UNIT 9 MEASURING THE CHARACTERISTICS OF MEASURES

DIAGNOSTIC TESTS

AIMS

To describe the basic ideas underlying diagnostic tests.

OBJECTIVES

At the end of this unit you should be able to:

Describe the function of a diagnostic test.

Explain the meaning of and be able to calculate the diagnostic test outcomes, e.g. true positive, false positive, etc.

Explain (with numeric examples) the meaning of and be able to calculate the main diagnostic test performance criteria, e.g. sensitivity, specificity, PPV, and NPV.

Describe the trade-off problem between sensitivity and specificity, and the clinical situations in which one criterion might be favoured over the other.

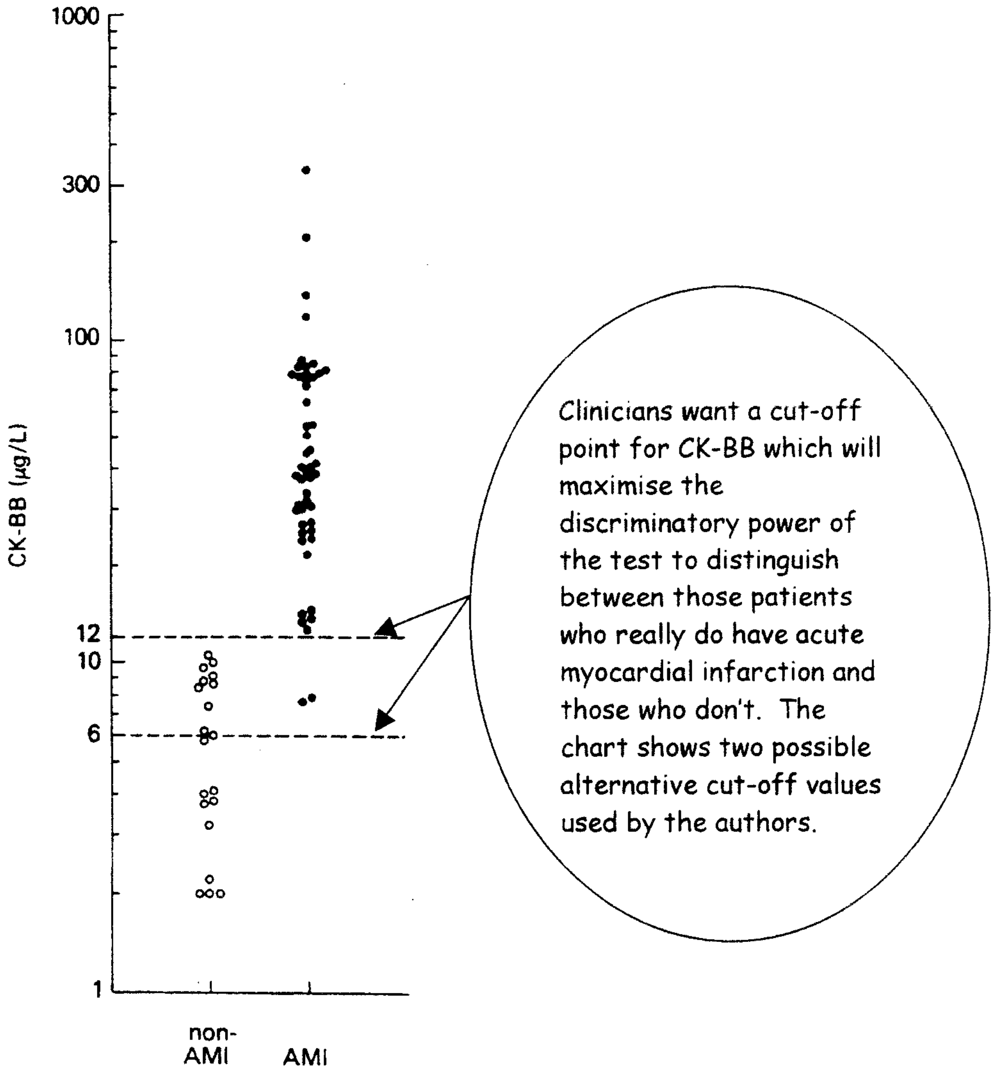
Explain the how the ROC and the AUC can be used to determine the optimal cut-off point for a diagnostic test, when the data is metric or ordinal. Plot a ROC.

