FEATURES: STATISTICS PRIMER

Understanding receiver operator characteristic (ROC) curves

Robin M Turner, Claire Cameron, Ari Samaranayaka

Receiver operating characteristic (ROC) curves summarise graphically the trade-off between sensitivity and specificity for diagnostic tests.[1] We will describe how to interpret these graphs, but first we need to understand how we assess diagnostic test accuracy and why we are interested in these concepts of sensitivity and specificity.

Diagnostic tests diagnose whether a person has a particular disease or not. They vary in how well they perform. We start by considering what we call the gold standard; this tells us whether the person truly has the disease or not. The gold standard may be imperfect (that is a whole other area of research),[2] but is the best test we have. For instance, histopathology might be considered the gold standard test for deciding if a person has cancer or not, or glycated haemoglobin (HbA1c) for testing for diabetes.

The diagnostic test that we want to estimate the accuracy of is compared to the gold standard test. For example, we may have a new (hypothetical) cancer test and we could take 100 people, some who have cancer and some who do not, and we would apply both the new cancer test and the gold standard of histopathology. We can group people into testing positive or negative with the new cancer test, versus which truly have cancer or not based on the gold standard. Table 1 shows the cross classification of people by the new test and gold standard. If the test is positive and they have cancer, this is considered a true positive result. If they test negative and do not have cancer, this is a true negative result. The new test may also give an incorrect result; if the test is positive but they do not have cancer, this is a false positive result and if the test is negative but they do have cancer, this is a false negative result. We can measure how many correct results there are by measuring the percentage correctly classified (the diagonal of the table): 100% x (45 + 40)/100 = 85%.

Table 1: The number of people testing positive or negative by cancer status for a total sample of 100 people undergoing both the new test and gold standard.

New test	Gold standard		
	Cancer	No Cancer	Total
Test positive	45	10	55
Test negative	5	40	45
Total	50	50	100

There are four measures of interest: sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Sensitivity measures the percentage of people who test positive out of all of those who truly have cancer. From Table 1, this is: 100% x 45/50 = 90%. Of all people who have cancer, the test will be positive for 90% of them (and negative for 10%). Specificity measures the percentage of people who test negative out of all of those who do not have cancer. From Table 1, this is: 100% x 40/50 = 80%. Of all people without cancer, 80% will test negative (and 20% will test positive). Sensitivity and specificity do not depend on the prevalence of the disease (in this example the prevalence is 50% as half have cancer and half do not) because they are estimated separately for those with cancer and those without cancer. Once there is a test result, the PPV and NPV are useful measures. The PPV estimates the percentage of people who have the disease out of all those who test positive, and similarly the NPV estimates the percentage of people who do not have the disease out of all those who test negative. For our example, the PPV is: 100% x 45/55 = 81.8% and the NPV is: 100% x 40/45 = 88.9%. The PPV and NPV vary as the prevalence of disease changes, so are less useful as measures of test accuracy. For example, if the test were used in a different population with an increased prevalence, the PPV would increase and the NPV would decrease, but sensitivity and specificity would remain the same.

The example above assumes a binary test, i.e. the new diagnostic test provides a positive or negative result, but many tests have a continuous result and a threshold is needed to define whether a test is positive or negative. The cancer test may measure the level of an antigen in the blood, which can take on any value between 0 and 10. Table 2 shows how many people have different antigen levels (the new hypothetical test) and whether or not they had cancer based on the gold standard test.

