Receiver Operating Characteristics

Receiver operating characteristics are a powerful statistical modeling tool which can be used in medical decision making, particularly when setting threshold values for tests. Receiver operating characteristic curves are a graphical plot of sensitivity of a test Vs 1-specificity for a binary classifier system as the test threshold varies.

Some Basic Statistics

Before we start, I will quickly explain some important basic statistical terms necessary to understanding and using ROC curves.

TPF – True positive fraction – The **proportion** of people **with** a disease who test **positive.** TPF = **Sensitivity**

FNF – False negative fraction – The **proportion** of people **with** a disease who test **negative.** TPF + FNF = 1

TNF – true negative fraction – the **proportion** of people **without** a disease who test **negative**.   TNF = **Specificity**

FPF – false positive fraction – the **proportion** of people **without** a disease who test **positive**.   TNF + FPF = 1

P = probability

| = given that

P (T+ **|** D+) – probability of positive test if patient has disease = sensitivity

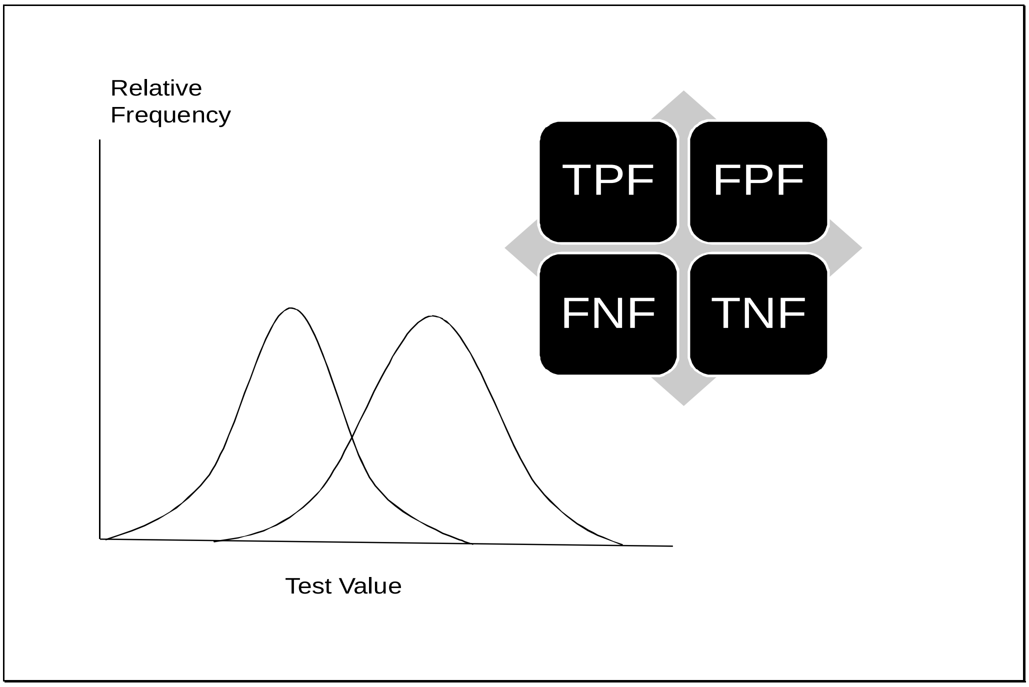
P (D+) = prevalence

Receiver Operating Characteristics

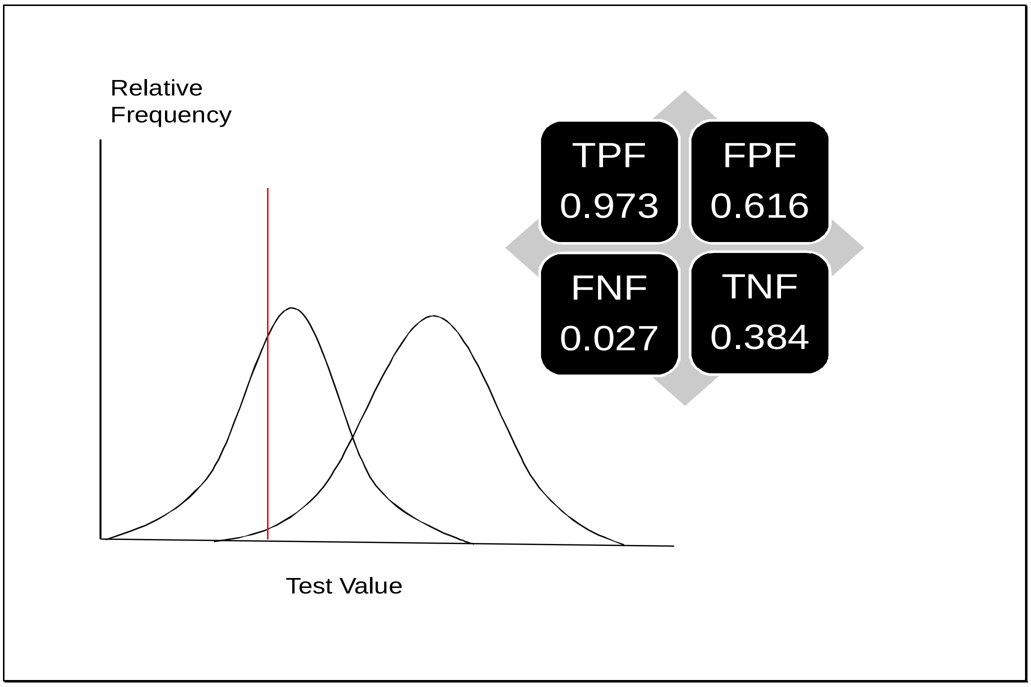
Receiver operating characteristics can help you decide on where to set a test threshold but to do this you need to be able to compare your test against a gold standard. We could then plot the results in a simple truth table

	Disease Positive	Disease Negative
Test Positive	**TPF**	**FPF**
Test Negative	**FNF**	**TNF**

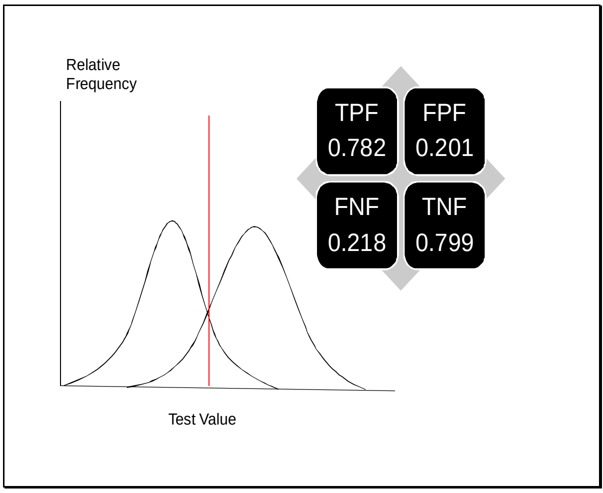
Imagine we have 2 populations – one with a disease and one without the disease. We apply our test to both populations and plot the test results as 2 overlapping histograms…

