CEBM

www.cebm.net

Teaching Tips for Diagnostic Studies

CEBM Teaching Course 2012

Matthew Thompson

Reader, Dept Primary Care Health Sciences

Director, Oxford Centre for Monitoring and Diagnosis

Deputy Director, Centre for Evidence Based Medicine

UNIVERSITY OF OXFORD

Financial Disclosure: Dr Leischow reports consulting and serving as a paid speaker for Pfizer and as a consultant for Johnson & Johnson.

REFERENCES

1. Vevers V. Obama says WH will remain smoke-free. http://www.cbsnews.com/blogs/2008/12/08/politics/politicalhotsheet/entry4654231.shtml. Accessed December 9, 2008.
2. Reinberg S. Cancer to surpass heart disease as world's leading killer. washingtonpost.com Web page. December 9, 2008. http://www.washingtonpost.com/wp-dyn/content/article/2008/12/09/AR2008120901814.html. Accessed December 10, 2008.
3. Centers for Disease Control and Prevention. Cigarette smoking among adults—United States, 2007 [published correction appears in *MMWR Morb Mortal Wkly Rep*. 2008;57(47):1281]. *MMWR Morb Mortal Wkly Rep*. 2008; 57(45):1221-1226.
4. Centers for Disease Control and Prevention. Smoking-attributable mortality, years of potential life lost, and productivity losses—United States, 2000-2004. *MMWR Morb Mortal Wkly Rep*. 2008;57(45):1226-1228.
5. American Nonsmokers' Rights Foundation. Summary of 100% smokefree state laws and population protected by 100% US smokefree laws. January 4, 2009. http://www.no-smoke.org/pdf/SummaryUSPopList.pdf. Accessed December 12, 2009.
6. Risbeck CA. ADHA smoking cessation initiative liaisons, II: partnering with tobacco quitlines. November 2007. http://findarticles.com/p/articles/mi_m1ANQ/is_/ai_n25015057. Accessed December 9, 2008.
7. WHO Framework Convention on Tobacco Control: third session of the conference of the parties to the WHO FCTC. http://www.who.int/fctc/en/. Accessed January 10, 2009.
8. Fiore M, Croyle RT, Curry SJ, et al. Preventing 3 million premature deaths and helping 5 million smokers quit: a national action plan for tobacco cessation. *Am J Public Health*. 2004;94(2):205-210.
9. Peto R, Lopez A. The future worldwide health effects of current smoking patterns. In: Boyle P, Gray N, Henningfield J, Seffrin J, Zatonski W, eds. *Tobacco and Public Health: Science and Policy*. New York, NY: Oxford University Press; 2004: 281-286.
10. Mackay J, Erikson M, Shafey O. *The Tobacco Atlas*. 2nd ed. Atlanta, GA: American Cancer Society; 2006.
11. Fiore MC, Jaén CR, Baker TB, et al. Treating tobacco use and dependence: 2008 update. Rockville, MD: US Dept of Health and Human Services; May 2008. http://www.surgeongeneral.gov/tobacco/. Accessed February 5, 2009.
12. National Institute of Allergy and Infectious Dieseases. Treatment of HIV infection. http://www.niaid.nih.gov/factsheets/treat-hiv.htm. Updated November 8, 2007. Accessed February 8, 2009.
13. CEO Roundtable on Cancer Web page. http://www.ceoroundtableoncancer.org. Accessed December 19, 2008.

Diagnostic Errors—The Next Frontier for Patient Safety

David E. Newman-Toker, MD, PhD
Peter J. Pronovost, MD, PhD

DURING THE PAST DECADE, AWARENESS AND UNDERstanding of medical errors have expanded rapidly, with an energetic patient safety movement promoting safer health care through "systems" solutions. Efforts have focused on translating evidence into practice, mitigating hazards from therapies, and improving culture and communication. Diagnostic errors have received relatively little attention. Although the science of error measurement is underdeveloped, diagnostic errors are an important source of preventable harm.[1-3] In this Commentary, we offer definitions for diagnostic error and misdiagnosis-related harm, present an overview of the magnitude of diagnostic errors, and give suggestions for how research can mature.

Distinguishing Errors From Harms

In considering diagnostic errors, it is important to distinguish between the error (a process) and the resulting harm (an outcome). *Diagnostic error* can be defined as a diagnosis that is missed, wrong, or delayed, as detected by some subsequent definitive test or finding.[1] However, not all misdiagnoses result in harm, and harm may be due to either disease or intervention. *Misdiagnosis-related harm* can be defined as preventable harm that results from the delay or failure to treat a condition actually present (when the working diagnosis was wrong or unknown) or from treatment provided for a condition not actually present.

An estimated 40 000 to 80 000 US hospital deaths result from misdiagnosis annually.[4] Roughly 5% of autopsies reveal lethal diagnostic errors for which a correct diagnosis coupled with treatment could have averted death.[5] In the Harvard Medical Practice Study, physician errors resulting in adverse events were more likely to be diagnostic than drug-related (14% vs 9%), and misdiagnoses were more likely to be considered negligent (75% vs 53%) and to result in serious disability (47% vs 14%).[6] Not surprisingly, tort claims for diagnostic errors are nearly twice as common as claims for medication errors and result in the largest payouts.[7] As with all types of medical error, the human toll of misdiagnosis on an individual or family can be tremendous, particularly when a healthy patient experiences an adverse event.

Diagnostic errors often are unrecognized or unreported, and the science of measuring these errors (and their effects) is underdeveloped.[1,2] Available statistics consider neither deaths due to misdiagnosis in outpatients nor misdiagnosis-related morbidity and associated costs. For example, stroke, the leading cause of serious, long-term disability in the United States, affects 780 000 Americans annually.[8] Opportunities to prevent disabling stroke are missed when patients experiencing mild or transient warning symptoms receive misdiagnoses. According to a recent systematic review, 9% of all cerebrovascular events are missed initially, and the odds of misdiagnosis increase at least 5-fold when symptoms are mild or transient.[9]

Author Affiliations: Departments of Neurology (Dr Newman-Toker) and Anesthesiology and Critical Care (Dr Pronovost), Johns Hopkins University School of Medicine, Baltimore, Maryland.
Corresponding Author: David E. Newman-Toker, MD, PhD, The Johns Hopkins Hospital, Pathology Bldg 2-210, 600 N Wolfe St, Baltimore, MD 21287 (toker@jhu.edu).

©2009 American Medical Association. All rights reserved.

Downloaded from www.jama.com at Oxford University Library Services on May 15, 2009

- 2/3 malpractice claims against GPs in UK
- 40,000-80,000 US hospital deaths from misdiagnosis per year
- Diagnosis uses <5% of hospital costs, but influences 60% of decision making

On the menu this morning

- Tests have multiple roles in health care
- The basic anatomy of diagnostic accuracy studies
- Using pictures to show biases
- Do tests make people better?
- Evaluating new diagnostic tests
- Making the numbers easy (sensitivity, specificity etc)
- Not just accuracy – other outcomes of diagnostic tests
- Systematic reviews
- Useful books and articles

“Diagnosis” means lots of things - tests can have many roles

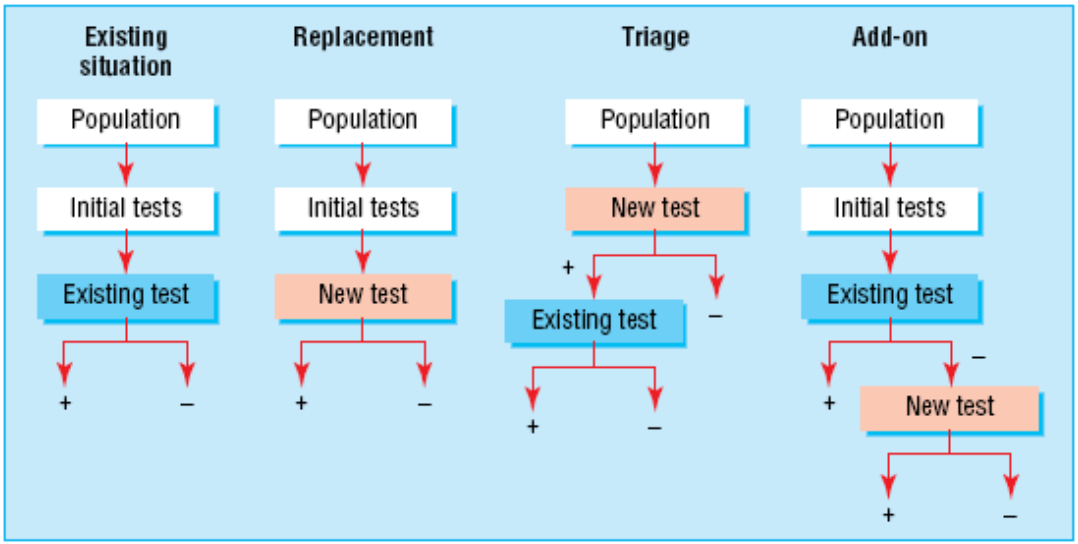
Roles of diagnostic tests in health care

Role	Description	Examples
Confirming or excluding a diagnosis	Used to confirm ("rule in") or exclude ("rule out") particular diagnoses. Most tests will be better at one than the other. May vary between different clinical settings / different spectrum of disease	Normal blood pressure measurement to exclude hypertension. Raised cardiac troponins to confirm cardiac ischaemia
Triage	An initial test in a clinical pathway, which usually directs the need (or not) for further (usually more invasive) testing. Ideal triage test is usually fairly rapid, and should not miss any patients (i.e. minimise false negatives)	Blood pressure and heart rate in initial triage of patients with multiple trauma to identify those with possible shock. D-dimer to screen for presence of pulmonary embolism in patients who have shortness of breath
Monitoring	Tests that are repeated at periodic intervals in patients with chronic conditions, or in those receiving certain treatments, in order to assess efficacy of interventions, disease progression, or need for changes in treatment	Haemoglobin A1c to monitor glucose control in patients with diabetes. Anticoagulation tests for patients taking oral anticoagulants (warfarin). HIV viral load and CD4 count
Prognosis	Provides information on disease course or progression, and individual response to treatment	CT scanning in patients with known ovarian cancer to determine the stage
Screening	Detecting conditions or risk factors for conditions in people who are apparently asymptomatic.	Mammography screening for breast cancer. Cholesterol testing to detect persons at greater risk of cardiovascular disease.

Roles of a new test

- Replacement – new replaces old
  - E.g., CT colonography for barium enema
- Triage – new determines need for old
  - E.g., B-natriuretic peptide for echocardiography
- Add-on – new combined with old
  - ECG and myocardial perfusion scan

Bossuyt et al BMJ 2006;332:1089–92

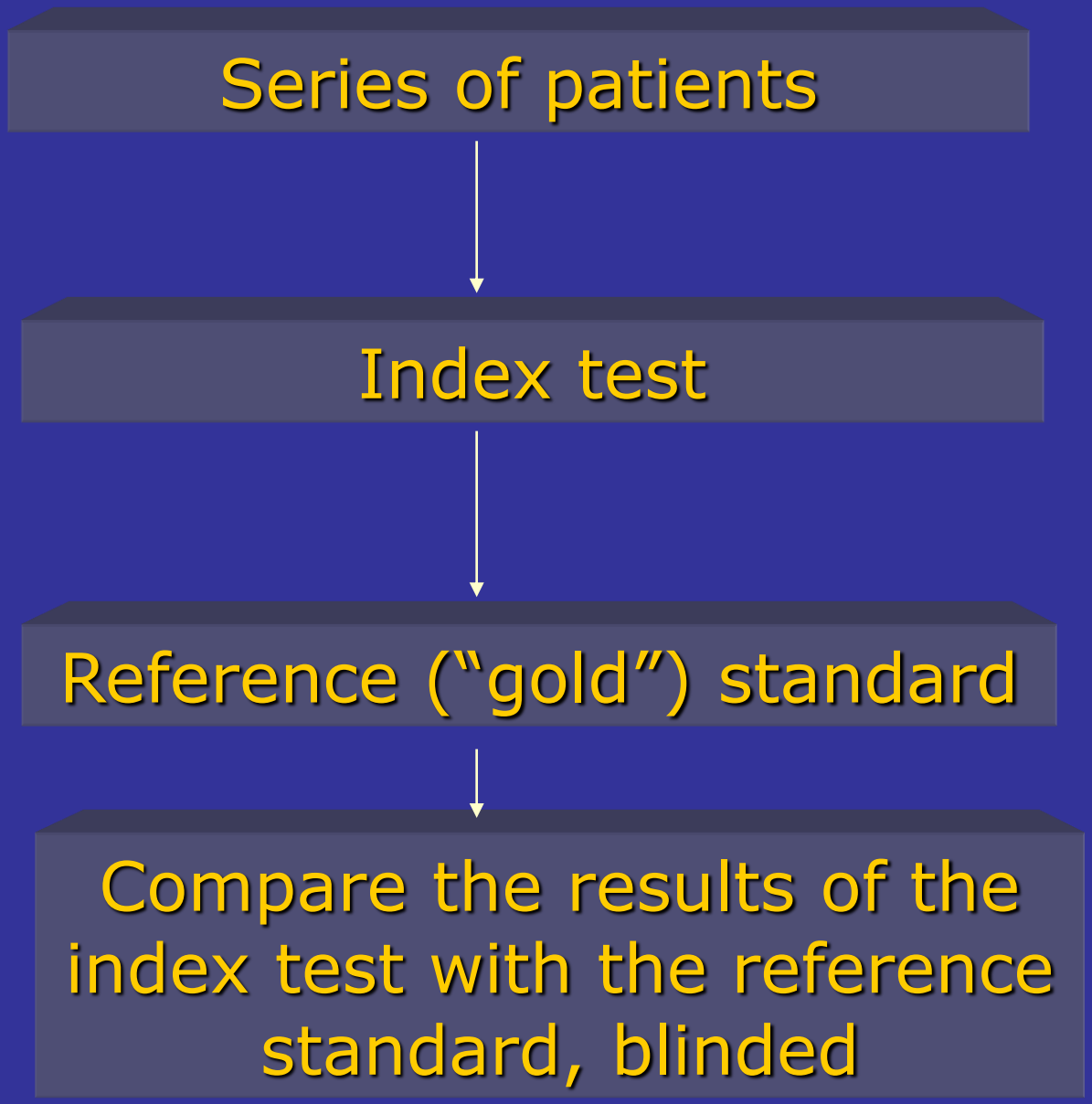


Roles of tests and positions in existing diagnostic pathways

Basic anatomy of Diagnostic Accuracy studies

Defining the clinical question: PICO or PIRT

- Patient/Problem
  - How would I describe a group of patients similar to mine?
- Index test
  - Which test am I considering?
- Comparator… or …Reference Standard
  - What is the best reference (gold) standard to diagnose the target condition?
- Outcome….or….Target condition
  - Which condition do I want to rule in or rule out?


Series of patients
Index test
Reference (“gold”) standard
Compare the results of the index test with the reference standard, blinded

read this abstract

Patient-Initiated Treatment of Uncomplicated Recurrent Urinary Tract Infections in Young Women

Kalpana Gupta, MD, MPH; Thomas M. Hooton, MD; Pacita L. Roberts, MS; and Walter E. Stamm, MD

Background: Recurrent urinary tract infections (UTIs) are a common outpatient problem, resulting in frequent office visits and often requiring the use of prophylactic antimicrobial agents. Patient-initiated treatment of recurrent UTIs may decrease antimicrobial use and improve patient convenience.

Objective: To determine the safety and feasibility of patient-initiated treatment of recurrent UTIs.

Design: Uncontrolled, prospective clinical trial.

Setting: University-based primary health care clinic.

Participants: Women at least 18 years of age with a history of recurrent UTIs and no recent pregnancy, hypertension, diabetes, or renal disease.

Intervention: After self-diagnosing UTI on the basis of symptoms, participating women initiated therapy with ofloxacin or levofloxacin.

Measurements: Accuracy of self-diagnosis determined by evidence of a definite (culture-positive) or probable (sterile pyuria and no alternative diagnosis) UTI on pretherapy urinalysis and culture. Women with a self-diagnosis of UTI that was not microbiologically confirmed were evaluated for alternative diagnoses. Post-therapy interviews and urine cultures were used to assess clinical and microbiological cure rates, adverse events, and patient satisfaction.

Results: 88 of 172 women self-diagnosed a total of 172 UTIs. Laboratory evaluation showed a uropathogen in 144 cases (84%), sterile pyuria in 19 cases (11%), and no pyuria or bacteriuria in 9 cases (5%). Clinical and microbiological cures occurred in 92% and 96%, respectively, of culture-confirmed episodes. No serious adverse events occurred.

Conclusion: Adherent women can accurately self-diagnose and self-treat recurrent UTIs.

Ann Intern Med. 2001;135:9-16. www.annals.org

For author affiliations, current addresses, and contributions, see end of text.

See related article on pp 41-50 and editorial comment on pp 51-52.

Patient-Initiated Treatment of Uncomplicated Recurrent Urinary Tract Infections in Young Women

Kalpana Gupta, MD, MPH; Thomas M. Hooton, MD; Pacita L. Roberts, MS; and Walter E. Stamm, MD

Background: Recurrent urinary tract infections (UTIs) are a common outpatient problem, resulting in frequent office visits and often requiring the use of prophylactic antimicrobial agents. Patient-initiated treatment of recurrent UTIs may decrease antimicrobial use and improve patient convenience.

Objective: To determine the safety and feasibility of patient-initiated treatment of recurrent UTIs.

Design: Uncontrolled, prospective clinical trial.

Setting: University-based primary health care clinic.

Participants: Women at least 18 years of age with a history of recurrent UTIs and no recent pregnancy, hypertension, diabetes, or renal disease.

Intervention: After self-diagnosing UTI on the basis of symptoms, participating women initiated therapy with ofloxacin or levofloxacin.

Measurements: Accuracy of self-diagnosis determined by evidence of a definite (culture-positive) or probable (sterile pyuria and no alternative diagnosis) UTI on pretherapy urinalysis and culture. Women with a self-diagnosis of UTI that was not microbiologically confirmed were evaluated for alternative diagnoses. Post-therapy interviews and urine cultures were used to assess clinical and microbiological cure rates, adverse events, and patient satisfaction.

Results: 88 of 172 women self-diagnosed a total of 172 UTIs. Laboratory evaluation showed a uropathogen in 144 cases (84%), sterile pyuria in 19 cases (11%), and no pyuria or bacteriuria in 9 cases (5%). Clinical and microbiological cures occurred in 92% and 96%, respectively, of culture-confirmed episodes. No serious adverse events occurred.

Conclusion: Adherent women can accurately self-diagnose and self-treat recurrent UTIs.

Ann Intern Med. 2001;135:9-16. www.annals.org

For author affiliations, current addresses, and contributions, see end of text.

See related article on pp 41-50 and editorial comment on pp 51-52.