Reading

Bland (2nd edit.), Section 15.4 in Chapter 15 "Clinical Measurement".

Introduction

Papers often contain studies which compare two or more diagnostic tests. By diagnostic test we mean some procedure which is designed to detect some clinical condition. For example, the breath test for the presence of helicobacter *pylorii*, or a tissue biopsy for cell malignancy, and so on. These studies often compare the performance of a new or improved procedure with a "gold standard" test - assumed to give the correct result. We need to say a few words about the ideas which underlie these studies. Look at Figure 9.1, which shows serum CK-BB concentrations (µg/l) in 70 patients admitted to emergency with typical chest pain suggestive of acute myocardial infarction (AMI).



Clinicians want a cut-off point for CK-BB which will maximise the discriminatory power of the test to distinguish between those patients who really do have acute myocardial infarction and those who don't. The chart shows two possible alternative cut-off values used by the authors.

Fig. 2. Serum CK-BB concentrations 16 h after onset of chest pain in patients with (*filled circles*) and without (*open circles*) an AMI

Figure 9.1 Serum CK-BB concentrations 16 hours after onset of symptoms in 70 patients presenting with chest pain typical of acute myocardial infarction. Clinical Chemistry, 28, 1982.

These data are metric continuous. The clinicians involved did not at the time actually know which patients had had an AMI and which not, although it was subsequently confirmed that 50 had and 20 had not, and these two outcomes are shown separately in the figure along with their CK-BB concentrations. Action needs to be taken if the patient really has experienced AMI, but not otherwise, and they could use serum CK-BB concentration as our **diagnostic test**, provided that they can arrive at an appropriate cut-off value.

If the physicians involved use as the cut-off a serum CK-BB level of ≥12 μg/l as indicative of AMI, then two patients with AMI will be missed, but no patients will be included who have *not* had an AMI. If a lower value is used, say ≥5 μg/l, then all of the patients with AMI will be detected, *but* so will 11 healthy patients. Clearly there is an optimum (but never perfect) cut-off value.

Q. 9.1 Explain the consequences if a serum CK-BB concentration of ≥ 8μg?/L is used as a cut-off in a diagnostic test for AMI.

Whatever cut-off value is used, there are four possible results when a test is applied to an individual patient:

True positive The test was **positive** and the patient **had** had an AMI (cell "a" in Table 9.1)

False positive The test was **positive** but the patient had **not** had an AMI (cell "b") (a Type I error - see Unit 7).

False negative The test was **negative** but the patient **had** had an AMI (cell "c") (a Type II error).

True negative The test was **negative** and the patient had **not** had an AMI (cell "d")

Q. 9.2 How many of each type of outcome (true positives, false positives, etc.) will occur if, in the serum CK-BB test for AMI, cut-off points of (a) 5μg/L, and (b) 12μg/L, are used?

We can describe the above possible outcomes in table form as in Table 9.1.

		Had AMI?		
		Yes	No	totals
Test result	Positive	a	b	a+b
	Negative	c	d	c+d
	totals	a+c	b+d	

Table 9.1 Table of the four possible outcomes from a diagnostic test

Researchers generally use four separate but interconnected measures of a test's efficacy:

- **Sensitivity**: the proportion (or %) of those patients *with* the condition whom the test correctly identifies as having it. So from Table 9.1, Sensitivity = a/(a+c)

- **Specificity**: the proportion (or %) of those patients *without* the condition whom the test correctly identifies as *not* having it. So from Table 9.1, Specificity = d/(b+d)

- **Positive predictive value (PPV):** the proportion (or %) of patients whom the test identifies as having the condition who *do* have it. So from Table 9.1, PPV = a/(a+b)

- **Negative predictive value (NPV):** the proportion (or %) of patients whom test does not identify as having the condition who do *not* have it. So from Table 9.1, NPV = d/(c+d)

Q. 9.3 Use Table 9.1 to show that (1 - specificity) = the false positive rate.

Clearly, the predictive diagnostics, PPV and NPV, are *clinically* more useful than knowledge of a tests sensitivity and specificity. Suppose you see a patient and you suspect, from the patient's description of her signs and symptoms, some particular condition. In these circumstances, you want to know the chances of the patient having the condition if they get a positive test result (PPV), rather than whether they will give a positive test result if they are known to have the condition (sensitivity).