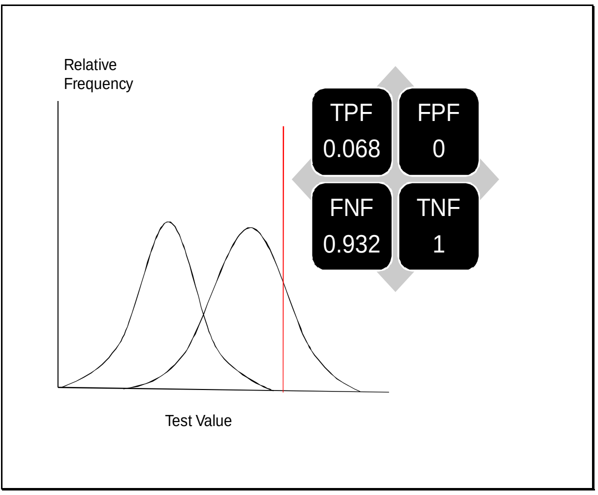


We then set a test threshold and generate our TPF, FPF, FNF and TNF fractions

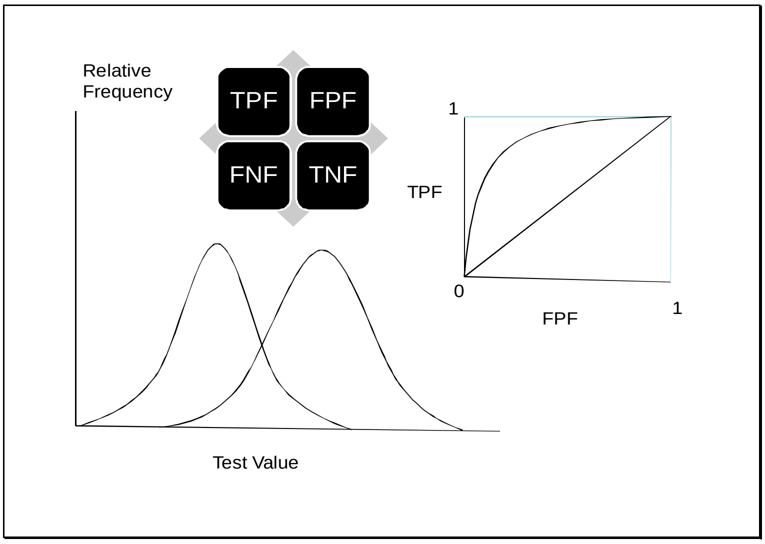


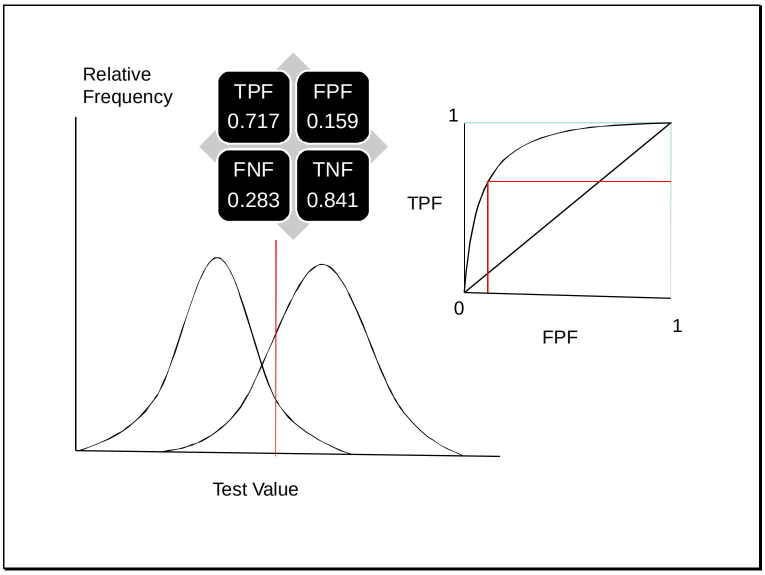
Or we could set a different test threshold. The higher we set the threshold the more true negatives we get BUT we also get more false negatives.



