The magnificent ROC
(Receiver Operating Characteristic curve)

"There is no statistical test, however intuitive and simple, which will not be abused by medical researchers"

Introduction - A statistical prelude

ROC curves were developed in the 1950's as a by-product of research into making sense of radio signals contaminated by noise. More recently it's become clear that they are remarkably useful in medical decision-making. That doesn't mean that they are always used appropriately! We'll highlight their use (and misuse) in our tutorial. We'll first try to move rapidly through basic stats, and then address ROC curves. We'll take a practical, medical approach to ROC curves, and give a few examples.

If you know all about the terms 'sensitivity', 'specificity', FPF, FNF, TPF and TNF, *as well as* understanding the terms 'SIRS' and 'sepsis', you can click here to skip past the basics, but we wouldn't advise it! Once we've introduced ROCs, we'll play a bit, and then look at two examples - procalcitonin and sepsis, and also tuberculosis and pleural fluid adenosine deaminase. Finally, in a footnote, we examine accuracy, and positive and negative predictive values - such discussion will become important when we find out about costing, and how to set a test threshold.

Consider patients in intensive care (ICU). One of the major causes of death in such patients is "sepsis". Wouldn't it be nice if we had a quick, easy test that defined early on whether our patients were "septic" or not? Ignoring for the moment what sepsis *is*, let's consider such a test. We imagine that we take a population of ICU patients, and do two things:

1. Perform our magical TEST and record the results;
2. Use some "gold standard" to decide who REALLY has "sepsis", and record this result (in a blinded fashion).

Here are the results:

Actuality v the TEST		
	SEPSIS	*NO sepsis*
*"high" TEST** (positive)	**TPF**	**FPF**
*"low" TEST** (negative)	**FNF**	**TNF**
** "high" and "low" refer to TEST value relative to some arbitrary cutoff level!*		

Please note (note this well) that we have represented our results as *fractions*, and that:

```
FNF + TPF = 1
```

In other words, given FNF, the False Negative Fraction, you can *work out* TPF, the True Positive Fraction, and vice versa. Similarly, the False Positive Fraction and True Negative Fraction must also add up to one - those patients who really have NO sepsis (in our example) must either be true negatives, or misclassified by the test as positives despite the absence of sepsis.

In our table, **TPF** represents the number of patients who have sepsis, and have this corroborated by having a "high" TEST (above whatever cutoff level was chosen). **FPF** represents *false positives* - the test has lied to us, and told us that non-septic patients are really septic. Similarly, true negatives are represented by **TNF**, and false negatives by **FNF**.

In elementary statistical texts, you'll encounter other terms. Here they are:

- The **sensitivity** is how good the test is at picking out patients with sepsis. It is simply the True Positive Fraction. In other words, sensitivity gives us the proportion of cases picked out by the test, relative to all cases who actually have the disease.
- **Specificity** is the ability of the test to pick out patients who do NOT have the disease. It won't surprise you to see that this is synonymous with the True Negative Fraction.

Probability and StatSpeak

Not content with the above terms and abbreviations, statisticians have further confused things using the following sort of terminology:

```
P( T+ | D- )
```

Frightening, isn't it? Well, not when one realises that the above simply reads "the probability of the test being positive, given that the disease is not present". T+ is simply an abbreviation for "a positive test", and "D-" is similarly a shorthand for "the disease isn't present". P(something) is a well-accepted abbreviation for "the probability of the event *something*", and the vertical bar means "given that". Not too difficult!

So here are the translations:

Statement	*Translation*
P(T+ | D+)	sensitivity, =true positive fraction, =TPF
P(T- | D-)	specificity, TNF
P(T+ | D-)	FPF
P(T- | D+)	FNF

Using similar notation, one can also talk about the prevalence of a disease in a population as "P(D+)". Remember (we stress this again!) that the false negative fraction is the same as one *minus* the true positive fraction, and similarly, FPF = 1 - TNF.

KISS

We'll keep it simple. From now on, we will usually talk about TPF, TNF, FPF and FNF. If you like terms like sensitivity, specificity, bully for you. Substitute them where required!

Truth

Consider our table again:

Actuality v the TEST		
	SEPSIS	*NO sepsis*
*"high" TEST**(positive)	**TPF**	**FPF**

*"low" TEST**(negative)	**FNF**	**TNF**

See how we've assumed that we have absolute knowledge of who has the disease (here, sepsis), and who doesn't. A good intensivist will probably give you a hefty swipe around the ears if you go to her and say that you have an infallible test for "sepsis". Until fairly recently, there weren't even any good definitions of sepsis! Fortunately, Roger Bone (and his committee) came up with a fairly reasonable definition. The ACCP/CCM consensus criteria [Crit Care Med 1992 20 864-74] first define something called the Systemic Inflammatory Response Syndrome, characterised by at least two of:

1. Temperature under 36^{o}C or over 38^{o}C;
2. Heart rate over 90/min;
3. Respiratory rate over 20/min *or* PaCO2 under 32 mmHg;
4. White cell count under 4000/mm^3 *or* over 12000/mm^3 *or* over 10% immature forms;

The above process is often abbreviated to "SIRS". The consensus criteria then go on to define **sepsis**:

When the systemic inflammatory response syndrome is the result of a confirmed infectious process, it is termed 'sepsis'.

Later, they define 'severe sepsis' (which is sepsis associated with organ dysfunction, hypoperfusion, or hypotension. "Hypoperfusion and perfusion abnormalities may include, but are not limited to lactic acidosis, oliguria, or an acute alteration in mental status"). Finally, 'septic shock' is defined as sepsis with hypotension, despite adequate fluid resuscitation, along with the presence of perfusion abnormalities. Hypotension is a systolic blood pressure under 90 mmHg *or* a reduction of 40(+) mmHg from baseline.

The above definitions have been widely accepted. Now, there are many reasons why such definitions can be criticised. We will not explore such criticism in detail but merely note that:

1. The definition of SIRS appears to be over-inclusive (Almost all patients in ICU will conform to the definition at some time during their stay);
2. Various modifications of the third criterion (respiratory rate) have been used to accommodate patients on mechanical ventilation;
3. The use of high *or* low values for temperature and white cell count appears to exclude patients who might be 'in transition' from low to high, or high to low values!
4. Proof that SIRS "is the result of an infectious process" may be difficult or impossible to achieve. '*Proof*' of anything in ICU (as opposed to 'showing an association') is particularly difficult because of the multiple problems experienced by patients. (Quite apart from the philosophical problems posed by 'proof')!
5. It may be difficult to establish whether infecting organisms are present. Even if adequate quantities of culture material have been collected *at the right time*, and before antibiotics have been started, *and* your microbiology laboratory maintains superb standards of quality control, infecting organisms may still be missed. Some have even claimed that organisms (or their toxic products) enter the portal vein and cause sepsis, but don't get into the systemic circulation!
6. Evidence of the presence of bacteria in an organ or tissue (say lung, or blood) is not evidence that the bacteria are causing the patient's systemic illness. Ventilated patients are often *colonised* by bacteria, without being infected; intravascular lines may likewise be colonised without the bacteria necessarily causing SIRS.

