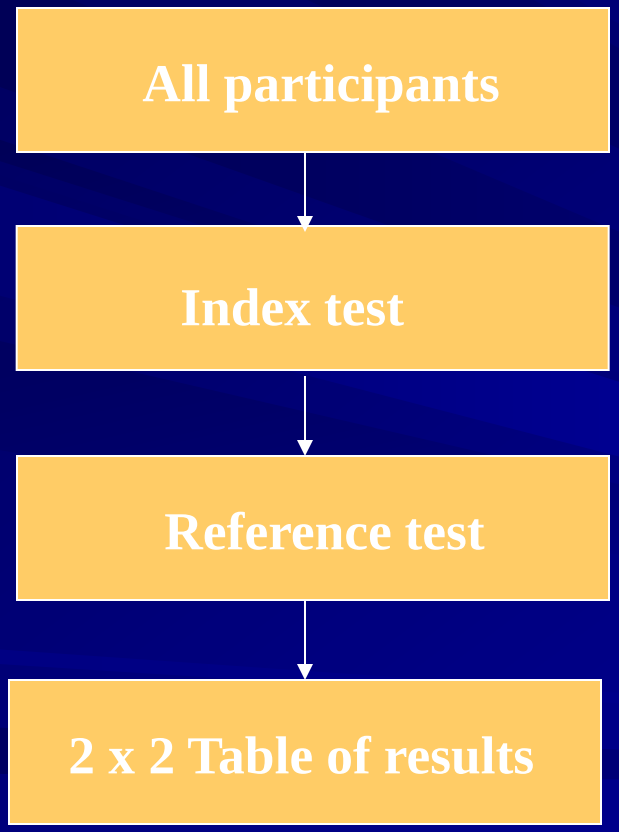
Diagnostic and screening tests

John Addison
Library Services Manager
Royal Oldham Hospital

Outline

- Design
- Sensitivity
- Specificity
- ROC curves
- Predictive values
- Likelihood ratios
- Eliminating bias

Ideal design of diagnostic studies



2 x 2 table of results

Index test	Reference test: Sick	Reference test: Well
Suspicious	90	60
Clear	10	240

Slide from Susan Mallett, Cancer Research UK Medical Statistics Group

Tests

- Index test

  New test

- Reference test

  Definitive result or ‘Gold standard’

  May be biopsy or long term follow up

  Existing test with known weaknesses is no use as a reference

Hypothetical example

	Actual +ve	Actual –ve	Total
Test +ve	95	45	140
Test -ve	5	855	860
Total	100	900	1000

- What proportion of people tested had the disease?

Another question…

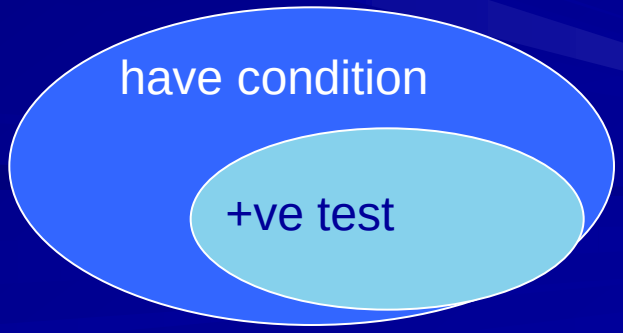
	Actual +ve	Actual –ve	Total
Test +ve	95	45	140
Test -ve	5	855	860
Total	100	900	1000

- How ‘accurate’ is the test? (proportion of correct results)

Sensitivity

	Actual +ve	Actual –ve	Total
Test +ve	95	45	140
Test -ve	5	855	860
Total	100	900	1000

- What proportion of people who have the condition are identified as positive by the test?



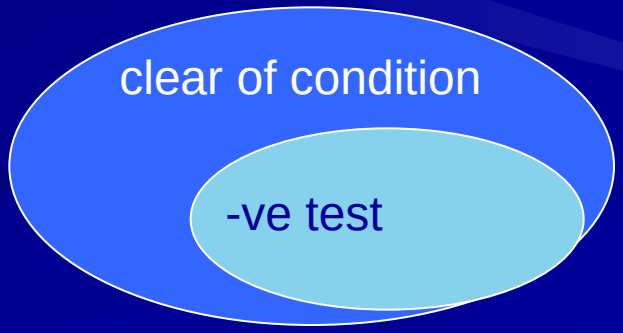
Sensitivity

- If a test has very high sensitivity
  - –most people with the condition are picked up by the test

Specificity

	Actual +ve	Actual –ve	Total
Test +ve	95	45	140
Test -ve	5	855	860
Total	100	900	1000

- What proportion of people who do not have the condition are identified as negative by the test?



Specificity

- If a test has very high specificity
  - –most people without the condition are ruled out by the test

Screening tests I

A new rapid urine test was evaluated as a screening tool for *Chlamydia trachomatis* infection in men. The test was compared with the gold standard diagnostic test for chlamydia infection—the polymerase chain reaction (PCR) assay. The rapid urine screening test was reported to have a sensitivity of 82.6% and specificity of 98.5%.

Which of the following statements, if any, are true?

a) Diagnostic ability: The rapid urine screening test does not provide a diagnosis of chlamydia infection.

b) Definition of sensitivity: out of all of the men with a “positive” result on the rapid urine screening test, 82.6% had a diagnosis of chlamydia infection on the PCR assay.

c) Definition of sensitivity: out of all of the men with a diagnosis of chlamydia infection according to the PCR assay, 82.6% had a “positive” result on the rapid urine screening test.

d) Definition of specificity: out of all of the men without a diagnosis of chlamydia infection on the PCR assay, 98.5% had a “positive” result on the rapid urine screening test.

Answers a and c are correct

Screening tests II

The performance of digital mammography as a screening tool for breast cancer was investigated in premenopausal and perimenopausal women. Digital mammography returned a “screen positive” or “screen negative” result. Breast cancer status was diagnosed by biopsy (gold standard). Digital mammography was reported to have a sensitivity of 47%, specificity of 97%, and positive predictive value of 10%.