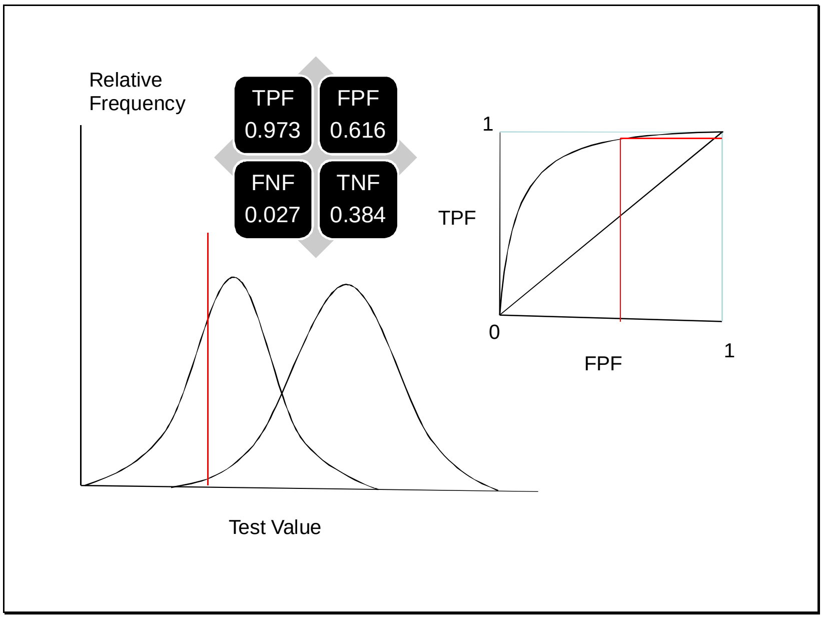


The receiver operating curve is simply a plot of TPF against FPF (or sensitivity against 1-specificity) as we move the test threshold. This gives us a graphical representation of the usefulness of the test across an entire range of possible thresholds.





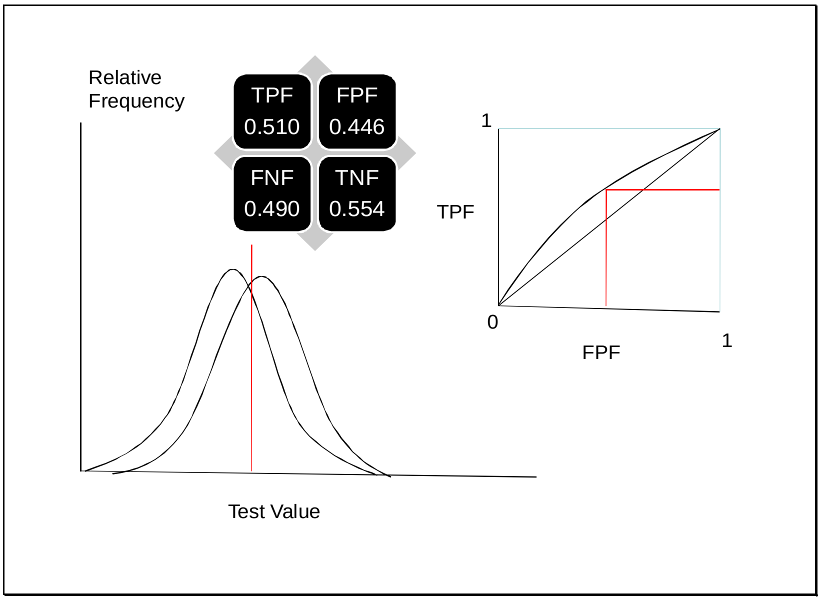
Now watch as we move the test threshold from right to left across the histogram – the ROC plot will move from left to right i.e. if the threshold is high, there will be very few false positives but also few true positives (TPF and FPF close to zero). As we move the threshold down (left on histogram), TPF increases (quickly at first), but the FPF also begins to increase and gathers speed as the threshold increases.



So all the receiver operating characteristic is, is a simple graphical way of modeling what happens to TPF and FPF (sensitivity and 1-specificity) as you move a test threshold. It’s as simple as that!!!

Making and Comparing Receiver Operating Curves

Imagine now we have 2 populations – with and without a disease – but this time our test is a lot less good at predicting the presence or absence of the disease. Our ROC curve might look more like this.



