Dr Mariska Leeflang
Dept. Clinical Epidemiology, Biostatistics and Bioinformatics
Academic Medical Center, University of Amsterdam
Room J1B – 210
PO Box 227700
1100 DE Amsterdam
m.m.leeflang@amc.uva.nl

Making the results understandable for end-users

Montreal, Monday May 25th, 2009

Mariska Leeflang
(with thanks to Chris Hyde and Rob Scholten)

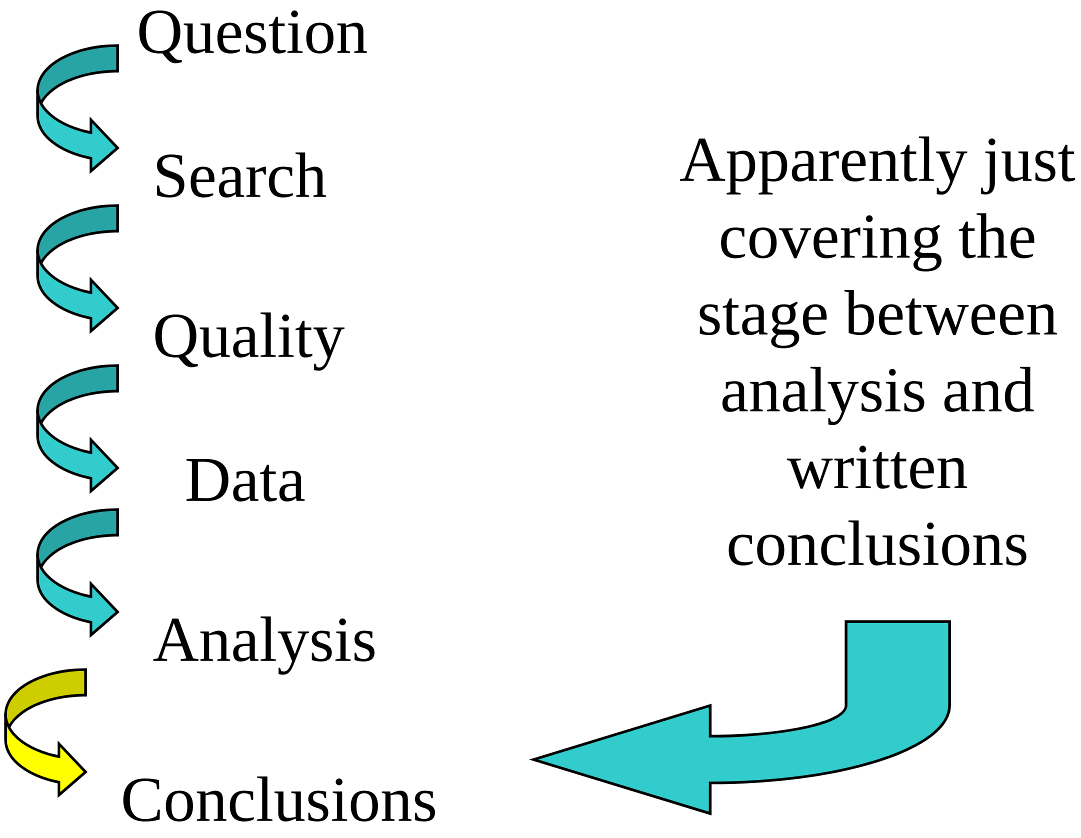
! There is more to life than hierarchical models !

Outline

- Ingredients of Discussion section
- Types of results of a DTA review
- Interpretation of results
- Presentation of results

- Small groups (if enough time)

Ground to be covered



Question

Search

Quality

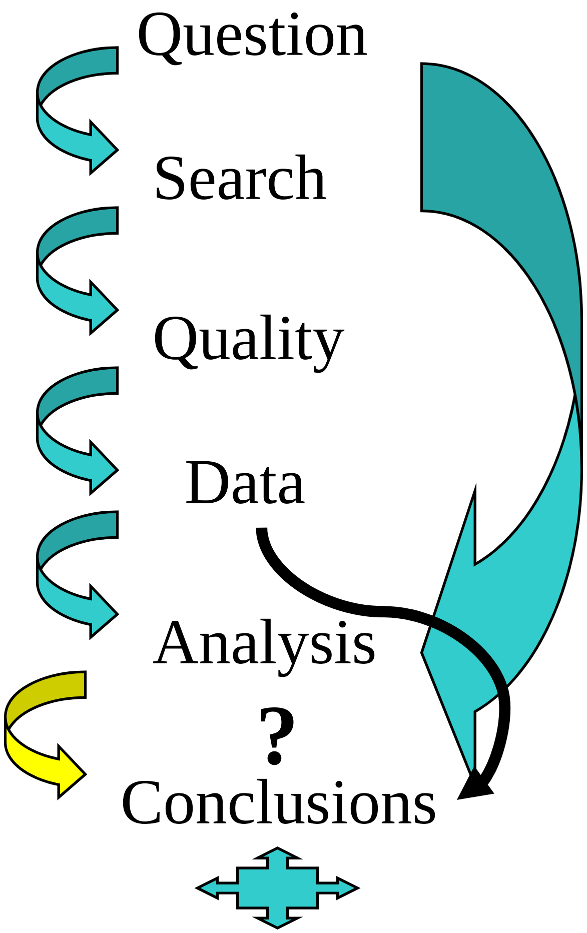
Data

Analysis

Conclusions

Apparently just covering the stage between analysis and written conclusions

But…



- Meta-analysis is not the only contributor to conclusions
- All stages of the review contribute to the conclusions
- Qualitative analysis still has an important contribution
- Output of meta-analysis is NOT a conclusion.
- Summary measures (even if you believe them!!) need interpretation

Interpreting results

- Its hard!
  - ? What proportion of review time is invested in considering results and writing conclusions which are truly supported by the data we present !
- Important
  - Many readers will rely authors conclusions
- Structure using standard sub-sections

Discussion section

Should contain the following issues:

- Summary of main results
- Strengths and weaknesses of the review
- Applicability of findings to practice and policy
- Implications for practice
- Implications for research

Summary of main results

- Restate the initial question(s)
- No. of included studies / patients / samples
- Characteristics of included studies *
- Quality *
- Study results, esp. summary sensitivity and specificity
- Consistent with summary of results table

* May want to defer to next section on "Strengths & weaknesses of the review"

Strengths & weaknesses

- Limitations of review method
  - Shortcomings in search
  - Studies not retrieved and translations pending
  - Not chasing missing data esp. quality and co-variates

- Limitations of included studies
  - Clinical spectrum esp. target condition, prevalence and clinical setting
  - Different versions of the index test, including use of different thresholds/cut-offs
  - Study quality

Strengths & weaknesses (2)

- Limitations in study results
  - Transferability of results to other settings
  - Sources of heterogeneity + implication
- Review results in context of other reviews
  - E.g. reviews on related treatments