Which of the following statements, if any, are true?

a) For every 100 women with diagnosed breast cancer, 47 will have had a “screen positive” result.

b) For every 100 women with a “screen positive” result, 47 will have diagnosed breast cancer.

c) For every 100 women without diagnosed breast cancer, 97 will have a “screen negative” result.

d) For every 100 women with a “screen positive” result, 10 will have diagnosed breast cancer.

Answers a, c, and d true.
Answer b is not true

Notes

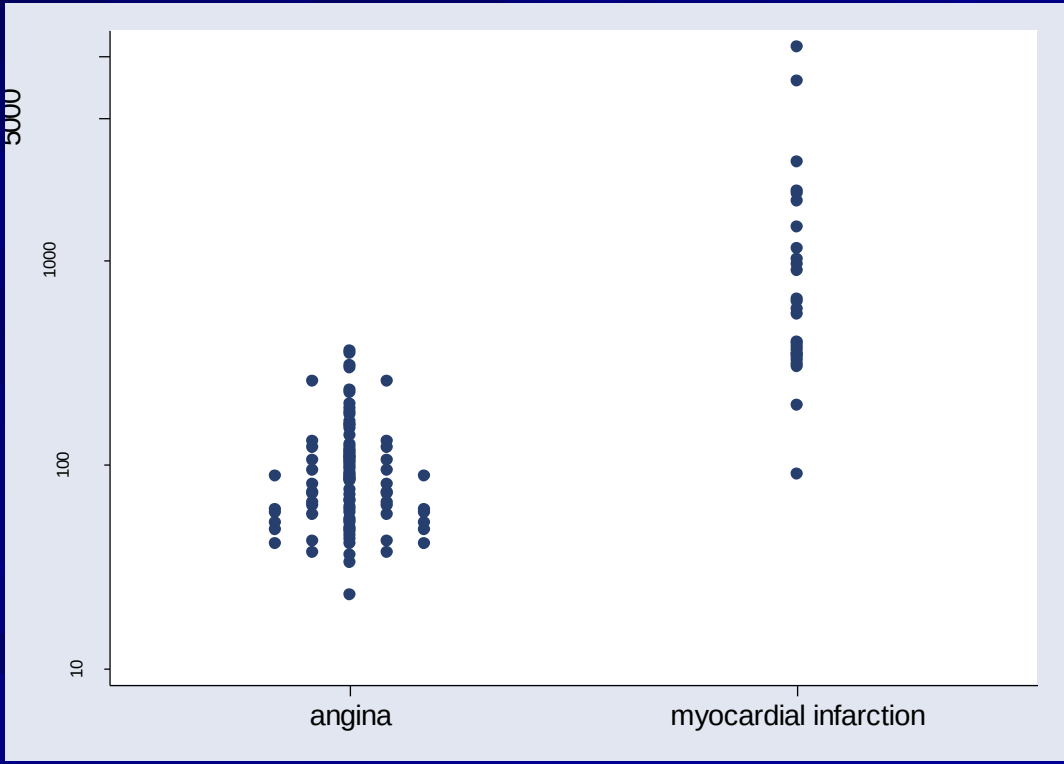
- It is essential to have a confirmed true diagnosis (+ve/-ve) for every patient to be able to judge the accuracy of a test (e.g. gold standard, long term follow up)
- Sensitivity and specificity should be accompanied by confidence intervals to convey the amount of uncertainty

Investigating the trade off between sensitivity and specificity

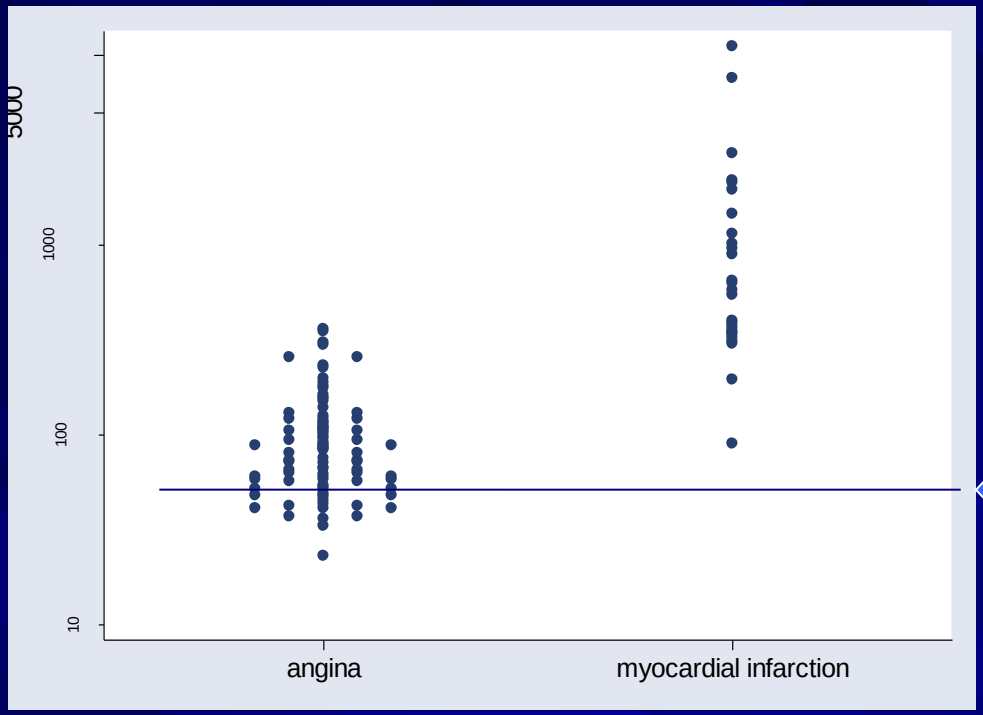
- Generally plot sensitivity v (1-specificity) ie true positive rate v false positive rate
- ROC curve (receiver operating characteristic)
- Look for cut off that gives us both high sensitivity and high specificity
  - Increase in sensitivity is at expense of specificity and vice versa

Tests based on continuous variables:

e.g. creatinekinase in patients with unstable angina or acute myocardial infarction



Data of Frances Boa, from ‘An introduction to Medical Statistics’ by Martin Bland