Table 2: Antigen level (new test) by cancer status for the sample of 100 people

	Number of people per antigen level	
Antigen level	Cancer	No Cancer
0	0	8
1	0	9
2	1	7
3	1	6
4	0	5
5	3	5
6	10	5
7	13	4
8	12	1
9	7	0
10	3	0
Total	50	50

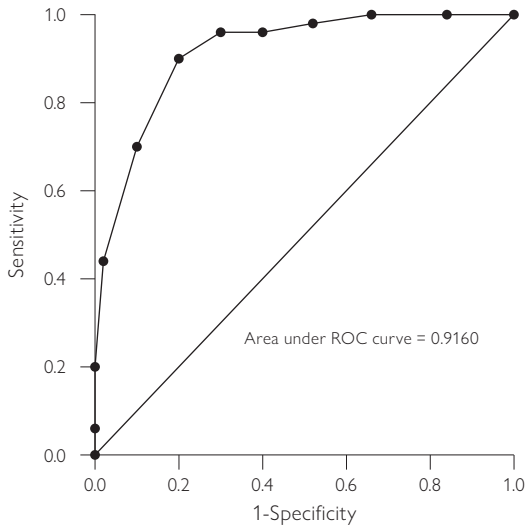
If we choose a threshold of greater than or equal to 6 to be considered a positive test, then we end up with the numbers shown in Table 1, i.e. anyone with an antigen level of 6 or more is considered to be positive. In our example, 45 people with cancer had a test result of 6 or more and only ten of those without cancer had a test result of 6 or more. We can vary the threshold for a positive test from

greater than or equal to 0 (where all people test positive), through to greater than 10 (where all people test negative) and calculate the sensitivity and specificity at each threshold. This is shown in Table 3. We can see that as the threshold increases, sensitivity decreases and specificity increases. Changing the threshold will always change one at the expense of the other. It is not possible to increase both sensitivity and specificity by altering the threshold.

Table 3: Sensitivity and specificity for each different test positivity threshold.

Threshold	Sensitivity (%)	Specificity (%)
(>= 0)	100	0
(>= 1)	100	16
(>= 2)	100	34
(>= 3)	98	48
(>= 4)	96	60
(>= 5)	96	70
(>= 6)	90	80
(>= 7)	70	90
(>= 8)	44	98
(>= 9)	20	100
(>= 10)	6	100
(> 10)	0	100

Figure 1: ROC curve for varying thresholds of antigen level compared to the gold standard histopathology.



The ROC curve plots sensitivity against 1-specificity to show this trade off. Figure 1 shows the ROC curve for our example. The points show the sensitivity and 1-specificity pairs from the different thresholds (note the graph is showing proportions not percentage as we have used previously, we use these interchangeably). The diagonal line represents if we decided randomly whether the test was positive or negative. Tests with poor accuracy will lie close to this line, while tests with high accuracy will be heading towards the upper left corner. A perfect test would have a sensitivity of 1 and a specificity of 1, which would lie in the top left corner.

The area under this curve can tell us how well the test is discriminating between those with the disease and those without the disease. A poor test lying on the diagonal line will have an area under the ROC curve of 0.5; a perfect test will have an area of 1. Most tests will lie in between. Our example cancer test has an area of 0.916, indicating it has very good discrimination between cancer and non-cancer.

The ROC curve allows us to make decisions about where a threshold might best be chosen. It is important to note that choosing a threshold is a clinical decision based on whether sensitivity or specificity is more important. If sensitivity is more important, a lower threshold might be chosen. This will minimise false negatives, but will come with worse specificity and thus an increased number of false positives. Increasing the threshold will do the opposite.

In summary, ROC curves have an important use in showing how a test performs against the gold standard across a range of thresholds. It allows easy assessment of which threshold might be better for a particular situation, and the area under the curve gives an estimate of the discriminative ability of the test.

References

1. Pepe, M S. The statistical evaluation of medical tests for classification and prediction. Oxford: Oxford University Press, 2003.

2. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. Health Technol Assess. 2007;11(50).

About the authors

› Associate Professor Robin Turner, BSc(Hons), MBiostat, PhD, is the Director, Centre for Biostatistics, Division of Health Sciences, University of Otago

› Dr Claire Cameron, BSc(Hons), DipGrad, MSc, PhD, is a Senior Research Fellow and Biostatistician, Centre for Biostatistics, Division of Health Sciences, University of Otago

› Dr Ari Samaranayaka, BSc, MPhil, PhD, is a Senior Research Fellow and Biostatistician, Centre for Biostatistics, Division of Health Sciences, University of Otago