Applicability to practice and policy

- Applicability to your own objectives?
- May bring in info from other sources (but remember it is not systematically reviewed)
    - Reliability of test
    - Direct harms and benefits of tests
    - Consequences of false positives and negatives
    - Costs
    - Other studies may indicate effects on diagnostic yield, changed decisions, patient outcome & cost-effectiveness

Authors conclusions

- Implications for practice
  - Implications for health care policy
  - Implications for clinical practice

    NB: present information rather than advice (review must be as relevant as possible to an international audience)

- Implications for research
  - "What" and "How"
  - Avoid bland statements like "more research is needed"

Outline

- Ingredients of Discussion section
- Types of results of a DTA review
- Interpretation of results
- Presentation of results

- Small groups (if enough time)

Types of results of a DTA SR

1. Quantitative results
2. sROC curve only
3. No quantitative results

1. Quantitative results

- What measure do we need?
  - Sensitivity / specificity?
  - Predictive values?
  - Likelihood ratios?
  - Proportion of false negatives?
  - Etc.

Sensitivity and specificity

Calculation of summary estimates of sensitivity and specificity sensible if

- clinical sensible
- not too much (statistical) heterogeneity
- no obvious threshold effect

Derive other measures (e.g. likelihood ratios, predictive values) from these

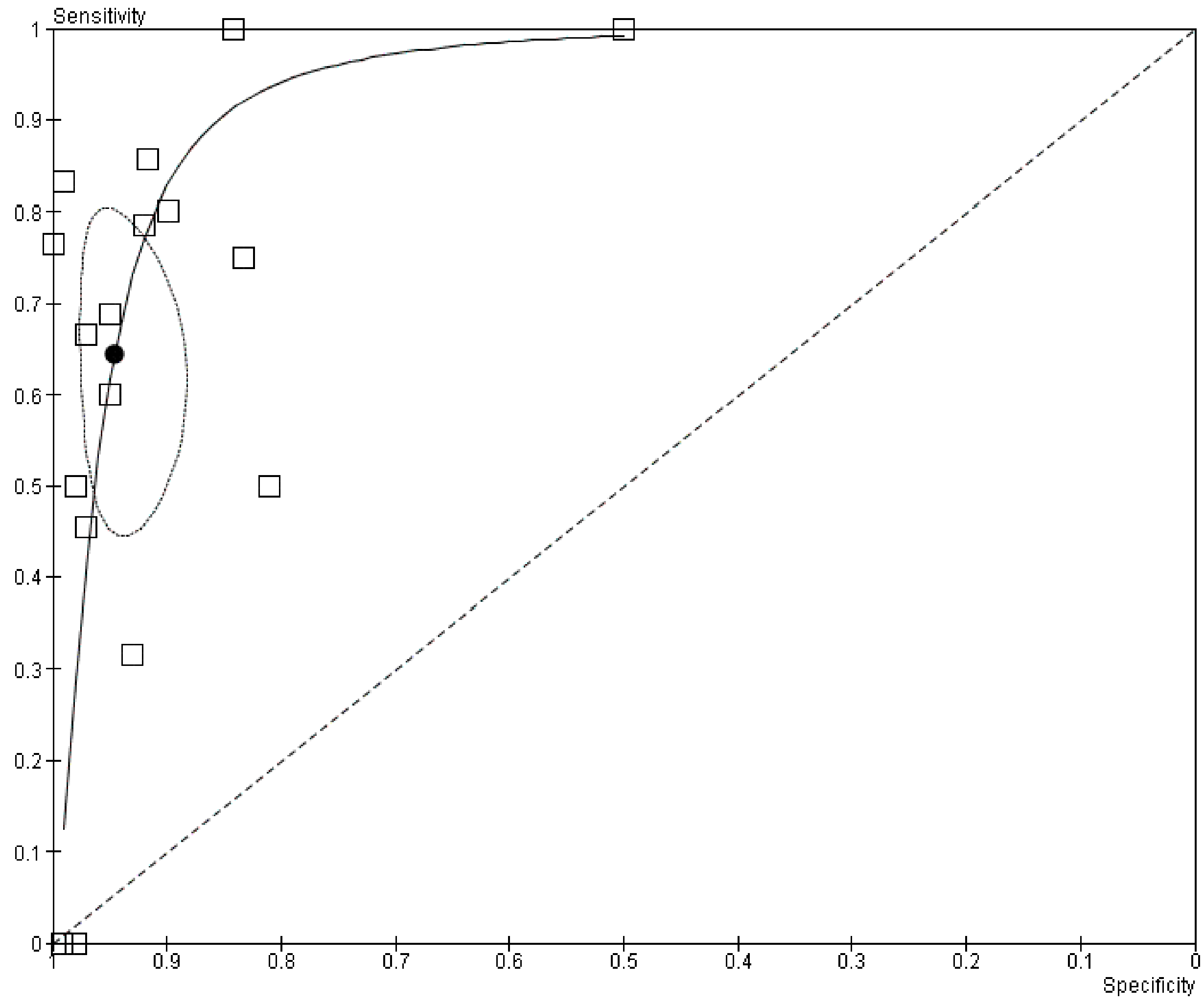
Interpretation of summary sensitivity and specificity

- Summary estimates are derived from random effects models

- Mean of a range of possible values for sens and spec (with a 95%-CE of the mean)

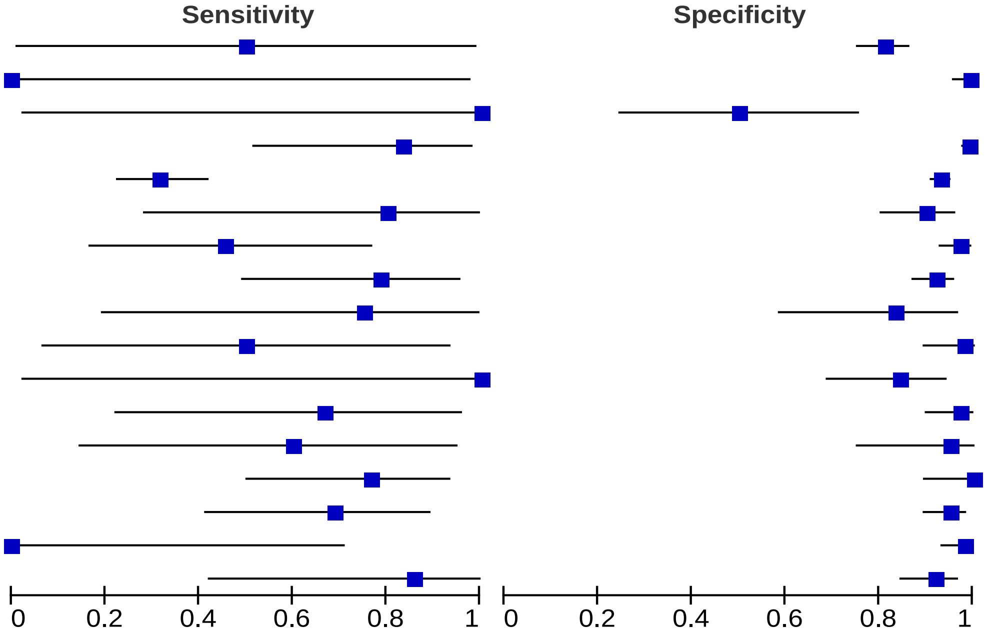
- Still many "real" values possible, including values outside the 95-CE range

Summary sensitivity and specificity



Apparent heterogeneity?