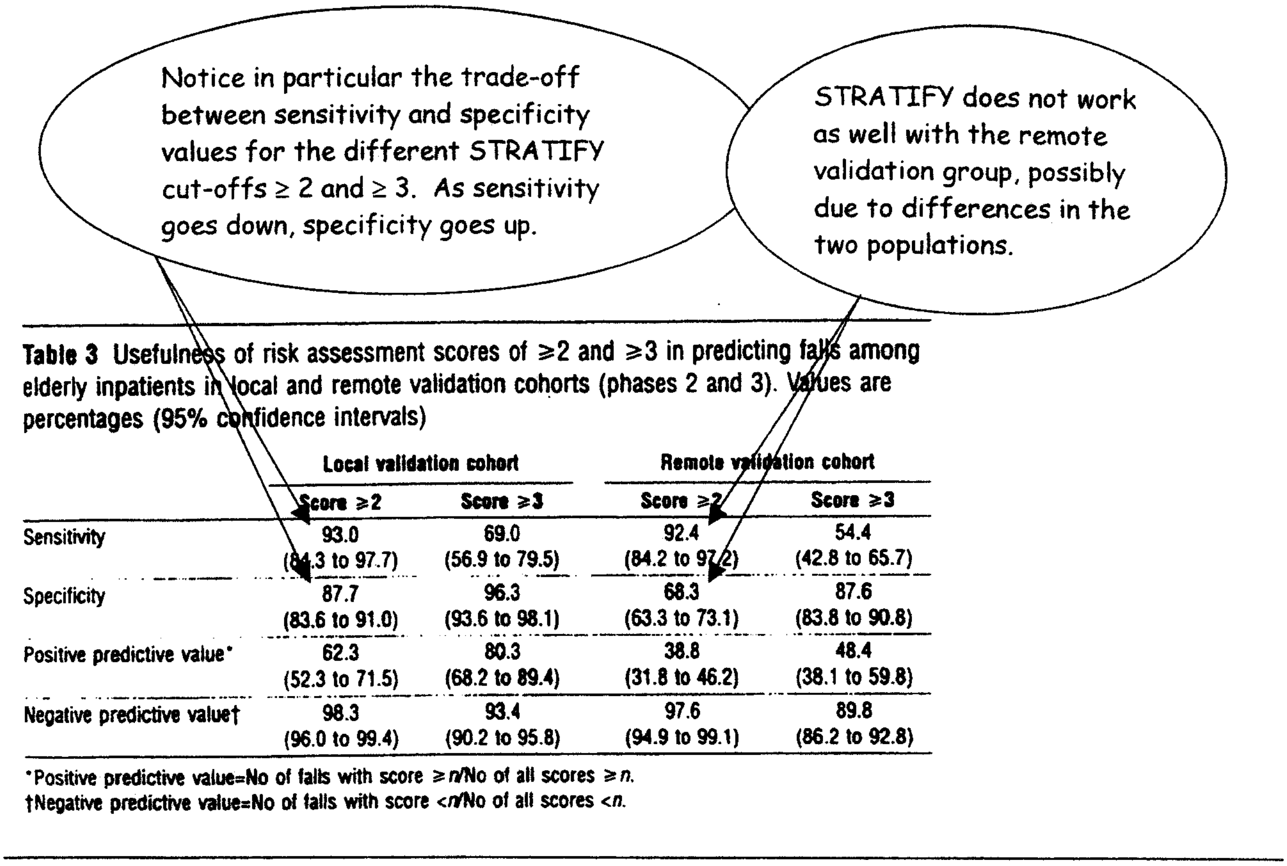
Q. 9.4 Construct a table like Table 9.1 for the AMI CK-BB test with cut-off values of: (a) ≥5μg/L; (b) ≥8μg/L; and (c) ≥12μg/L. (You should be able to see the correspondence with the answer to Q. 9.2 for cut-offs of ≥ 5 and ≥ 12). Calculate sensitivity, specificity, PPV and NPV for each cut-off. Comment briefly on the results.