Reference standard

Accuracy

Series of patients

Index test

More than just diagnostic accuracy - other outcomes are important

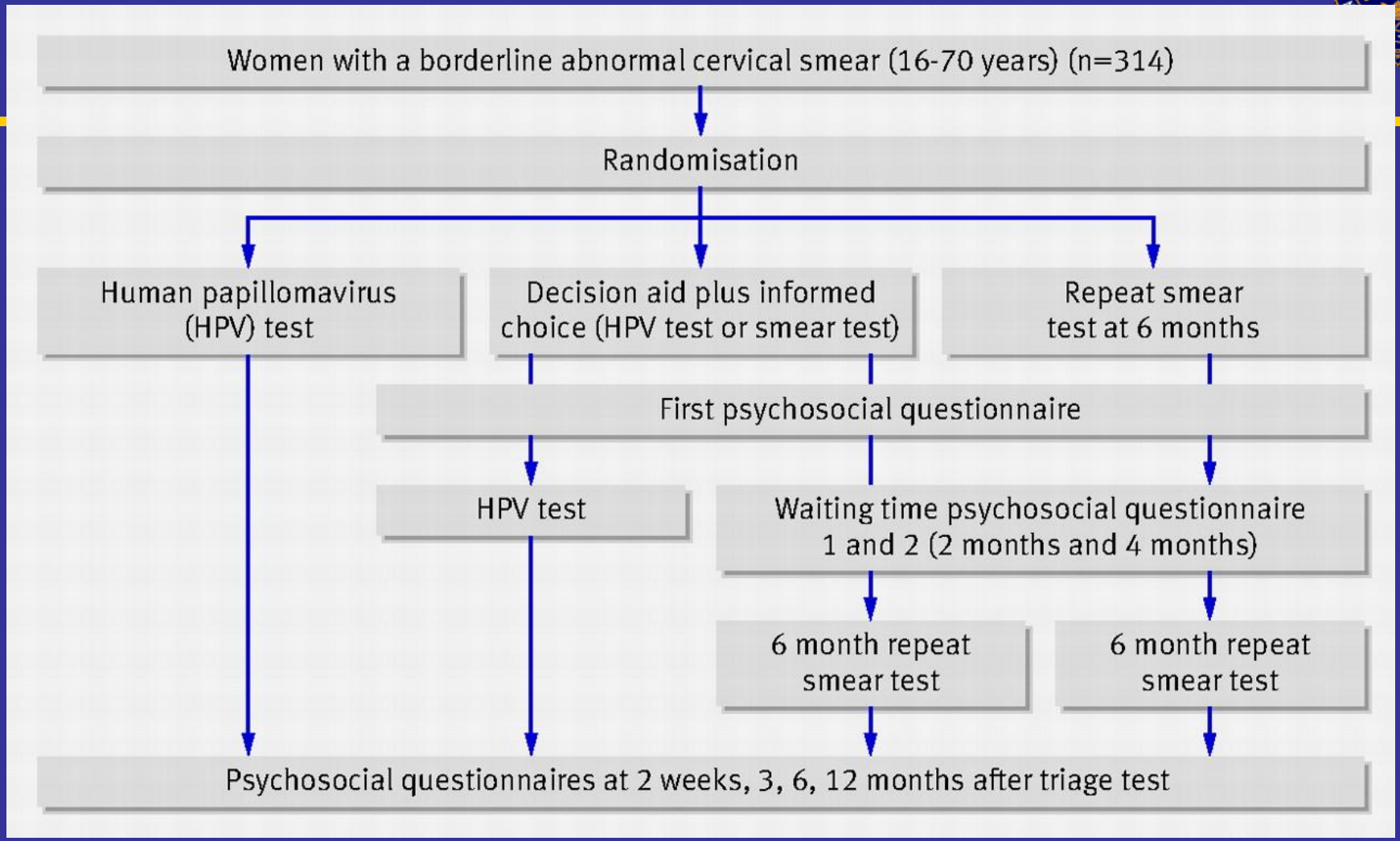
Other outcomes of tests

Effects of testing	What this means	Effects on health
Emotional	Test causes harmful or beneficial changes in anxiety levels, mood, depression, stress, psychological well being.	Increased anxiety and stress occur after a positive test on screening that has not been confirmed with a reference standard. Reassurance and improved overall well-being after negative test.
Social	Effects of testing on social roles, social functions, sexual relationships, social relationship.	Social isolation and stigmatisation after a positive test. Problems with employment or insurance coverage. Genetic testing results may cause guilt about passing on a genetic predisposition.
Cognitive	Patients' beliefs, perceptions and understanding about the test result and the condition.	May understand disease better – what causes it, how long it lasts etc., or affect adherence to therapy.
Behavioural	The combinations of emotional, social and cognitive effects can affect patient behaviour. Positive and negative tests can prompt change in behaviour.	Adherence to clinical intervention may be increased or decreased. Greater or less engagement with other health related behaviours, e.g. increased exercise after having cholesterol measured. Perceptions of risks from screening and repeated screening.

Example

Psychosocial outcomes of 3 triage methods for the management of borderline abnormal cervical smears: an open randomised trial. McCaffery BMJ 2010

Fig 1 Randomised trial design and psychosocial assessment



McCaffery K J et al. BMJ 2010;340:bmj.b4491

©2010 by British Medical Journal Publishing Group

Results

- At 12 months, distress about the abnormal cervical smear was lowest in women allocated to HPV testing compared with those allocated to repeat smear testing
- Satisfaction with care highest in women allocated to HPV testing

Steps in evaluating new tests

Evaluating new diagnostic tests
What are the key steps?

1. **Technical accuracy**
   "Can it work?"

2. **Place in the clinical pathway**
   "Where does the test fit in the existing clinical pathway?"

3. **Ability of the test to diagnose or exclude the target condition**
   "Does it work in patients?"

4. **The effect of the test on patient outcomes**
   "Are patients better off?"

5. **Cost-effectiveness**
   "Is it worth the cost"?

Evaluating new diagnostic tests

What are the key steps?

Information type	Question	Output	Study designs
Technical accuracy	Is the test reliable under standardised, artificial conditions?	Analytical sensitivity and specificity. Reproducibility, i.e., accuracy, precision and observer variation	Accuracy studies using standardised material, such as bloodbank samples
Place in clinical pathway	Where does the new test fit in existing clinical pathways?	Identification of current diagnostic pathway for a condition. Problems with current pathway (e.g time, costs, side effects of tests) Opportunities for new test to improve clinical outcomes	Reviews of existing diagnostic pathways. Descriptions of attributes of new tests.
Diagnostic accuracy	How good is this test at confirming or excluding a target condition?	Sensitivity and specificity Likelihood ratios Odds ratio Area under the curve	Diagnostic accuracy studies including real patients, comparing the new test to a reference standard.
Impact on patient outcome	After introducing this test to the clinical pathway, do patients fare better?	Mortality Morbidity Functional status Quality of life	Randomised controlled trials Clinical non-randomised trials Before-after studies
Cost-effectiveness	Is this test good value for money?	Cost per life year gained Cost per QALY	Economic modelling

Explaining bias in diagnostic studies with pictures

Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

Jeroen G. Lijmer, MD
Ben Willem Mol, MD, PhD
Siem Heisterkamp, PhD
Gouke J. Bonsel, MD, PhD
Martin H. Prins, MD, PhD
Jan H. P. van der Meulen, MD, PhD
Patrick M. M. Bossuyt, PhD

Context The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would invalidate the clinical application of most study results.

Objective To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy.

Design and Setting Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through a systematic search of the literature using MEDLINE, EMBASE, and DARE databases and the Cochrane Library (1996-1997). Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model.

Main Outcome Measures Relative diagnostic odds ratio (RDOR), which compared the diagnostic odds ratios of studies of a given test that lacked a particular methodological feature with those without the corresponding shortcomings in design.

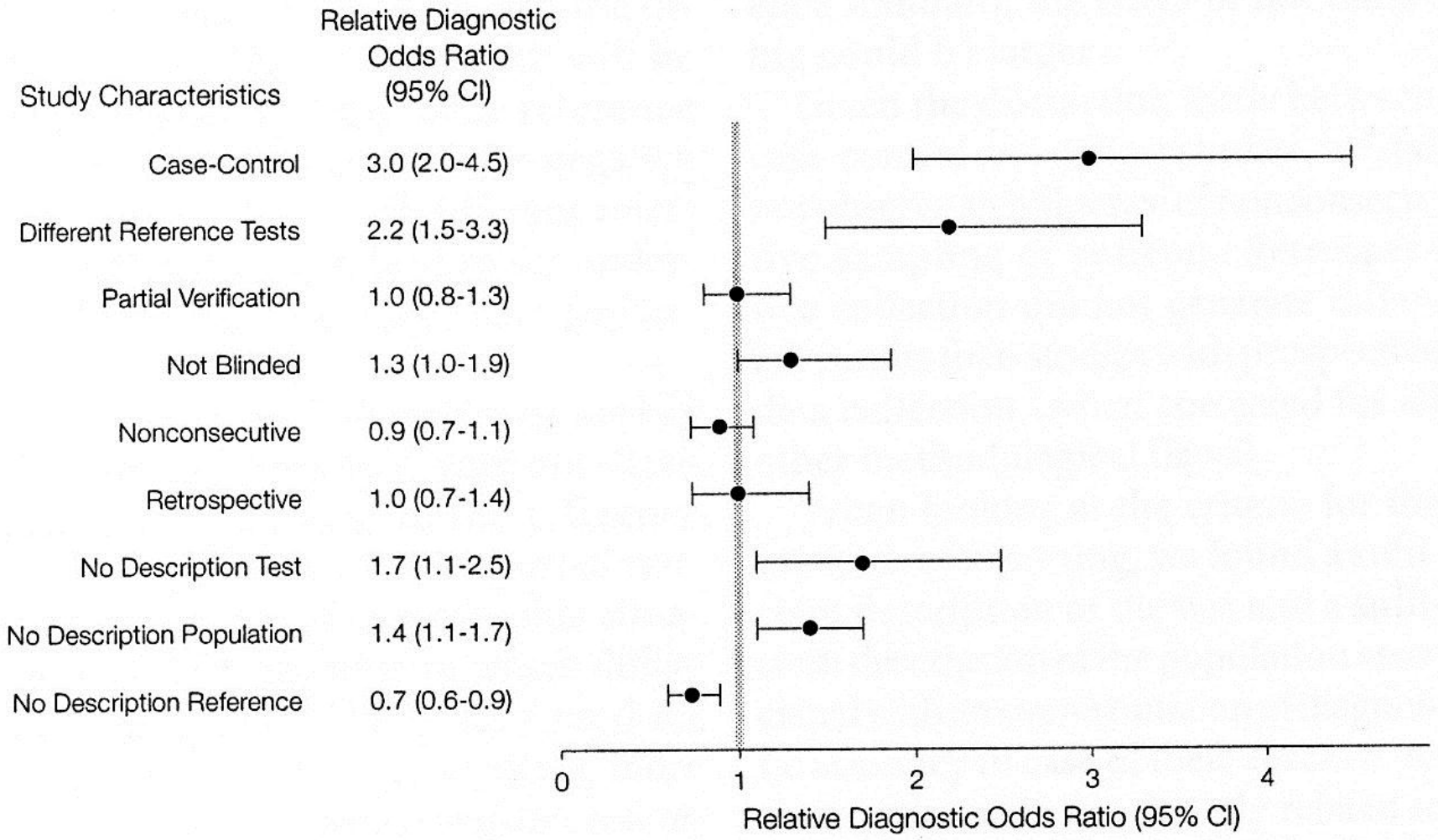
Results Fifteen (6.8%) of 218 evaluations met all 8 criteria; 64 (30%) met 6 or more. Studies evaluating tests in a diseased population and a separate control group overestimated the diagnostic performance compared with studies that used a clinical population (RDOR, 3.0; 95% confidence interval [CI], 2.0-4.5). Studies in which different reference tests were used for positive and negative results of the test under study overestimated the diagnostic performance compared with studies using a single reference test for all patients (RDOR, 2.2; 95% CI, 1.5-3.3). Diagnostic performance was also overestimated when the reference test was interpreted with knowledge of the test result (RDOR, 1.3; 95% CI, 1.0-1.9), when no criteria for the test were described (RDOR, 1.7; 95% CI, 1.1-2.5), and when no description of the population under study was provided (RDOR, 1.4; 95% CI, 1.1-1.7).

Conclusion These data provide empirical evidence that diagnostic studies with methodological shortcomings may overestimate the accuracy of a diagnostic test, particularly those including nonrepresentative patients or applying different reference standards.

JAMA. 1999;282:1061-1066 www.jama.com

DURING RECENT DECADES, THE number of available diagnostic tests has been rapidly increasing. As for all new medical technologies, new diagnostic tests should be thoroughly evaluated prior to their introduction into daily practice. The number of test evaluations in the literature is increasing but the methodological quality of these studies is on average poor. A survey of the diagnostic literature (1990-1993) showed that only 18% of the studies satisfied 5 of the 7 methodological standards examined.[1] Different guidelines have been written to help physicians with the critical appraisal of the diagnostic literature consisting of lists of criteria for the assessment of study quality.[2-4] Criteria enable readers to check whether studies fulfill

Figure. Relative Diagnostic Odds Ratios and 95% Confidence Intervals (CIs) of the 9 Study Characteristics Examined With a Multivariate Regression Analysis



Assessing bias – what is most important for diagnostic studies?

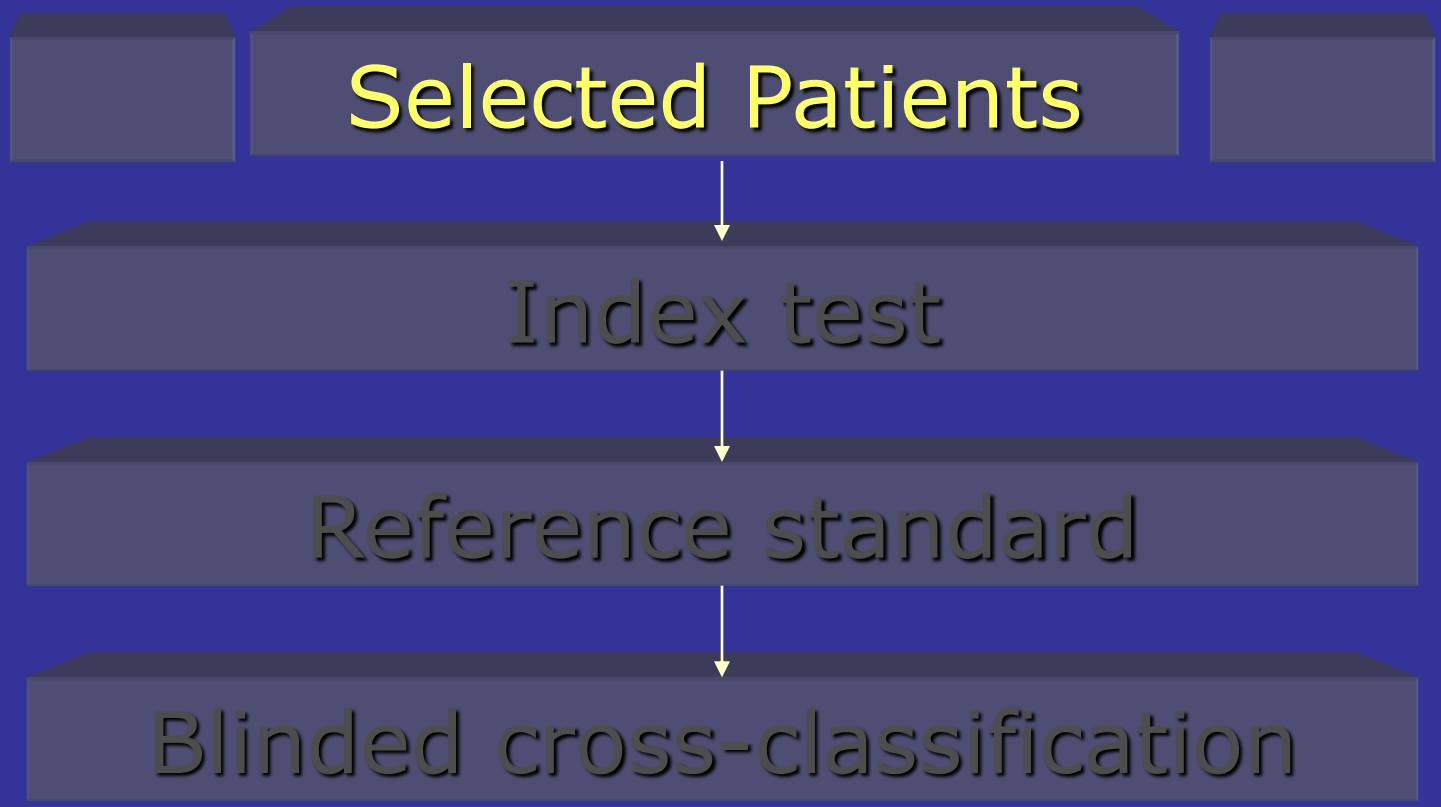
- **Appropriate spectrum of patients selected?**
- **Was the index test performed on all patients?**
- **Is the same reference test performed on all patients, regardless of the result of the index test? How objective is the reference test?**
- **Were the index and reference tests compared in independent, blind ?**

A***ppropriate spectrum*** of patients?