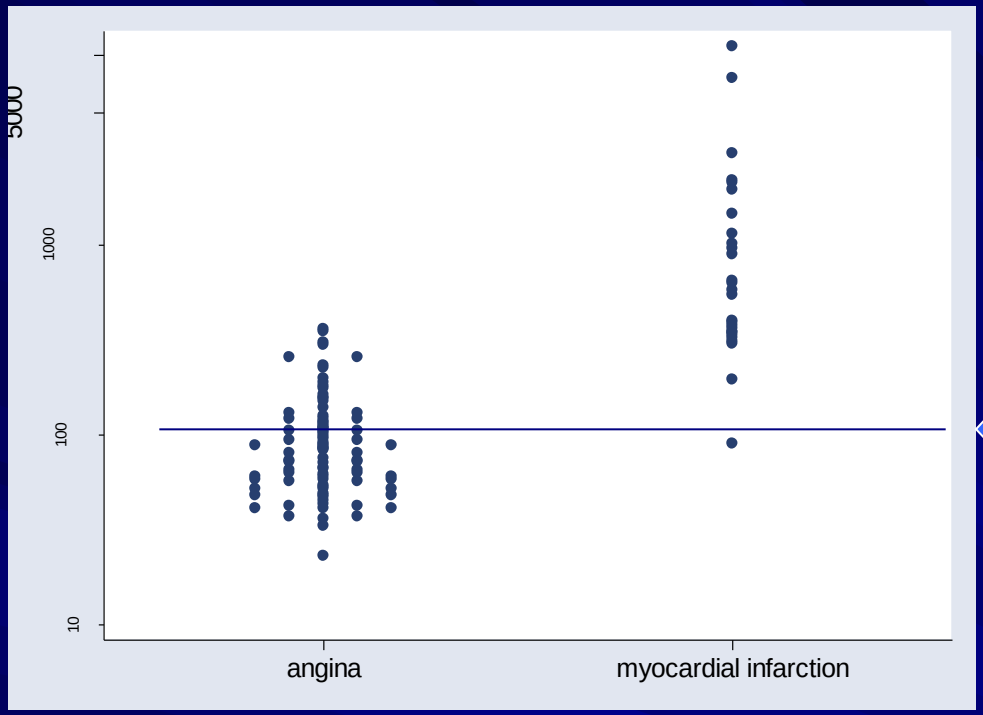


cut-off at 80

	Actual +ve	Actual –ve	
Test +ve	27	54	81
Test -ve	0	39	39
	27	93	120

Sensitivity=27/27=100%

Specificity=39/93=42%



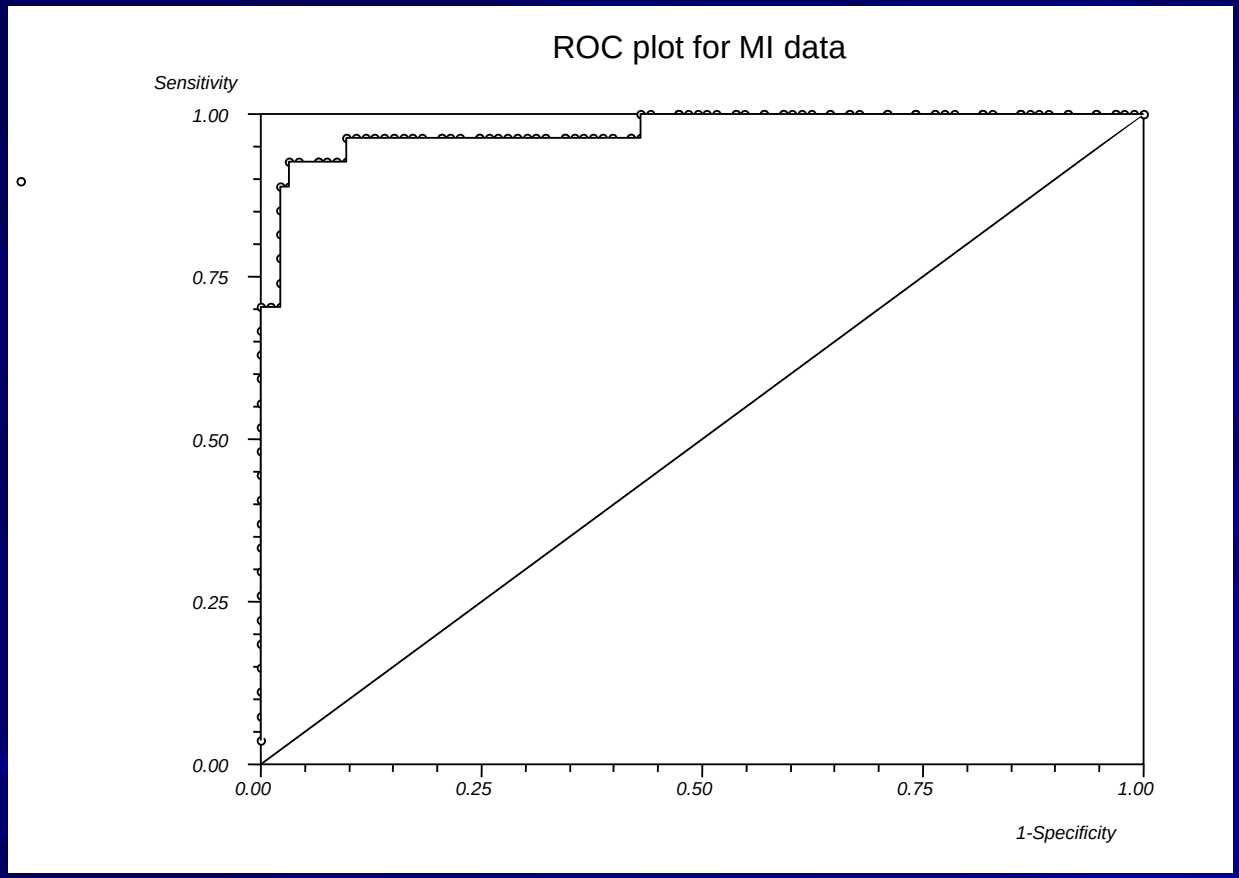
cut-off at 100

	Actual +ve	Actual –ve	
Test +ve	26	35	61
Test -ve	1	58	59
	27	93	120

Sensitivity=26/27=96%

Specificity=58/93=62%

ROC curve



'Optimum' cut-off point selected = 302

sensitivity (95% CI) = 0.93 (0.76 to 0.99)

specificity (95% CI) = 0.97 (0.91 to 0.99)