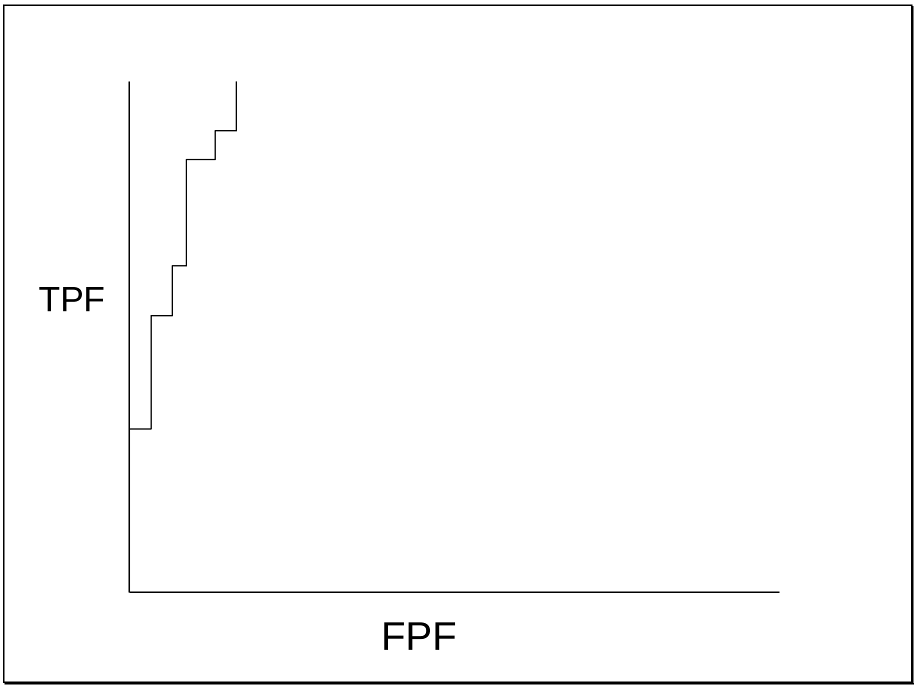
As populations overlap more and more in terms of test result, so the ROC curve flattens towards a diagonal line. This closeness to the midline on an ROC curve is a useful property of your test to know. It is traditionally measured as the **Area Under Curve** or **AUC**. A completely useless test (i.e. the diagonal midline) has an AUC of 0.5. A perfect test would have an AUC of 1.0.

How to make an ROC Curve

Rank your data according to test result from highest to lowest and compare against gold standard…

Patient number	D Dimer	DVT (y/n)
1	5012	y
2	4989	y
3	4590	y
4	4434	n
5	3946	y
6	3922	y
7	3860	n

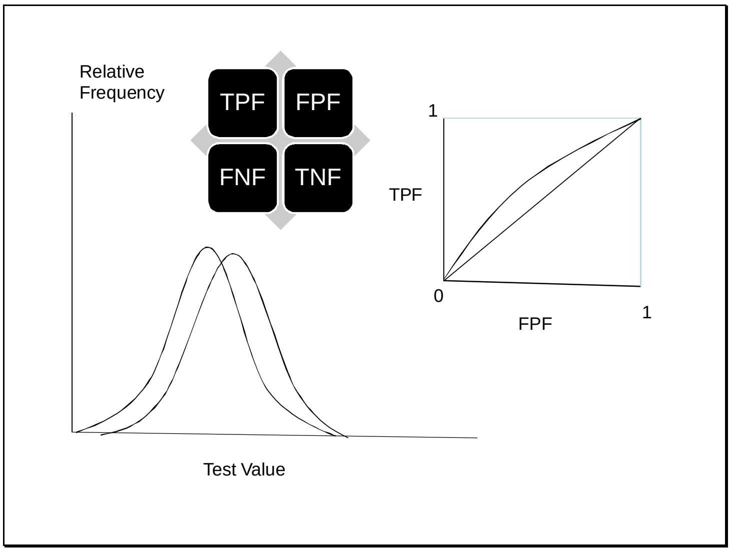
Examine the largest result. We set the threshold *just below* this large result (the red marker moves left on the histogram). If this first result belongs to a patient with the disease then this is a *true positive*. The TPF on the ROC graph must now increase so we plot the first ROC point by moving up the screen and plotting a point. Now set the test threshold to just below the second point and again plot a point on the ROC curve. If this patient does not have the disease we have a *false positive* and we plot a point on the ROC graph directly right of the previous point. We then continue this process for all results. The very first part of our ROC curve may look like this...