- Ideally, test should be performed on group of patients in whom it will be applied in the real world

- Spectrum bias = study using only highly selected patients…….*perhaps those in whom you would really suspect have the diagnosis*

Spectrum Bias

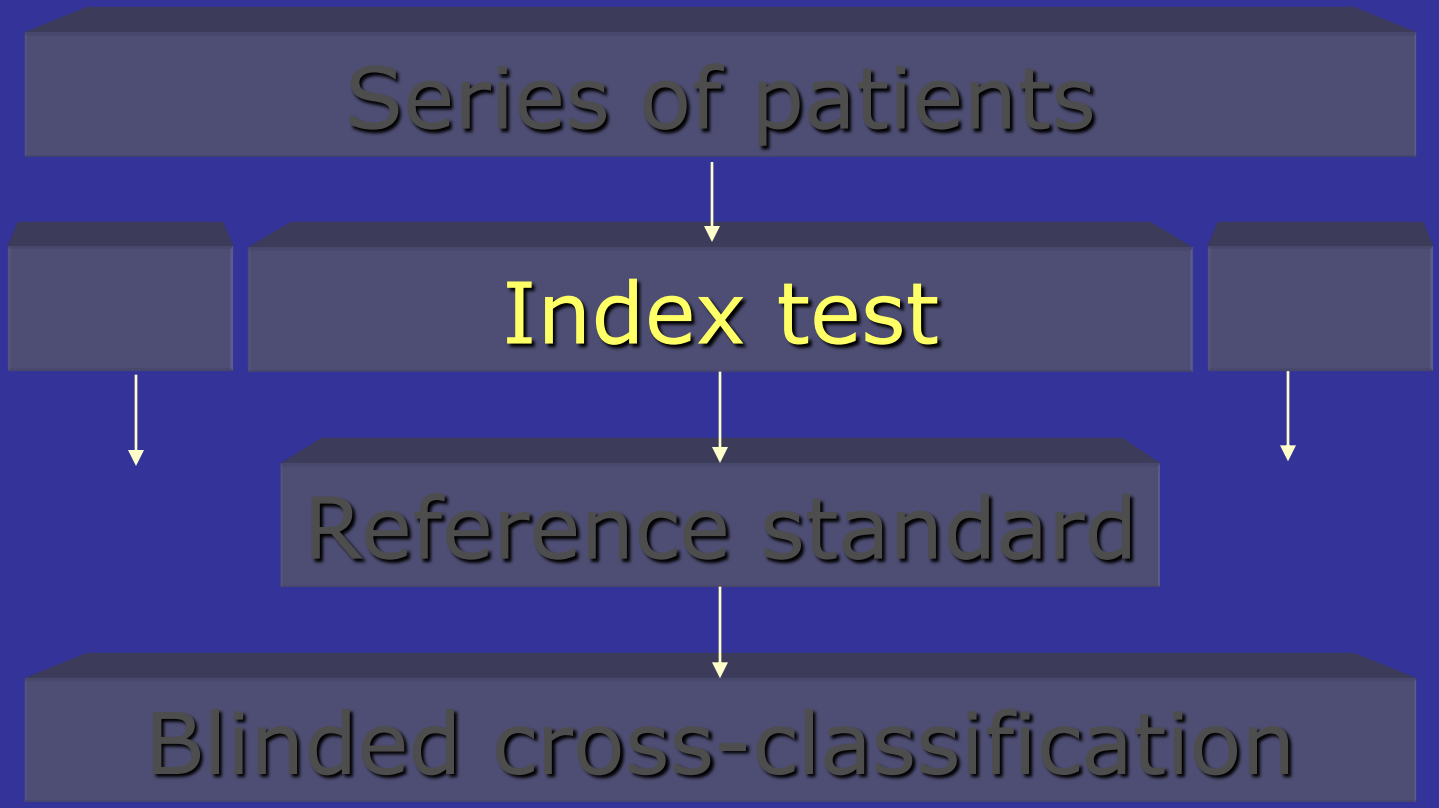


2. Do A***LL*** patients get the gold standard test?

- Ideally all patients get the reference (“gold”) standard test

- Verification/work-up bias = only **some** patients get the gold standard…..(*probably the ones in whom you really suspect have the disease*)

Verification (work-up) bias

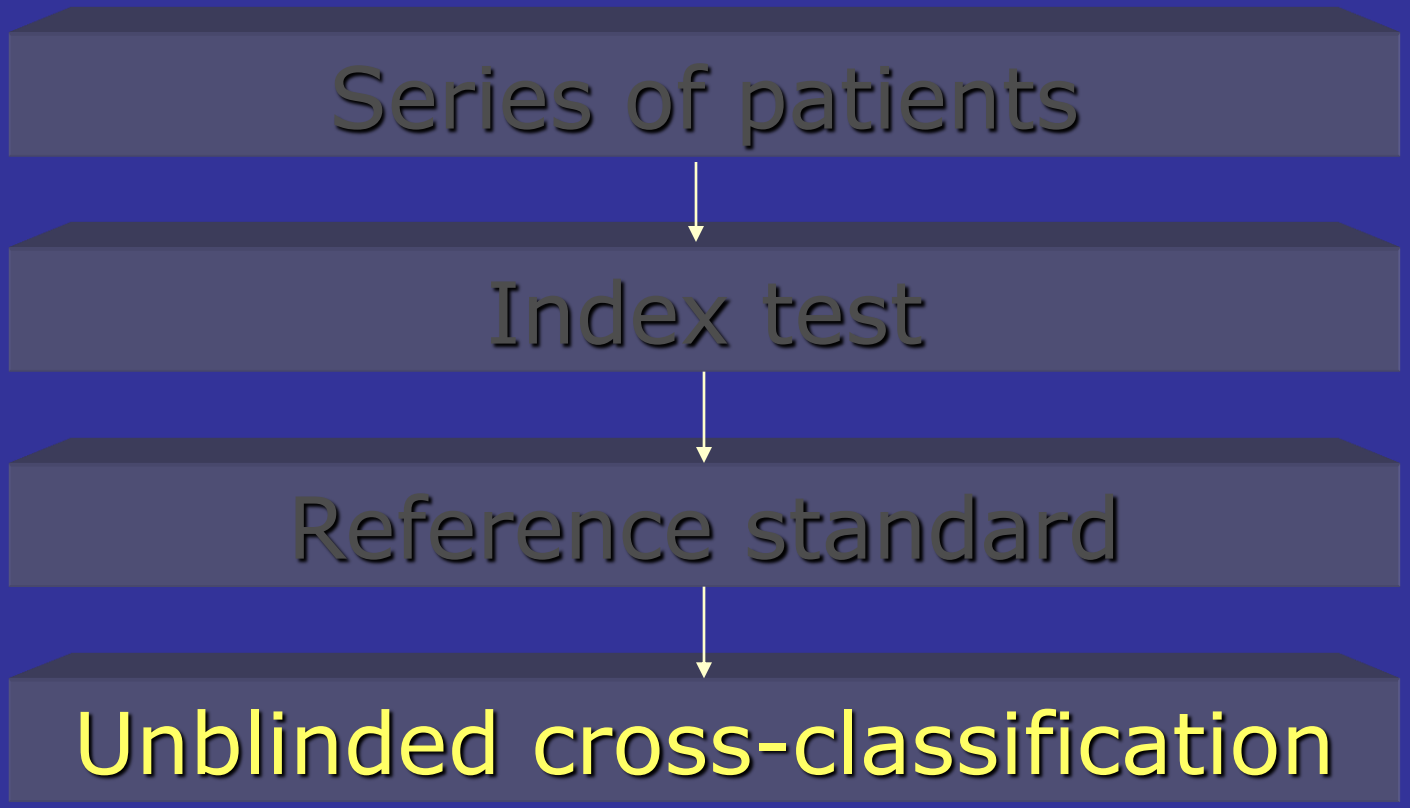


3. ***Independent, blind or objective comparison*** with the gold standard?

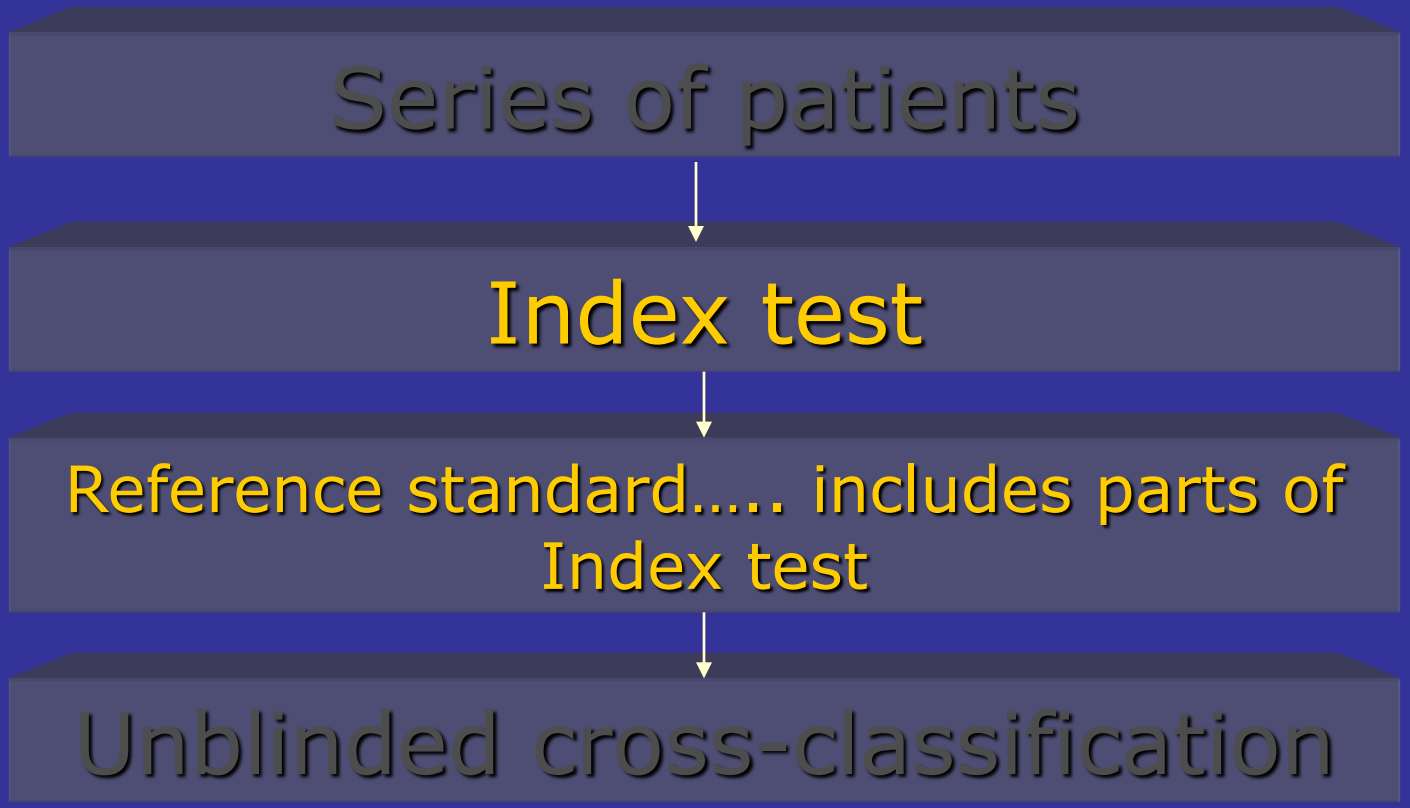
- Ideally, the gold standard is independent, blind and objective

- Observer bias = test is very subjective, or done by person who knows something about the patient

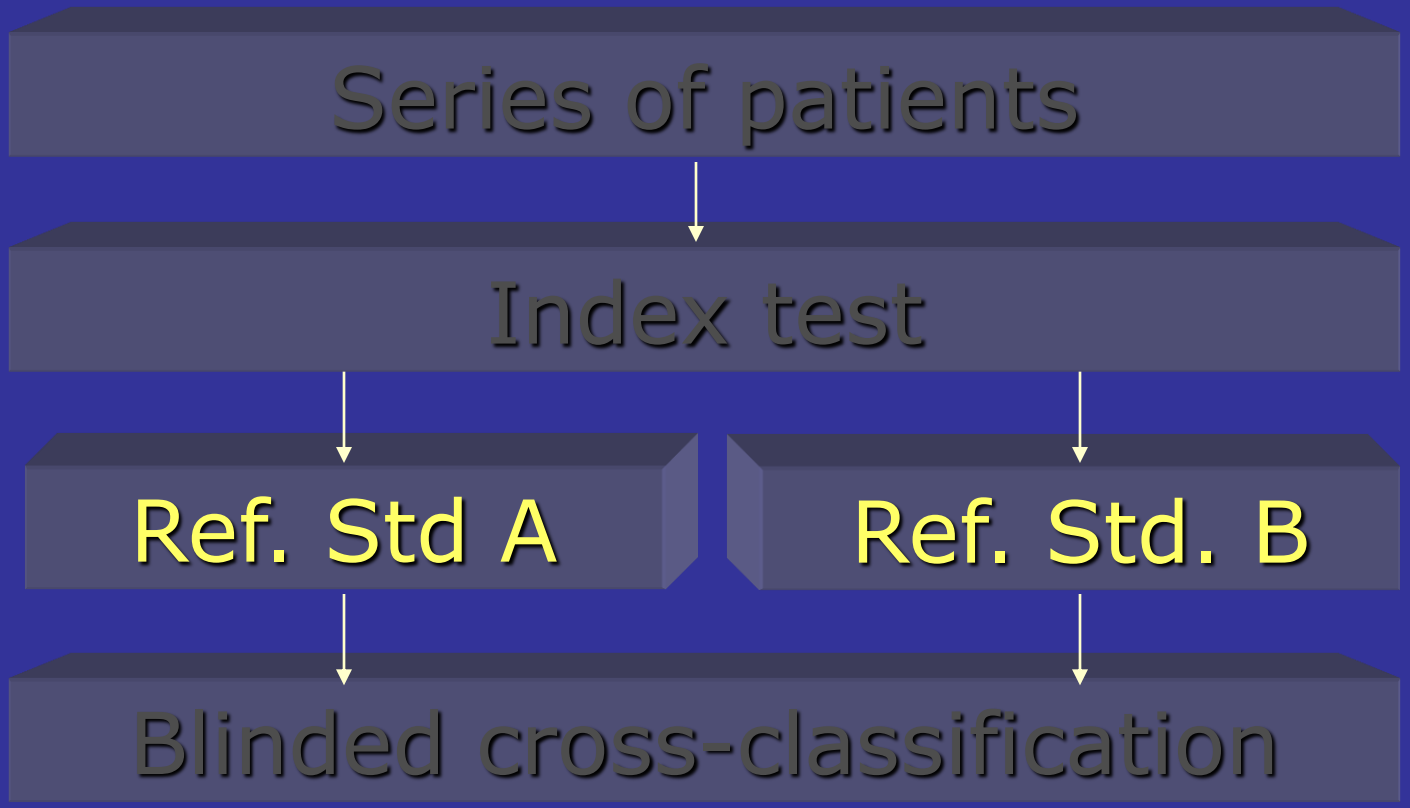
Observer/test review Bias



Incorporation Bias



Differential reference bias



Which bias matters the most?

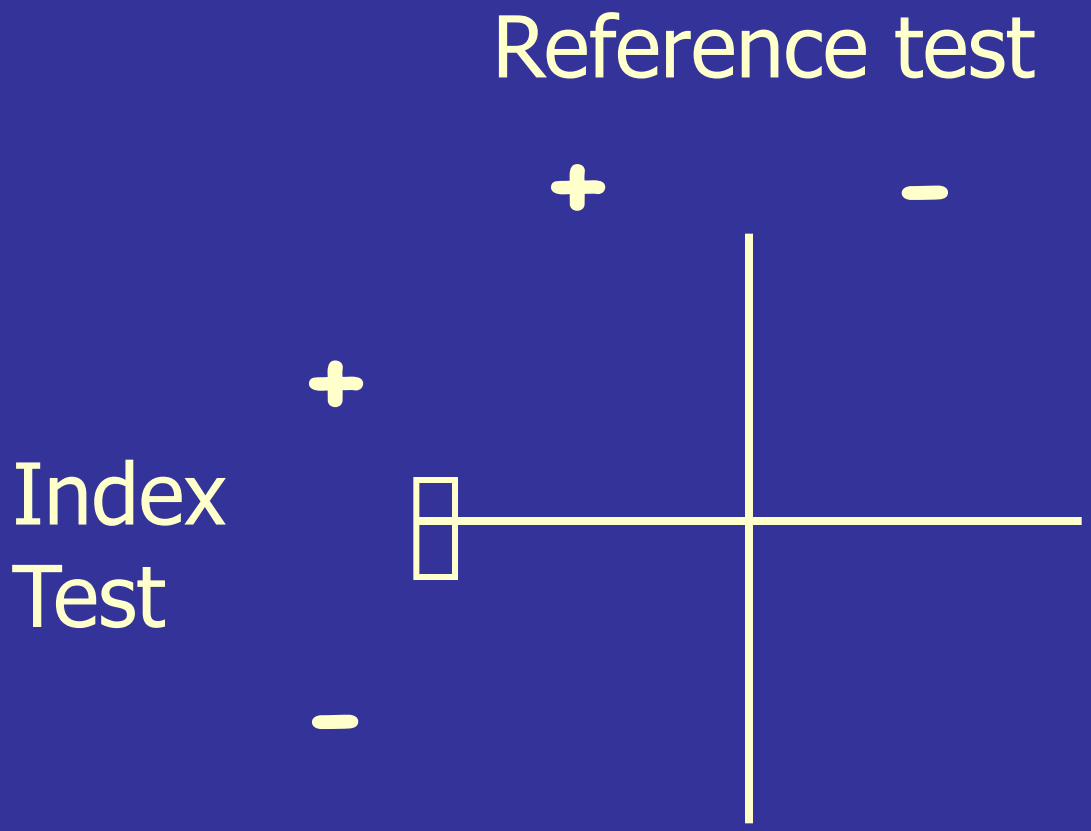
- Many diagnostic studies will have biases, does not mean you discard them, but decide what effects may have on results
- Some design features/biases more important than others
- Biggest overestimation of diagnostic accuracy
  - Selection of patients (spectrum bias) most important ie case control studies
  - Differential verification

How to explain results of diagnostic accuracy

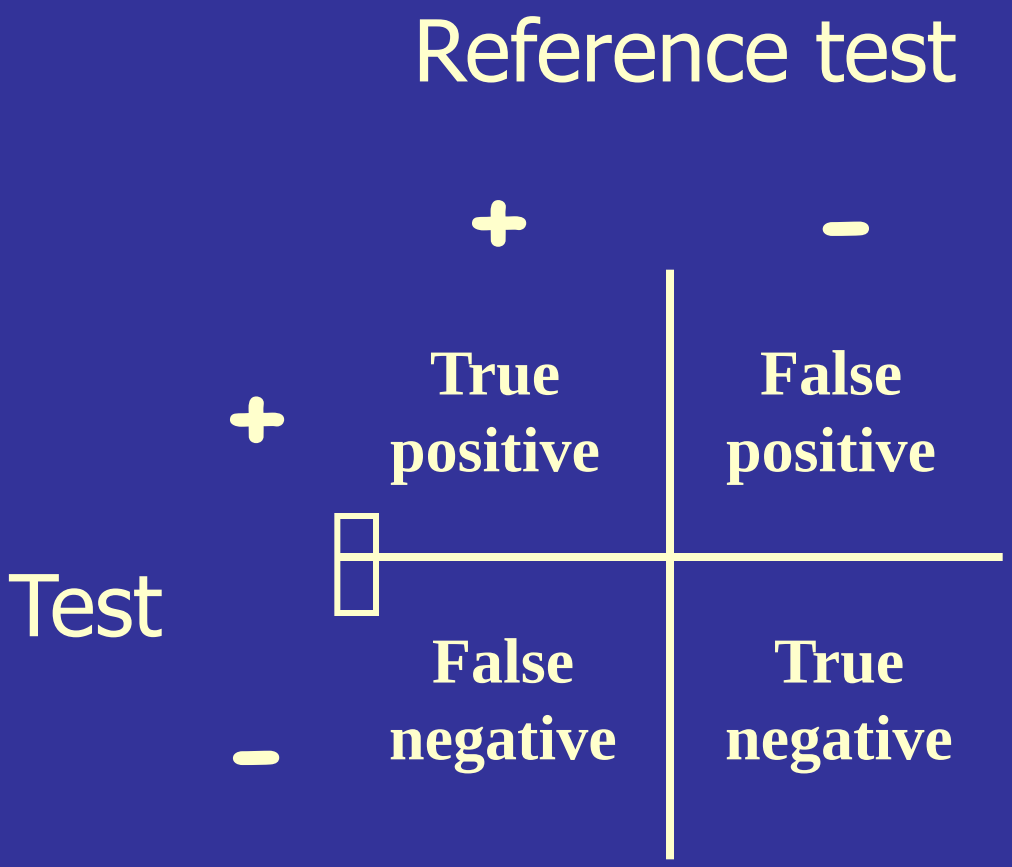
What's the problem?