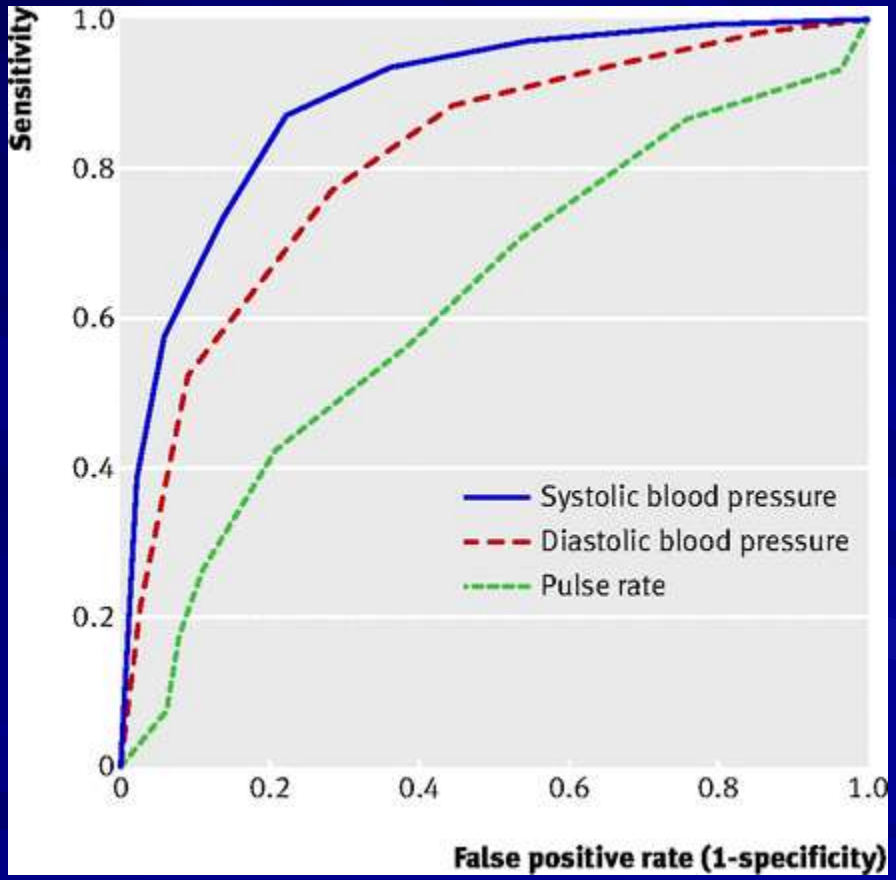
Note: 'optimum' assumes sensitivity and specificity of equal concern

Receiver operating characteristic curves (1)

Researchers investigated the performance of vital signs as screening tests for identifying brain lesions in patients with impaired consciousness on arrival at an emergency department. A total of 529 consecutive patients presenting with impaired consciousness, as assessed by a score of less than 15 on the Glasgow coma scale, were included. The vital signs of systolic and diastolic blood pressure plus pulse rate were recorded on arrival. All patients were followed until discharge, and the final diagnosis of a brain lesion was determined after brain imaging and neurological examination were performed. In total, 312 patients were diagnosed with a brain lesion.

The performance of each vital sign as a screening tool for diagnosed brain lesions was evaluated separately. The measurement scale for each vital sign was categorised using equal sized strata. Each stratum for a vital sign was taken successively as the cut off between a “negative” and “positive” screening test result; all measurements with values greater than the categorised strata were considered a “positive” result, and all others were considered “negative.” For each stratum of a vital sign unique sensitivity and specificity values were derived, from which a receiver operating characteristic curve for each vital sign was derived

The ROC curves for each of the three vital signs as screening tools for diagnosed brain lesions



Which vital sign showed the greatest discrimination as a screening tool for diagnosed brain lesions?

- a) Diastolic blood pressure
- b) Pulse rate
- c) Systolic blood pressure

Systolic blood pressure (answer *c*) showed the greatest discrimination as a screening tool for diagnosed brain lesions.

Receiver operating characteristic curves 2

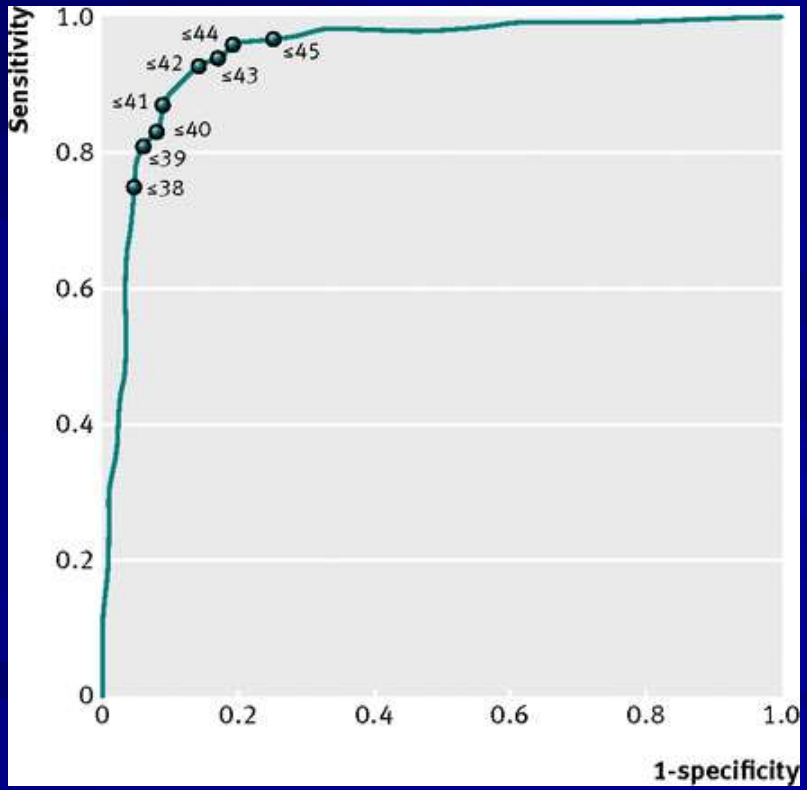
Researchers evaluated the performance of a cognitive test called “test your memory” as a screening test for Alzheimer’s disease. The screening test is designed to use minimal operator time and to be suitable for non-specialist use. It is self administered under medical supervision. The test has a minimum score of zero and a maximum score of 50; lower scores indicate greater cognitive impairment.