Despite the above limitations, one needs some starting point in defining sepsis, and we will use the ACCP/SCCM criteria. Our problem then becomes one of differentiating between patients with SIRS *without* evidence of bacterial infection, and patients who "truly" have sepsis. (We will **not** here examine whether certain patients have severe systemic infection *without* features of SIRS).

The magnificent ROC!

Remember that, way back above, we said that our TEST is "positive" if the value was above some arbitrary cutoff, and "negative" if below? Central to the idea of ROC curves (receiver operating characteristic, otherwise called 'relative operating characteristic' curves) is this idea of a cutoff level. Let's imagine that we have two populations - septic and non-septic patients with SIRS, for example. We have a TEST that we apply to each patient in each population in turn, and we get numeric results for each patient. We then plot histograms of these results, for each population, thus:

If you can read this message, then your browser is almost certainly not Java enabled. To view the acid-base calculator, get a Java-enabled browser!

Play around with the above simple applet - move the (green) demarcating line from low to high (left to right), and see how, *as you move the test threshold from left to right*, the proportion of false positives decreases. Unfortunately, there is a problem - as we decrease the false positives, so the true positives also decrease! As an aside, note how we have drawn the curve such that where the curves overlap, we've shaded the overlap region. This is ugly, so in future, we'll leave the overlap to your imagination, thus:

Now we introduce the magnificent ROC! All an ROC curve is, is an exploration of what happens to TPF and FPF as we vary the position of our arbitrary TEST threshold. (AUC refers to the Area under the curve and will be discussed later).

Watch how, as you move the test threshold from right to left using the 'slider' bar at the bottom, so the corresponding point on the ROC curve moves across from *left to right*! Why is this? Simple. If our threshold is very high, then there will be almost no *false* positives .. but we won't really identify many *true*positives either. Both TPF and FPF will be close to zero, so we're at a point low down and to the left of the ROC curve.

As we move our test threshold towards a more reasonable, lower value, so the number of true positives will increase (rather dramatically at first, so the ROC curve moves steeply up). Finally, we reach a region where there is a remarkable increase in *false positives* - so the ROC curve slopes off as we move our test threshold down to ridiculously low values.

And that's really that! (We will of course explore a little further).

Playing with ROCs

In this section we will fool around with ROCs. We will:

1. Create ROC curves;
2. Find out why the area under the ROC curve is non-parametric, and why this is important;
3. Learn to calculate required sample sizes;
4. Compare the areas under two ROC curves;
5. Examine the effects of noise, a bad 'gold standard', and other sources of error.

Let's play some more. In the following example, see how closely the two curves are superimposed, and how flat the corresponding ROC curve is! This demonstrates an important property of ROC curves - the greater the overlap of the two curves, the smaller the area under the ROC curve.

Vary the curve separation using the upper "slider" control, and see how the ROC curve changes. When the curves overlap almost totally the ROC curve turns into a diagonal line from the bottom left corner to the upper right corner. What does this mean?

Once you've understood what's happening here, then the true power of ROCs will be revealed. Let's think about this carefully..

Let's make an ROC curve

Consider two populations, one of "normal" individuals and another of those with a disease. We have a test for the disease, and apply it to a mixed group of people, some with the disease, and others without. The test values range from (say) zero to a very large number - we rank the results in order. (We have rather arbitrarily decided that patients with bigger test values are more likely to be 'diseased' but remember that this is not necessarily the case. Of the thousand possibilities, consider patients with low serum calcium concentrations and hypoparathyroidism - here the low values are the abnormal ones). Now, here's how we construct our curve..

1. Start at the bottom left hand corner of the ROC curve - here we know that both FPF and TPF must be zero (This corresponds to having the green 'test threshold' line in our applet way over on the right);
2. Now examine the largest result. In order to start constructing our ROC curve, we set our test threshold at *just* below this large result - we move the green marker slightly left. Now, if this, the first result, belongs to a patient *with* the disease, then the case is a *true positive*, the TPF must now be bigger, and we plot our first ROC curve point by moving UP on the screen and plotting a point. Conversely, if the disease is absent, we have a *false positive*, the FPF is now greater than zero, and we move RIGHT on the screen and plot our point.
3. Set the test threshold lower, to just below the second largest result, and repeat the process described in (2).
4. .. and so on until we've moved the threshold down to below the lowest test value. We will now be in the upper right hand corner of the ROC curve - because our green threshold marker is below the lowest value, all results will be classified as positive, so the TPF and FPF will both be 1.0 !

Consider two tests. The first test is *good* at discriminating between patients with and without the disease. We'll call it test A. The second test is lousy - let's call it test Z. Let's examine each:

- **Test Z**. Because this is a lousy test, as we move our green marker left, picking off either false or true positives, our likelihood of encountering either is much the same. For every true positive (that moves us UP) we are likely to encounter a *false* positive that moves us to the RIGHT, as we plot the graph. You can see what will happen - we'll get a more-or-less diagonal line from the bottom left corner of the ROC curve, up to the top right corner.

- **Test A**. This is a good test, so we're initially more likely to encounter *true*positives as we move our green marker left. This means that initially our curve will move steeply UP. Only later, as we start to encounter fewer and fewer true positives, and more and more false positives, will the curve ease off and become more horizontal!

From the above, you can get a good intuitive feel that the closer the ROC curve is to a diagonal, the less useful the test is at discriminating between the two populations. The more steeply the curve moves up and then (only later) across, the better the test. A more precise way of characterising this "closeness to the diagonal" is simply to look at the **AREA** under the ROC curve. The closer the area is to 0.5, the more lousy the test, and the closer it is to 1.0, the better the test!

The Area under the ROC curve is non-parametric!

The real beauty of using the area under this curve is its simplicity. Consider the above process we used to construct the curve - we simply *ranked* the values, decided whether each represented a true or false positive, and then constructed our curve. It didn't matter whether result number 23 was a zillion times greater than result number 24, or 0.00001% greater. We certainly didn't worry about the 'shapes of the curves', or any sort of curve parameter. From this you can deduce that the area under the ROC curve is not significantly affected by the shapes of the underlying populations. This is most useful, for we don't

have to worry about "non-normality" or other curve shape worries, and can derive a single parameter of great meaning - the area under the ROC curve!

> We're about to get rather technical, so you might wish to skip the following, and move on to the nitty gritty!

In an authoritative paper, Hanley and McNeil [Radiology 1982 143 29-36] explore the concept of the area under the ROC curve. They show that there is a clear similarity between this quantity and well-known (at least, to statisticians) Wilcoxon (or Mann-Whitney) statistics. Considering the specific case of randomly paired normal and abnormal radiological images, the authors show that the area under the ROC curve is a *measure of the probability* that the perceived abnormality of the two images will allow correct identification. (This can be generalised to other uses of the AUC). Note that ROC curves can be used even when test results don't necessarily give an accurate number! As long as one can *rank* results, one can create an ROC curve. For example, we might rate x-ray images according to degree of abnormality (say 1=normal, 2=probably normal, and so on to 5=definitely abnormal), check how this ranking correlates with our 'gold standard', and then proceed to create an ROC curve.

Hanley and McNeil explore further, providing methods of working out *standard errors* for ROC curves. Note that their estimates for standard error (SE) depend to a degree on the shapes of the distributions, but are conservative so even if the distributions are not normal, estimates of SE will tend to be a bit too large, rather than too small. (If you're unfamiliar with the concept of standard error, consult a basic text on statistics).