- Pairs of numbers usually
- The 2 numbers depend on each other
- The consequences of false positive and false negative results are different
- Most people don't understand what the numbers actually mean

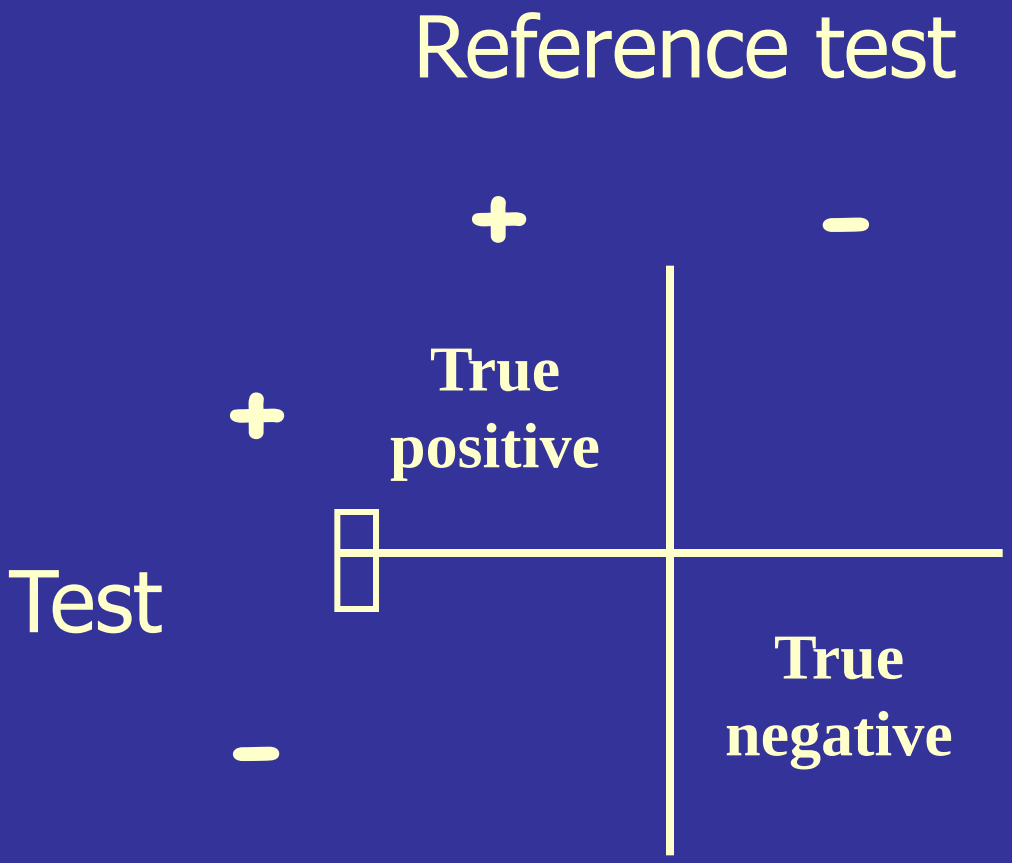
2 by 2 table



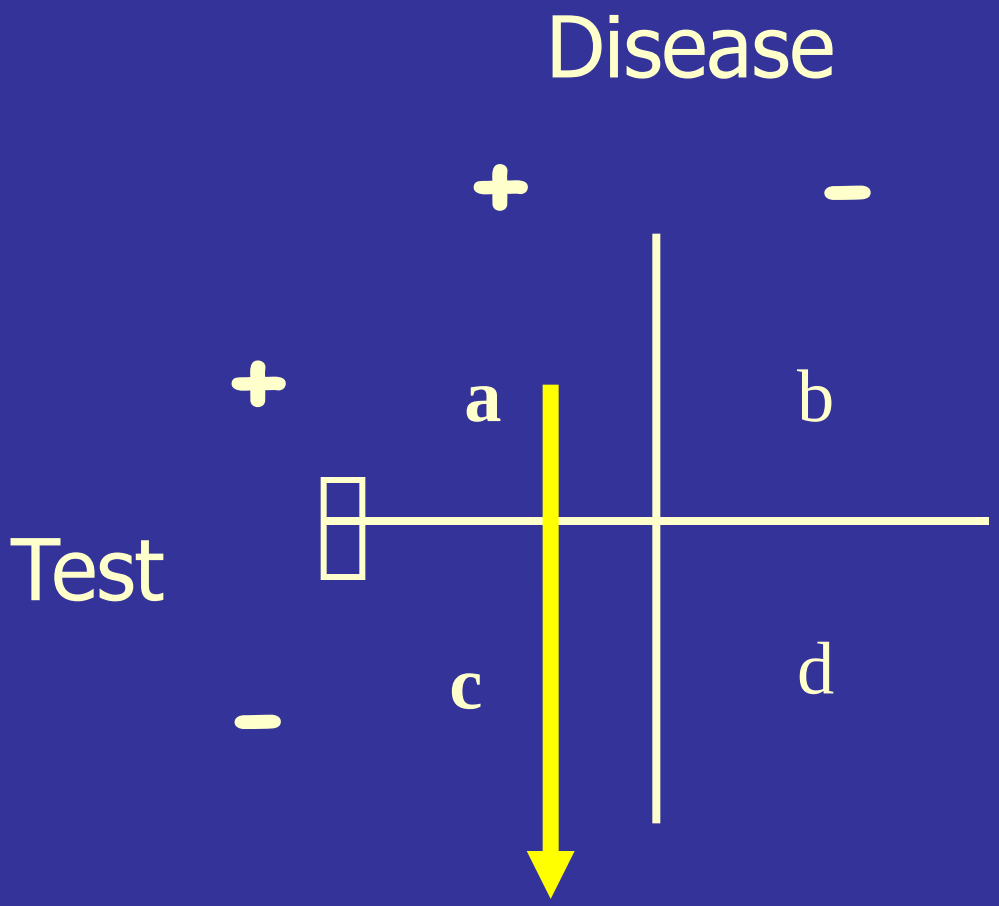
2 by 2 table


Reference test
+
−
Test
+
−
True positive
False positive
False negative
True negative

IF only a test had perfect discrimination…



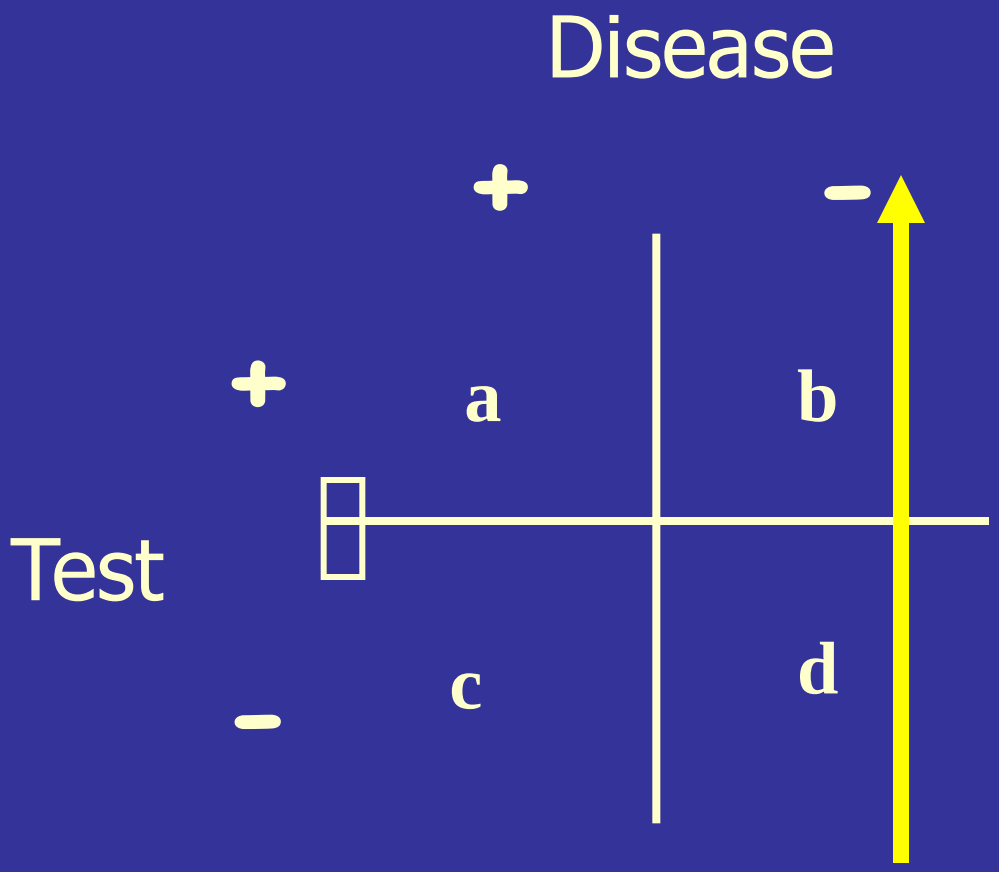
Sensitivity



Sensitivity = a / a + c

Proportion of people <u>with</u> the disease who have a <u>positive</u> test.

Specificity



Proportion of people <u>without</u> the disease who have a negative test.

Specificity = d / b + d

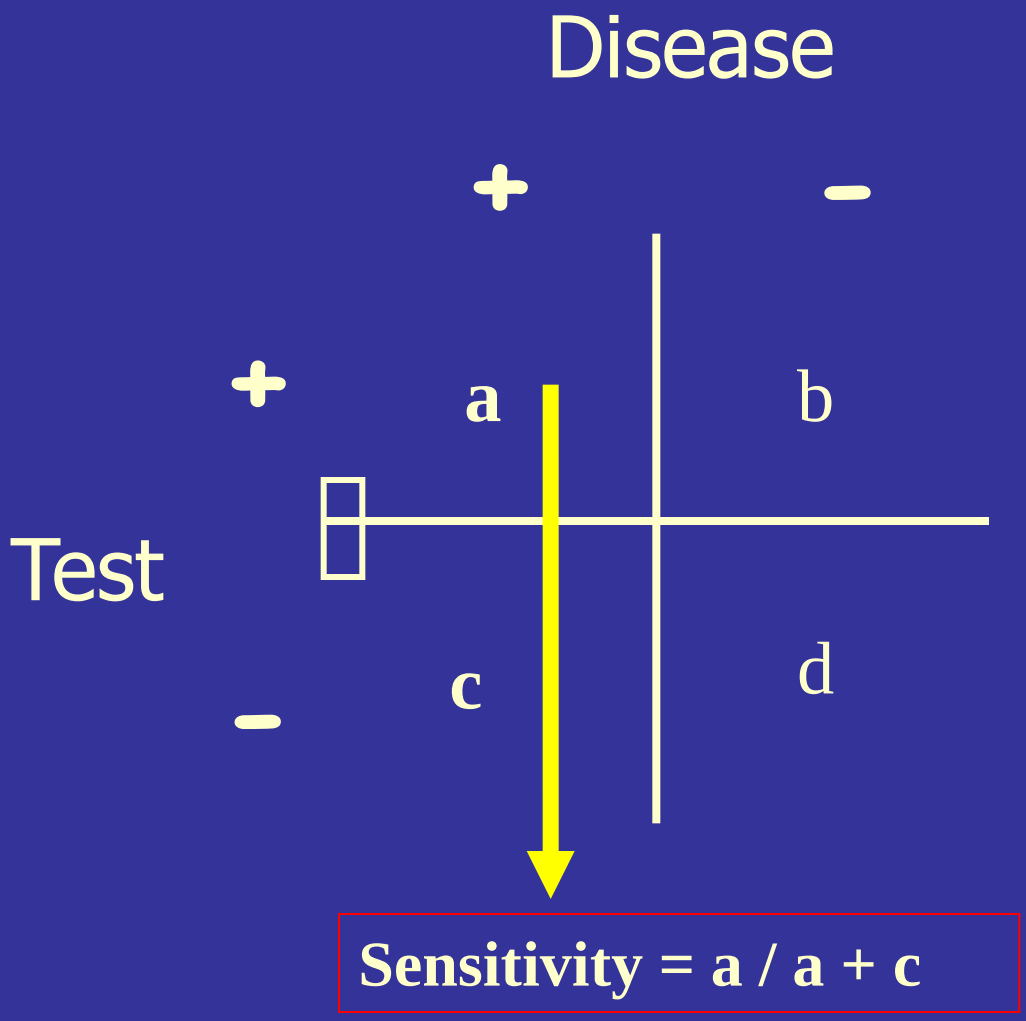
Tip…..

- **Sensitivity** is useful to me
  - *'The new chlamydia test was positive in 47 out of 56 women with chlamydia (sensitivity =83.9%)'*
- Specificity seems a bit confusing
  - *'The new chlamydia test was negative in 600 of the 607 women who did not have chlamydia (specificity = 98.8%)'*
- So…**false positive rate** is sometimes easier
  - False positive rate = 1 – specificity
  - *So a specificity of 98.8% means that the new test is wrong (or falsely positive) in 1.2% of women*

Maybe forget sensitivity and specificity?..

- **True positive rate (= Sensitivity)**
- **False positive rate (= 1 – Specificity)**

How about this? SnNOUT



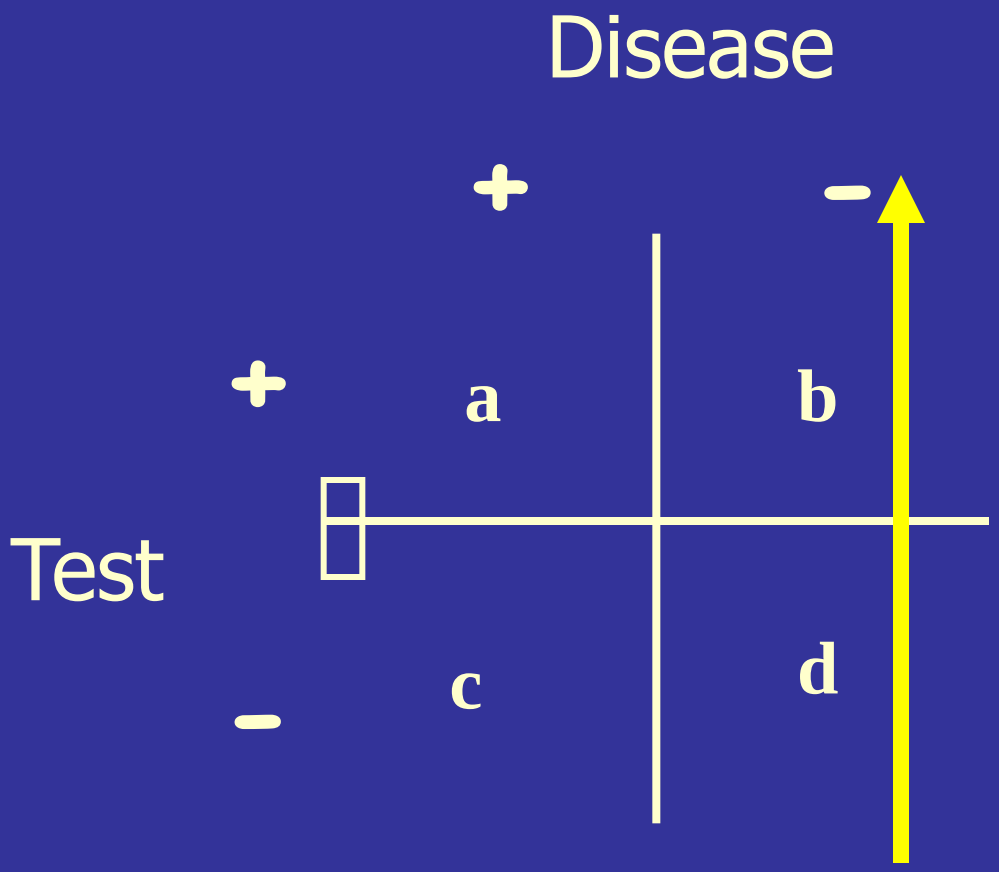
Highly sensitive tests = good for screening

or

SnNOUT

Highly sensitive test, negative result rules out.

SpPIN



Highly <u>specific</u> tests = good for ruling in

or

SpPIN

Highly specific test, positive result rules in.

Specificity = d / b + d

Using natural frequencies to explain results of diagnostic accuracy

Using natural frequencies

You return home from the CEBM course. Your father telephones you and tells you that he went to his doctor and was told that his test for a disease was positive. He is really worried, and asks you for help!

- After doing some reading, you find that for men of his age:
  - The prevalence of the disease is 30%
  - The test has a sensitivity of 50% and specificity of 90%
- *"Son, tell me what's the chance I have this disease?"*

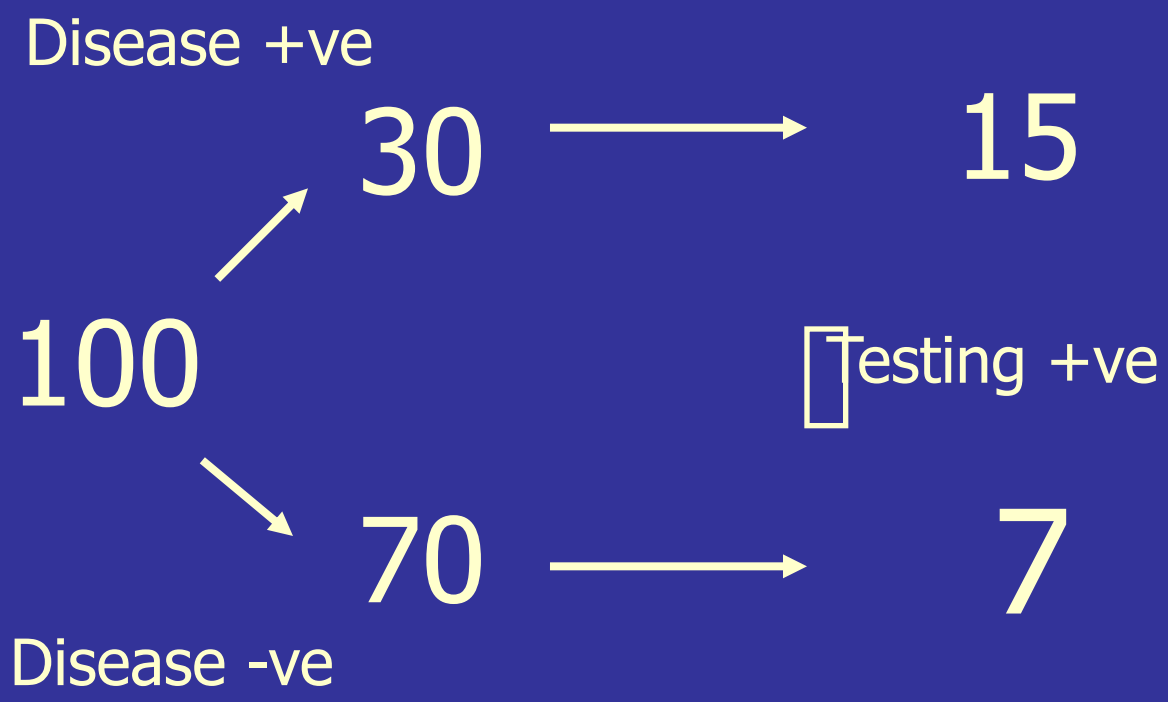
A disease with a prevalence of 30% must be diagnosed.

The test has a sensitivity of 50% and a specificity of 90%.

Given a positive test, what's the chance he has the disease?