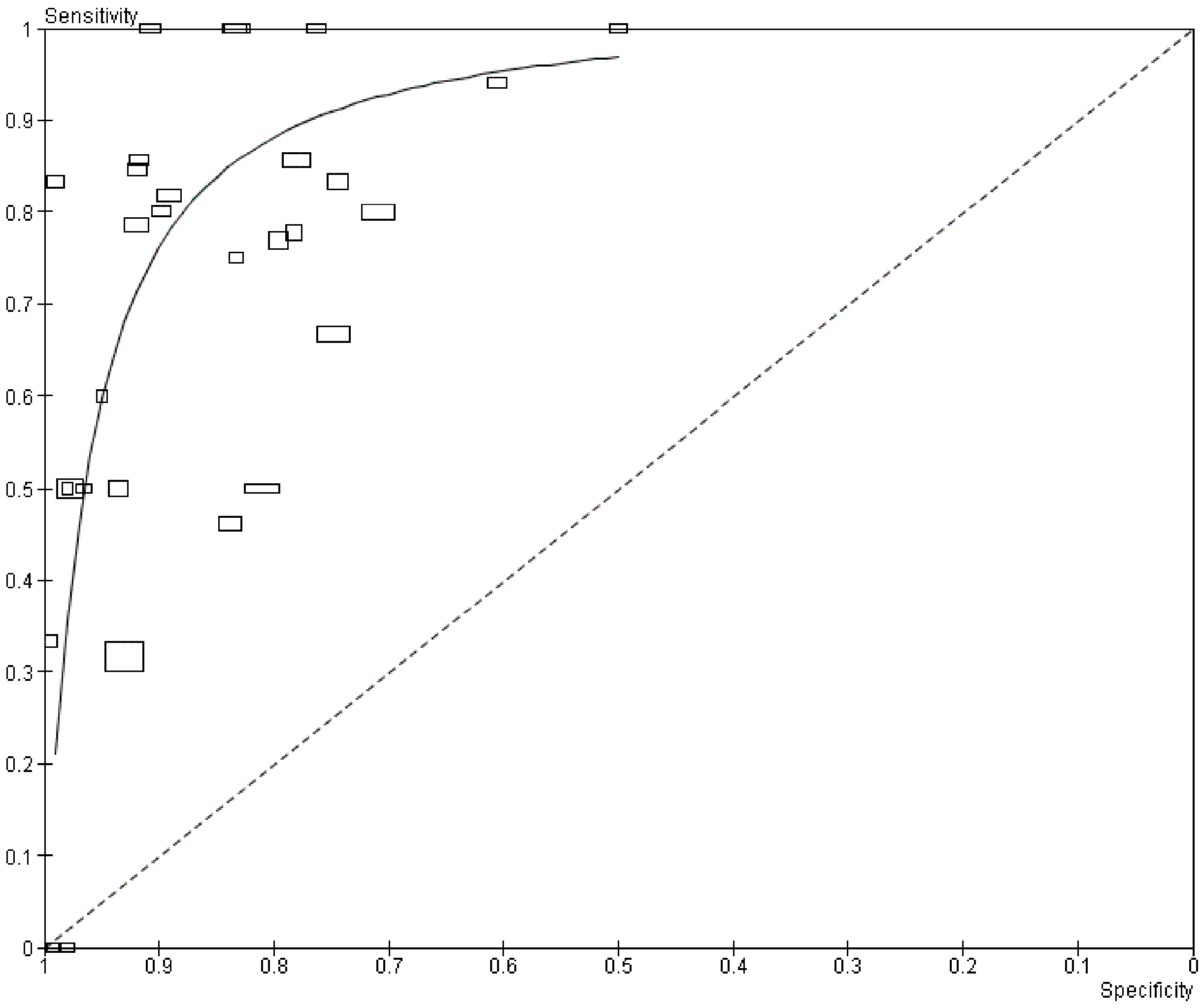
Study	TP	FP	FN	TN	Sensitivity	Specificity
Adam 2004	1	41	1	175	0.50 [0.01, 0.99]	0.81 [0.75, 0.86]
Allan 2005	0	1	1	123	0.00 [0.00, 0.97]	0.99 [0.96, 1.00]
Bialek 2002	1	8	0	8	1.00 [0.03, 1.00]	0.50 [0.25, 0.75]
Doermann 2002	10	4	2	407	0.83 [0.52, 0.98]	0.99 [0.98, 1.00]
Herbrecht 2002	31	49	67	650	0.32 [0.23, 0.42]	0.93 [0.91, 0.95]
Kallel 2003	4	7	1	62	0.80 [0.28, 0.99]	0.90 [0.80, 0.96]
Kawazu 2004	5	4	6	134	0.45 [0.17, 0.77]	0.97 [0.93, 0.99]
Lai 2007	11	14	3	161	0.79 [0.49, 0.95]	0.92 [0.87, 0.96]
Machetti 1998	3	3	1	15	0.75 [0.19, 0.99]	0.83 [0.59, 0.96]
Moragues 2003	2	1	2	49	0.50 [0.07, 0.93]	0.98 [0.89, 1.00]
Pereira 2005	1	6	0	32	1.00 [0.03, 1.00]	0.84 [0.69, 0.94]
Rovira 2004	4	2	2	66	0.67 [0.22, 0.96]	0.97 [0.90, 1.00]
Scotter 2005	3	1	2	19	0.60 [0.15, 0.95]	0.95 [0.75, 1.00]
Suankratay 2006	13	0	4	33	0.76 [0.50, 0.93]	1.00 [0.89, 1.00]
Ulusakarya 2000	11	6	5	113	0.69 [0.41, 0.89]	0.95 [0.89, 0.98]
White 2005	0	2	3	100	0.00 [0.00, 0.71]	0.98 [0.93, 1.00]
Williamson 2000	6	8	1	89	0.86 [0.42, 1.00]	0.92 [0.84, 0.96]



2. sROC curve only

- Multiple cut-off values
- Vast heterogeneity
- ?

Multiple cut-offs



Relevant subgroups?

- Subgroups according to
  - Cut-off value
  - Prevalence
  - Spectrum of disease
  - Patient characteristics
  - Setting
  - Etc.