In short, they calculate standard error as

$$SE = \sqrt{\frac{A(1-A) + (n_a-1)(Q1 - A^2)+(n_n-1)(Q2 - A^2)}{n_a n_n}}$$

Where A is the area under the curve, n_a and n_n are the number of abnormals and normals respectively, and Q1 and Q2 are estimated by:

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

Note that it is extremely silly to rely on Gaussian-based formulae to calculate standard error when the number of abnormal and normal cases in a sample are not the same. One should use the above formulae.

Sample Size

Now that we can calculate the standard error for a particular sample size, (given a certain AUC), we can plan sample size for a study! Simply vary sample size until you achieve an appropriately small standard error. Note that, to do this, you *do* need an idea of the area under the ROC curve that is anticipated. Hanley and McNeil even provide a convenient diagram (Figure 3 in their article) that plots number against standard error for various areas under the curve. As usual, standard errors vary with the square root of the number of samples, and (as you might expect) numbers required will be smaller with greater AUCs.

Planning sample size when comparing two tests

ROC curves should be particularly valuable if we can use them to compare the performance of two tests. Such comparison is also discussed by Hanley and McNeil in the above mentioned paper, and a

subsequent one [Hanley JA & McNeil BJ, Radiology 1983 148 839-43] entitled *A method of comparing the areas under Receiver Operating Characteristic curves derived from the same cases*.

Commonly in statistics, we set up a null hypothesis (that there is *no* statistically significant difference between two populations). If we reject such a hypothesis when it should be accepted, then we've made a *Type I* error. It is a tradition that we allow a one in twenty chance that we have made a type I error, in other words, we set our criterion for a "significant difference" between two populations at the 5% level. We call this cutoff of 0.05 "alpha".

Less commonly discussed is "beta", (ß) the probability associated with committing a *Type II* error. We commit a type II error if we accept our null hypothesis when, in fact, the two populations *do* differ, and the hypothesis should have been rejected. Clearly, the smaller our sample size, the more likely is a type II error. It is common to be more tolerant with beta - to accept say a one in ten chance that we have missed a significant difference between the two populations. Often, statisticians refer to the *power* of a test. The power is simply (1 - ß), so if ß is 10%, then the power is 90%.

In their 1982 paper, Hanley & McNeil provide a convenient table (Table III) that gives the numbers of normal and abnormal subjects required to provide a probability of 80%, 90% or 95% of detecting differences between various ROC areas under the curve (with a one sided alpha of 0.05). For example, if we have one AUC of 0.775 and a second of 0.900, and we want a power of 90%, then we need 104 cases *in each group* (normals and abnormals). Note that generally, the greater the areas under both curves, the smaller the **difference**between the areas needs to be, to achieve significance. The tables are however ***not*** applicable where two tests are applied to the *same set of cases*.

The approach to two different tests being applied to the same cases is the subject of Hanley & McNeil's second (1983) paper. This approach is discussed next.

Actually comparing two curves

This can be non-trivial. Just because the areas are similar doesn't necessarily mean that the curves are not different (they might cross one another)! If we have two curves of similar area and still wish to decide whether the two curves differ, we unfortunately have to use complex statistical tests - bivariate statistical analysis.

In the much more common case where we have different areas *derived from two tests applied to different sets of cases*, then it is appropriate to calculate the standard error of the difference between the two areas, thus:

$$SE(A1 - A2) = \sqrt{SE^2(A1) + SE^2(A2)}$$

Such an approach is *NOT* appropriate where two tests are applied to the same set of patients. In their 1983 paper, Hanley and McNeil show that in these circumstances, the correct formula is:

$$SE(A1 - A2) = \sqrt{SE^2(A1) + SE^2(A2) - 2r.SE(A1)SE(A2)}$$

where *r* is a quantity that represents the correlation induced between the two areas by the study of the same set of cases. (The difference may be non-trivial - if *r* is big, then we will need far fewer cases to demonstrate a difference between tests on the same subjects)!

Once we have the standard error of the difference in areas, we can then calculate the statistic:

$$z = (A1 - A2) / SE(A1-A2)$$

If z is above a critical level, then we accept that the two areas are different. It is common to set this critical level at 1.96, as we then have our conventional one in twenty chance of making a type I error in rejecting the hypothesis that the two curves are similar. (Simplistically, the value of 1.96 indicates that the areas of the two curves are two standard deviations apart, so there is only an ~5% chance that this occurred randomly and that the curves are in fact the same).

In the circumstance where the same cases were studied, we still haven't told you how to calculate the magic number *r*. This isn't that simple. Assuming we have two tests T1 and T2, that classify our cases into either normals (n) or abnormals (a), and we have already calculated the ROC AUCs for each test (Let's call these areas A1 and A2). The procedure is as follows:

1. Look at (n), the non-diseased patients. Find how the two tests correlate for these patients, and obtain a value r_n for this correlation. (We'll soon reveal how to obtain this value);
2. Look at (a), the abnormals, and similarly derive r_a, the correlation between the two tests for these patients;
3. Average out r_n and r_a;
4. Average out the areas A1 and A2, in other words, calculate (A1+A2)/2;
5. Use Hanley and McNeil's Table I to look up a value of *r*, given the average areas, and average of r_n and r_a.

You now have *r* and can plug it into the standard error equation. But wait a bit, how do we calculate r_n and r_a? This depends on your method of scoring your data - if you are measuring things on an *interval* scale (for example, blood pressure in millimetres of mercury), then something called the Pearson product-moment correlation method is appropriate. For *ordinal* information (e.g. saying that 'this image is definitely abnormal and that one is probably abnormal'), we use something called the *Kendall tau*. Either can be derived from most statistical packages.

Sources of Error

The effect of noise

Let's consider how "random noise" might affect our curve. Still assuming that we have a 'gold standard' which confirms the presence or absence of disease, what happens as 'noise' confuses our test, in other words, when the test results we are getting are affected by *random variations* over which we have no control. If we start off by assuming our test correlates perfectly with the gold standard, then the area under the ROC curve (AUC) will be 1.0. As we introduce noise, so some test results will be mis-classified - false positives and false negatives will creep in. The AUC will diminish.

What if the test is already pretty crummy at differentiating 'normals' from 'abnormals'? Here things become more complex, because some false positives or false negatives might accidentally be classified as true values. You can see however, that on average (provided sample numbers are sufficient and the test has some discriminatory power), noise will in general degrade test performance. It's unlikely that random noise will lead you to believe that the test is performing *better* than it really is - a most desirable characteristic!

Independence from the gold standard

The one big catch with ROC curves is where the test and gold standard are not independent. This interdependence *will* give you spuriously high area under the ROC curve. Consider the extreme case where the gold standard is compared to itself (!) - the AUC will be 1.0, regardless. This becomes extremely worrying where the "gold standard" is itself a bit suspect - if the test being compared to the standard now also varies as does the standard, but both have a poor relationship to the disease you want to detect, then you might believe you're doing well and making appropriate diagnoses, but be far from the truth! Conversely, if the gold standard is a bit shoddy, but independent from the test, then the effect

will be that of 'noise' - the test characteristics will be underestimated (often called "nondifferential misclassification" by those who wish to confuse you)!