- 100% likely
- 50% maybe
- 0% unlikely

Prevalence of 30%
Sensitivity of 50%
Specificity of 90%



22 people test positive……

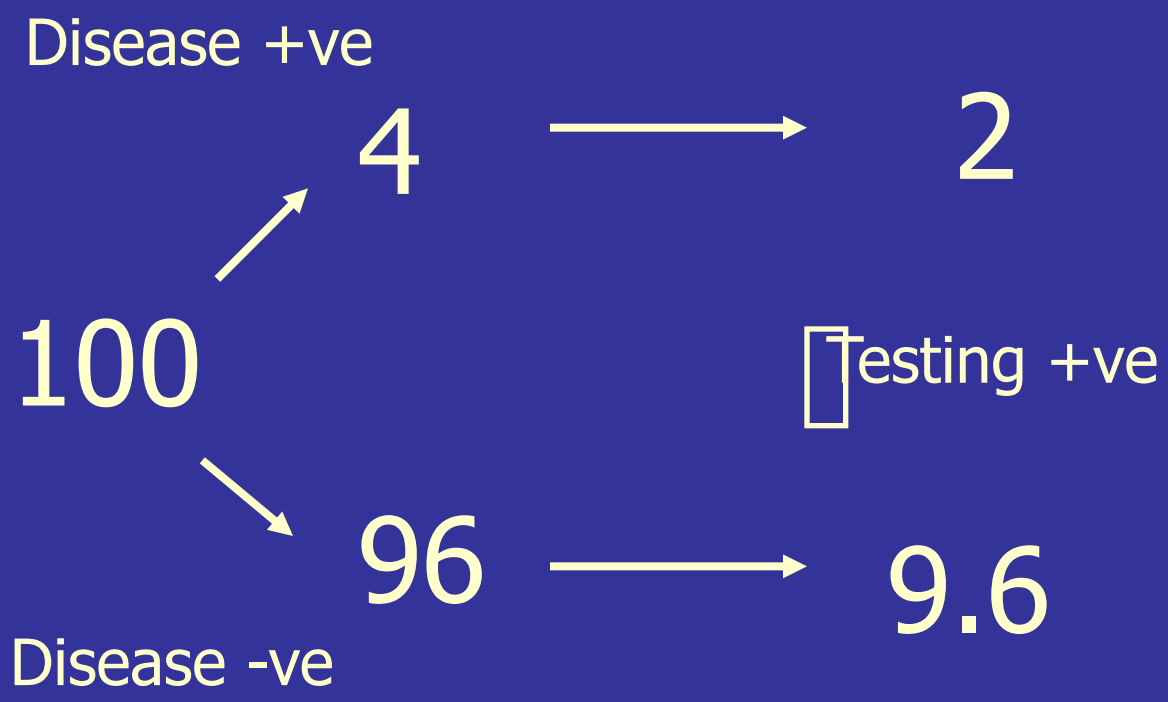
of whom 15 have the disease

About 70%

Try it again

- A disease with a prevalence of 4% must be diagnosed.
- It has a sensitivity of 50% and a specificity of 90%.
- If the patient tests positive, what is the chance they have the disease?

Prevalence of 4%
Sensitivity of 50%
Specificity of 90%



11.6 people test positive…

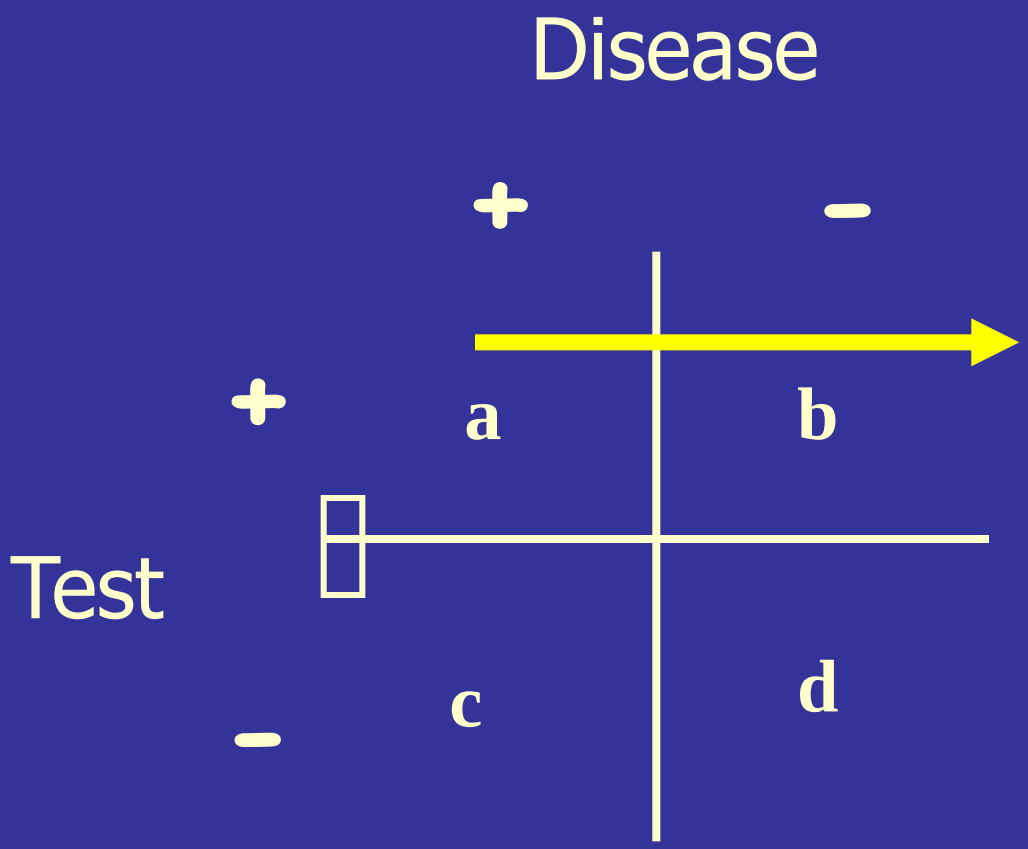
of whom 2 have the disease

About 17%

Doctors with an average of 14 yrs experience
Answers ranged from 1% to 99%
….half of them estimated the probability as 50%
Gigerenzer G BMJ 2003;327:741-744

What about positive and negative predictive values?

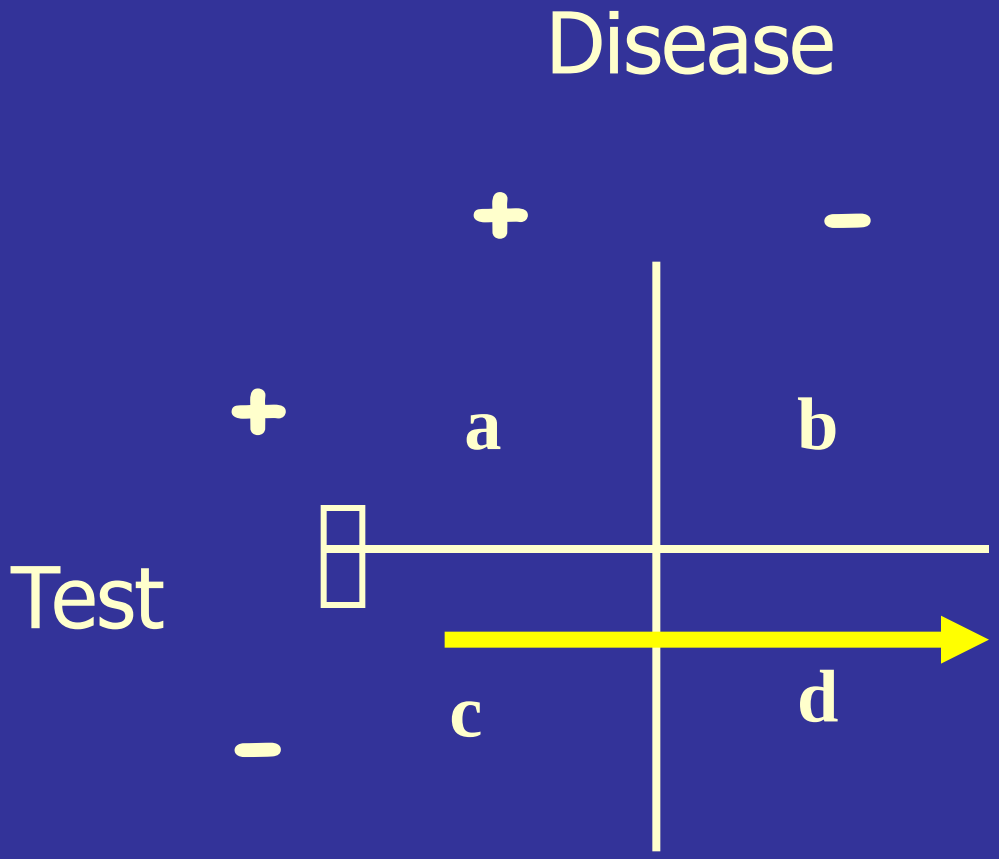
positive predictive value (PPV)



PPV = a / a + b

Proportion of people with a positive test who have the disease

negative predictive value (NPV)



NPV = d / c + d

Proportion of people with a negative test who do not have the disease

Sensitivity/specificity

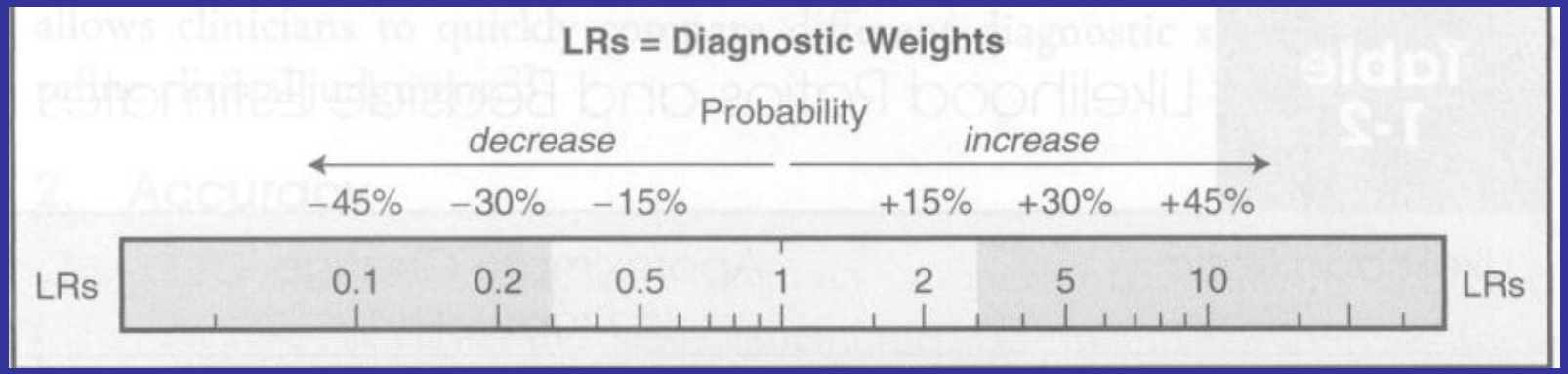
- Disease status known
- Not as dependent on prevalence
- but can be affected by disease spectrum eg selection of patients

Positive/Negative predictive values

- Test result known
- Depend on prevalence

Likelihood Ratios and Bayesian reasoning

- Can use in situations with more than 2 test outcomes
- Direct link from pre-test probabilities to post-test probabilities

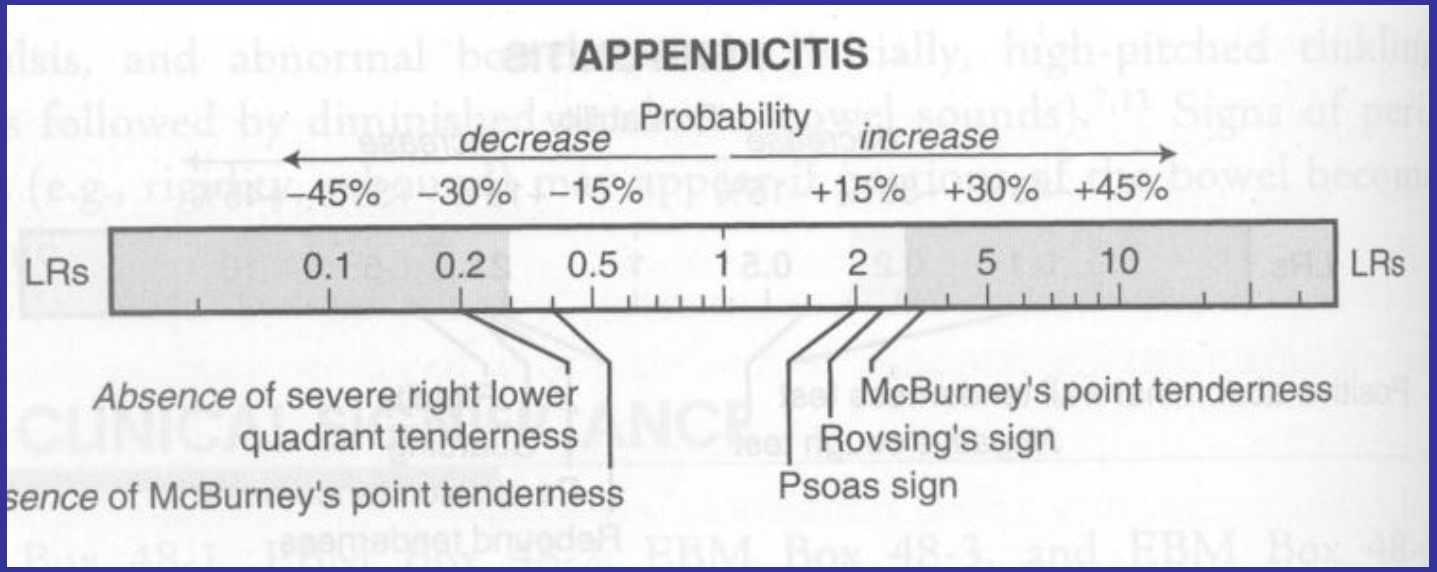


LR<0.1…. strong negative test result

LR=1

No diagnostic value

LR>10 …. strong positive test result



McGee: Evidence based Physical Diagnosis (Saunders Elsevier)

Positive and negative likelihood ratios

LR+ How much more often a positive test occurs in people <u>with</u> compared to those <u>without</u> the disease

LR+ = a/a+c / b/b+d

Or

LR+ = sens/(1-spec)

LR- How less likely a negative test result is in people with the disease compared to those without the disease

LR- = c/a+c / d/b+d

Or

LR- = (1-sens)/(spec)

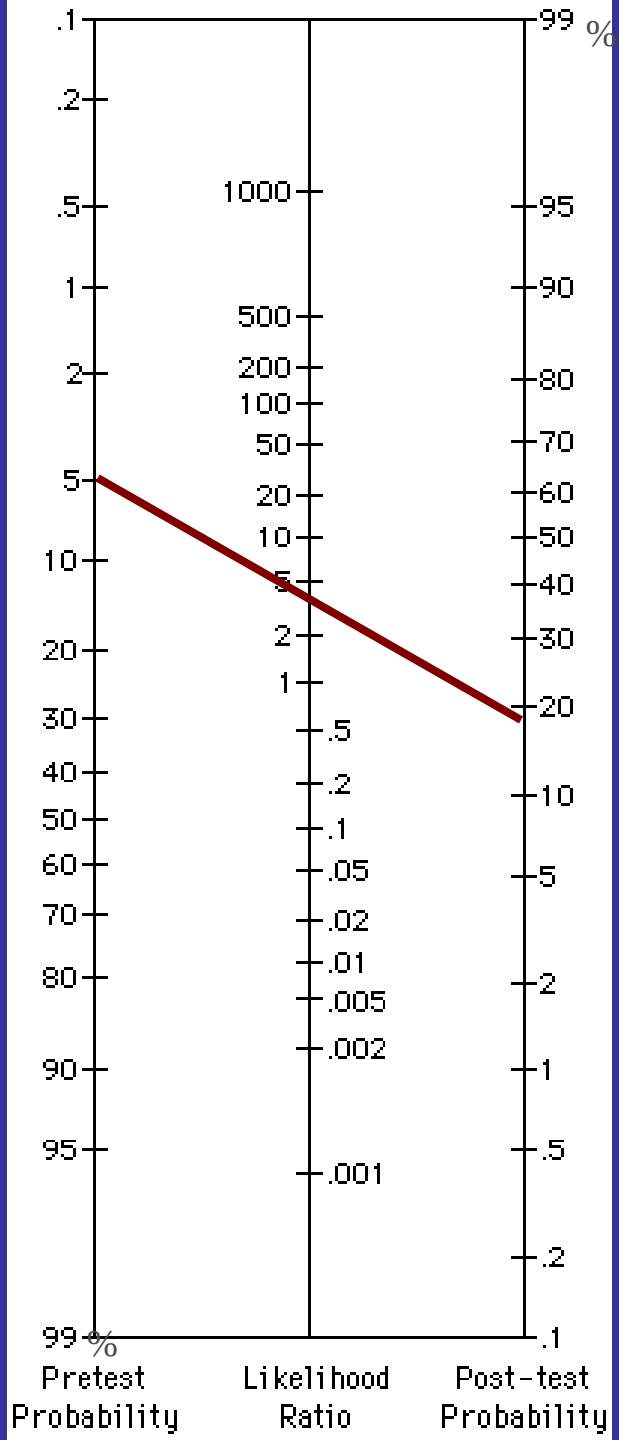
Bayesian reasoning

Post-test odds = Pre-test odds x Likelihood ratio

- Post-test odds for disease after *one* test become pre-test odds for *next* test etc

Bayesian reasoning using Fagan Nomogram

Pre test 5%



Post test 20%

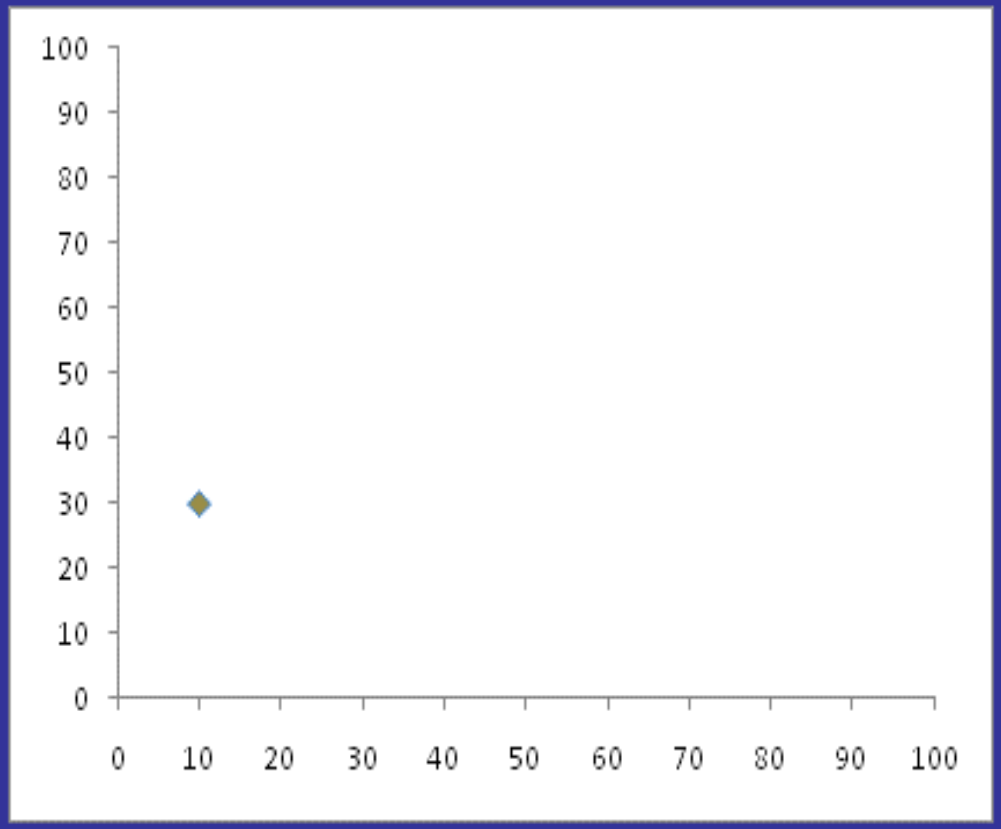
? Appendicitis:

McBurney tenderness
LR+ = 3.4

ROC curves (Receiver Operating Characteristic curves) – What are they and what aren’t they?