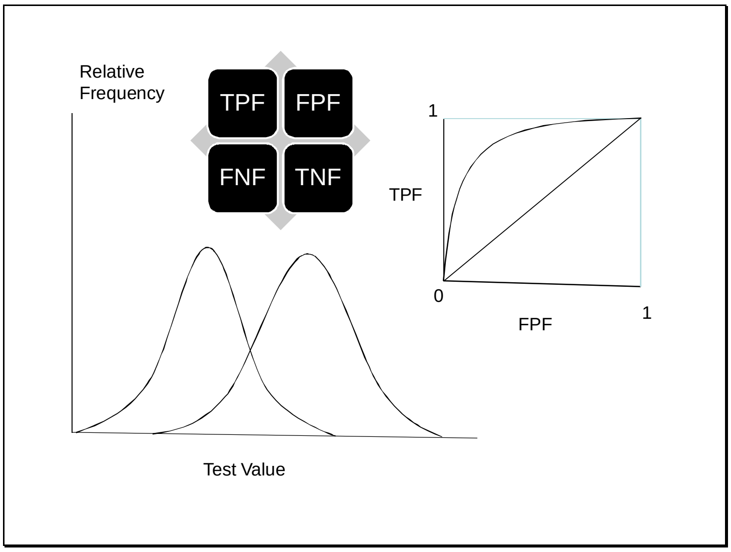
But when all the results are plotted it will resemble a curve with (hopefully) a steep gradient at first which later approaches the horizontal gradient.

Comparing ROC curves

Consider 2 tests A and Z. Test A is good at discriminating between populations with and without a particular disease. Test Z is a poor discriminator. We plot the ROC curve of test Z. As we move our threshold left picking off true positives and false positives, the likelihood of encountering a true positive is roughly the same as encountering a false positive. Our curve goes up in a more or less diagonal line.



Whereas for test A, the early likelihood of encountering a true positive is much higher so the curve begins to climb much more steeply. Only later as we run out of true positives do we start to pick off the false positives and the curve approaches the diagonal.



The area under curve (AUC) is obviously much higher for test A.

As the curve is in fact a long series of small vertical and horizontal steps, the AUC is easily calculated. Every time the curve takes a horizontal step, simply calculate the height of curve (i.e. number of TP's so far/number of people with the condition) and multiply this by 1/number of people without disease). Or in other words you need to know the step width and multiply this by the height each time you hit a FP result on your ROC curve. Of course a stats program will do this for you!

Some Technical Aspects of ROC Curves

ROC curves are constructed by simply ranking the population according to their test result. Therefore the difference in test result between subjects 23 and 24 may not be the same as the difference between subjects 24 and 25. This means that the area under the ROC curve does not reflect the shapes of the underlying populations (i.e. normal or not normal) and the area under the curve is non parametric. This means that the AUC is a useful parameter regardless of the distribution of the underlying populations. It also means that AUC can be used even when a test result does not give an accurate number – as long as one can rank the results one can construct the curve. The test may be something as simple as a radiologist giving an opinion on whether an X ray shows a malignancy on a scale of 1-5 (1 = definitely normal 5 = definitely abnormal). Questionnaire based rating scales can also be used.

It is also possible to calculate the standard error of the Area Under Curve which can give you an idea of whether your AUC approximates to 0.5 or not. It can also be used to estimate necessary sample sizes and calculate confidence intervals. (although SE does depend on the shapes of the populations to some extent)

Standard error = the square root of:

$[A\ (1-A) + (n_a-1)(Q1 - A^2)+(n_n-1)(Q2 - A^2)]\ /\ n_a n_n$

n_a and n_n are the number of abnormals and normals respectively. A = Area under curve.