3. No quantitative results

- Flawed studies
- Very poor quality
- No data
- Too much heterogeneity
- ..

Outline

- Ingredients of Discussion section
- Types of results of a DTA review
- Interpretation of results
- Presentation of results

- Small groups (if enough time)

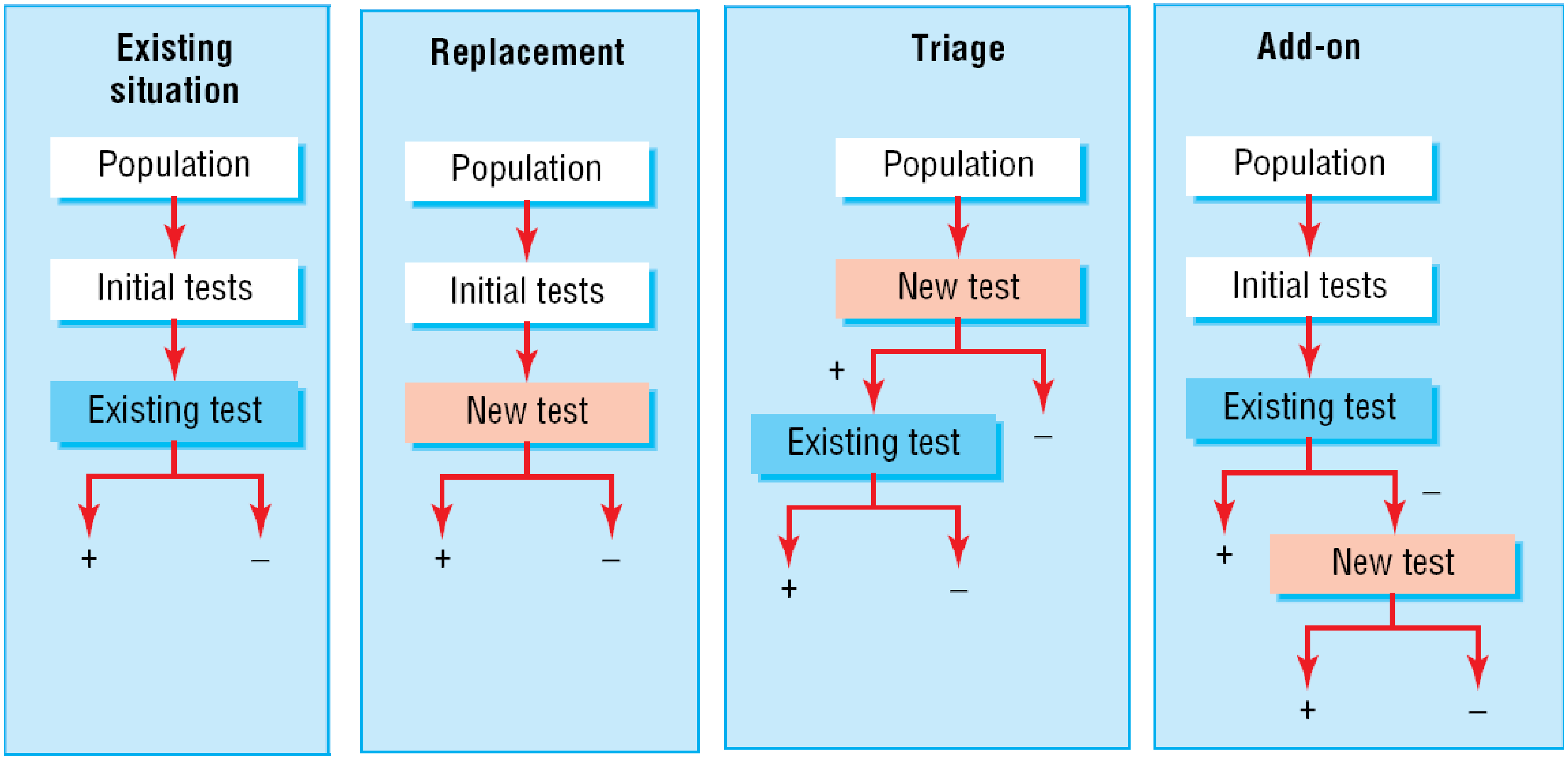
Purpose of test and test features

- Remember the purpose of your test
  1. Replacement
  2. Triage / screening
  3. Add-on

- Each situation may require different test features

Bossuyt et al. BMJ 2006

Test comparisons



1. Replacement

Replace test A with test B, because test B

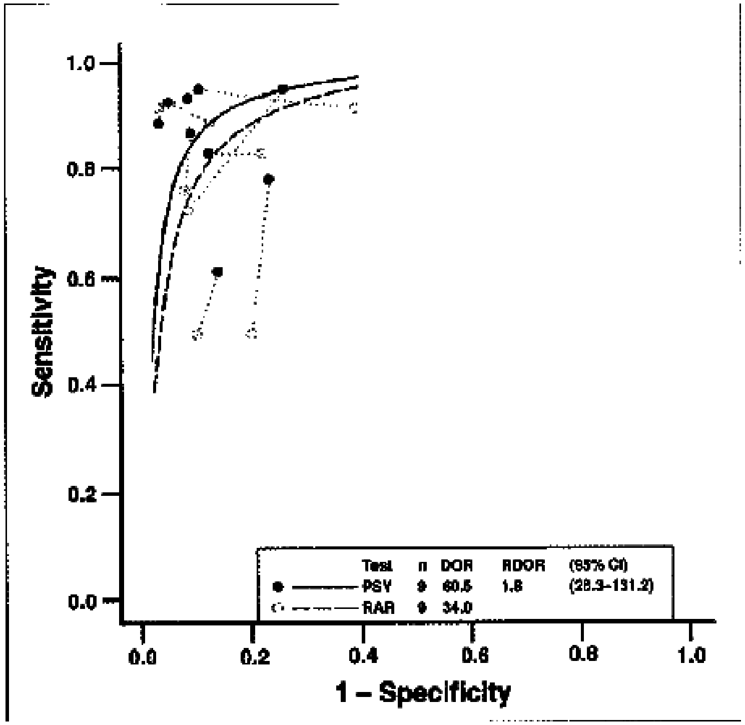
- more accurate
- less invasive, easier to do, less risky
- less uncomfortable for patients
- quicker to yield results
- technically less challenging
- more easily interpreted
- etc.