ROC curves – provide accuracy results over a ***range*** *of thresholds*

Sensitivity

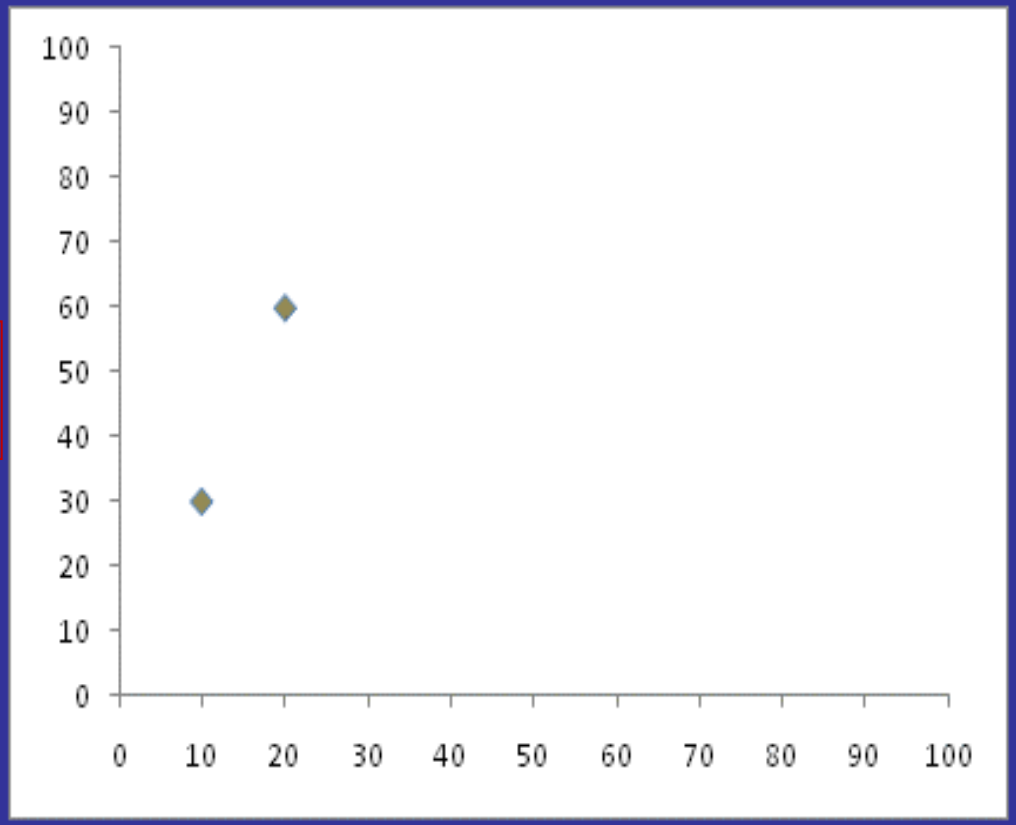


1-Specificity or false positive rate

A test with 30% sensitivity and 90% specificity (10% false positive rate) at one cut-point is plotted in the lower left corner.

ROC curves

Sensitivity

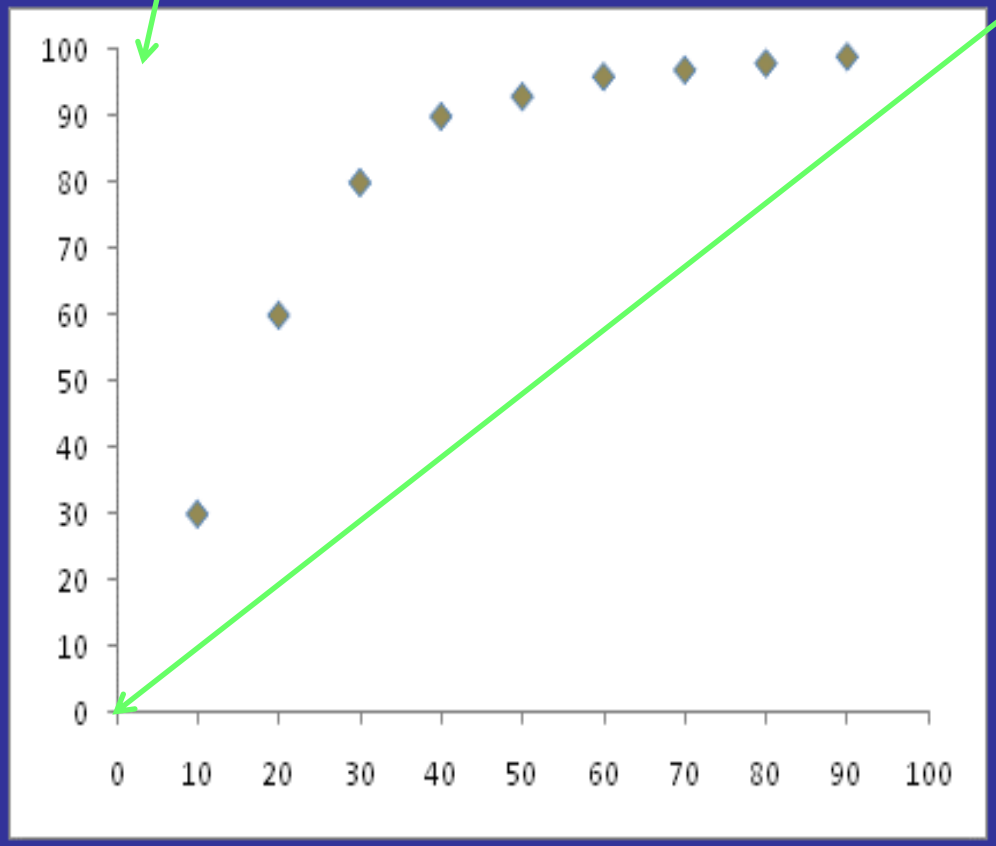


1-Specificity

It has another cut-point with a sensitivity of 60% and specificity of 80%

Perfect test = upper left hand corner

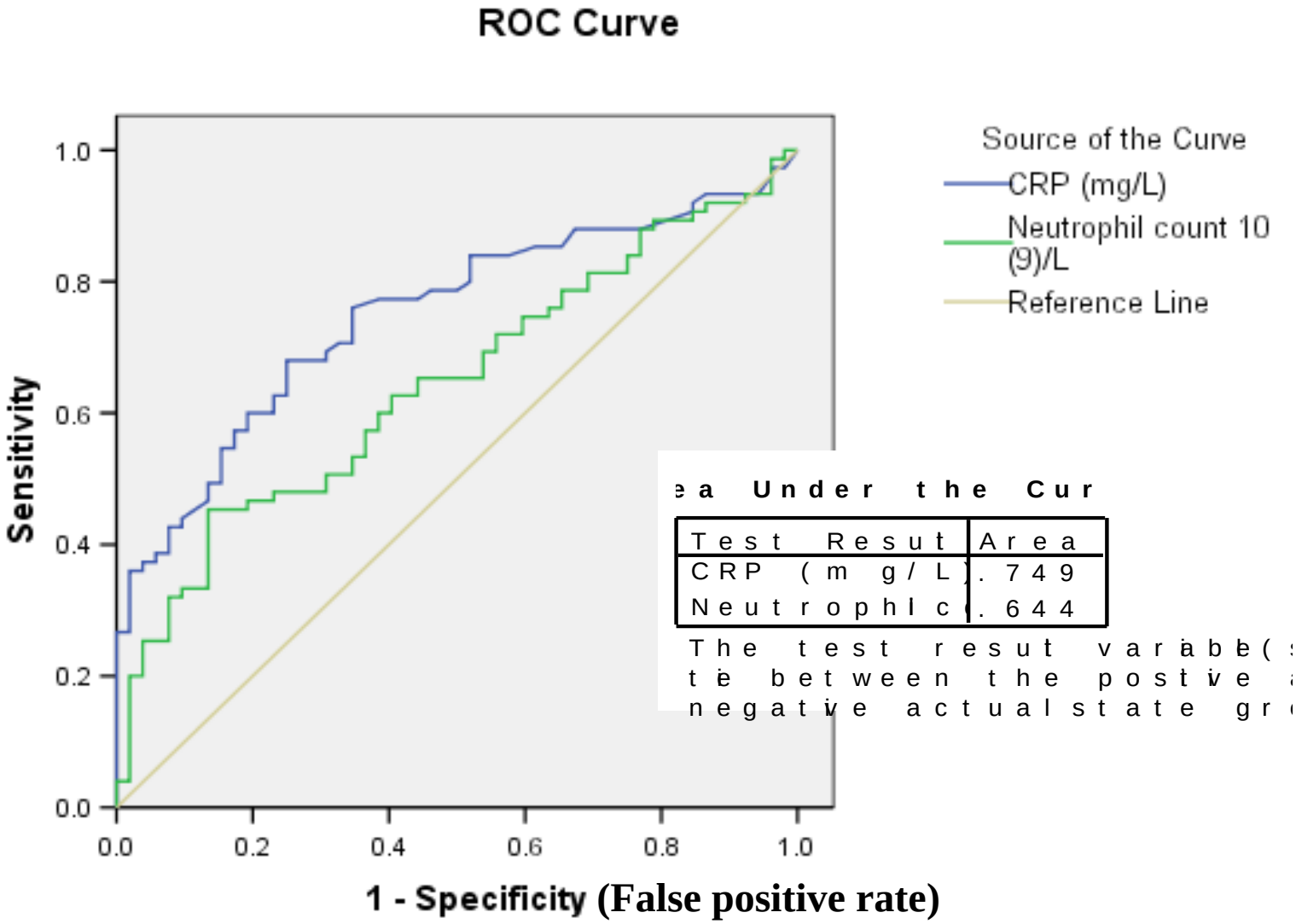
Diagonal = no discrimination



Sensitivity

1-Specificity

Area under the curve (AUC)
0.5 = useless
1.0 = perfect



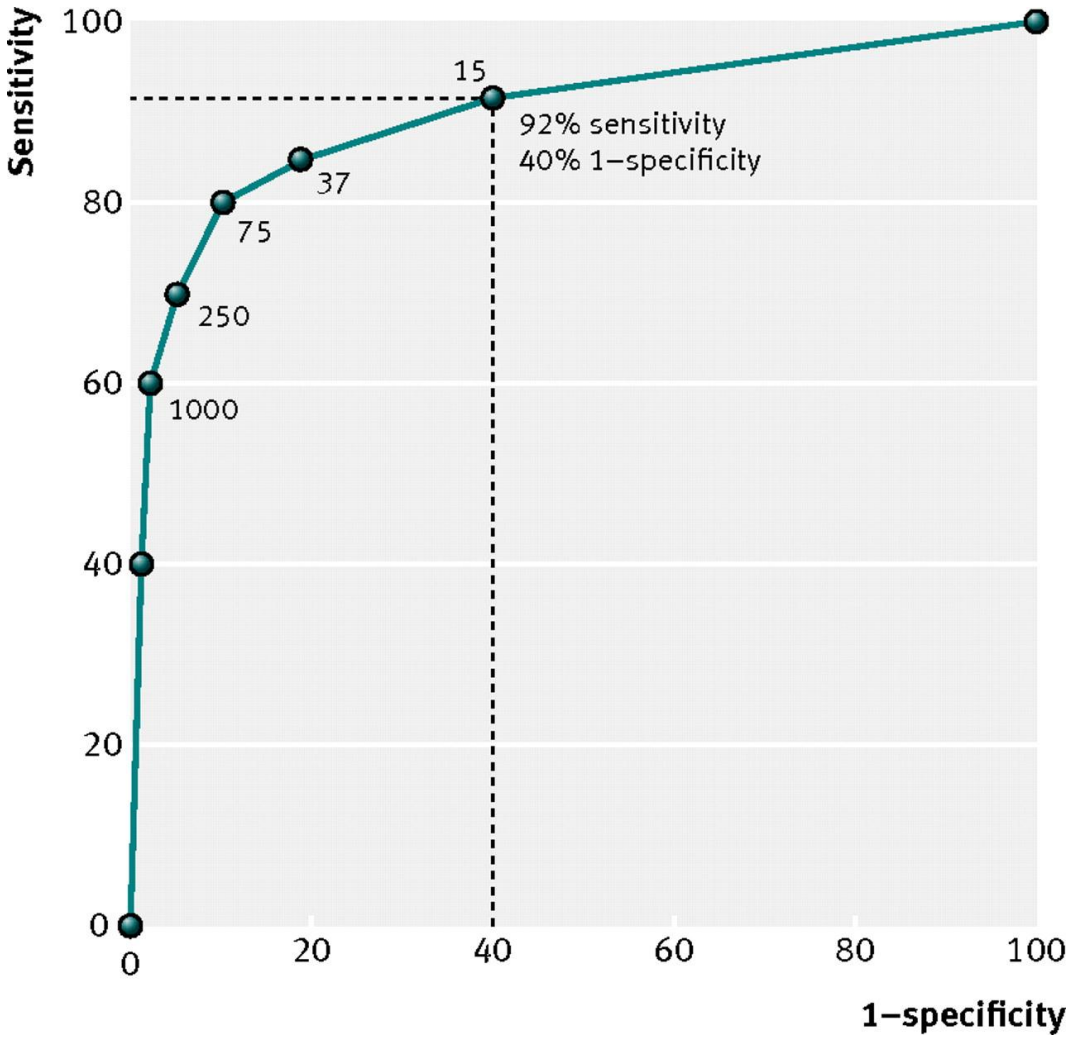
ea Under the Cur

Test Resut	Area
CRP (mg/L)	.749
Neutrophil c	.644

The test resut variable(s
te between the postive a
negative actual state gr

Diagonal segments are produced by ties.

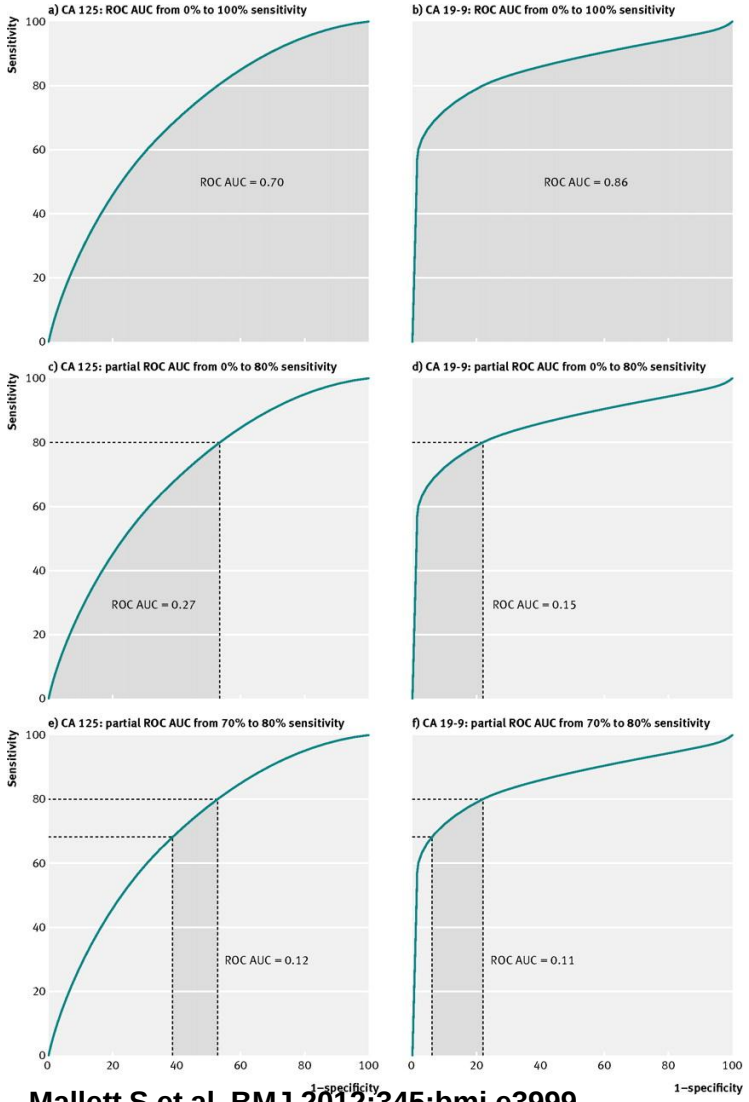
Fig 2 ROC plot of test accuracy at different thresholds.



Mallett S et al. BMJ 2012;345:bmj.e3999

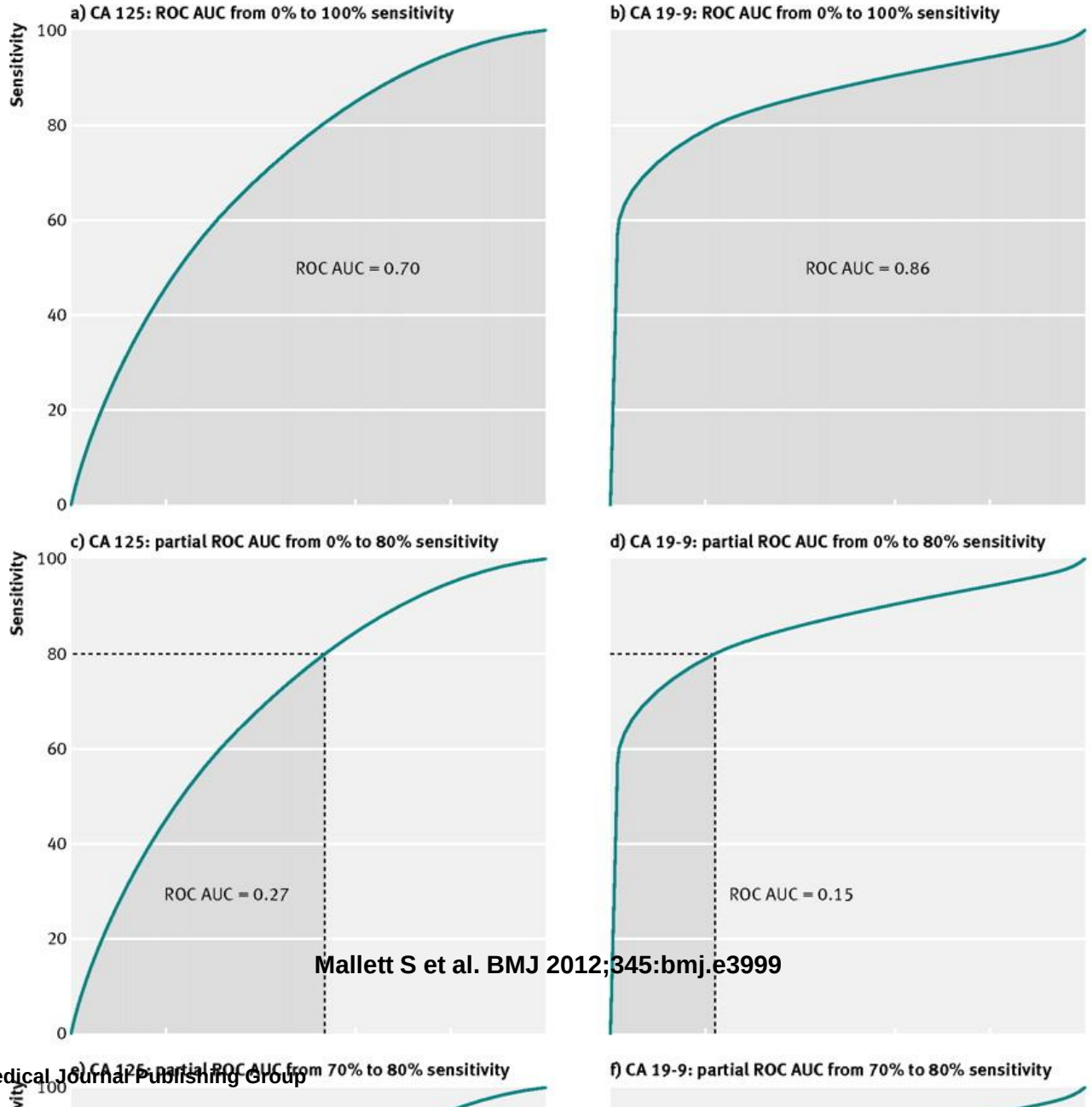
©2012 by British Medical Journal Publishing Group

Fig 3 Use of ROC AUC to compare two tests: CA 19-9 and CA 125.



