HOME

ROC Curve

Many full term births in a hospital require induction of labor. Induction succeeds in most cases but fails in a few. In case the induction fails, a Cesarean is done for delivery. This involves pain, time and money but also requires mental preparedness.

It would be nice for both the woman and family and the attending obstetrician to anticipate a cesarean delivery on the basis of patient characteristics. The traditional method is to compute Bishop score based on dilatation, effacement, consistency and position. Other parameters that influence the success of induction of labor are maternal age, parity, BMI and amniotic fluid index. A study was carried out in *n* =166 cases with pre labor rupture of membrane to find if the duration since rupture can help in predicting the cesarean delivery. Although the study was prospective but sensitivity and specificity were calculated for different durations of rupture. The data obtained are shown in Table 1.

Table 1: Sensitivity and (1 – specificity) for Cesarean delivery at different duration of rupture of membrane

Duration (hr) greater than or equal to	Sensitivity	1 - specificity
.00	1.000	1.000
.63	1.000	.976
.88	1.000	.969
1.25	1.000	.890
1.75	1.000	.866
2.13	1.000	.819
2.38	1.000	.811
2.75	1.000	.780
3.25	1.000	.717
3,75	1.000	.709
4.50	1.000	.646
5.13	.971	.583
5.38	.971	.575
5.75	.971	.551
6.25	.914	.378
6.75	.914	.346
7.13	.857	.291
7.38	.857	.283
7.75	.857	.276
8.25	.800	.189

8.75	.800	.181
9.25	.743	.110
9.75	.743	.102
10.25	.543	.039
10.75	.543	.031
11.50	.457	.024
12.50	.400	.008
13.50	.343	.000
14.50	.286	.000
15.50	.257	.000
16.50	.200	.000
17.50	.171	.000
18.50	.143	.000
19.50	.114	.000
20.25	.057	.000
21.50	.000	.000

The ROC curve obtained by plot at different cut-offs is shown in Figure 1. A statistical software found that the area under the curve is *C* = 0.898 with SE = 0.029 and 95% CI from 0.841 to 0.956.

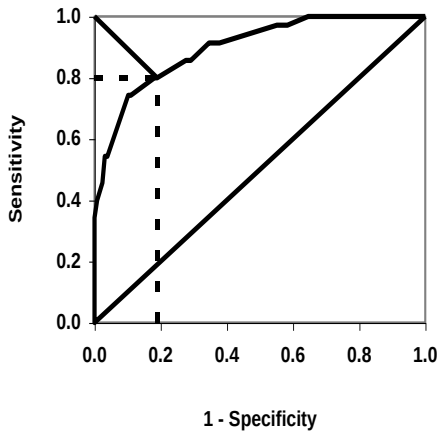

FIGURE 1: ROC Curve of duration since rupture membrane for Cesarean delivery

It seems from the ROC that duration since rupture of membrane itself is a good indicator to anticipate Cesarean delivery. The best cut-off that maximizes (sensitivity + specificity) is 8.25 hours. At this duration, the sensitivity is 0.80 and specificity is 0.81 (1 – specificity = 0.19).

Although not shown above, the Bishop score was not found to be as good an indicator of impending cesarean as was duration since rupture in this example. Otherwise also,

Bishop's score is subjective and nonreproducible because of higher inter- and intra-observer variability.

The example illustrates how a statistical tool such as ROC can be effectively used for medical decisions. The example is illustrative only and should not be construed to mean that duration since rupture can solely be used to anticipate cesarean. For this, studies in different locales are needed.

For those interested, the details of ROC curve are as follows.

DETAILS

Medical tests play a vital role in modern medicine not only for confirming the presence of disease but also to rule out the disease in individual patient. A test test with two outcome categories such as test+ and test- is known as dichotomous, whereas more than two categories such as positive, indeterminate and negative called polytomous test. The validity of a dichotomous test compared with the gold standard is determined by sensitivity and specificity. These two are components that measure the inherent validity of a test. For details of sensitivity and specificity, see http://www.medicalbiostatistics.com/Sensitivity-specificity.pdf. These require that the disease status is already known and the ability of the test to correctly identify positives and negatives is assessed. Thus, sensitivity and specificity do not assess diagnostic efficiency of the tests as made out in certain texts.