Other sources of error

It should also be clear that any bias inherent in a test is not transferred to bias the ROC curve. If one is biased in favour of making a diagnosis of abnormality, this merely reflects a position on the ROC curve, and has no impact on the overall shape of the curve.

Other errors may still creep in. A fine article that examines sources of error (and why, after initial enthusiasm, so many tests fall into disfavour) is that of Ransohoff and Feinstein [New Engl J Med 1978 299(17) 926-30]. With *every*examination of a test one needs to look at:

1. Whether the full spectrum of a disease process is being examined. If only severe cases are reported on, then the test may be useless in milder cases (both pathologic and clinical components of the disease should represent its full spectrum). A good example is with malignant tumours - large, advanced tumours will be easily picked up, and a screening test might also perform well in this setting, but miss early disease!

2. Comparative ('control') patients. These should be similar - for example, "the search for a comparative pathological spectrum should include a different process in the same anatomical location .. and the same process in a different anatomical location" (citing the case of a test for say, cancer of the colon);

3. Co-morbid disease. This may affect the positivity or negative status of a test.

4. Verification bias. If the clinician is not blinded to the result of the test, a positive may make him scrutinise the patient very carefully and find the disease (which he missed in the other patient who had a negative test). Another name for verification bias is *work-up bias*. Verification bias is common and counter-intuitive. People tend to get rather angry when you say it might exist, for they will reply along the lines of "We confirmed all cases at autopsy, dammit!" (The positive test may have influenced the clinicians to send the patients to autopsy). A good test will be *more* likely to influence selection for 'verification', and thus introduce a stronger bias! (Begg & McNeil describe this bias well, and show how it can be corrected for).

5. Diagnostic review bias. If the test is first performed, and then the definitive diagnosis is made, knowledge of the test result may affect the final 'definitive' diagnosis. Similar is "test-review bias", where knowledge of the 'gold standard' diagnosis might influence interpretation of the test. Studies in radiology have shown that provision of *clinical information* may move observers along an ROC curve, or even to a new curve entirely! ('Co-variate analysis' may help in controlling for this form of bias).

6. "Incorporation bias". This has already been mentioned above under "independence from the gold standard". Here, the test is incorporated into the evidence used to diagnose the disease!

7. Uninterpretable test results. These are *infrequently* reported in studies! Such results should be considered 'equivocal' if the test is not repeatable. However, if the test is repeatable, then correction (and estimation of sensitivity and specificity) may be possible, provided the variation is random. Uninterpretable tests may have a positive association with the disease state (or even with 'normality').

8. Interobserver variation. In studies where observer abilities are important, different observers may perform on different ROC curves, or move along the same ROC curve.

An Example: Procalcitonin and Sepsis

Let's see how ROC curves have been applied to a particular TEST, widely promoted as an easy and quick method of diagnosing sepsis. As with all clinical medicine, we must first state our problem. We will simply repeat our SIRS/sepsis problem from above:

The Problem

Some patients with SIRS have underlying bacterial infection, whereas others do not. It is generally highly inappropriate to empirically treat everyone with SIRS as if they had bacterial infection, so we need a reliable diagnostic test that tells us early on whether bacterial infection is present.

Waiting for culture results takes days, and such delays will compromise infected patients. Although positive identification of bacterial infection is our gold standard, the delay involved (1 to 2 days) is too great for us to wait for cultures. We need something quicker. The test we examine will be **serum procalcitonin**.

Clearly what we now need is to perform a study on patients with SIRS, in whom bacterial infection is suspected. These patients should then have serum PCT determination, and adequate bacteriological investigation. Knowledge of the presence or absence of infection can then be used to create a receiver operating characteristic curve for the PCT assay. We can then examine the utility of the ROC curve for distinguishing between plain old SIRS, and sepsis. (We might even compare such a curve with a similar curve constructed for other indicators of infection, such as C-reactive protein).

(Note that there are other requirements for our PCT assay, for example, that the test is reproducible. In addition, we must have reasonable evidence that the 'gold standard' test - here interpretation of microbiological data - is reproducibly and correctly performed).

PCT - a look at the literature

Fortunately for us, there's a 'state of the art' supplement to *Intensive Care Medicine* (2000 **26** S 145-216) where most of the big names in procalcitonin research seem to have had their say. Let's look at those articles that seem to have specific applicability to intensive care. Interestingly enough, most of these articles make use of ROC analysis! Here they are:

1. **Brunkhorst FM, et al** (pp 148-152) *Procalcitonin for the early diagnosis and differentiation of SIRS, sepsis, severe sepsis and septic shock*
2. **Cheval C. et al** (pp 153-158) *Procalcitonin is useful in predicting the bacterial origin of an acute circulatory failure in critically ill patients*
3. **Rau B. et al** (pp 158-164) *The Clinical Value of Procalcitonin in the prediction of infected necro[s]is in acute pancreatitis*
4. **Reith HB. et al** (pp 165-169) *Procalcitonin in patients with abdominal sepsis*
5. **Oberhoffer M. et al** (pp170-174) *Discriminative power of inflammatory markers for prediction of tumour necrosis factor-alpha and interleukin-6 in ICU patients with systemic inflammatory response syndrome or sepsis at arbitrary time points*

Quite an impressive list! Let's look at each in turn:

1. **Brunkhorst FM, et al** (pp 148-152)
   Procalcitonin for the early diagnosis and differentiation of SIRS, sepsis, severe sepsis and septic shock

   The authors recruited 185 consecutive patients. Unfortunately, only seventeen patients in the study had uncomplicated 'SIRS' - the rest had sepsis (n=61), 'severe sepsis' (n=68) or septic shock (n=39). The authors then indulge in intricate statistical manipulation to differentiate between sepsis, severe sepsis, and septic shock - they even construct ROC curves (although we are not told, when they construct an ROC curve for 'prediction of severe sepsis' *what those with severe sepsis are being differentiated from* - presumably the rest of the population)! The authors

do not address why, in their ICU, so many patients had sepsis, and so few had SIRS without sepsis. The bottom line is that the results of this study, with an apparently highly selected group of just seventeen 'non-septic' SIRS patients, seem useless for addressing our problem of differentiating SIRS and sepsis! Their ROC curves seem irrelevant to *our* problem. (Parenthetically one might observe that if you walk into their ICU and find a patient with SIRS, there would appear to be an over 90% chance that the patient has sepsis - who needs procalcitonin in such a setting)?

2. **Cheval C. et al** (pp 153-158)
Procalcitonin is useful in predicting the bacterial origin of an acute circulatory failure in critically ill patients

This study looked at four groups:

1. septic shock (n=16);
2. shock without infection(n=18);
3. SIRS related to proved infection(n=16);
4. ICU patients without shock or infection(n=10).

The choice of groups is somewhat unfortunate! Where are the patients we really want to know about - those with SIRS but *no* infection? Reading on, we find that only four of the patients in the fourth group met the criteria for SIRS! This study too does not appear to help us in our quest! (The authors use ROC curves to analyse their patients in shock, comparing those with and without sepsis. The numbers look impressive - an AUC of 0.902 for procalcitonin's ability to differentiate between septic shock and 'other' causes of shock. But hang on - let's look at the 'other' causes of shock. We find that in these cases, shock was due to haemorrhage(n=8), heart failure(n=7), anaphylaxis(n=2), and 'hypovolaemia' (n=1). One doesn't need a PCT level to decide whether a patient is in heart failure, bleeding to death, etc. A study whose title promises more than is delivered)!