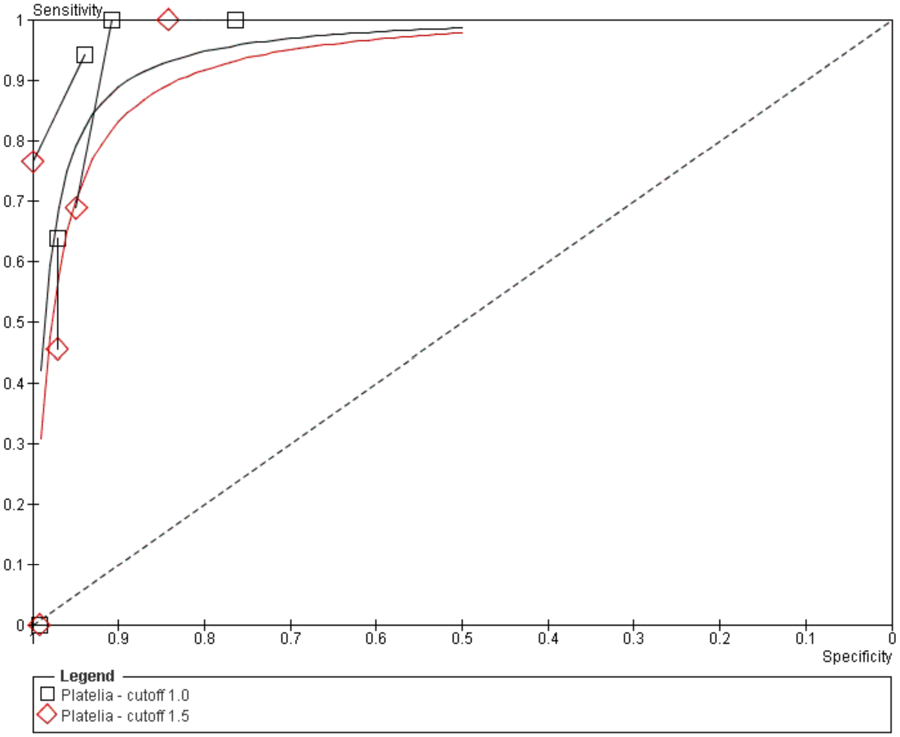
Replacement: preferred design

- Both tests tested in same patients (paired design)
  - All patients undergo A, B and reference standard
  - Direct comparisons

- RCT
  - Patients randomly allocated to either A or B
  - Both groups undergo reference standard
  - Valid comparisons

Direct comparisons





Often only indirect comparisons

- Comparisons may then be biased due to
  - Subgroups
  - Differences in methodological quality
  - Etc.

- Be cautious with conclusions

Multiple sROCs

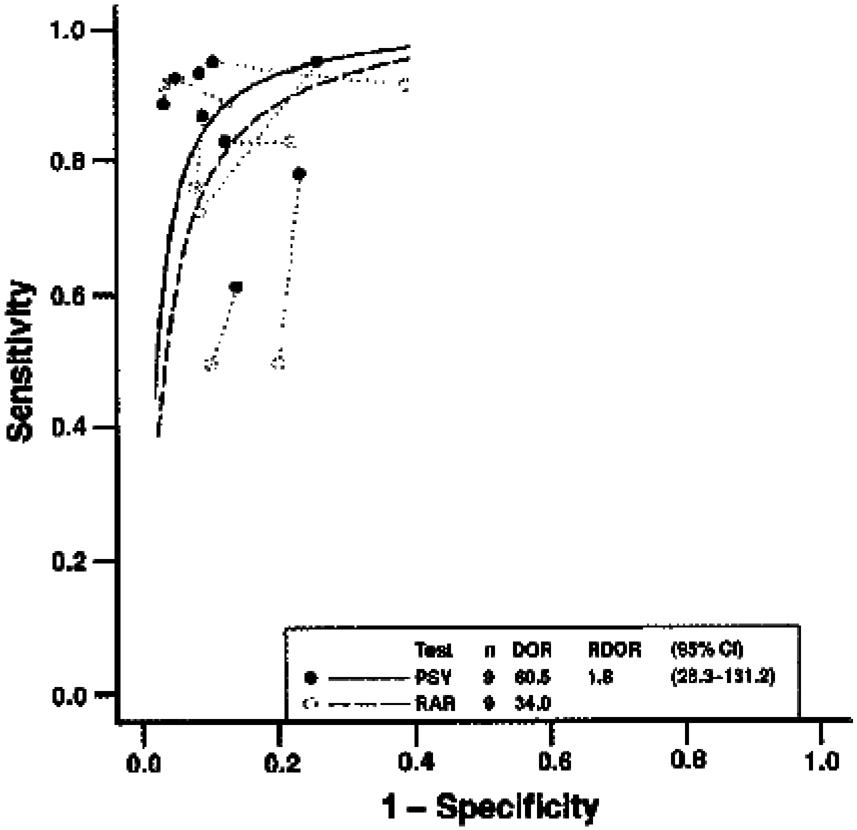
a. Curve B “Northwest” of curve A

b. Curves cross

c. Curves in different areas

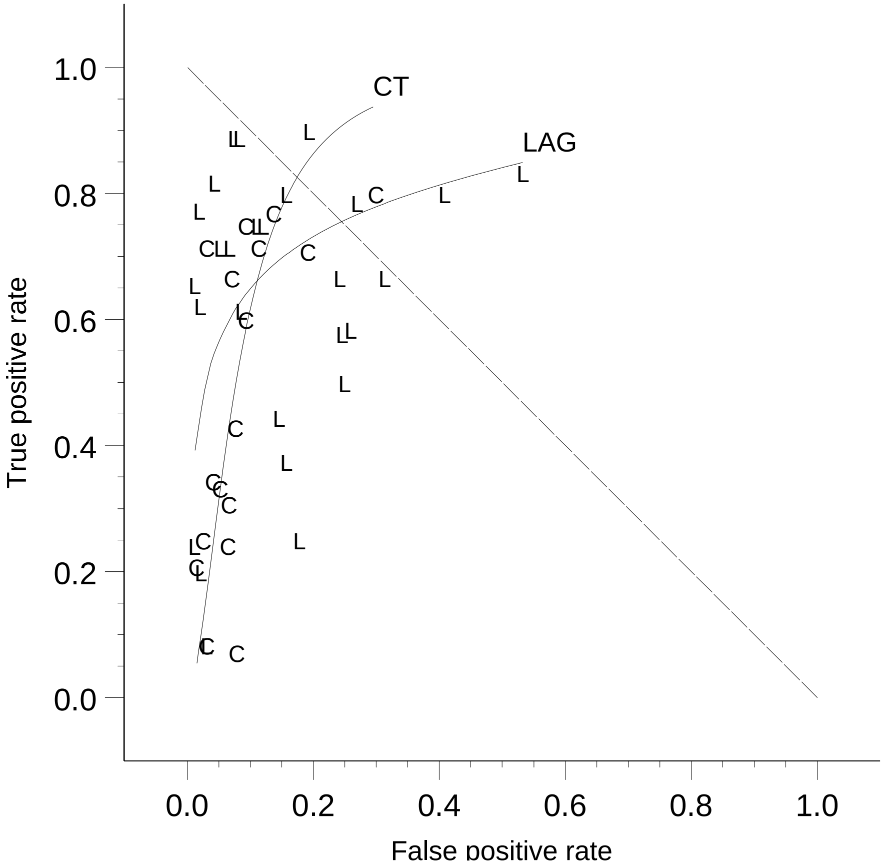
a. B more accurate than A

- Trade-off
- Assess other aspects
  - Costs
  - Burden
  - Complexity
  - Etc.