A test is called continuous when it yields numeric values such as bilirubin level and nominal when it yields categories such as Mantoux test. Sensitivity and specificity can be calculated in both cases but ROC curve is applicable only for continuous test or at least ordinal with many categories.

The receiver operating characteristic (ROC) curve is the plot that displays the full picture of trade-off between the sensitivity (true positive rate) and (1- specificity) (false positive rate) across a series of cut-off points. Area under the ROC curve is considered as an effective measure of inherent validity of a diagnostic test. This curve is useful in (i) evaluating the discriminatory ability of a test to correctly pick up diseased and non-diseased subjects; (ii) finding optimal cut-off point to least misclassify diseased and non-diseased subjects; (iii) comparing efficacy of two or more medical tests for assessing the same disease; and (iv) comparing two or more observers measuring the same test (inter-observer variability).

Non-parametric and parametric methods to obtain area under the ROC curve

Statistical softwares provide non-parametric and parametric methods for obtaining the area under ROC curve. The user has to make a choice. The following details may help.

Non-parametric methods are distribution-free and the resulting area under the ROC curve is called empirical. First such method uses trapezoidal rule. If sensitivity and specificity are denoted by s_n and s_p, respectively, the trapezoidal rule calculates the area by joining the points (s_n, 1 – s_p) at each interval value of the continuous test and draws a straight line joining the x-axis. This forms several trapezoids and their area can be easily calculated and summed. Another non-parametric method uses Mann-Whitney statistics, also known as Wilcoxon rank-sum statistic and the c-index for calculating area. Both these non-parametric methods of estimating AUC estimate have been found equivalent (1).

Parametric methods are used when the statistical distribution of test values in diseased and non-diseased is known. Binormal distribution is commonly used for this purpose. This is applicable when both diseased and non-diseased test values follow normal distribution. If data are actually binormal or a transformation such as log, square or Box-Cox makes the data binormal then the relevant parameters can be easily estimated by the means and variances of test values in diseased and non-diseased subjects. For details, see (2).

The choice of method to calculate AUC for continuous test values essentially depends upon availability of statistical software. Binormal method produces the smooth ROC curve, further statistics can be easily calculated but gives biased results when data are degenerate and distribution is bimodal (3-4). When software for both parametric and non-parametric methods is available, conclusion should be based on the method that yields greater precision of estimate of inherent validity, namely, of AUC.

Interpretation of ROC curve

Total area under ROC curve is a single index for measuring the performance a test. The larger the AUC, the better is overall performance of the medical test to correctly identify diseased and non-diseased subjects. Equal AUCs of two tests represents similar overall performance of tests but this does not necessarily mean that both the curves are identical. They may cross each other.

Figure 1 depicts three different ROC curves. Considering the area under the curve, test A is better than both B and C, and the curve is closer to the perfect discrimination. Test B has good validity and test C has moderate.

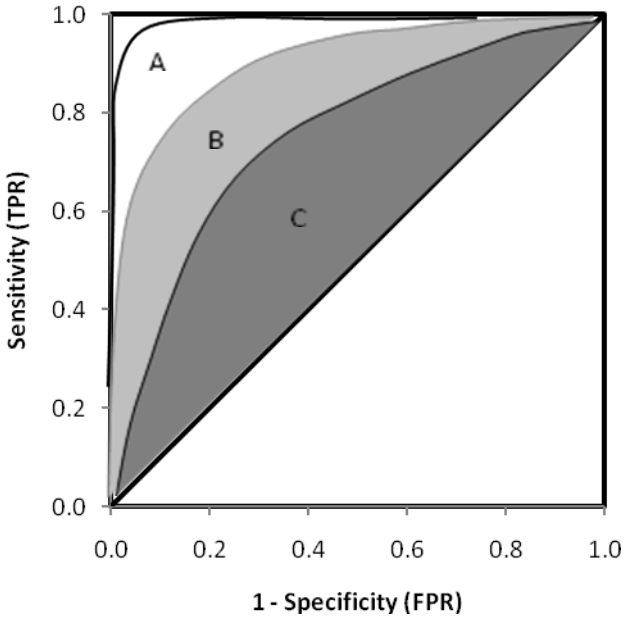

Figure 1: Three ROC curves with different areas under the curve

Figure 2(a) has hypothetical ROC curves of two medical tests A and B applied on the same subjects to assess the same disease. Test A and B have nearly equal area but cross each other. Test A performs better than test B where high sensitivity is required, and test B performed better than A when high specificity is needed.