3. **Rau B. et al** (pp 158-164)
The Clinical Value of Procalcitonin in the prediction of infected necro[s]is in acute pancreatitis

Sixty one patients were entered into this study. Twenty two had oedematous pancreatitis, 18 had sterile necrosis, and 21 had infected necrosis. Serial PCT levels were determined over a period of fourteen days. The 'gold standard' used to determine whether infected necrosis was present was fine needle aspiration of the pancreas, combined with results of intra-operative bacteriology. We learn that

"PCT concentrations were significantly higher from ***day 3-13*** *after onset of symptoms in patients with [infected necrosis, compared with sterile necrosis]".*{The emphasis is ours}.

The authors then inform us that

"ROC analysis for PCT and CRP has been calcul[a]ted on the basis of at least two maximum values reached during the total observation period. By comparison of the areas under the ROC curve (AUC), PCT was found to have the closest correlation to the presence and severity of bacterial/fungal infection of necrosis and was clearly superior to CRP in this respect (AUC for PCT: 0.955, AUC for CRP: 0.861; p<0.02)."

Again, the numbers look impressive. Hold it! Does this mean that we have to do daily PCT levels on all of our patients, and then take the two maximum values, and average them in order to decide who has infected necrosis?? Even more tellingly, we are *not* provided with information about how PCT might have been used in prospectively differentiating between those who developed sepsis and those who didn't, before bacterial cultures became available. In other words, *was PCT useful in identifying infected necrosis* ***early on***? If I have a sick patient with pancreatitis, can I base my management decision on a PCT level? This vital question is left unanswered, but the lack of utility of PCT in the first two days is of concern!

4. **Reith HB. et al** (pp 165-169)
   Procalcitonin in patients with abdominal sepsis

   A large study compared 246 patients with "infective or septic episodes confirmed at laparotomy" with 66 controls. And this is where the wheels fall off, for the sixty six controls were undergoing elective operation! Clearly, any results from such a study are irrelevant to the problem ICU case where you are agonizing over whether to send the patient for a laparotomy - "is there sepsis or not"?

5. **Oberhoffer M. et al** (pp170-174)
   Discriminative power of inflammatory markers for prediction of tumour necrosis factor-alpha and interleukin-6 in ICU patients with systemic inflammatory response syndrome or sepsis at arbitrary time points

   The authors reason that TNF and IL-6 levels predict mortality from sepsis. Strangely enough, they do not appear to have looked at actual mortality in the 243 patients in the study! This is all very well if you're interested in deciding whether the TNF and IL-6 levels in your patients are over their cutoff levels of 40pg/ml and 500pg/ml respectively, but perhaps of somewhat less utility unless such levels themselves absolutely predict fatal outcome (they don't). From a clinical point of view, this study suffers from use of a 'gold standard' that may not be of great overall relevance. A hard end point (like death) would have been far better. (In addition, the authors are surprisingly coy with their AUCs. If you're really keen, you might try and work these out from their Table 4).

A Summary

Four of the five papers above used ROC analysis. In our opinion, this use provides us with little or no clinical direction. If the above articles reflect the 'state of the art' as regards use of procalcitonin in distinguishing between the systemic inflammatory response syndrome and sepsis, we can at present find no justification in using the test on our critically ill patients! (This does not mean that the test is of no value, simply that we have no substantial evidence that it is of use).

What would be most desirable is a study that conformed to the requirements we gave above - a study that examines a substantial number of patients with either:

- SIRS not complicated by sepsis; OR
- sepsis;

and demonstrates unequivocally that serum procalcitonin is useful in differentiating between the two *early on*, before blood cultures become positive. Clearly a substantial area under an appropriately constructed ROC curve would be powerful evidence in support of using the test.

A second example - Tuberculosis, ADA, and pleural fluid

For our second example, we'll use some data on Adenosine Deaminase (ADA) levels determined on pleural effusions. It is well known that ADA levels in empyemas may be high, (we might explore this later), so at first we will concentrate on data for pleural fluid obtained from patients with either neoplasms, or those with documented tuberculosis (TB). The data and ROC curve can be downloaded as a self-extracting Microsoft Excel spreadsheet. To derive full benefit from this example, some knowledge of spreadsheets (specifically, Excel) is desirable but probably not vital. The data are the property of Dr Mark Hopley of Chris-Hani Baragwanath Hospital (CHB, the largest hospital in the world).

The spreadsheet contains three important columns of data:

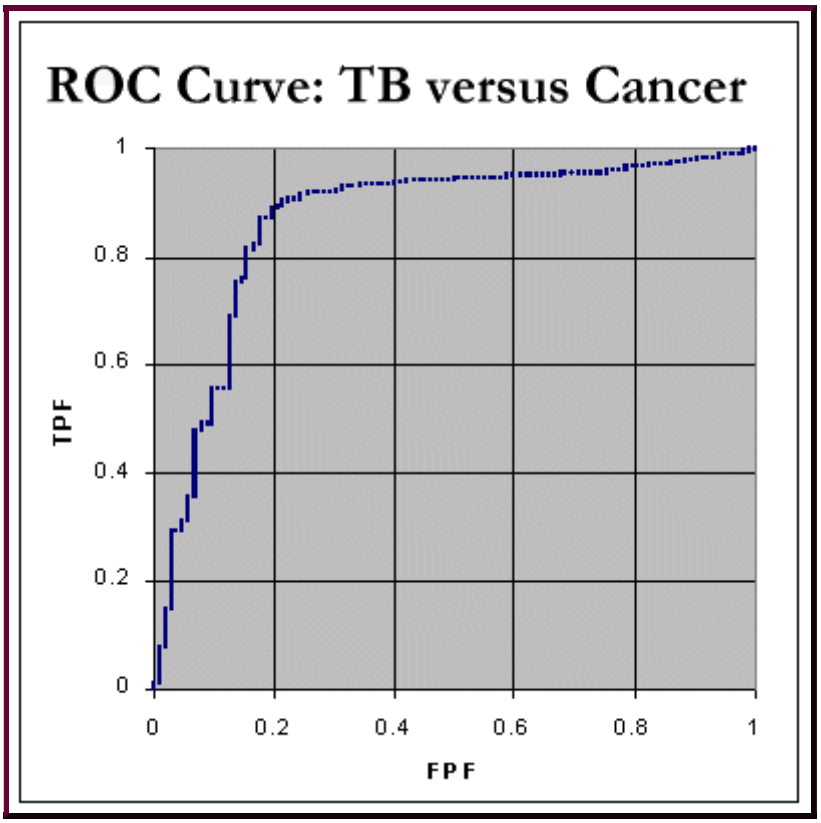
1. The leftmost column contains ADA levels;
2. The next column contains a '1' if the patient had documented tuberculosis, and otherwise a zero;

3. The third column contains a '1' only if the patient had documented carcinoma. There were six patients who had both carcinoma and tuberculosis - these have been excluded from analysis.

