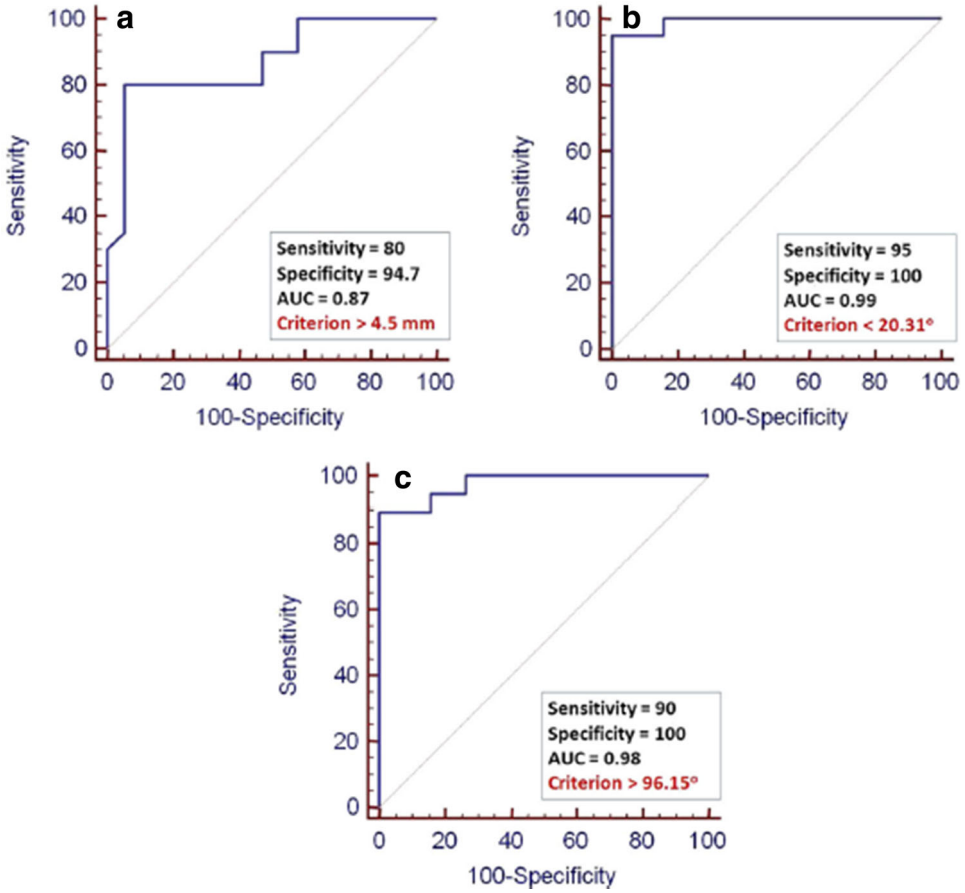


Fig. 5 ROC curves for **a** *left* common carotid to left subclavian artery distance, **b** ascending aorta–descending aorta angle, **c** transverse aorta–descending aorta angle reproduced with permission from the publishers [6]

paper, the authors have tried to assess the diagnostic potential of two other tests, namely the angle between ascending aorta and descending aorta (AAo–DAo angle) and the angle between transverse aorta and descending aorta (TAo–DAo angle). The areas under the curve for these two tests were 0.99 and 0.98, respectively, while the AUC for LSCA was 0.87 (Fig. 5). Therefore, we can easily conclude that the LSCA test performs least among the three.

Conclusion

The concept of pretest and post-test probability is inherent in medical training. While most physicians understand it in a tacit way, understanding it explicitly and objectively will help in their choosing, evaluating, and applying the appropriate diagnostic test in the appropriate clinical situation using the appropriate decision models. Apart from sensitivity and specificity, other useful test result related parameters such as positive and negative likelihood ratios and ROC curve should be utilized to optimize clinical effectiveness.

Compliance with Ethical Standards

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Cicero S, Sonek JD, McKenna DS, Croom CS, Johnson L, Nicolaides KH. Nasal bone hypoplasia in trisomy 21 at 15–22 weeks' gestation. Ultrasound Obstet Gynecol. 2003;21(1):15–8.
2. Griffiths AJ, Lowry RB, Renwick DH. Down's syndrome and maternal age in British Columbia, 1972–1975. Environ Health Perspect. 1979;31:9–11.
3. Agathokleous M, Chaveeva P, Poon LC, Kosinski P, Nicolaides KH. Meta-analysis of second-trimester markers for trisomy 21. Ultrasound Obstet Gynecol. 2013;41(3):247–61.
4. Fagan TJ. Letter: nomogram for Bayes theorem. N Engl J Med. 1975;293(5):257.
5. McGee S. Simplifying likelihood ratios. J Gen Intern Med. 2002;17(8):646–9.
6. Arya B, Bhat A, Vernon M, Conwell J, Lewin M. Utility of novel fetal echocardiographic morphometric measures of the aortic arch in the diagnosis of neonatal coarctation of the aorta. Prenat Diagn. 2016;36(2):127–34.