(a) (b)

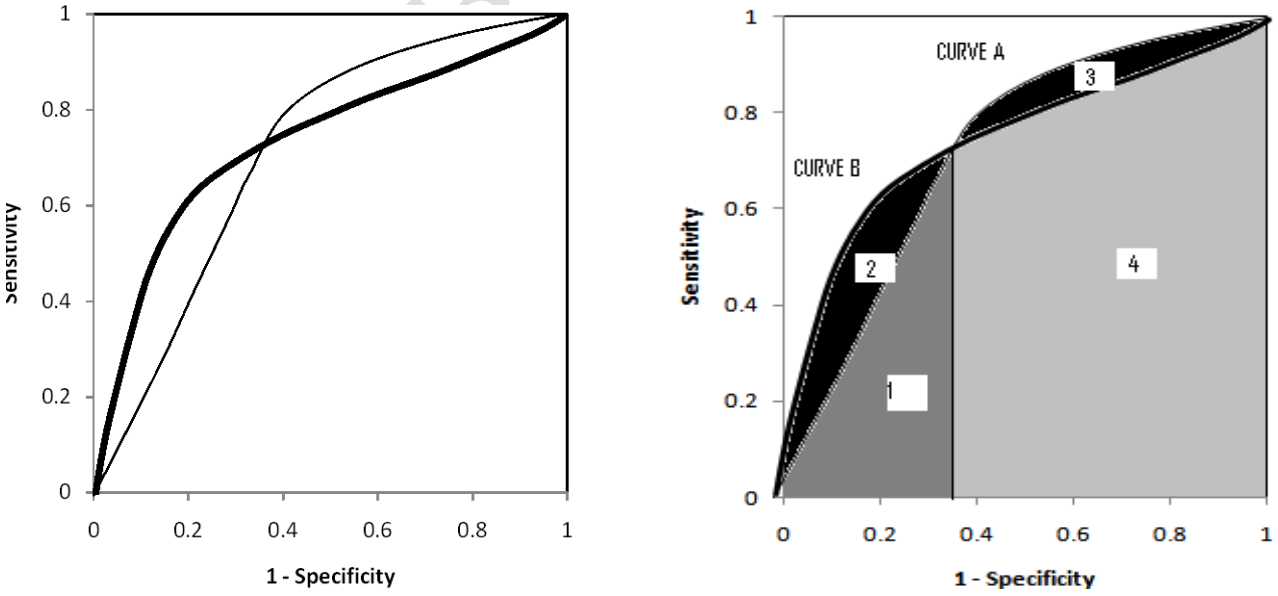

Figure 2: (a) Two ROC curves crossing each other but with nearly same area
(b) Illustration of partial area under the ROC curve

In such cases and in some other situations, the interest may be restricted to specific values of sensitivity or specificity. You may be interested in a test with high specificity as for a disease with grave prognosis (cancer). Then the interest will be in test B and that too for specificity ≥ 0.65 or (1 – specificity) < 0.35. In that case, the area of interest is 1+2 as shown in Figure 2(b). This is called partial area under the curve. Software such as STATA calculates this also and, if you want for easy interpretability, you can standardize it to 1 by considering total area of box upto (1 – specificity) equal to 1.

Methods to find the 'optimal' threshold point

Three criteria are used to find optimal threshold point from ROC curve. First two methods give equal weight to sensitivity and specificity and impose no ethical, cost, and no prevalence constraints. The third criterion considers cost which mainly includes financial cost for correct and false diagnosis, cost of discomfort to person caused by treatment, and cost of further investigation when needed. This method is rarely used in medical literature because it is difficult to estimate the respective costs and prevalence is often difficult to assess. These three criteria are known as points on curve closest to the (0, 1), Youden index, and minimize cost criterion, respectively.