There were eight hundred and twelve tuberculosis patients, and one hundred and two patients with malignant pleural effusion. How do we go about creating an ROC curve? The steps, as demonstrated in the worksheet, are:

1. Sort the data according to the ADA level - largest values first;
2. Create a column where each row gives the total number of TB patients with ADA levels greater than or equal to the ADA value for that row;
3. Create a similar column for patients with cancer;
4. Create two new columns, containing the TPF and FPF for each row. In other words, we position our 'green marker' (remember our ROC applet!) *just below* the current ADA level for that row, and then work out a TPF and an FPF at that cutoff level. We work out the TPF by taking the number of TB cases identified at or above the ADA level for the current row, and dividing by the total number of TB cases. We determine the FPF by taking the number of "false alarms" (cancer patients) at or above that level, and dividing by the total number of such non-TB patients.

We now have sufficient data to plot our ROC curve. Here it is:



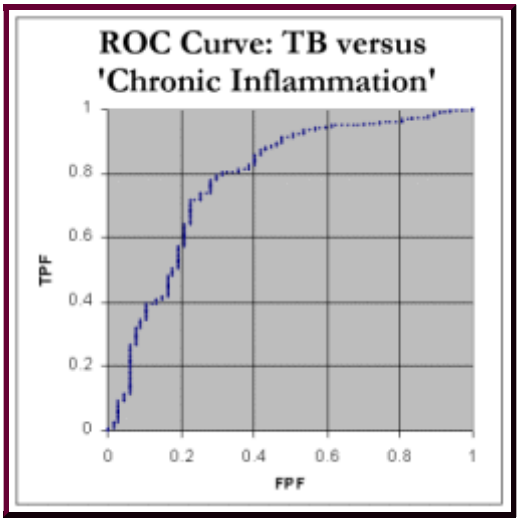
We still need to determine the Area Under the Curve (AUC). We do this by noting that every time we move RIGHT along the x-axis, we can calculate the increase in area by finding:

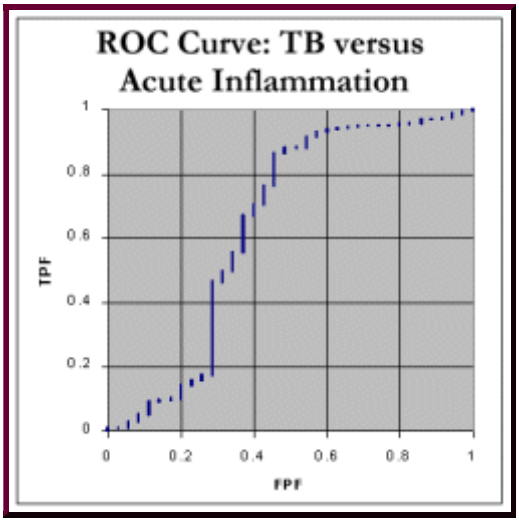
```
(how much we moved right) * (the current y value)
```

We can then add up all these tiny areas to get a final AUC. As shown in the spreadsheet, this works out at 85.4%, which indicates that, in distinguishing between tuberculosis and neoplasia as a cause of pleural effusion, ADA seems to be a fairly decent test!

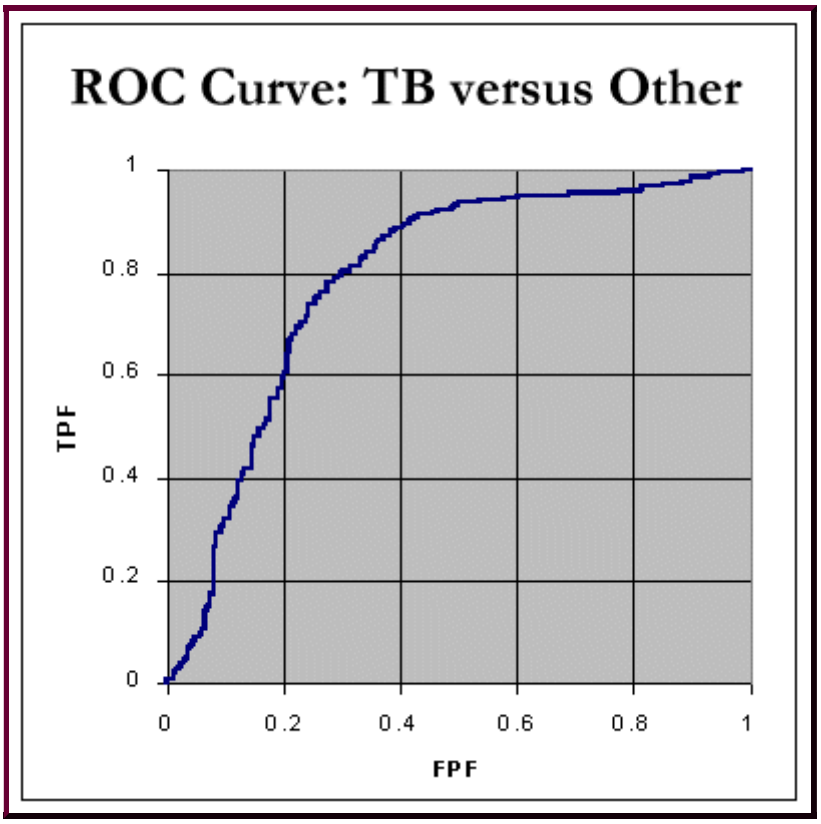
Here are the corresponding ROC curve for tuberculosis compared with inflammatory disorders. As expected, the AUC is less for chronic inflammatory disorders, about 77.9%, and pretty poor at 63.9%

for 'acute inflammation' which mainly represents empyemas.





Note that there were only 67 cases of "chronic inflammatory disorders", and thirty five with "acute inflammation". Finally, let's look at TB versus "all other" effusion data - there were 393 "non-tuberculous" cases. The data include the above 'cancer' and 'inflammatory' cases. The AUC is still a respectable 78.6%.



Is the above credible?

Through our analysis of ADA in pleural fluid, we've learnt how to create an ROC curve. But we still *must* ask ourselves questions about error and bias! Here are a few questions you have to ask - they will profoundly influence your interpretation and use of the above ROC curves:

- Are the data selected, or were *all* samples of pleural fluid subject to analysis?
- Does the hospital concerned have a peculiar case spectrum, or will *your* case profile be similar?

- How severe were the cases of tuberculosis - is the full spectrum of pleural effusions being examined?
- Should the cases who had two diseases (that is, carcinoma and tuberculosis) have been excluded from analysis?
- What co-morbid diseases were present (for example, Human Immunodeficiency Virus infection)?
- Was there verification bias introduced by, for example, a high ADA value being found, and the diagnosis of tuberculosis therefore being aggressively pursued?
- Were any test results uninterpretable?
- In how many cases was the diagnosis known *before* the test was performed? How many of the cases were considered by the attending physician to be "really problematical diagnoses"? One could even ask "How good were the physicians at clinically diagnosing the various conditions - did the ADA *add* to diagnostic sensitivity and specificity?"

(Makes you think, doesn't it?)

{ Just as an aside, it's perhaps worth mentioning that the above ADA results are *not* normally distributed, for either the 'tuberculosis' or the 'neoplasia' samples. Even taking the logarithms of the values (although it decreases the skewness of the curves dramatically) doesn't quite result in normal distributions, so any ROC calculations that assume normality are likely to give spurious results. Fortunately our calculations above make no such assumption.}

Working out Standard Errors

You can calculate Standard Errors for the Areas Under the Curves we've presented, using the following JavaScript calculator. It's based on the formulae from above.