Mallett S et al. BMJ 2012;345:bmj.e3999

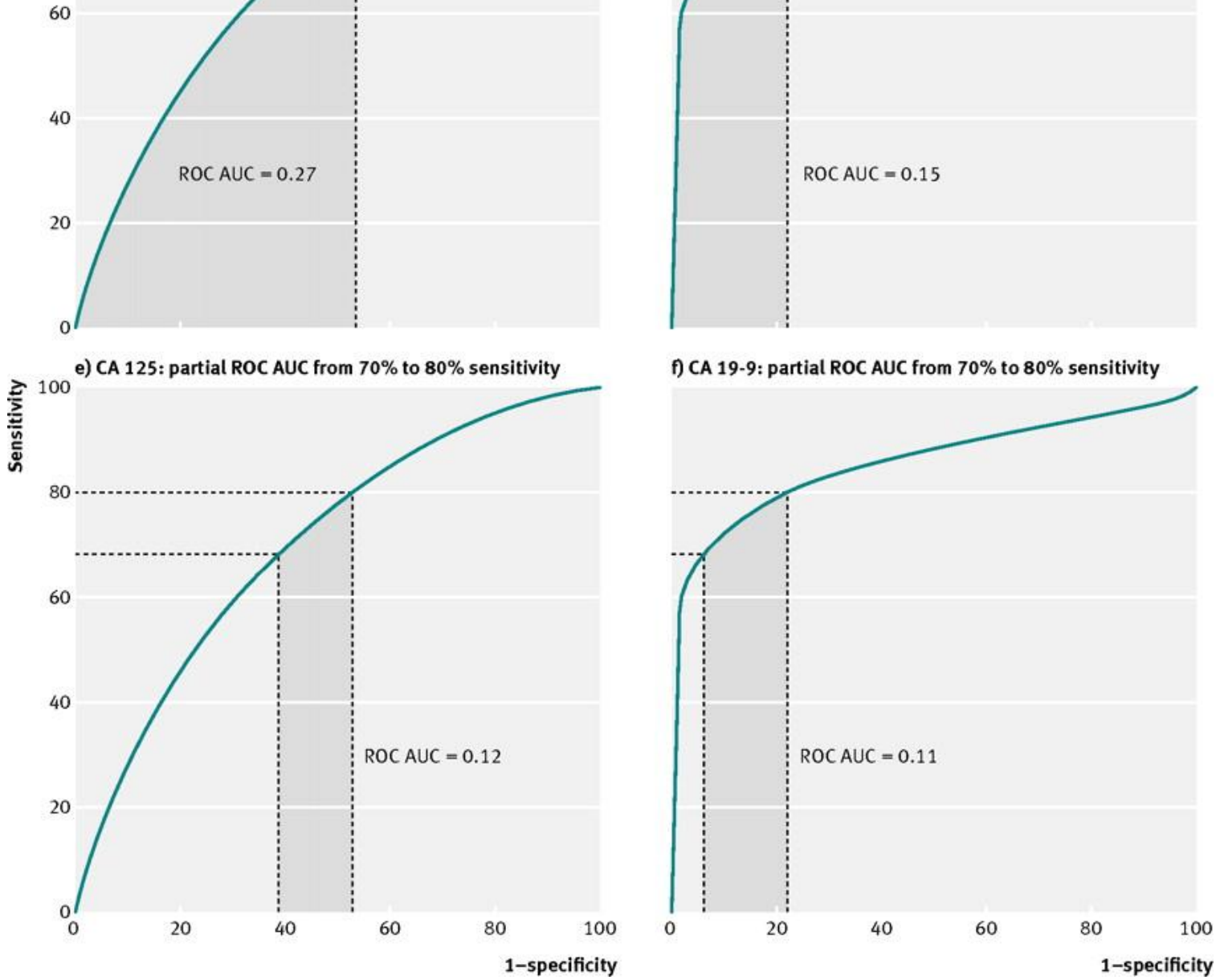
©2012 by British Medical Journal Publishing Group



Mallett S et al. BMJ 2012;345:bmj.e3999

©2012 by British Medical Journal Publishing Group

BMJ

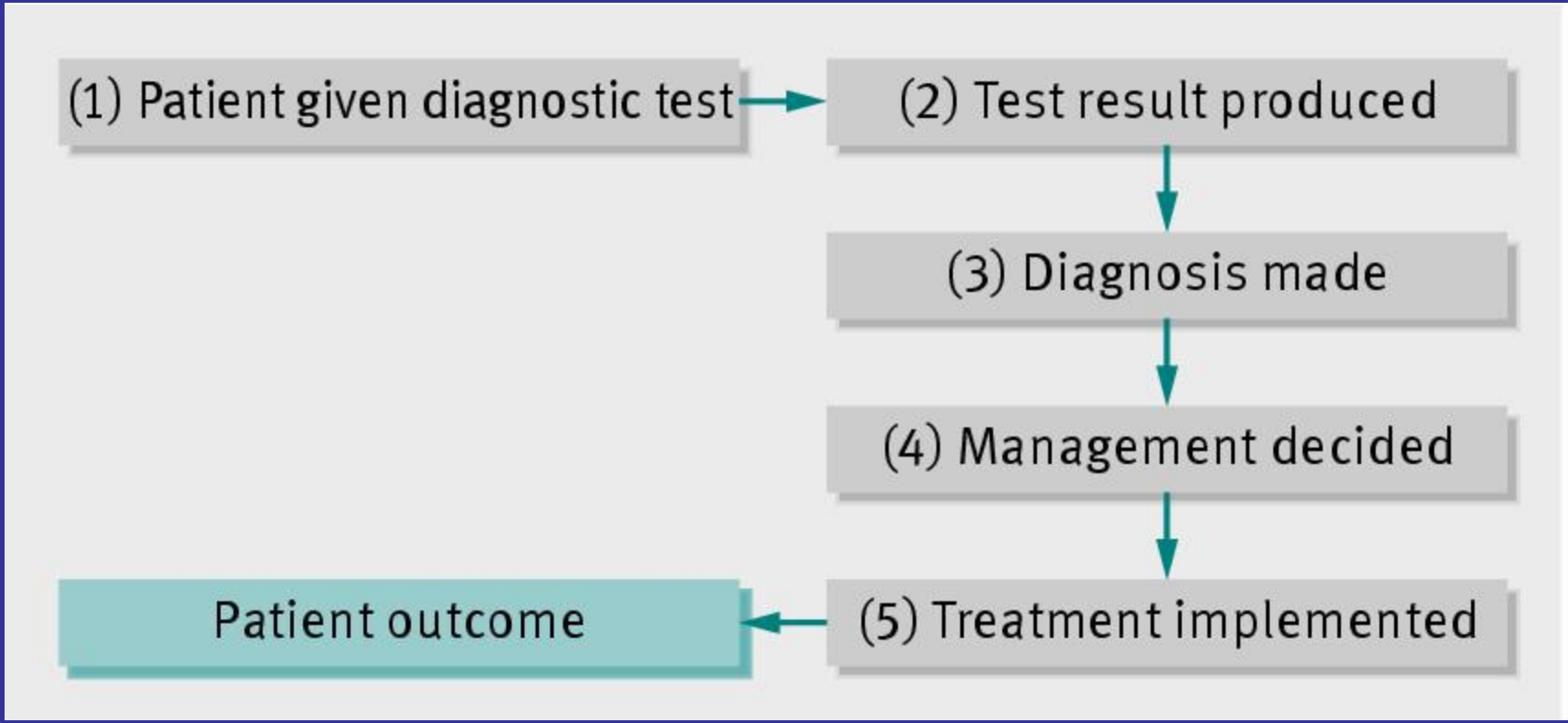


Mallett S et al. BMJ 2012;345:bmj.e3999

©2012 by British Medical Journal Publishing Group

Diagnostic tests don’t make patients better!

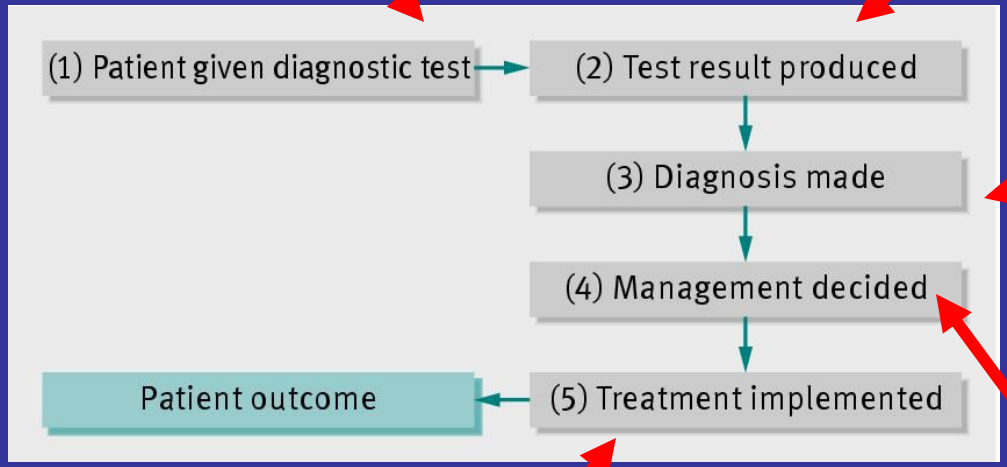
Pathway from test to outcome



Ferrante di Ruffano. BMJ 2012

Timing of test
Feasibility
Test process

Interpretability
Accuracy
Speed of results



Speed of diagnosis
Diagnostic yield
Diagnostic confidence

Speed of receiving treatment
Treatement efficacy
Adherence

Therapeutic yield
Therapeutic confidence

Systematic reviews of diagnostic test accuracy studies

Systematic reviews of diagnostic test accuracy studies: How to rapidly appraise?

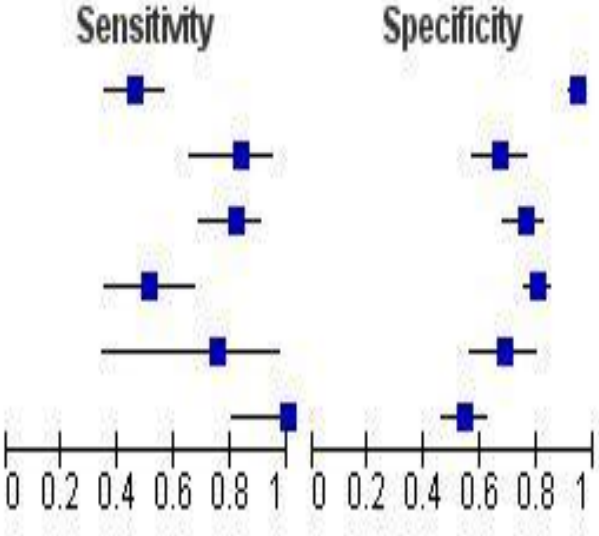
- Well formatted question
- Find all the studies
- Appraisal (use QUADAS-2 tool!)
- Summarise
- *Sometimes* meta-analysis

Table of Study Characteristics is always the most important table

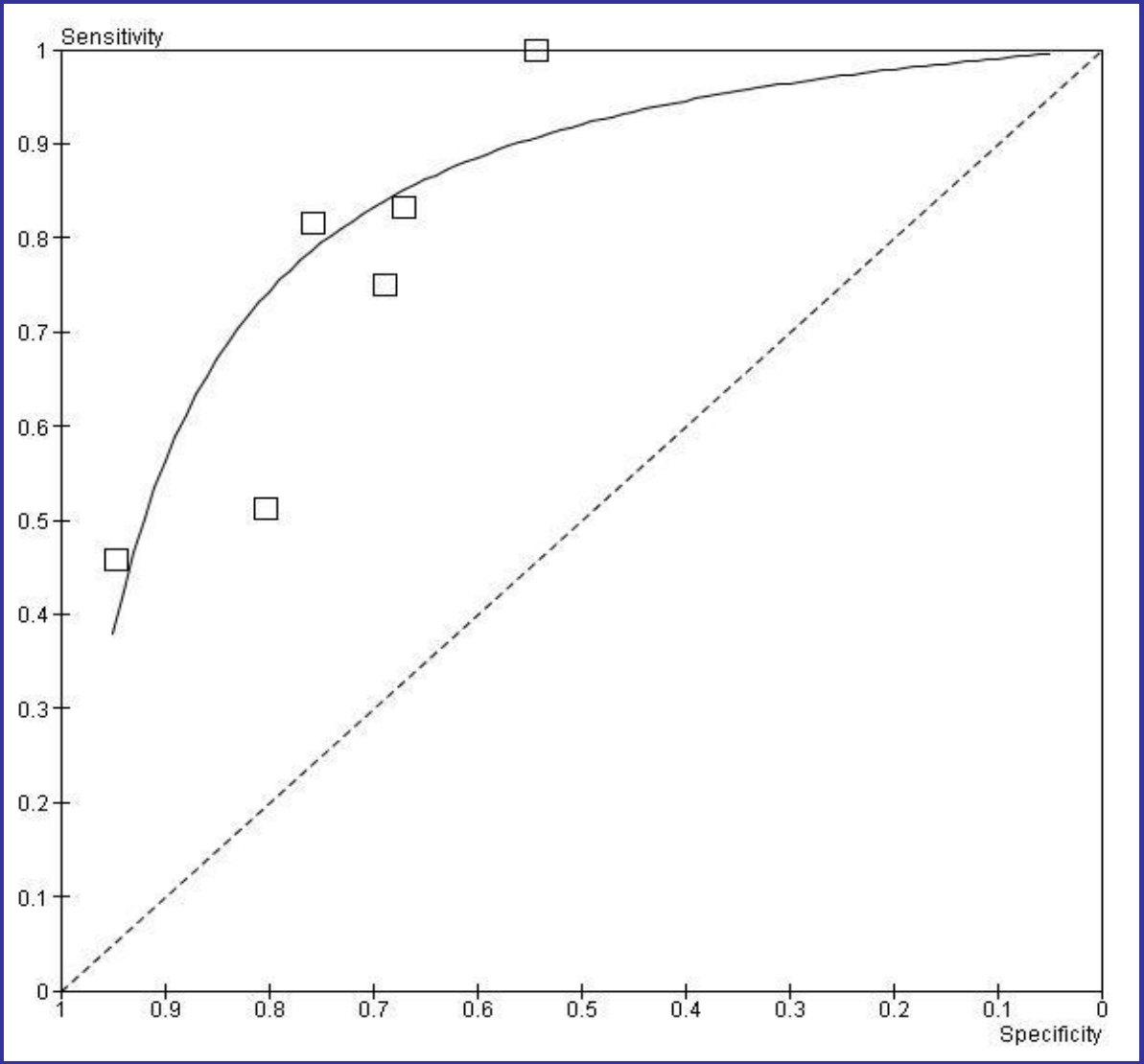
- design features (e.g. prospective/retrospective),
- Recruitment (e.g. consecutive/case-control)
- setting (e.g. country, health care setting)
- participants (e.g. inclusion & exclusion criteria, age)
- details of the index test (e.g. how was it done, cut-offs used)
- details of the reference standard (e.g. may vary between studies)
- target condition (e.g. prevalence, severity)

Presenting results: “Forest plot” (but it is not really!)

Study	TP	FP	FN	TN	Sensitivity	Specificity
Andreola 2007	43	17	51	297	0.46 [0.35, 0.56]	0.95 [0.91, 0.97]
Berger 1996	25	32	5	65	0.83 [0.65, 0.94]	0.67 [0.57, 0.76]
Galetto-Lacour 2008	44	36	10	112	0.81 [0.69, 0.91]	0.76 [0.68, 0.82]
Hsiao 2006	21	68	20	278	0.51 [0.35, 0.67]	0.80 [0.76, 0.84]
Thayyil 2005	6	20	2	44	0.75 [0.35, 0.97]	0.69 [0.56, 0.80]
Wells 2001	17	76	0	90	1.00 [0.80, 1.00]	0.54 [0.46, 0.62]

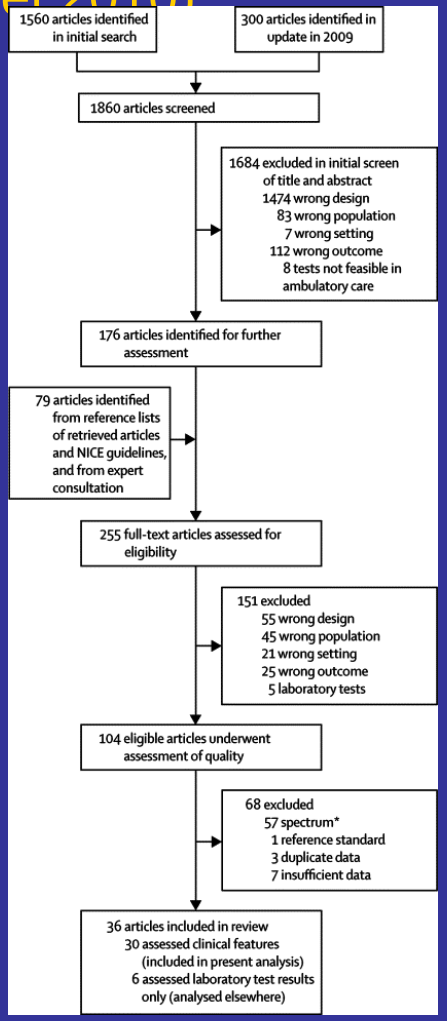


Presenting results in ROC space - each point is a different study



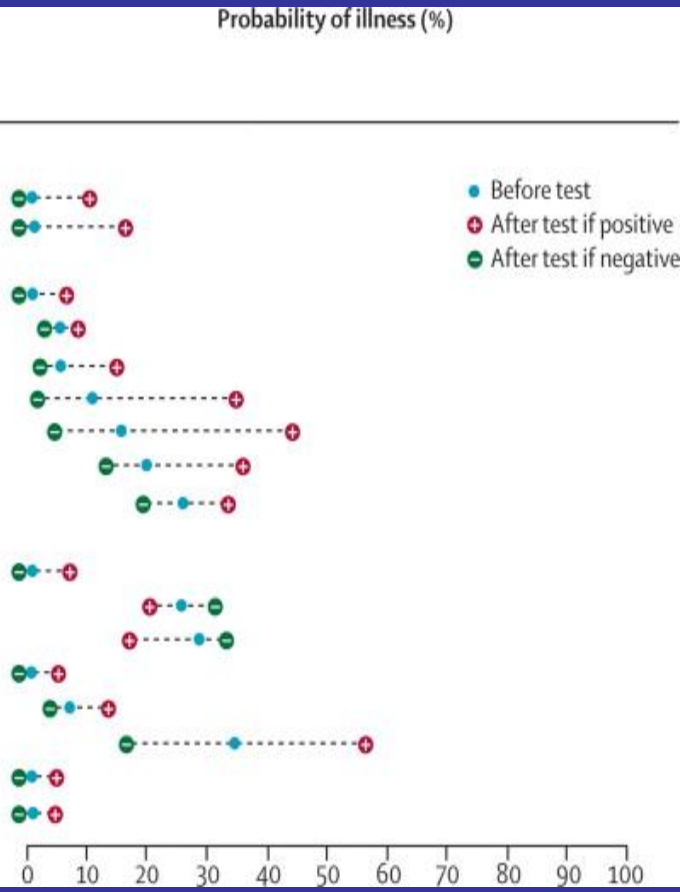
Systematic review of clinical features & lab tests to identify serious infection in children in ambulatory care (Van den Bruel, Haj-Hassan, Thompson et al. Lancet 2010)