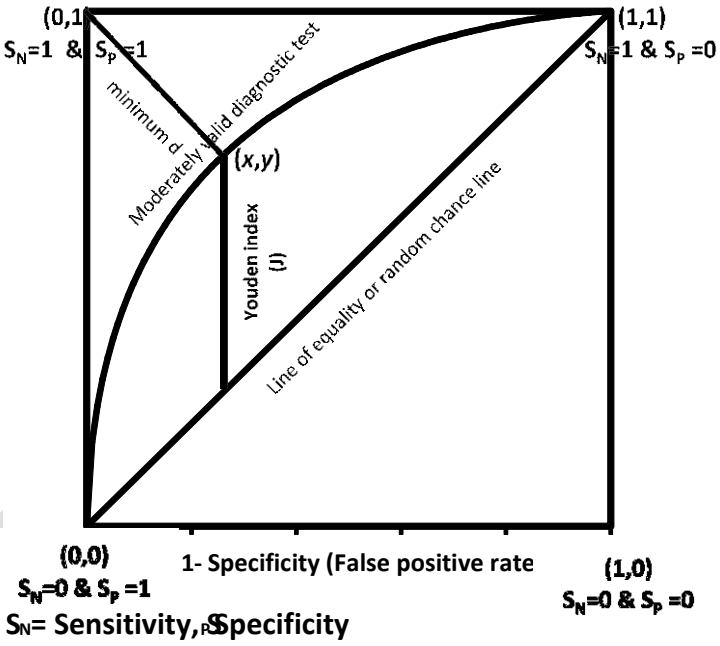

Figure 3: Finding best cut-off from the ROC curve

If s_n and s_p denote sensitivity and specificity, respectively, the distance between the point (0, 1) and any point on the ROC curve is $d = \sqrt{[(1 - s_n)^2 + (1 - s_p)^2]}$. To obtain the

optimal cut-off point to discriminate the disease with non-disease subject, calculate this distance for each observed cut-off point, and locate the point where the distance is minimum. Most of the ROC analysis softwares calculate the sensitivity and specificity at all the observed cut-off points allowing you to do this exercise.

The second is Youden index that maximizes the vertical distance from line of equality to the point [*x*, *y*] as shown in Figure 3. The *x* represents (1- specificity) and *y* represents sensitivity. In other words, the Youden index J is the point on the ROC curve which is farthest from line of equality (diagonal line). The main aim of Youden index is to maximize the difference between TPR (s_n) and FPR (1 – s_p) and little algebra yields J = max[s_n+s_p]. The value of J for continuous test can be located by doing a search of plausible values where sum of sensitivity and specificity can be maximum. Youden index is more commonly used criterion because this index reflects the intension to maximize the correct classification rate and is easy to calculate. Many authors advocate this criterion.

Biases that can affect the ROC curve results

We describe more prevalent biases in this section that affect the sensitivity and specificity, and thereby the ROC curve.

1. Gold standard: Validity of gold standard is important— ideally it should be error free and the medical test under review should be independent of the gold standard as this can increase the area under the curve spuriously. The gold standard can be clinical follow-up, surgical verification, biopsy or autopsy or in some cases opinion of panel of experts. When gold standard is imperfect, such as peripheral smear for malaria parasites, sensitivity and specificity of the test are would be affected.
2. Verification bias: This occurs when all disease subjects do not receive the same gold standard for some reason such as economic constraints and clinical considerations. For example, in evaluating the breast bone density as screening test for diagnosis of breast cancer and only those women who have higher value of breast bone density are referred for biopsy, and those with lower value but suspected are followed clinically. In this case, verification bias would overestimate the sensitivity of breast bone density test.
3. Selection bias: Selection of right patients with and without diseased is important because some tests produce prefect results in severely diseased group but fail to detect mild disease.
4. Test review bias: The clinician should be blind to the actual diagnosis while evaluating a test. A known positive disease subject or known non-disease subject may influence the test result.
5. Inter-observer bias: In the studies where observer abilities are important in diagnosis, such as for bone density assessment through MRI, experienced radiologist and junior radiologist may differ. If both are used in the same study, the observer bias is apparent.