Notice that as we increase the cut-off point in this example, sensitivity decreases but specificity increases. Clearly there is an optimal value for the cut-off between sensitivity and specificity which jointly maximises both measures, although this will also often be influenced by the nature of the condition. For example, a diagnostic test for AMI needs as high a sensitivity as possible so that immediate action (e.g. thrombolysis) can be taken. A high specificity is not so crucial since counter-measures wouldn't harm patients (although it might cause them some alarm). On the other hand, if emergency surgery is the action taken for those identified as having some condition, we would obviously want a very high specificity (preferably 100%). We don't want to perform invasive procedures on healthy individuals. In situations like this high sensitivity, although desirable, is not as important, even though we run the risk of missing some individuals who require treatment, in view of the alternative.

We'll come back to this trade-off problem shortly, but for now have a look at Figure 9.2, which is from a validation study for "STRATIFY", a Risk of Falling scale proposed for use with elderly hospitalised patients. This scale produces an ordinal risk-of-falling score ranging from 0 to 5 (smaller value means lower risk).

Note that the values for the diagnostic measures in Figure 9.2 are sample estimates of the true population values. The confidence interval shown with each measure enable us to assess the precision of each estimate.

Q. 9.5 In Figure 9.2, (a) For a cut-off ≥ 2 in the STRATIFY scores for the *local* validation cohort, sensitivity is 93.0% and for PPV is 62.3%. Explain what these values mean. (b) Which diagnostic measure is estimated with the least precision in the *local* validation cohort?



Notice in particular the trade-off between sensitivity and specificity values for the different STRATIFY cut-offs ≥ 2 and ≥ 3. As sensitivity goes down, specificity goes up.

STRATIFY does not work as well with the remote validation group, possibly due to differences in the two populations.

Table 3 Usefulness of risk assessment scores of ≥2 and ≥3 in predicting falls among elderly inpatients in local and remote validation cohorts (phases 2 and 3). Values are percentages (95% confidence intervals)

	Local validation cohort		**Remote validation cohort**	
	Score ≥2	**Score ≥3**	**Score ≥2**	**Score ≥3**
Sensitivity	93.0 (84.3 to 97.7)	69.0 (56.9 to 79.5)	92.4 (84.2 to 97.2)	54.4 (42.8 to 65.7)
Specificity	87.7 (83.6 to 91.0)	96.3 (93.6 to 98.1)	68.3 (63.3 to 73.1)	87.6 (83.8 to 90.8)
Positive predictive value*	62.3 (52.3 to 71.5)	80.3 (68.2 to 89.4)	38.8 (31.8 to 46.2)	48.4 (38.1 to 59.8)
Negative predictive value†	98.3 (96.0 to 99.4)	93.4 (90.2 to 95.8)	97.6 (94.9 to 99.1)	89.8 (86.2 to 92.8)