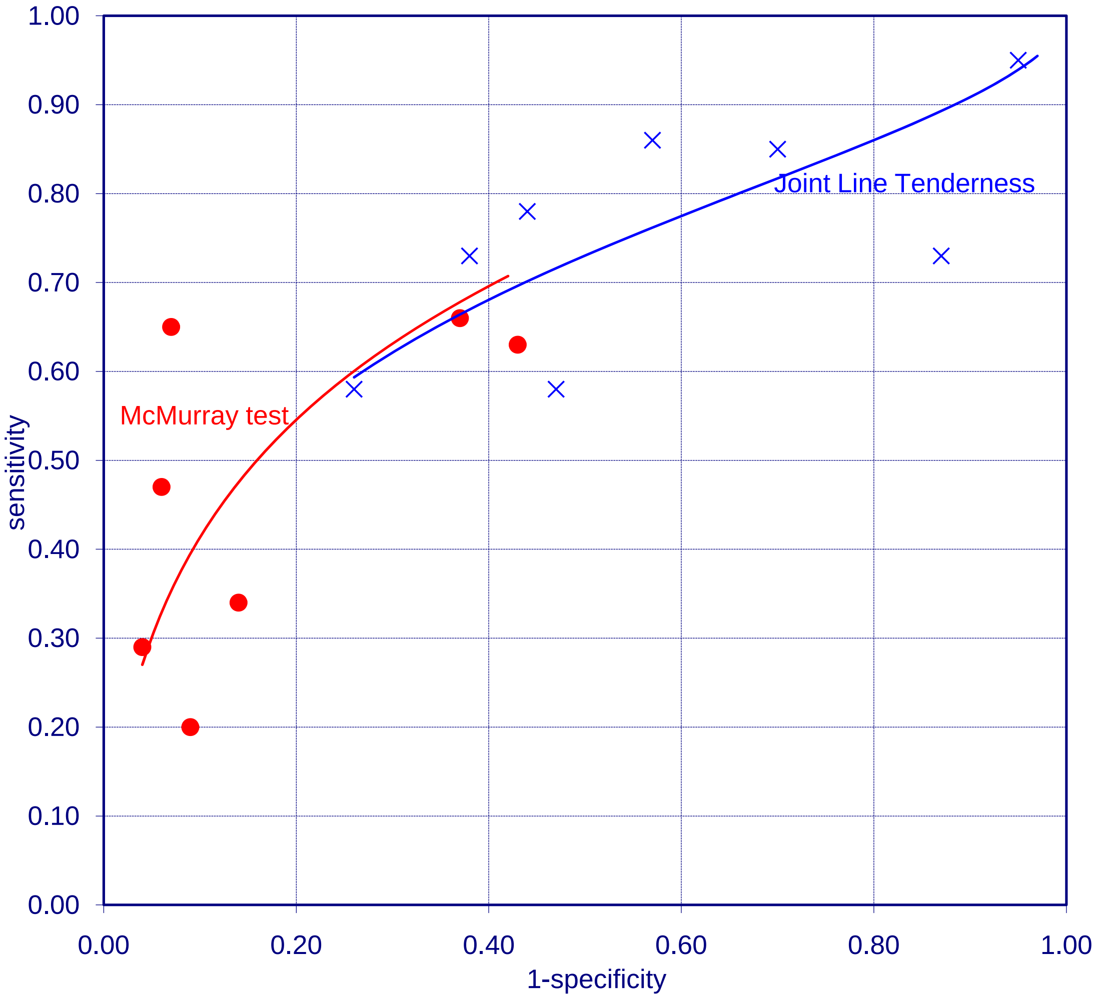
b. Curves cross

- Summary Sens and Spec B > A …
- but the curves cross
  - Interpretation will depend on place on curve
- Where would you be on the curve?



c. Curves in different areas

- In this case:
  - Sens B < A
  - Spec B > A
- Assess consequences of FN and FP
- What’s worse?



Replacement: results

- Direct vs indirect comparisons
- Location of sROC curves:
    - Test B more accurate than Test A
    - Curves cross
    - Curves in different areas

2. Triage

- New test positioned before the existing test pathway
- Purpose: to select patients for further testing (or not)
- Triage tests may be less accurate than existing tests
- They may have other advantages (like simplicity or low cost)

2. Triage

Requirements for triage test depend on purpose

- Triage test positive: further testing with very specific existing test to filter out FPs
- Triage test must be very sensitive to detect all diseased (low no. of FNs)

- Triage test negative: further testing with very sensitive existing test to filter out FNs
- Triage test must be very specific to detect all non-diseased (low no. of FPs)

3. Add-on

- New test positioned after the existing test pathway
- Purpose: to detect patients not identified by existing test(s)
- New test limited to subgroup of patients
- New test more accurate but otherwise less attractive than existing tests
  - Costs
  - Invasiveness
  - Etc.

3. Add-on

- Previous test(s) negative: add-on test
  - Add-on test to filter out all FNs of previous tests
  - Add-on test must be highly sensitive (low no. of FNs)

- Previous test(s) positive: add-on test
  - Add-on test to detect all FPs of previous tests
  - Add-on test must be highly specific (low no. of FPs)

Outline

- Ingredients of Discussion section
- Types of results of a DTA review
- Interpretation of results
- Presentation of results

- Small groups (if enough time)

Summary of Results Table

- Something new for Cochrane reviews
- Should show in one glance what the review is about, what the most important results were and what the conclusions are (including quality of evidence)
- GRADE Working Group in process of developing SoR template
- Input from authors more than welcome!

Summary of Results [Please note this table contains sample data]

What is the diagnostic accuracy of the Platelia© Aspergillus test for invasive aspergillosis?