6. Co-morbidity bias: Sometimes patients have other types of known or unknown diseases which may affect the positivity or negativity of test. For example, NESTROFT (Naked eye single tube red cell osmotic fragility test), used for screening of thalassaemia in children, shows good sensitivity in patients without any other hemoglobin disorders but also produces positive results when other hemoglobin disorders are present.
7. Uninterpretable test results: This bias occurs when test provide results which can not be interpreted and clinician excludes these subjects from the analysis. This results in overestimation of validity of the test.

It is difficult to rule out all the biases but you should be aware and try to minimize them.

Sample size

Adequate power of the study depends upon the sample size. Power is probability that a statistical test will indicate significant difference where certain pre-specified difference is actually present. In a survey of eight leading journals, only two out of 43 studies reported a prior calculation of sample size in diagnostic studies (5). In estimation set-up, adequate sample size ensures the study will yield the estimate with desired precision. Small sample size produces imprecise or inaccurate estimate, while large sample size is wastage of resources especially when the test is expensive.

Table 2: Sample size formula for estimating sensitivity and specificity and area under the ROC curve

Sl.no.	Problem	Formula	Description of symbol used
1	Estimating the sensitivity of test	$\frac{Z^2_{1-\alpha/2} S_N (1-S_N)}{\varepsilon^2 \times Prev}$	S_N = Anticipated sensitivity Prev = Prevalence of disease in population can be obtained from previous literature or pilot study ε = required absolute precision on either side of the sensitivity
2	Estimating the specificity of test	$\frac{Z^2_{1-\alpha/2} S_p (1-S_p)}{\varepsilon^2 \times (1-Prev)}$	S_N = Anticipated specificity Prev = Prevalence of disease in population can be obtained from previous literature or pilot study ε = required absolute precision on either side of the specificity.
3	Estimating the area under the ROC curve	$n_D = \frac{Z^2_{1-\alpha/2} \times V(AUC)}{\varepsilon^2}$ n_D = number of diseased subjects $n = n_D(1+k)$, k is ratio of prevalence of non-disease to disease subjects	V(AUC) = Anticipated variance function for area under ROC curve (**This is not the variance**. For large samples, Variance of AUC = V(AUC)/No. of positives) ε = required absolute precision on either side of the area under the curve.

$Z_{1-\alpha/2}$ is a standard normal value and α is the confidence level. $Z_{1-\alpha/2}$ = 1.645 for α=0.10 and $Z_{1-\alpha/2}$= 1.96 for α=0.05.

The sample size formula depends upon whether interest is in estimation or in testing of the hypothesis. Table 2 provides the required formula for estimation of sensitivity, specificity and AUC. These are based on the binormal distribution or asymptotic assumption (large sample theory) which is generally used for sample size calculation.

Variance of AUC can be obtained by using parametric and non-parametric methods for inserting into formula 3 in Table 2. This may also be available in literature on previous studies. If no previous study is available, a pilot study is done to get some workable estimates to calculate sample size. For pilot study data, appropriate statistical software can provide estimate of this variance.

Formulas of sample size for testing hypothesis on sensitivity-specificity or the AUC with a pre-specified value and for comparison on the same subjects or different subjects are complex. Refer (2) for details.

There are many more topics for interested reader to explore such as combining the multiple ROC curve for meta-analysis, ROC analysis to predict more than one alternative, ROC analysis in the clustered environment, and for tests repeated over time, etc. For these see (2,6).

Acknowledgement

This is an abridged version of the article “Introduction to ROC Curve for Medical Researchers” submitted to Indian Pediatrics for publication. For full details, see http://www.indianpediatrics.net/apr2011/277.pdf.

References

1. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143:29-36.
2. Zhou Xh, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York: John Wiley and Sons, Inc, 2002.
3. Faraggi D, Reiser B. Estimating of area under the ROC curve. Stat Med 2002; 21:3093-3106.
4. Hajian Tilaki KO, Hanley JA, Joseph L, Collet JP. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnosis tests. Med Decis Making 1997; 17:94-102.
5. Bachmann LM, Puhan MA, ter Riet G, Bossuyt PM. Sample sizes of studies on diagnostic accuracy: literature survey. BMJ 2006; 332:1127-1129.
6. Kester AD, Buntinx F. Meta analysis of curves. Med Decis Making 2000; 20:430-439.