$Q1 = A/(2-A)$ and $Q2 = 2A^2 / (1 + A)$

It is possible to compare the differences between tests using ROC analysis using bivariate statistical analysis. The calculation depends on whether the tests were run on the same or different populations. Hanley and McNeil’s papers contain the statistical tools necessary to perform these calculations.

Sources of Error

Receiver operating characteristics are as prone to error as any other statistical tool, although there are a few things that are quite specific to this kind of analysis. ‘Random noise’ will result in misclassifications (i.e. of TP’s as FP’s or vice versa). The result of this is generally to degrade test performance rather than falsely inflate it.

Receiver operating characteristic calculations are crucially dependant on the fact that your test and the gold standard test are completely independant. Any interdependance will falsely inflate the AUC. Consider comparing the gold standard against itself – AUC = 1.0. But what if the gold standard is actually slightly imperfect – the AUC will remain 1.0 regardless. It is therefore extremely important that the two tests are performed and analysed completely seperately.

For other sources of error one should look for standard sources of error – was the full spectrum of disease considered, were the populations truly similar, was there any other comorbid factors influencing results, was there verification or diagnostic review bias, what was done with results that did not easily fit with the rest, or were uninterpretable. Was there any interobserver variation etc.

Some uses of ROC Curves

It is possible to use economic analysis to help decide on a test threshold using ROC curves. Plotting an ROC curve is basically plotting ‘hits’ against ‘false alarms’. We can limit the false alarms at the expense of fewer hits but the decision of where to set a threshold will depend on the relative costs of each (monetary and other costs).

When setting a threshold one should consider the following: Financial costs of treating a disease (present or not) or failing to treat a disease (present or not). Costs of further investigations, discomfort to patient of investigation and treatment and mortality and morbidity of treating or not treating. If cost of missing a disease is great and treatment relatively harmless one would set the threshold towards the right of the ROC curve (i.e. high TPF and FPF). If treatment risk is grave or treatment effect is limited, we would want to spare well subjects from the treatment and set the threshold towards te left.

The average cost resulting from the use of a diagnostic test could be said to be:
Cavg = C_0 + CTP*P(TP) + CTN*P(TN) + CFP*P(FP) + CFN*P(FN)

C_0 = overhead cost per test. CTP is cost associated with a true positive result and P(TP) is the probability of a true positive result

P(TP) = P(D+) x P(T+|D+) or in other words P(TP) = P(D+) x TPF (probability of a true positive equals prevalence x true positive fraction). Therefore:

Cavg = Co + CTP*P(D+)*P(T+|D+) + CTN*P(D-)*P(T-|D-) + CFP*P(D-)*P(T+|D-) +

CFN*P(D+)*P(T-|D+) or:

Cavg = Co + CTP*P(D+)*TPF + CTN*P(D-)*TNF + CFP*P(D-)*FPF +

CFN*P(D+)*FNF

But we can substitute TNF for 1 – FPF and FNF for 1 – TPF.