The study was based in hospital outpatient departments. Participants included 94 patients diagnosed as having Alzheimer’s disease. For each patient, three age matched healthy controls without Alzheimer’s disease (n=282) were recruited from accompanying relatives. All patients and controls completed the screening test.

Receiver operating characteristic curves

The optimal test score for discriminating between patients with Alzheimer's disease and controls was investigated. Each score from 50 down to zero was taken successively as the cut-off point between a "negative" and "positive" screening test result; all scores less than or equal to the cut-off score were considered positive and others were considered negative. For each cut-off score the sensitivity and specificity of the screening test was calculated, and these values were used to derive a receiver operating characteristic curve. The area under the receiver operating characteristic curve was 0.95.

Receiver operating characteristic curve for “test your memory” scores differentiating patients with Alzheimer’s disease (n=94) and age matched controls (n=282). Numbers on curve refer to a range of selected cut-off scores between “negative” and “positive” results



Which of the following statements, if any, are true?

a) For successive cut-off scores, as the sensitivity of the screening test decreases in value the specificity increases

b) The value of (1 minus specificity) represents the proportion of controls identified as positive (high risk) on screening

c) A screening test with an area under the curve equal to one half (0.5) would discriminate perfectly between patients and controls

Statements *a* and *b* are true, whereas *c* is false.

Note

- Should always check sensitivity and specificity of cut off in a different sample to be sure

Predictive value

- Positive predictive value

  If a person tests positive, what is the probability that they have the disease?

- Negative predictive value

  If a person tests negative, what is the probability that they do not have the disease?

Rule of thumb

- Sensitivity and Specificity

  tell you about the test in general

- Positive/Negative Predictive Value

  tell you what a test result means for the patient in front of you

Positive predictive value

	Actual +ve	Actual –ve	Total
Test +ve	95	45	140
Test -ve	5	855	860
Total	100	900	1000

- If a person tests positive, what is the probability that they have the disease?

Negative predictive value

	Actual +ve	Actual –ve	Total
Test +ve	95	45	140
Test -ve	5	855	860
Total	100	900	1000

- If a person tests negative, what is the probability that they do NOT have the disease?

Practice exercise (5)

1000 people over 60 submit faecal samples as part of the bowel cancer screening programme. 100 of them truly have cancer, of whom the test correctly identifies 90. The test also indicates a positive result for another another 60 people who are actually disease free.

Cancer screening: complete the cells in this table

	Bowel cancer	No bowel cancer	Total
Test positive			
Test negative			
Total			1000

Practice exercise (6)

Lord Snooty will be disinherited unless he persuades a suitable debutante to marry him by the end of the year. He decides to propose to as many debs as he has time to approach at the last ball of the season. He reasons that he will maximise his chances of becoming betrothed by only approaching those debs who are not wearing rings.

Given the following data is Snooty's reasoning sound?

200 debs are wearing rings, of whom 100 are married or engaged. Of the 800 ringless debs another 50 are secretly married or engaged

Debs at the end of season ball

	Married or engaged	Not married or engaged	Total
Wear a ring			
No ring			
Total			1000

Scenario 1: trouble in Cairo

The prevalence of West Nile disease amongst symptomatic patients presenting at the Nasser General Hospital is 10%. The sensitivity of the The IgM antibody capture enzyme-linked immunosorbent assay is 85% and its specificity is 90%. What is the positive predictive value of the test?

Sudoku for medics: fill in the missing values

	West Nile disease	No West Nile disease	Total
Test positive			
Test negative			
Total			1000

Scenario 2: pregnancy testing in ancient times

An ancient Egyptian pregnancy test involved watering bags of wheat and barley with the urine of a possibly pregnant woman. Germination indicated pregnancy. If 15% of women who take the test are actually pregnant, and the positive predictive value of this test was 60%, and the sensitivity was 40% what was the specificity of the test?

Pregnancy testing in Ancient Egypt

	Pregnant	Not pregnant	Total
Test positive			
Test negative			
Total			1000

Scenario 3: whooping cough

The prevalence of pertussis amongst a tested population during an outbreak of the disease is 5%. Following polymerase chain reaction testing 1% of patients tested receive a false negative diagnosis, and 4% receive a false positive diagnosis. What is the negative predictive value of the PCR test?

Whooping cough

	Got pertussis	Not got pertussis	Total
PCR test positive			
PCT test negative			
Total			1000

Scenario 4: breast cancer screening

Breast cancer amongst women over 55 has a prevalence of 10%. If only 2% of routine mammogram screenings result in false negative results what is the sensitivity of the test?

Breast cancer screening

	Breast cancer	No breast cancer	Total
Positive mammogram			
Negative mammogram			
Total			1000

PROBLEM

- Sensitivity and specificity of a test should be constant
- However, positive and negative predictive values will vary depending on the prevalence

Test with 95% sensitivity and 95% specificity in a population with a diseases prevalence of 1%

	True Positive	True Negative	Total
Screen Positive	19	99	118
Screen Negative	1	1881	1882
Total	20	1980	2000

Positive predictive value = 0.16

Test with 95% sensitivity and 95% specificity in a population with a diseases prevalence of 5%

	True Positive	True Negative	Total
Screen Positive	96	96	190
Screen Negative	5	1805	1810
Total	100	1900	2000

Positive predictive value = 0.5

Test with 95% sensitivity and 95% specificity in a population with a diseases prevalence of 10%

	True Positive	True Negative	Total
Screen Positive	190	90	280
Screen Negative	10	1710	1720
Total	200	1800	2000

Positive predictive value = 0.67

Test with 95% sensitivity and 95% specificity in a population with a diseases prevalence of 25%

	True Positive	True Negative	Total
Screen Positive	475	75	550
Screen Negative	25	1425	1450
Total	500	1500	2000

Positive predictive value = 0.86

Partial solution: Likelihood ratio

- Likelihood Ratio of a Positive Test (LR+)

  how many times more likely is a positive test to be found in a patient with, as opposed to without, the disease?