Patients/population	Immunocompromized patients, mostly haematology patients
Prior testing	Varied, mostly physical examination and history (fever, neutropenia)
Settings	Mostly inpatients in hematology or cancer departments
Index test	Platelia© Aspergillus test, a sandwich ELISA for galactomannan
Importance	Non-invasive test needed to guide therapy; currently 70% of invasive aspergillosis patients die
Reference standard	Gold standard would have been autopsy, but this is nearly never done. Actual reference used: clinical and microbiological criteria

Test/ Subgroup	Pooled accuracy (95% CI)	No. of participants (studies)	Prevalence	Quality and Comments
All studies, All cut-offs	**Sensitivity 0.75** (0.66-0.84) **Specificity 0.91** (0.86-0.95)	2209 (19)	Median 8.1% (range 0.9-43.9%)	If studies show severe flaws, this can be reported here. E.g. 75% of the studies had partial verification.
Cut-off 0.5	**Sensitivity 0.85** (0.69-1.00) **Specificity 0.77** (0.50-1.00)	195 (2)	11% in one and 36% in another study	Only 2 studies
Cut-off 1.0	**Sensitivity 0.75** (0.62-0.89) **Specificity 0.89** (0.80-0.97)	1279 (7)	Median 13% (range 2.6-43.9%)	
Cut-off 1.5	**Sensitivity 0.72** (0.55-0.88) **Specificity 0.94** (0.89-0.98)	1116 (10)	Median 6.8% (range 0.9-20%)	

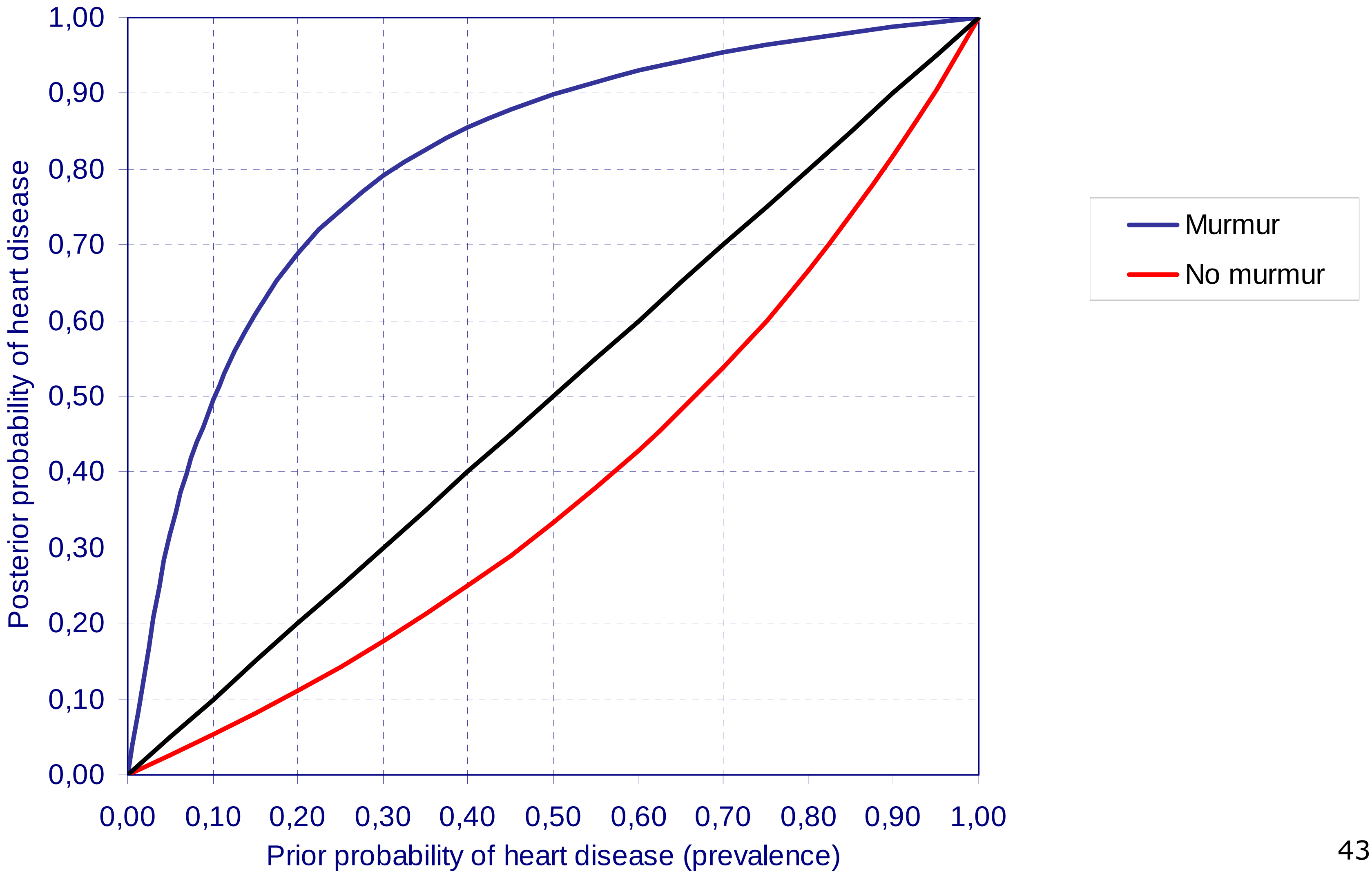
CAUTION:
The results on this table should not be interpreted in isolation from the results of the individual included studies contributing to each summary test accuracy measure. These are reported in the main body of the text of the review