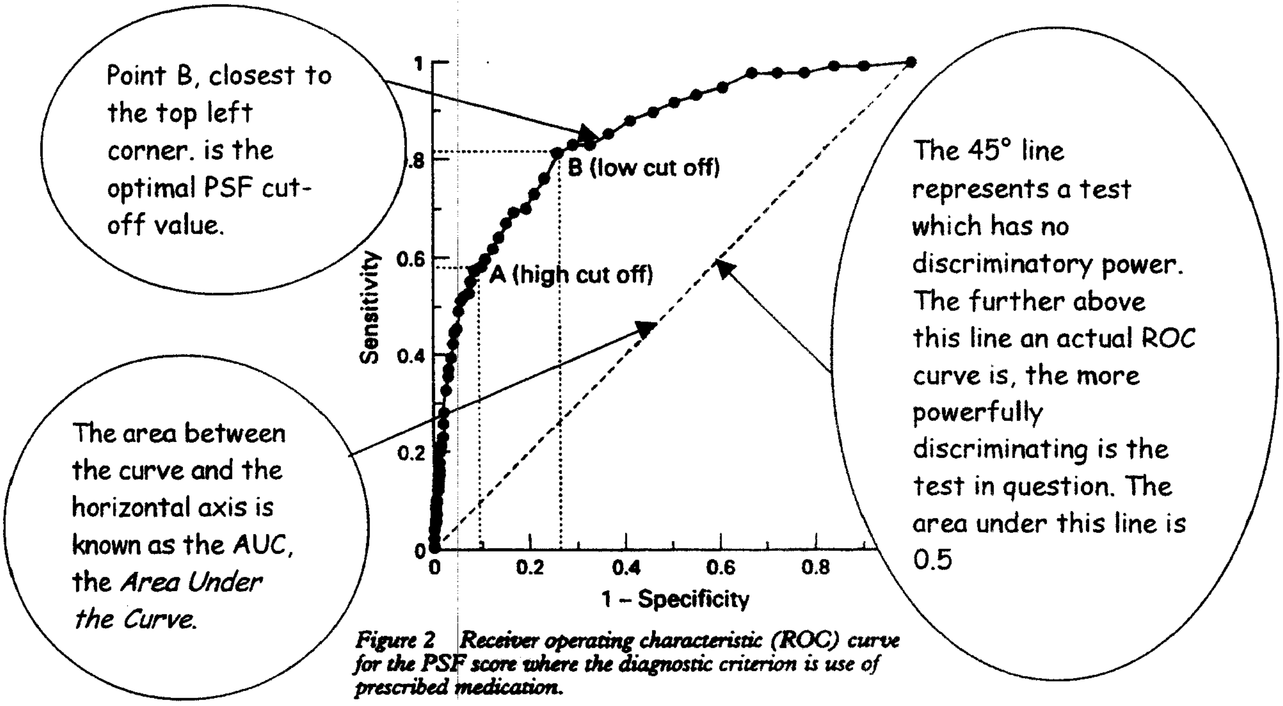
*Positive predictive value=No of falls with score ≥*n*/No of all scores ≥*n*.
†Negative predictive value=No of falls with score <*n*/No of all scores <*n*.

Figure 9.2 Diagnostic tests performance estimates, using two alternative cut- offs (with 95% confidence intervals) for STRATIFY, a risk assessment tool for detecting potential fallers in an elderly hospital population. BMJ, 315, 1997.

The ROC curve

We now want to return to the trade-off between sensitivity and specificity. One popular method for finding the optimum cut-off point, is to draw a Receiver Operating Curve or **ROC**. This is a plot, for each cut-off point, of sensitivity, the true positive rate, on the vertical axis, versus (1 - specificity), the false positive rate, on the horizontal axis. The optimal cut-off is that point on the curve which lies closest to the top left corner. This is also the point which maximises the **area under the curve** (or **AUC**). In practice, the AUC is calculated (along with its 95% confidence interval) for each cut-off point and the largest value indicates the optimum cut-off.

To see how this works, consider Figure 9.3 which is a ROC diagram from a study proposing a new scale - the Psychiatric Symptom Frequency (or PSF) scale - to measure symptoms of anxiety and depression in the UK population. The scale has a range from 0 to 100 (low scores good, high scores bad). The scale was validated with a sample of 3262 subjects taken from the MRC's National Survey of Health and Development, which has followed a cohort of 5374 men and women from birth in 1946.



Point B, closest to the top left corner. is the optimal PSF cut-off value.

The area between the curve and the horizontal axis is known as the AUC, the *Area Under the Curve.*

The 45° line represents a test which has no discriminatory power. The further above this line an actual ROC curve is, the more powerfully discriminating is the test in question. The area under this line is 0.5

Figure 2 Receiver operating characteristic (ROC) curve for the PSF score where the diagnostic criterion is use of prescribed medication.

Figure 9.3 The ROC curve for each point on the PSF. J of Epid & Community Health, 51, 1997.

The 45° diagonal represents a diagnostic test that does not discriminate between those with the condition and those without. The AUC for this diagonal line is 0.5. The ROC shown here has been plotted for a range of cut-off points in the PSF scale. Point A represents cut-off PSF between 22 and 23. The optimal cut-off, Point B, is between 13 and 14.

As a final point, note that if a test uses a nominal (yes/no) measure, for example, blood in stool (Y/N), pain when urinating (Y/N), etc., then there can clearly be no trade-off between sensitivity and specificity.