Calculate Standard Error![8]

(Enter data, press 'Calculate')

Area Under Curve	
Number *without* disease	
Number WITH disease	

Calculate | Clear Form | Standard Error:

Footnotes

1. Exploring Accuracy

Accuracy, PPV and NPV

It would be great if we could lump things together in some way, and come up with a single number that could tell us how well a test performs. One such number is represented by the area under the ROC. Another more traditional (and far more limited) number is *accuracy*, commonly given as:

accuracy = number of correct diagnoses / number in total population

While we're about it, let's also consider a few other traditional terms:

- **Positive predictive value** (PPV) is of some interest to clinicians. It answers the question "*How likely is the patient to have the disease, given that the test is positive?*". You can work out that this is given by:

$$\frac{\text{true positives}}{\text{all positive tests}}$$

You'll find that positive (and negative) predictive values depend on the frequency of the disease in the population, which is one reason why you cannot just blindly apply tests, without considering whom you are applying them to!

- In a completely analogous fashion, we calculate the **negative predictive value**, which tells us how likely it is that the disease is NOT present, given that the test is negative. We calculate:

```
true negatives / all negative tests
```

(Yet another name for the PPV is *accuracy for positive prediction*, and the negative predictive value, *accuracy for negative prediction*).

KISS(2)

We will refer to positive predictive value as PPV, and negative predictive value as NPV. Accuracy we'll refer to as 'accuracy' (heh).

An examination of 'accuracy'

Let's consider two tests with the same accuracy. Let's say we have a population of 1000 patients, of whom 100 have a particular disease (D+). We apply our tests (call them T1 and T2) to the population, and get the following results.

NOTE that our tables now contain *actual numbers* of cases, and **NOT** fractions. The four values are actual numbers of true positives, false positives, false negatives and true negatives!

	D+	*D-*
T+	**a**	**b**
T-	**c**	**d**

('a' represents true positives, 'd' true negatives, 'b' false positives, and 'c' false negatives).

Test performance: T1 (n=1000)		
	D+	*D-*
T+	60	**5**
T-	**40**	895
PPV = 92.3% NPV = 95.7%		

Test performance: T2 (n=1000)		
	D+	*D-*
T+	95	**40**
T-	**5**	860
PPV = 70.3% NPV = 99.4%		

See how the two tests have the same accuracy (a + d)/1000 = 95.5%, but they do remarkably different things. The first test, T1, misses the diagnosis 40% of the time, but makes up for this by providing us with few *false positives* - the TNF is 99.4%. The second test is quite different - impressive at picking up the disease (a sensitivity of 95%) but relatively lousy performance with false positives (a TNF of 95.5%). At first glance, if we accept the common medical obsession with "making the diagnosis", we would be tempted to use T2 in preference to T1, (the TPF is after all, 95% for T2 and only 60% for T1), but surely this depends on the disease? If the consequences of missing the disease are relatively minor, and

the costs of work-up of the false positives are going to be enormous, we might just conceivably favour T1.

Now, let's drop the prevalence of the disease to just ten in a thousand, that is P(D+) = 1%. Note that the TPF and TNF (or sensitivity and specificity, if you prefer) are of course the same, but the positive predictive and negative predictive values have altered substantially.

Test performance: T1 (n=1000)		
	D+	*D*-
T+	6	**5.5**
T-	**4**	984.5
PPV = 52.2% NPV = 99.6%		

Test performance: T2 (n=1000)		
	D+	*D*-
T+	9.5	**44**
T-	**0.5**	946
PPV = 17.8% NPV = 99.9%		

(Okay, you might wish to round off the "fractional people")! See how the PPV and NPV have changed for both tests. Now, almost five out of every six patients reported "positive" according to test T2, will in fact be false positives. Makes you think, doesn't it?

Another example

Now let's consider a test which is 99% sensitive and and 99% specific for the diagnosis of say, Human Immunodeficiency Virus infection. Let's look at how such a test would perform in two populations, one where the prevalence of HIV infection is 0.1%, another where the prevalence is 30%. Let's sample 10 000 cases:

Test performance: Population A (n=10 000, prevalence 1/1000)		
	D+	*D*-
T+	**10**	**100**
T-	**0**	**9890**
PPV = 9.1% NPV = almost 100%		

Test performance: Population B (n=10 000, prevalence 300/1000)		
	D+	*D*-
T+	**2970**	**70**
T-	**30**	**6930**
PPV = 97.7% NPV = 99.5%		

If the disease is rare, use of even a very specific test will be associated with many false positives (and all that this entails, especially for a problem like HIV infection); conversely, if the disease is common, a positive test is likely to be a true positive. (This should really be common sense, shouldn't it?)

You can see from the above that it's rather silly to have a fixed test threshold. We've already played around with our applet where we varied the test threshold, and watched how the TPF/FPF coordinates moved along the ROC curve. The (quite literally) million dollar question is "Where do we set the threshold"?

2. Deciding on a test threshold

The reason why we choose to plot **FPF** against **TPF** when we make our ROC is that all the information is contained in the relationship between just these two values, and it's awfully convenient to think of, in the words of Swets, "hits" and "false alarms" (in other words, TPF and FPF). We can limit the false alarms, but at the expense of fewer "hits". What dictates where we should put our cutoff point for diagnosing a disease? The answer is not simple, because we have many possible criteria on which to base a decision. These include:

- Financial costs both direct and indirect of treating a disease (present or not), and of failing to treat a disease;
- Costs of further investigation (where deemed appropriate);
- Discomfort to the patient caused by disease treatment, or failure to treat;
- Mortality associated with treatment or non-treatment;

Soon we will explore the mildly complex maths involved, but first let's use a little common sense. It would seem logical that if the cost of missing a diagnosis is great, and treatment (even inappropriate treatment of a normal person) is safe, then one should move to a point on the *right* of the ROC, where we have a high TPF (most of the true positives will be treated) at the cost of many false positives. Conversely, if the risks of therapy are grave, and therapy doesn't help much anway, we should position our point far to the left, where we'll miss a substantial number of positives (low TPF) but not harm many unaffected people (low FPF)!

More formally, we can express the average cost resulting from the use of a diagnostic test as:

$$C_{avg} = C_o + C_{TP}*P(TP) + C_{TN}*P(TN) + C_{FP}*P(FP) + C_{FN}*P(FN)$$

where C_{avg} is the average cost, C_{TP} is the cost associated with management of true positives, and so on. C_o is the "overhead cost" of actually doing the test. Now, we can work out that the probability of a true positive P(TP) is given by:

$$\begin{aligned} P(TP) &= P(D+) * P(T+|D+) \\ &= P(D+) * TPF \end{aligned}$$

In other words, P(TP) is given by the product of the prevalence of the disease in the population, P(D+), multiplied by the true positive fraction, for the test. We can similarly substitute for the three other probabilities in the equation, to get:

$$\begin{aligned} C_{avg} = C_o &+ C_{TP}*P(D+)*P(T+|D+) + C_{TN}*P(D-)*P(T-|D-) \\ &+ C_{FP}*P(D-)*P(T+|D-) + C_{FN}*P(D+)*P(T-|D+) \end{aligned}$$