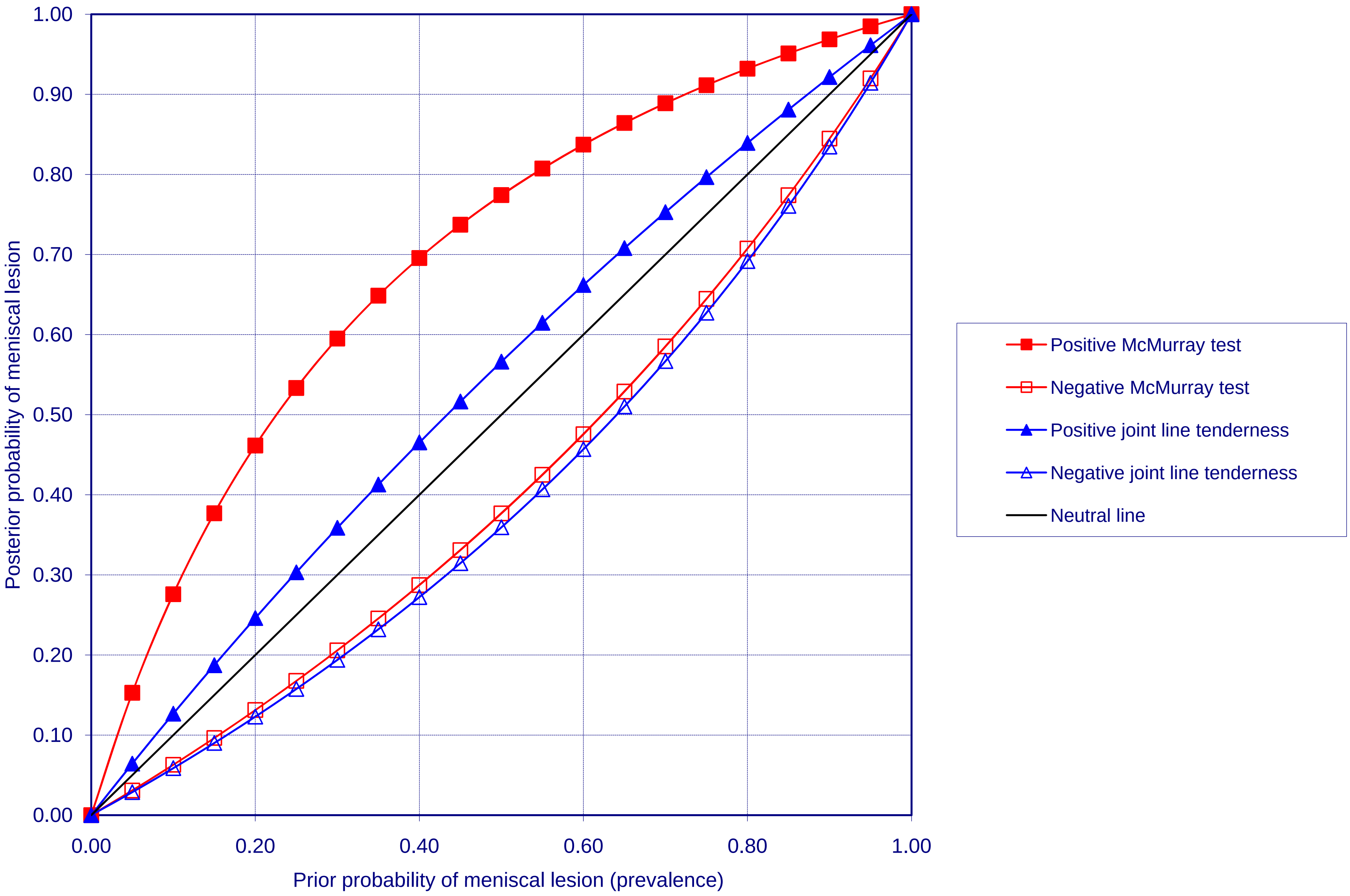
Pre-test post-test graph

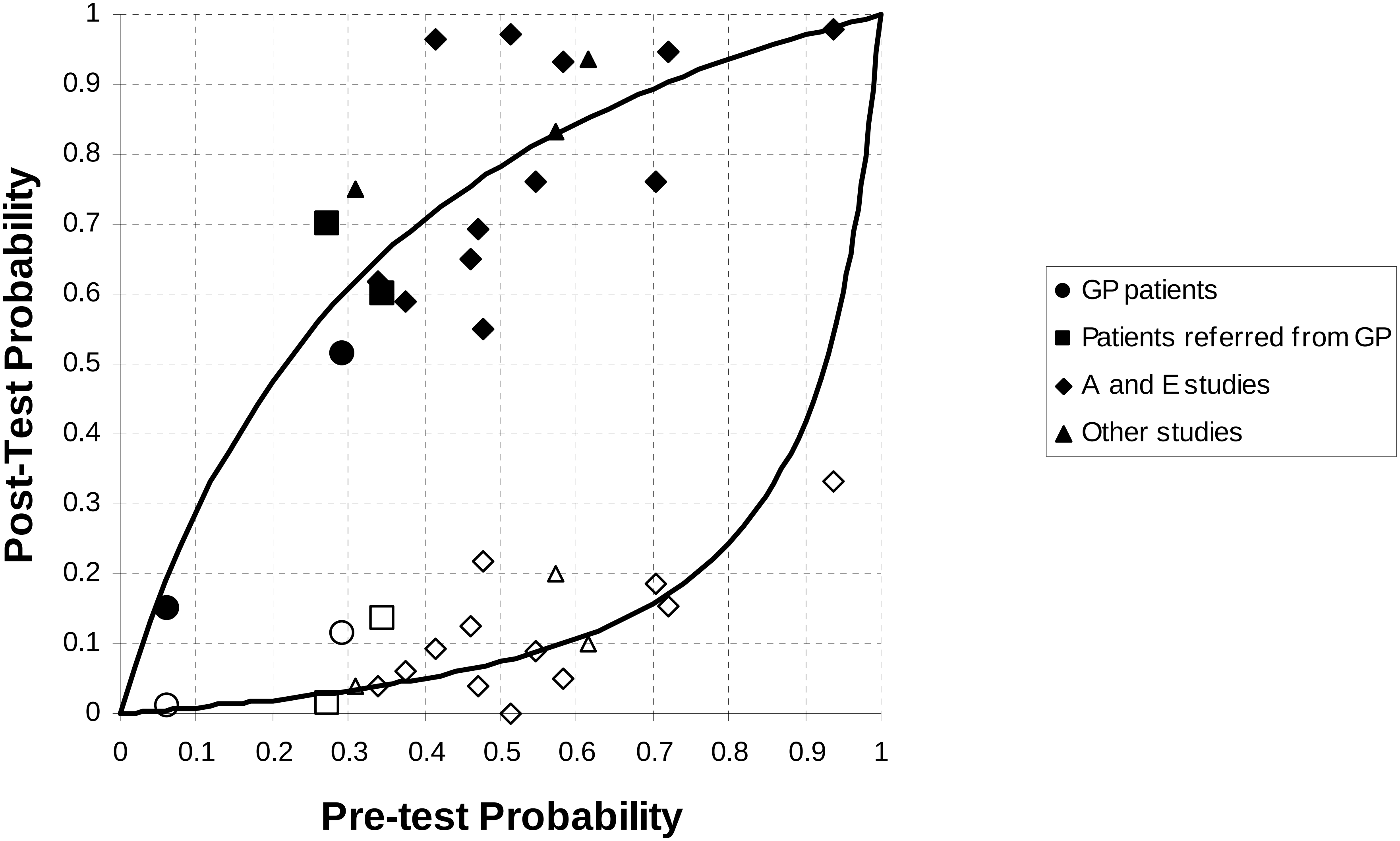
- X-axis: prevalence (pre-test probability)
- Y-axis: post-test probability
- Uses summary sens and spec of index test

- Visualises “effect” of test(s)

Pre-test post-test graph




Posterior probability of meniscal lesion
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.00
0.20
0.40
0.60
0.80
1.00
Prior probability of meniscal lesion (prevalence)
■ Positive McMurray test
□ Negative McMurray test
▲ Positive joint line tenderness
△ Negative joint line tenderness
Neutral line


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Post-Test Probability
Pre-test Probability
● GP patients
■ Patients referred from GP
◆ A and E studies
▲ Other studies

Small Group Exercise?

- Read the Introduction / Background
  - What is their question?
  - What would you expect from the results?
- Have a look at the results
  - Are these the results that you expect / want?
- Read the discussion and the conclusions
  - Do you agree with these conclusions?
  - Why / why not?
  - Do you miss some info in the discussion?
- **Extra:** *What info is needed to summarize this review in one A4?*

Glas et al. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. J Urol. 2003.