Q. 9.6 STASH (Spurious Test Against Systemic Hash) is a new roadside diagnostic test for the presence of marihuana in drivers,developed by a public health department. It involves scoring drivers on 10 items, each with a score of 0 or 1. The total score thus ranges from 0 to 10 (high scores is baaad). The researchers want to establish the optimal cut-off point. Drivers with a STASH larger than this cut-off will receive an on-the-spot fine.

In a validation exercise in a new area, 14 drivers whom biochemical assay showed had used marihuana and 14 who similarly hadn't, were given the STASH. The Stash score for each driver is shown in the table below.

(a) Plot these points on a graph similar to that shown in Figure 9.1.

Previous similar studies in other road authority areas has suggested that the optimal cut-off point for the STASH is either ≥ 4, ≥ 5, or ≥ 6. (b) Draw the two axes of a ROC (both axes from 0 to 100%) and plot the ROC values for these three points (and points ≥ 9, ≥ 2 and ≥ 1, if you wish to see the whole curve). (c) Which cut-off is optimal here? Mark this as a horizontal line on your figure, as in Figure 9.1. What are sensitivity and specificity at the optimal cut-off? (d) What is the PPV at the optimal cut-off point? What does your result mean? (e) What are the risks for drivers in this area who have *not* used marihuana if the optimal cut-off point determined above is put into use? (Note: all of this is hypothetical.)

Driver	1	2	3	4	5	6	7	8	9	10	11	12	13	14
User	10	9	9	8	8	7	7	6	6	5	5	4	3	1
Non-user	6	5	4	4	3	3	2	2	2	1	1	1	0	0

* * *

There are many other areas of medical statistics that time considerations prevent us from discussing. Prime among these is *meta-analysis* - the merging of several smaller, less-precise, studies into a single, larger, and (hopefully therefore) more precise study, and *survival analysis* - the study of comparative survival times by groups of patients.

Unit 9 Diagnostic Tests - Solutions to Questions

Q. 9.1 Two patients who've *had* an AMI will *not* be detected; while seven patients who have *not* had an AMI will be *incorrectly* detected.

Q. 9.2

		TP	FP	FN	TN
Cut-off	(a) ≥ 5μg/L	50	11	0	9
	(b) ≥ 12μg/L	48	0	2	20

Q. 9.3 (1 - specificty) = (1 - d/(b+d)) = (b+d-d)/(b+d) = b/(b+d), which is the false positive rate, i.e. those who are identified as having the condition who do not in fact have it.

Q. 9.4

(a) Cut-off ≥ 5μg/L.

		AMI		
		Yes	No	totals
CK-BB Test result	≥ 5μg/L	50	11	61
	< 5μg/L	0	9	9
	totals	50	20	70

Sensitivity = 50/50 = 1.00, or 100%
Specificity = 9/20 = 0.45 or 45%
PPV = 50/61 = 0.82 or 82%
NPV = 9/9 = 1.0 or 100%

(b) Cut-off ≥ 8μg/L

		AMI		
		Yes	No	totals
CK-BB Test result	≥ 8μg/L	48	7	55
	< 8μg/L	2	13	15
	totals	50	20	70

Sensitivity = 48/50 = 0.96, or 96%
Specificity = 13/20 = 0.65 or 65%
PPV = 48/55 = 0.87 or 87%
NPV = 13/15 = 0.87 or 87%

(c) Cut-off of 12μg/L

		AMI		
		Yes	No	totals
CK-BB Test result	≥ 12μg/L	48	0	48
	< 12μg/L	2	20	22
	totals	50	20	70