- LR+ = sensitivity/(1 – specificity) or true+ rate/false+ rate

See Diagnostic test 4: likelihood ratios by Deeks and Altman; BMJ 2004 329 p168

- Likelihood Ratio indicates by how much the test result raises or lowers the pretest probability of the disease
- LR = 1.0 means post-test probability is same as pretest probability

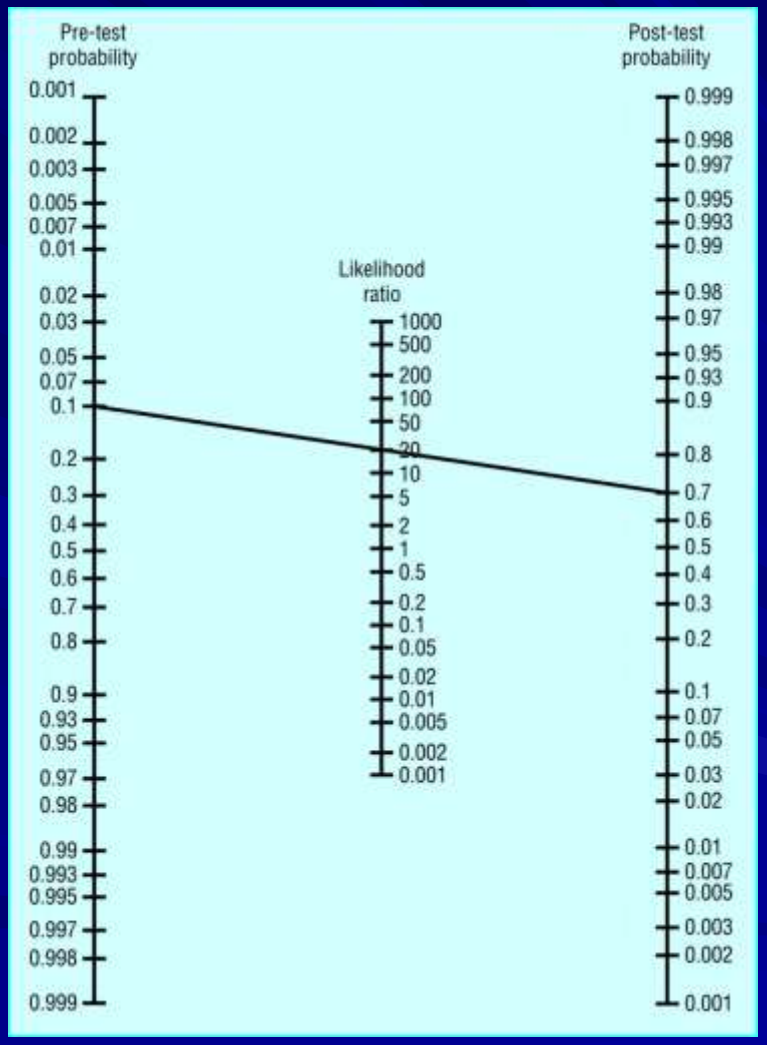
rough guide:

LR	Interpretation
> 10	Large and often conclusive increase in the likelihood of disease
5 - 10	Moderate increase in the likelihood of disease
2 - 5	Small increase in the likelihood of disease
1 - 2	Minimal increase in the likelihood of disease
1	No change in the likelihood of disease
0.5 - 1.0	Minimal decrease in the likelihood of disease
0.2 - 0.5	Small decrease in the likelihood of disease
0.1 - 0.2	Moderate decrease in the likelihood of disease
< 0.1	Large and often conclusive decrease in the likelihood of disease

See: Guyatt G, Users' guides to the medical literature, AMA Press, 2002, p.199

- Likelihood ratio can be used to calculate the probability of individual patient having condition based on test results

Use of Fagan's nomogram for calculating post-test probabilities



Deeks, J. J et al. BMJ 2004;329:168-169

How do you know the pre-test probability?