Cavg = Co + CTP*P(D+)*TPF + CTN*P(D-)*(1-FPF) + CFP*P(D-)*FPF +

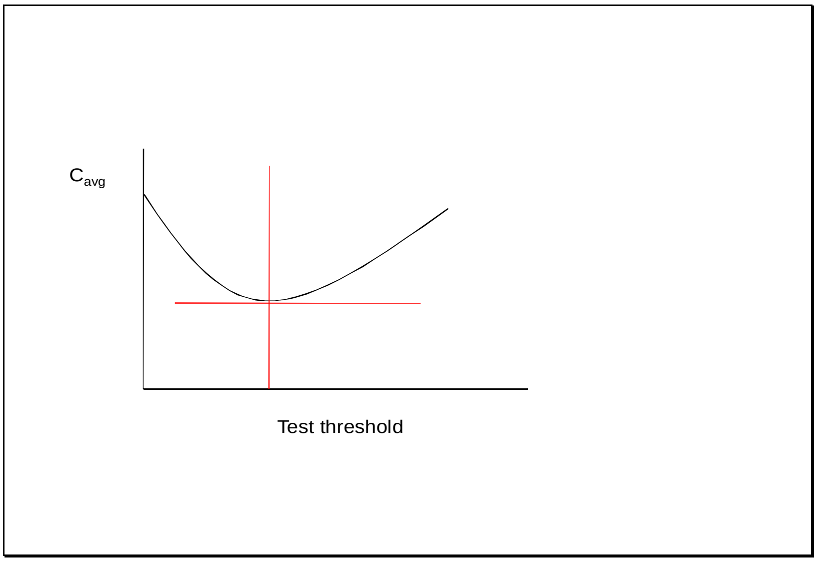
CFN*P(D+)*(1-TPF) or:

Cavg = TPF * P(D+) * { CTP - CFN } +FPF * P(D-) * { CFP - CTN } + Co +

CTN*P(D-) + CFN*P(D+)

So C_{avg} in fact depends on TPF and FPF – the co-ordinates on the ROC curve. Therefore average cost depends on the threshold set on the ROC curve. Varying the threshold alters the cost. We want to get C_{avg} as low as is possible

Using calculus we would predict that the average cost is lowest when the derivative (gradient) of the cost equation is 0. If threshold is too low, cost is high (too many expensive FP's). If threshold is high, cost is also high (too many expensive FN's). We would expect the cost equation to look something like this: (C_{avg} Vs threshold)



The red lines mark the spot where cost is least (and gradient = 0)

Using the ROC we can express TPF as a function of FPF.

Cavg = ROC(FPF) * P(D+) * { CTP - CFN } + FPF * P(D-) * { CFP - CTN } + Co

+ CTN*P(D-) + CFN*P(D+)

Differentiate this equation with respect to FPF:

dC/dFPF = dROC/dFPF * P(D+) * { CTP - CFN } + P(D-) * { CFP - CTN }

- then set dC/dFPF to zero and we get

  dROC/dFPF * P(D+) * { CTP - CFN } = - P(D-) * { CFP - CTN }

dROC/dFPF =

P(D-) * { CFP - CTN }

P(D+) * { CFN - CTP}

dROC/dFPF is the gradient of the ROC when costs are optimal

When costs are optimal dROC/dFPF (gradient of ROC curve) =

$$\frac{P(D-) * \{ CFP - CTN \}}{P(D+) * \{ CFN - CTP\}}$$

If a disease is very rare P (D-) / P (D+) will be very large. We should set our threshold near left hand side of ROC graph where gradient is large. This minimises false positives which can very quickly exceed the number of true positives (low PPV of test because of high ration of D- to D+ population). Conversely if disease is common, set threshold towards right (a more lenient threshold), otherwise we get high numbers of false negatives.

Disease prevalence is not the only factor having a profound effect on the optimal threshold. The curve slope will also be large if the cost difference is far greater for CFP – CTN than for CFN – CTP. Consider a test for a brain cancer – the cost of a FP may be very high (neurosurgery/rehabilitation and care) and the cost of a TP relatively low (they may die whether or not you operate). The reverse of this scenario also holds – if a treatment is cheap and harmless even if treating false positives it makes sense that the test threshold will be very lenient.

References

Hanley JA, McNeil BJ. **Radiology** 1982 143 29-36. *The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve*

Metz CE. **Semin Nuclear Med** 1978 VIII(4) 283-298. *Basic principles of ROC analysis*

Begg CB, McNeil BJ **Radiology** 1988 167 565-9. *Assessment of radiologic tests: control of bias and other design considerations*.

Swets JA. **Science** 1988 240 1285-93. *Measuring the accuracy of diagnostic systems*.

http://www.clinchem.org/cgi/reprint/39/4/561

http://www.anaesthetist.com/mnm/stats/roc/Findex.htm

Matthew Smith 2009