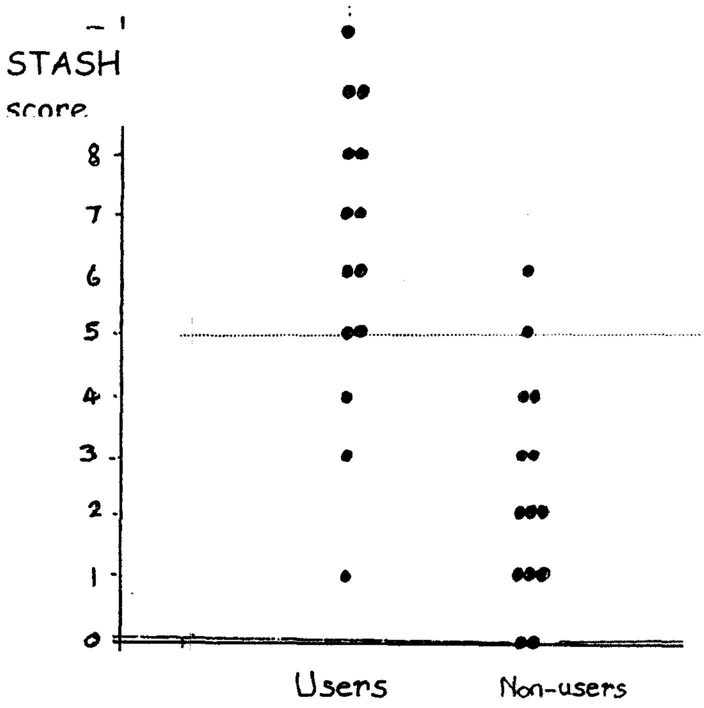
Sensitivity = 48/50 = 0.96, or 96%
Specificity = 20/20 = 1.00 or 100%
PPV = 48/48 = 1.00 or 100%
NPV = 20/22 = 0.91 or 91%

It is helpful to collect these results together, as in the table below:

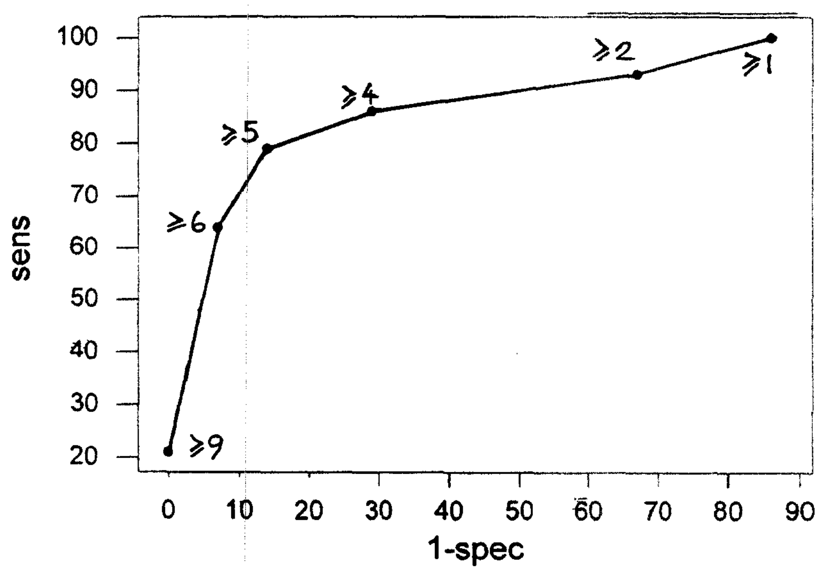
	CK-BB cut-off (μg/L)		
	5	8	12
Sensitivity	100	96	96
Specificity	45	65	100
PPV	82	87	100
NPV	100	87	91

Q. 9.5 (a) A sensitivity of 93% means that 93% of those likely to be fallers will be identified but 7% *won't* be identified. A PPV of 62.3% means that 62.3% of those whom STRATIFY identifies as potential fallers *will* be potential fallers but 37.7% won't be. (b) Sensitivity with a cut-off value of ≥ 3 (has the widest confidence interval).

Q. 9.6 (a) The figure shows the values of the STASH scores for the two groups, users and non-users:



(b) The ROC



(c) Optimal cut-off is ≥ 5, the point on the ROC nearest to the top left-hand corner. This is the point for which sensitivity = 0.78 or 78% and specificity = 0.86 or 86%.

(d) PPV when cut-off is ≥ 5 is: 11/13 = 0.85 or 85%. This means that 85% of those who test positive will have used marihuana.

(e) The number of false positives is equal to (1-specificity) which is (1 - 0.86) = 0.14 or 14%. So 14% of those identified as marihuana users will *not* be users.