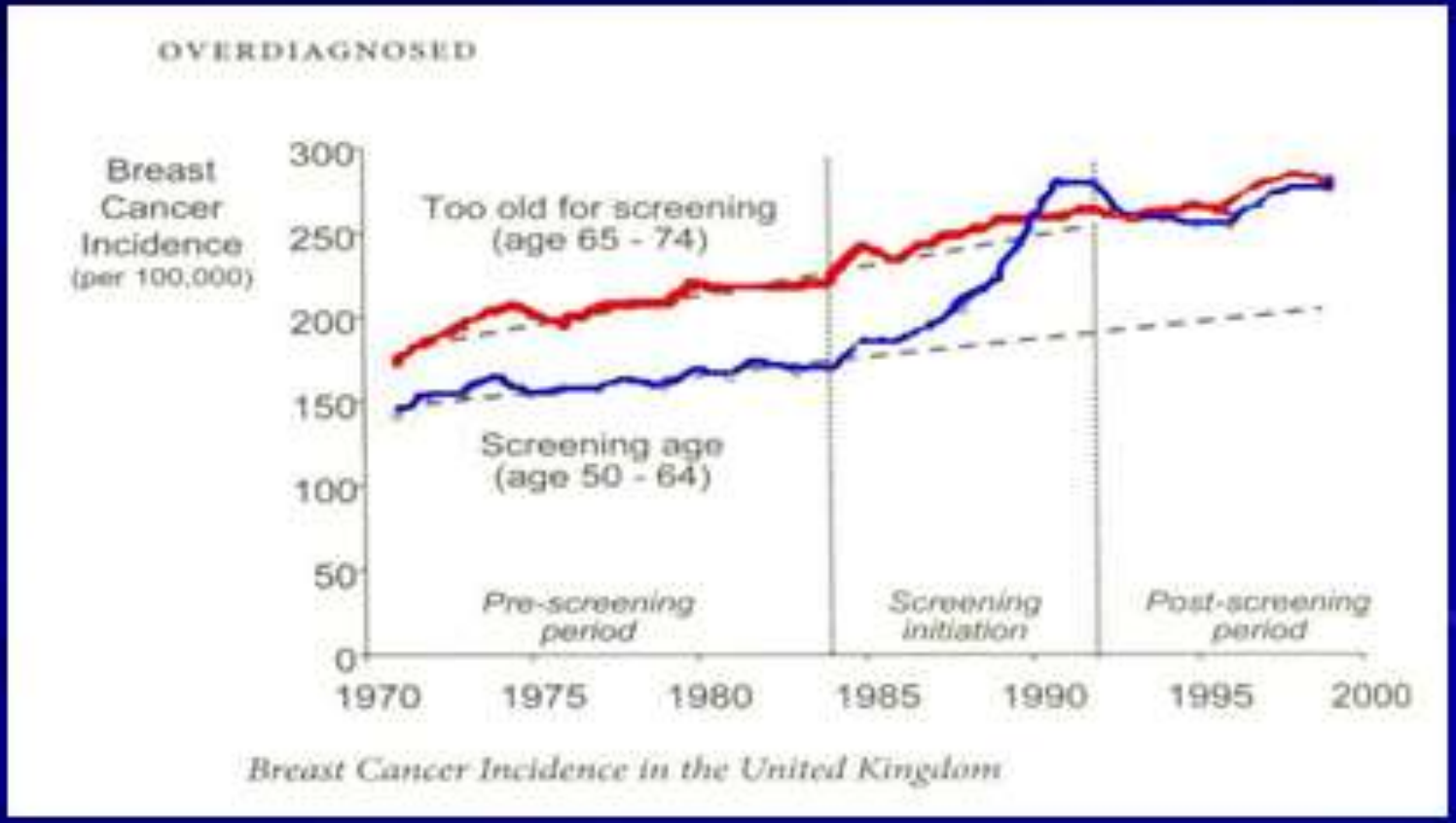
- Own experience (use with caution)
- Regional/national prevalence statistics (but you may be interested in patients with a particular symptom)
- Local/regional/national databases reporting frequency of disorders diagnosed with certain symptoms (a thing of the future?)
- Use the pre-test probabilities observed in a research paper on the diagnostic test: can it be applied to your patients?
- (most ideal) Research reports devoted to documenting pre-test probabilities for the possible diagnoses for a patient with a specific set of symptoms

Straus S E et al, Evidence-Based Medicine: how to practice and teach EBM, Elsevier, 2005, pp.80-81

Which test for which purpose?

- Important that tests are developed in population for which they will be used
- Good diagnostic test not necessarily a good screening test

…and mass screening programmes may create as many problems as they solve



The benefits and harms of breast cancer screening: an independent review

A report jointly commissioned by Cancer Research UK and the Department of Health (England) October 2012.

Putting together benefit and overdiagnosis …the panel estimates that for 10 000 UK women invited to screening from age 50 for 20 years, about 681 cancers will be found of which 129 will represent overdiagnosis, and 43 deaths from breast cancer will be prevented. In round terms, therefore, for each breast cancer death prevented, about three overdiagnosed cases will be identified and treated. Of the ~307 000 women aged 50–52 who are invited to screening each year, just >1% would have an overdiagnosed cancer during the next 20 years. Given the uncertainties around the estimates, the figures quoted give a spurious impression of accuracy.

Evidence from a focus group conducted by Cancer Research UK … in line with previous similar studies, was that this was an offer many women will feel is worth accepting: the treatment of overdiagnosed cancer may cause suffering and anxiety, but that suffering is worth the gain from the potential reduction in breast cancer mortality. Clear communication of these harms and benefits to women is of utmost importance and goes to the heart of how a modern health system should function.

2013 Cochrane review: Screening for breast cancer with mammography

The review includes seven trials that involved 600,000 women in the age range 39 to 74 years who were randomly assigned to receive screening mammograms or not. The studies which provided the most reliable information showed that screening did not reduce breast cancer mortality. Studies that were potentially more biased (less carefully done) found that screening reduced breast cancer mortality. However, screening will result in some women getting a cancer diagnosis even though their cancer would not have led to death or sickness.

Currently, it is not possible to tell which women these are, and they are therefore likely to have breasts or lumps removed and to receive radiotherapy unnecessarily. If we assume that screening reduces breast cancer mortality by 15% after 13 years of follow-up and that overdiagnosis and overtreatment is at 30%, it means that for every 2000 women invited for screening throughout 10 years, one will avoid dying of breast cancer and 10 healthy women, who would not have been diagnosed if there had not been screening, will be treated unnecessarily. Furthermore, more than 200 women will experience important psychological distress including anxiety and uncertainty for years because of false positive findings.

The 2008 Cochrane leaflet summary

It may be reasonable to attend for breast cancer screening with mammography, but it may also be reasonable not to attend, as screening has both benefits and harms.

If 2000 women are screened regularly for 10 years, one will benefit from the screening, as she will avoid dying from breast cancer.

At the same time, 10 healthy women will, as a consequence, become cancer patients and will be treated unnecessarily. These women will have either a part of their breast or the whole breast removed, and they will often receive radiotherapy, and sometimes chemotherapy.

Furthermore, about 200 healthy women will experience a false alarm. The psychological strain until one knows whether or not it was cancer, and even afterwards, can be severe.

But what do women think?

- 68% of women believe that mammography lowers their risk of getting breast cancer
- 62% believe that screening at least halves the rate of breast cancer
- 75% believe that 10 years of screening would prevent 10 breast cancer deaths per thousand women.