Another way of writing this is:

$$\begin{aligned} C_{avg} = C_o &+ C_{TP}*P(D+)*TPF + C_{TN}*P(D-)*TNF \\ &+ C_{FP}*P(D-)*FPF + C_{FN}*P(D+)*FNF \end{aligned}$$

Remembering that TNF = 1 - FPF, and FNF = 1 - TPF, we can write:

$$\begin{aligned} C_{avg} = C_o &+ C_{TP}*P(D+)*TPF + C_{TN}*P(D-)*(1-FPF) \\ &+ C_{FP}*P(D-)*FPF + C_{FN}*P(D+)*(1-TPF) \end{aligned}$$

and, rearrange to ..

$$\begin{aligned} C_{avg} = &\quad TPF * P(D+) * \{ C_{TP} - C_{FN} \} \\ &+ FPF * P(D-) * \{ C_{FP} - C_{TN} \} \\ &+ C_o + C_{TN}*P(D-) + C_{FN}*P(D+) \end{aligned}$$

As Metz has pointed out, even if a diagnostic test improves decision-making, it may still increase overall costs if C_o is great. Of even more interest is the dependence of C_{avg} on TPF and FPF - the coordinates on an ROC curve! Thus average cost depends on the test threshold defined on an ROC curve, and varying this threshold will vary costs. The best cost performance is achieved when C_{avg} is minimised.

We know from elementary calculus that this cost will be minimal when the derivative of the cost equation is zero. Now because we can express TPF as a function of FPF using the curve of the ROC, thus:

$$C_{avg} = ROC(FPF) * P(D+) * \{ C_{TP} - C_{FN} \} + FPF * P(D-) * \{ C_{FP} - C_{TN} \} + C_o + C_{TN}*P(D-) + C_{FN}*P(D+)$$

we can differentiate this equation with respect to FPF, and obtain:

$$dC/dFPF = dROC/dFPF * P(D+) * \{ C_{TP} - C_{FN} \} + P(D-) * \{ C_{FP} - C_{TN} \}$$

Setting dC/dFPF to zero, we get:

$$dROC/dFPF * P(D+) * \{ C_{TP} - C_{FN} \} = - P(D-) * \{ C_{FP} - C_{TN} \}$$

or, rearranging:

$$dROC/dFPF = \frac{P(D-) * \{ C_{FP} - C_{TN} \}}{P(D+) * \{ C_{FN} - C_{TP} \}}$$

In other words, we have found a differential equation that gives us the slope of the ROC curve at the point where costs are optimal. Now let's look at a few circumstances:

- Where the disease is rare, P(D-)/P(D+) will be enormous, and so we should shift our test threshold down to the lower left part of the ROC curve, where dROC/dFPF, the slope of the curve, is large. This fits in with our previous simple analysis, where with uncommon diseases, we found that false positives are a very bad thing. We must minimise our false positives, even at the expense of missing true positives!

- Conversely, with a common disease, we move our threshold to a lower, more lenient level, (and our position on the ROC curve necessarily moves right). Otherwise, most of our negatives are false negatives!

- Also notice that the curve slope is great if the cost difference is far greater for $C_{FP} - C_{TN}$ than for $C_{FN} - C_{TP}$. Let's consider a practical scenario - assume for a particular disease (say a brain tumour) that if you get a positive test, you have to open up the patient's skull and cut into the brain to find the presumed cancer. If you have a negative, you do nothing. Let's also assume that the operation doesn't help those who have the cancer - many die, regardless. Then the cost of a false positive (operating on the brains of normal individuals!) is indeed far greater than the cost of a true negative (doing nothing), *and* the cost of a false negative (not doing an operation that doesn't help a lot) is similar to the cost of a true positive (doing the rather unhelpful operation). The curve slope is steep, so we move our test threshold down on the left of the ROC curve.

- The opposite is where the consequences of a false positive are minimal, and there is great benefit if you treat sufferers from the disease. Here, you must move up and to the right on the ROC curve.

Fine Print - Old fashioned assumptions of Normality

Earlier literature on ROC curves often seems to have made the unfortunate assumption that the underlying distributions are *normal curves*. (The only reason we used normal curves in our applet is their convenience - perhaps the same reason that others have 'assumed normality'). Under this assumption, one trick that has been used is to create special 'graph paper' where axes are transformed according to the normal distribution. ('double normal probability co-ordinate scales'). Using such coordinates, ROC curves become linear (!), and one can read off slope and axis, which correspond to the two parameters that contain the mean and standard deviation. Curve fitting can be done (using special techniques, NOT least squares) to work out the line that best fits the plotted coordinates. Such methods appear to have been applied mainly in studies of experimental psychology.

Note that if one uses double normal probability plots, the slope of the straight line obtained by plotting TPF against FPF will give us the ratio of standard deviations of the two distributions (assuming normality). In other words, if the standard deviations of the populations D+ and D- are s_{D+} and s_{D-}, the line slope is s_{D-} / s_{D+}. In the particular case where this value is 1, we can measure the distance between the plotted line and the 'chance line' (connecting the bottom left and top right corners of the graph). This distance is a normalised measure of the distance between the means of the two distributions where m refers to mean, and **s**, standard deviation:

$$d' = (m_{D+} / m_{D-}) / s$$

References

1. Hanley JA, McNeil BJ. **Radiology** 1982 143 29-36. *The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve*. An excellent paper, but not an easy read! Their follow-up paper is also good [Radiology 1983 148 839-43].

2. Metz CE. **Semin Nuclear Med** 1978 VIII(4) 283-298. *Basic principles of ROC analysis*. A really good introduction, on which we've based a lot of the above text. Things have however come a long way since 1978. See also the paper by Dennis Patton in the same issue (p273) which has quite a bit on Bayesian decision making, as does the paper by McNeil et al in the New England Journal of Medicine [1975, 293 211-5].

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

4. Swets JA. **Science** 1988 240 1285-93. *Measuring the accuracy of diagnostic systems*. A fascinating and wide-ranging article.

5. There's quite an attractive ROC applet athttp://acad.cgu.edu/wise/sdt/sdt.html. The box on the right demonstrates double normal probability co-ordinate scales, for the curious.

6. Good ROC software (not freeware, ~1 Meg) is available for download (IBM PC version, MAC also available) at ftp://random.bsd.uchicago.edu/roc/ibmpc/. Note that this 'ROCKIT' is Metz's implementation of software that assumes the data can be *monotonically transformed into binormal distributions*. Do **not** use it unless you've verified that the data meet this criterion! There is extensive associated documentation.

   Thanks to Richard Zur for pointing out that the curves don't have to be binormally distributed, they just 'have to be able to be transformed into normal distributions without changing the rank-ordering'.

7. Try http://brighamrad.harvard.edu/research/topics/vispercep/ROC.htmlfor a little note on ROC curves that is linked to an interesting paper where they are applied.

8. Thanks to Koen Vermeer for pointing out an error in our JavaScript Standard Error calculator that resulted in a significant miscalculation (2003/10/13). Sharp!

9. And we'd really like to thank AB Meijer for correcting a really silly error involving the use of the word `incidence'!

10. ... and Wolfgang Hackl for (we hope) improving our accuracy (3/2007).

11. ... and Abder-Rahman Ali for spotting a small typo (28/7/2011).

Date of First Publication: 2001/9/21 **Date of Last Update**: 2011/07/28 **Web page author:** Click here