- 36 studies included in review
  - 30 clinical features
  - 6 lab tests only
- 1 study from general practice (Belgium), rest from ED or ambulatory paediatrics
- **Red flags = where feature reported to have positive LR > 5.0 in at least one study**



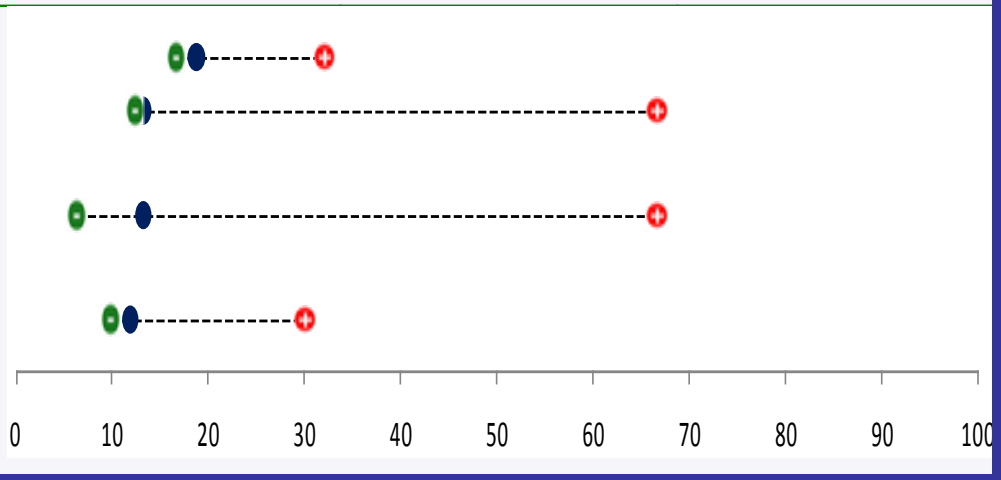
Results: child assessment and behaviour features

	Study reference	Prevalence*	Age range	Likelihood ratio (95% CI)	
				Positive	Negative
Global assessment					
Parental concern†	5	Low	<17 years	14·40 (9·30–22·10)	0·55 (0·39–0·78)
Clinician instinct that something wrong	5	Low	<17 years	23·50 (16·80–32·70)	0·38 (0·24–0·60)
Clinical impression	5	Low	<17 years	8·30 (6·25–11·10)	0·37 (0·23–0·62)
	36	Intermediate	3–36 months	1·05 (0·15–7·48)	1·00 (0·90–1·11)
	40	Intermediate	<24 months	2·75 (1·56–4·86)	0·64 (0·41–1·00)
	49‡	Intermediate	≤15 years	4·27 (2·98–6·11)	0·26 (0·12–0·56)
	42§	Intermediate	1 month to 5 years	4·14 (2·33–7·35)	0·28 (0·10–0·77)
Child appears ill	27	Intermediate	0–36 months	2·20 (1·78–2·78)	0·65 (0·55–0·77)
	24	High	1–36 months	1·40 (1·15–1·71)	0·67 (0·50–0·88)
Child behaviour					
Changed crying pattern	5	Low	<17 years	10·50 (4·62–13·20)	0·67 (0·51–0·89)
	24	High	1–36 months	0·74 (0·56–0·96)	1·30 (1·07–1·60)
	45‡	High	1 month to 15 years	0·49 (0·25–0·96)	1·16 (1·03–1·31)
Child drowsy	5	Low	<17 years	6·60 (4·17–10·50)	0·65 (0·49–0·86)
	44¶	Intermediate	3 months to 6 years	1·99 (1·29–3·08)	0·65 (0·42–1·00)
	45¶	High	1 month to 15 years	2·43 (1·82–3·26)	0·37 (0·25–0·56)
Child moaning	5	Low	<17 years	5·90 (1·97–17·70)	0·92 (0·81–1·03)
Child inconsolable	5	Low	<17 years	5·50 (2·66–11·50)	0·83 (0·69–0·99)



Presenting results: Dumbbell plots

Study	Setting	Cut-off used	Likelihood ratios	
			LR+	LR-
Index test 1				
Study a	Int	prolonged	2.05 (1.01-4.19)	0.87 (0.72-1.04)
Study b	Int	≥1.18	13.1 (1.23-38.8)	0.92 (0.82-1.04)
Index test 2				
Study b	Int	≥1.2	13.1 (5.88-29.0)	0.44 (0.27-0.70)
Index test 3		**(1000/mm³)**		
Study b	Int	≤150	3.20 (1.36-7.53)	0.81 (0.64-1.03)



Metaanalysis- simple pooling?

- Simply pooling together sensitivity or specificity gives an estimate of this “average” effect.
- But too simplistic - ignores some details of diagnostic accuracy studies eg different thresholds, heterogeneity between studies, correlation between sensitivity and specificity

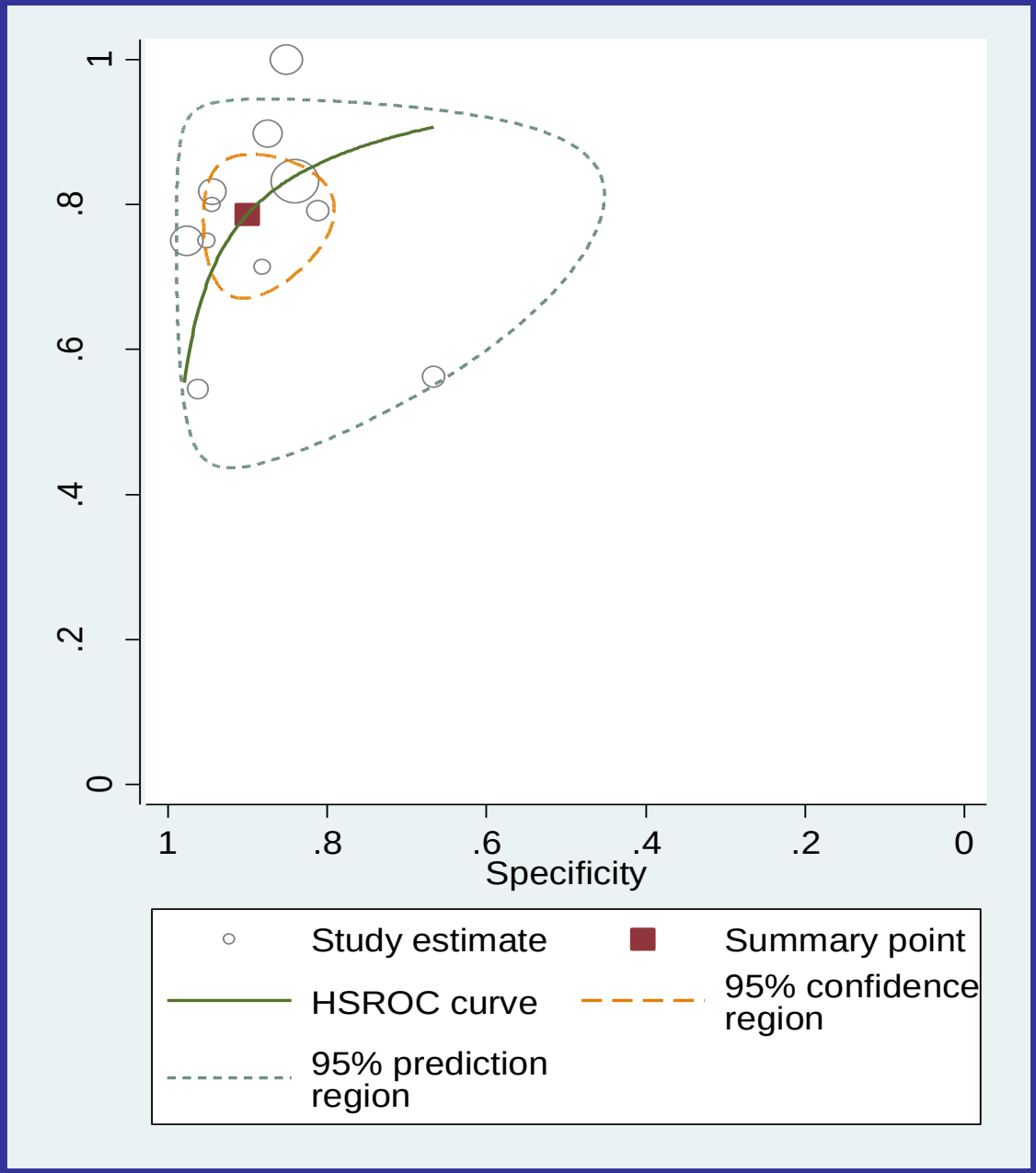
- For example in a meta-analysis of 3 studies which had different values of sensitivity and specificity;
  - Study 1: 10% & 90%,
  - Study 2: 80% and 80%, and
  - Study 3: 90% and 10%.
- Simply averaging these, gives sensitivity of 60% and specificity of 60% - which does not really tell us anything useful about these data!

Meta analysis: Hierarchical summary ROC curves and bivariate random-effects models

- Two statistical models are used to incorporate the variation between studies (both use random effects) to give a summary ROC curve or an “average” sensitivity and specificity.

  Hierarchical summary ROC curves

  The bivariate random-effects model


1
.8
.6
.4
.2
0
1
.8
.6
.4
.2
0
Specificity
Study estimate
Summary point
HSROC curve
95% confidence region
95% prediction region

Diagnostic reasoning

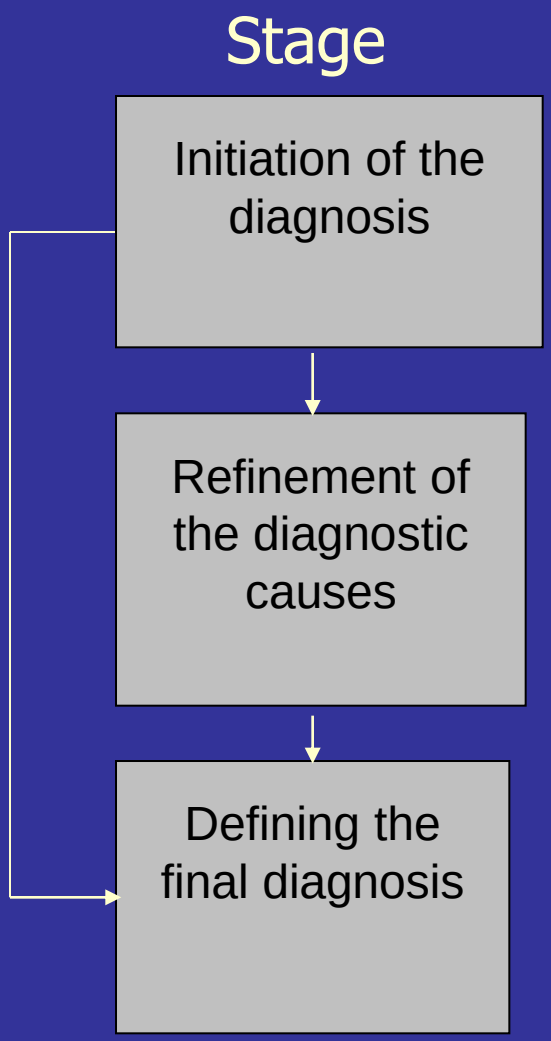
- Clinicians use many different methods to make diagnostic decisions
- Too much emphasis on 2x2 tables and numbers can seem irrelevant

Diagnostic strategies

Diagnostic strategies used in primary care. Heneghan et al BMJ 2009

- Aim: identify types and frequency of diagnostic strategies used in primary care
  - 6 GPs collected and recorded strategies used on 300 patients.
  - Identified and refined diagnostic strategies

Diagnostic stages & strategies used by GPs



Strategy

- **Spot diagnoses**
- **Self-labelling**
- **Presenting complaint**
- **Pattern recognition**
- Restricted Rule Outs
- Stepwise refinement
- Probabilistic reasoning
- Pattern recognition fit
- Clinical Prediction Rule

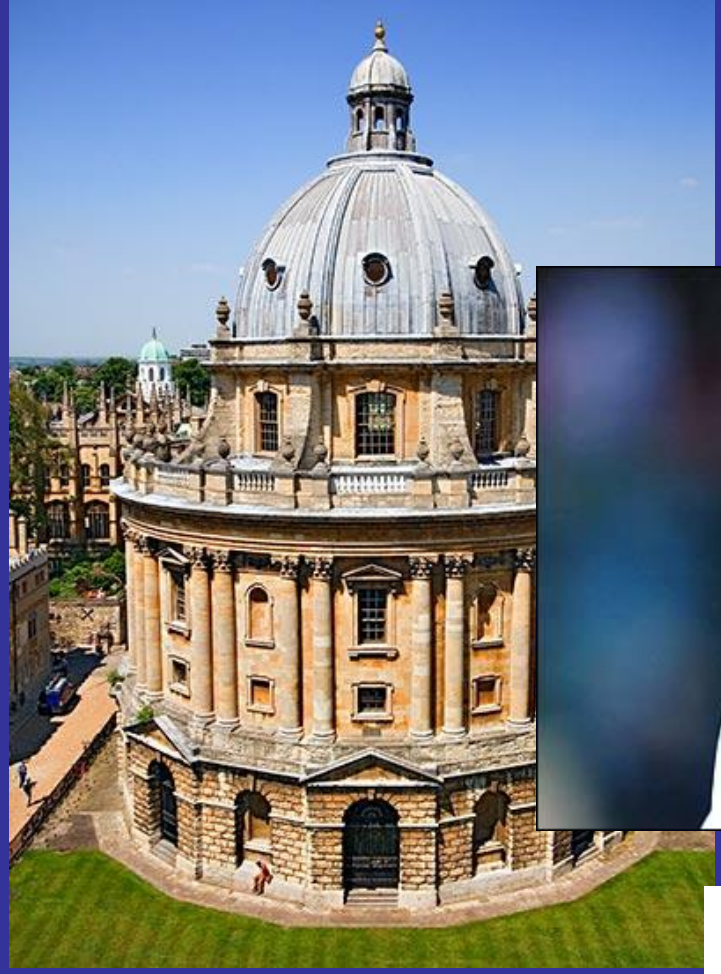
- Known Diagnosis
- Further tests ordered
- Test of treatment
- Test of time
- No label

Some examples of diagnostic strategies clinicians might use

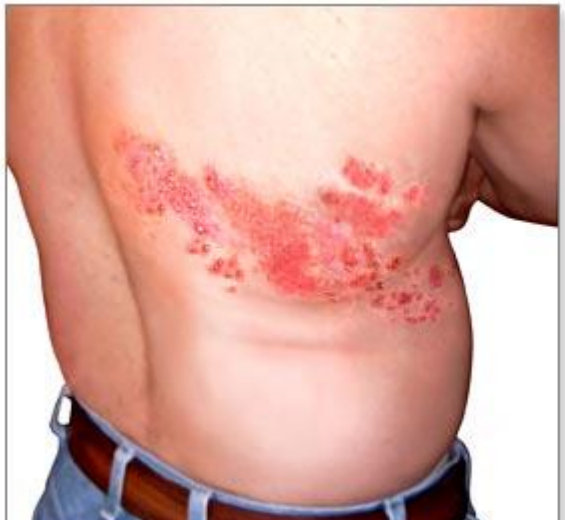
Spot diagnosis

- Unconscious recognition of non-verbal pattern, e.g.:
  - visual (skin condition)
  - auditory (barking cough with croup)
- Fairly instantaneous, no further history needed.

*Brooks LR. Role of specific similarity in a medical diagnostic task. J Exp Psychol Gen 1991;220:278-87




10



Useful books

- Diagnostic Tests Toolkit. Thompson & Van den Bruel. Wiley-Blackwell.
- Evidence base of Clinical Diagnosis. Knottnerus & Buntinx. Wiley-Blackwell
- Evidence-based Diagnosis. Newman & Cohn. Cambridge Univ Press
- The Diagnostic Process. John Balla. Cambridge Univ Press
- Evidence based Physical Diagnosis. Steven McGee. Saunders

Useful journal articles on diagnostics

- Bossuyt. Additional patient outcomes and pathways in evaluations of testing. Med Decis Making 2009
- Heneghan et al. Diagnostic strategies used in primary care. BMJ 2009
- Ferrante di Ruffano. Assessing the value of diagnostic tests: a framework for designing and evaluating trials. BMJ 2012
- Mallett et al. Interpreting diagnostic accuracy studies for patient care. BMJ 2012
- Bossuyt et al. STARD initiative. Ann Int Med 2003
- Lord et al. Using priniciples of RCT design to guide test evaluation. Med Decis Making 2009
- Rutjes et al. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006
- Lijmer et al. Proposals for phased evaluation of medical tests. Med Decis Making 2009
- Whiting et al. QUADAS-2: revised tool for quality assessment of diagnostic accuracy studies. Ann Int Med 2011

Thanks!

Centre for Monitoring and Diagnosis
www.madox.org

Centre for Evidence Based Medicine
www.cebm.net

Matthew.thompson@phc.ox.ac.uk

Clinical prediction rules

Clinical prediction rules

- Incorporates information from diagnostic studies into clinical practice.
- A formal way of bringing together combinations of predictor variables, which may include clinical features, examination findings, results of laboratory or imaging investigations.
- Why have them?
  - Simplify/streamline the diagnostic process, by identifying the clinical features which are the most useful (or predictive) of a particular outcome
  - teaching aid to help clincians learn which clinical features are most important

Clinical prediction rules

Clinical prediction rules most useful:

- The clinical decision is particularly complex
- Clinical outcome of interest is rare and/or very serious
- To guide the need (or not) for more invasive/ expensive further diagnostic investigations
- As screening tests,
- Determine prognosis

Selecting Clinical prediction rules

- How valid is it?
  - Consider how the rule was derived and the level of validation (see above)
- How sensible is it?
  - Predictors should be those that are routinely collected, can be measured in the same way
  - Some rules exclude predictors that clinicians are used to using. A rule that does not include these are not likely to be used /believed by clinicians.
- What is its possible impact?
  - Change patient outcomes?
  - How easy will it be to use it?
  - How often is the rule likely to be overruled in clinical practice?