Domenighetti G, D'Avanzo B, Egger M, et al. Women's perception of the benefits of mammography screening: population-based survey in four countries. Int J Epidemiol *2003*; *32*: *816*-2

Bias in studies

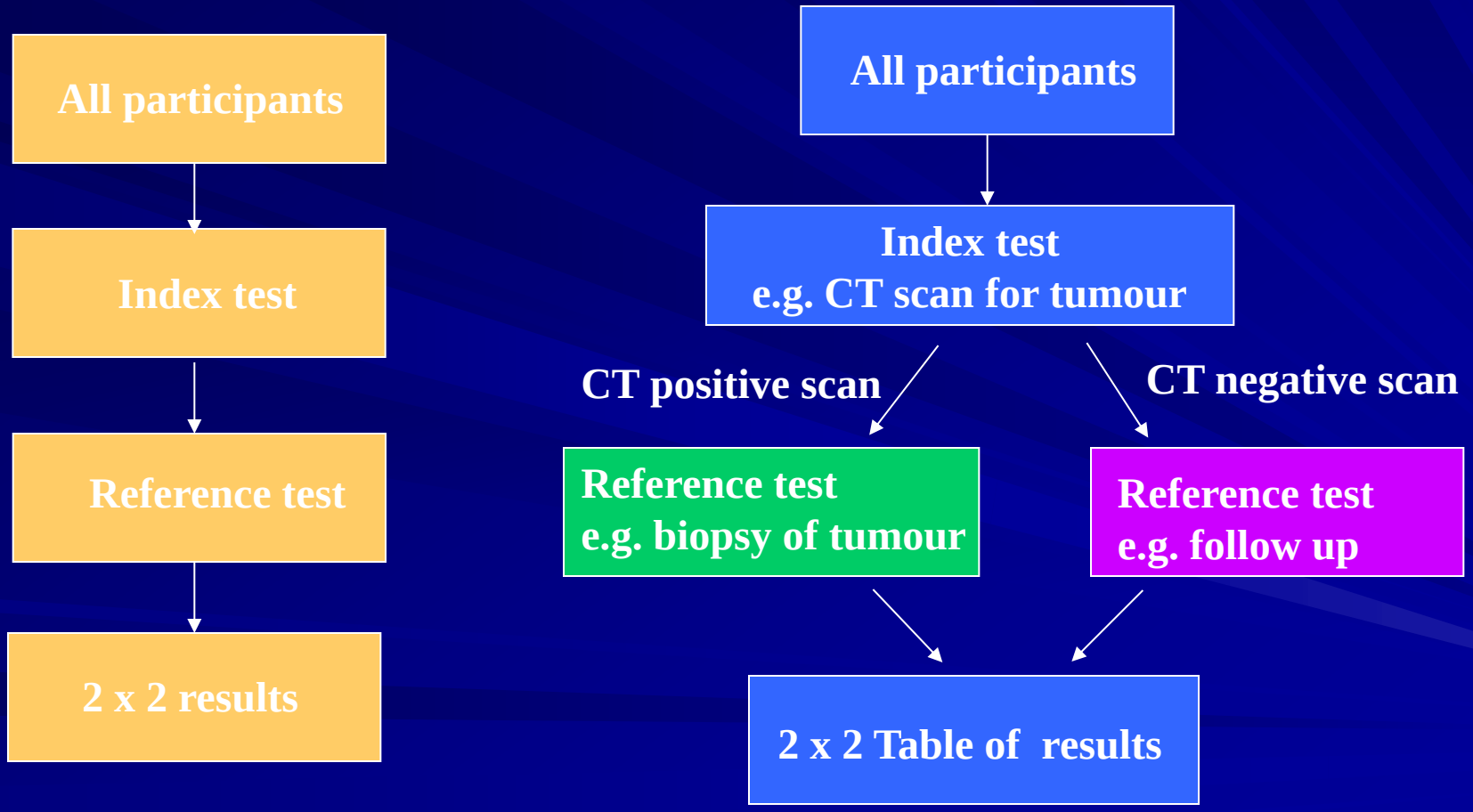
- Is the reference appropriate?
- Was the same reference used for all patients (verification bias)?
- Were assessors blind to case details?
- Was it a ‘diagnostic case-control study’?

See: Trisha Greenhalgh, How to read a paper: Papers that report diagnostic or screening tests, BMJ 1997 315 p540

Differential Verification

Differential verification often inevitable

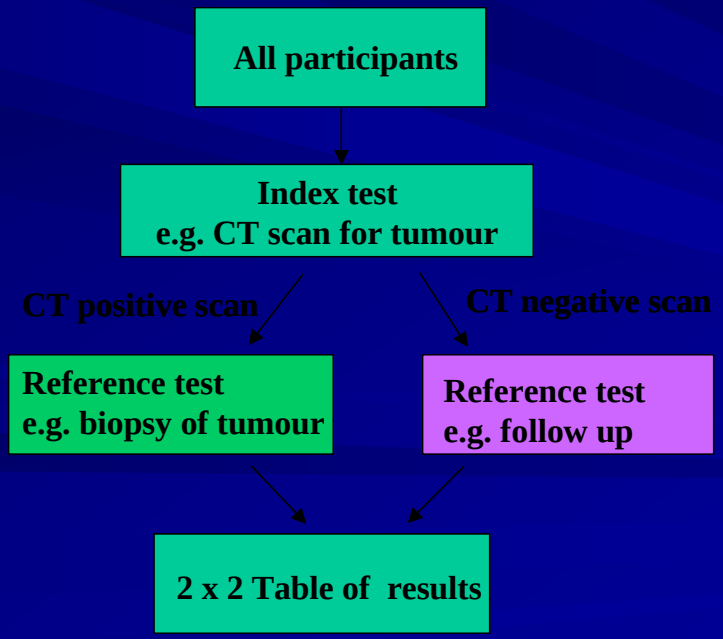
- **biopsy on detected lumps, but follow-up if normal**



Slide from Susan Mallett, Cancer Research UK Medical Statistics Group

Verification bias

- Are the two reference tests as accurate as each other?
- If not, then get verification bias.
- Different accuracies can be due to different time frames e.g. biopsy today vs follow-up over 2 years. Same cancer?



Slide from Susan Mallett, Cancer Research UK Medical Statistics Group

Best evidence

- Reporting using STARD guidelines (Standards for Reporting of Diagnostic Accuracy)
- Systematic reviews (Cochrane)
- Use of QUADAS quality checklist
- RCTs that look at effect of test on patient outcome (rare)

Summary

- All patients must have both new test + reference (gold standard)
- Report 2x2 table and give sensitivity, specificity with precision
- Test cut-offs in independent sample
- Predictive values vary according to prevalence
- Consider all potential sources of bias