THE COCHRANE COLLABORATION®

Cochrane Diagnostic test accuracy reviews

Presenting and interpreting results

Chris Hyde	UK Support Unit
Rob Scholten	Continental Europe Support Unit

Presenting and Interpreting results

- Chapter 11 of the Handbook
  - Still under development
  - Need for input from future review-authors
- 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 on authors conclusions

Outline

- Types of results of a DTA review
- Interpretation of results
- Small groups
- Presentation of results / Summary of Results (SoR) Table(s)

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

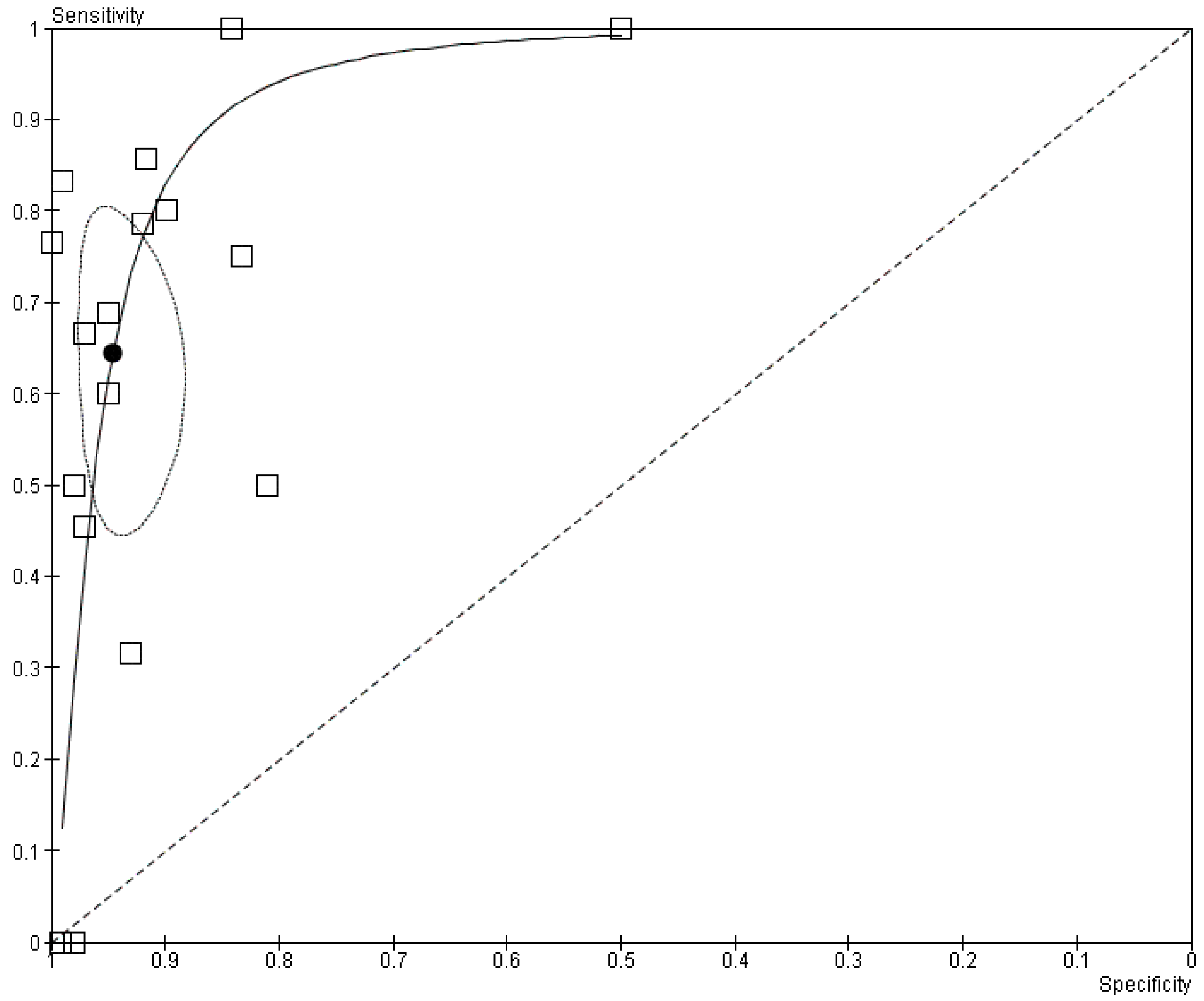
- clinically sensible
- not too much (statistical) heterogeneity
- no obvious threshold effect

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

Interpretation of summary sensitivity and specificity

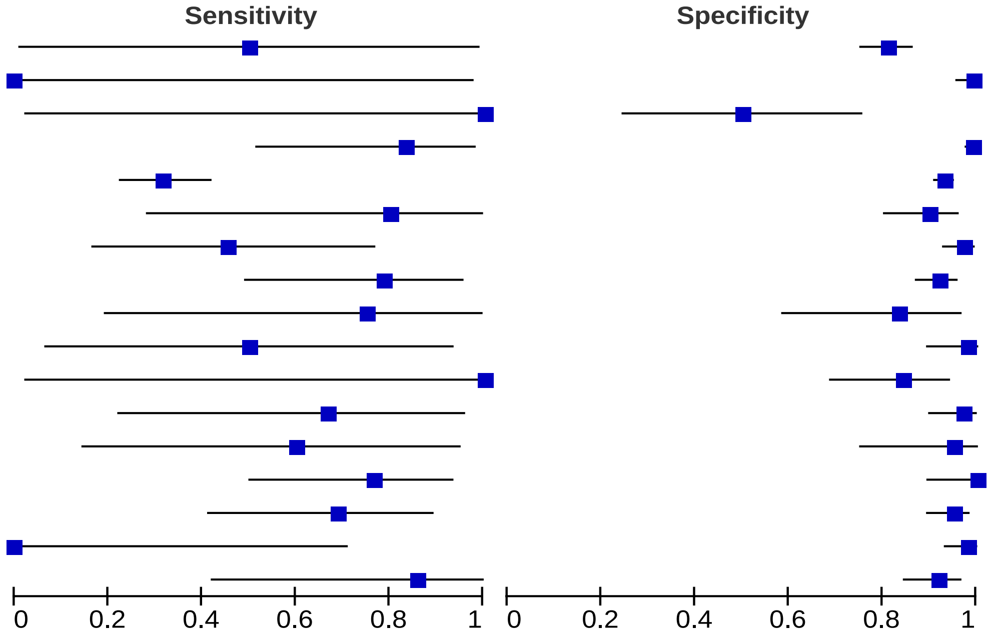
- Summary estimates are derived from random effects models
- Mean of a range of possible values for sens and spec (with a 95%-CE of the mean)
- Still many “real” values possible, including values outside the 95-CE range

Summary sensitivity and specificity



Apparent heterogeneity?

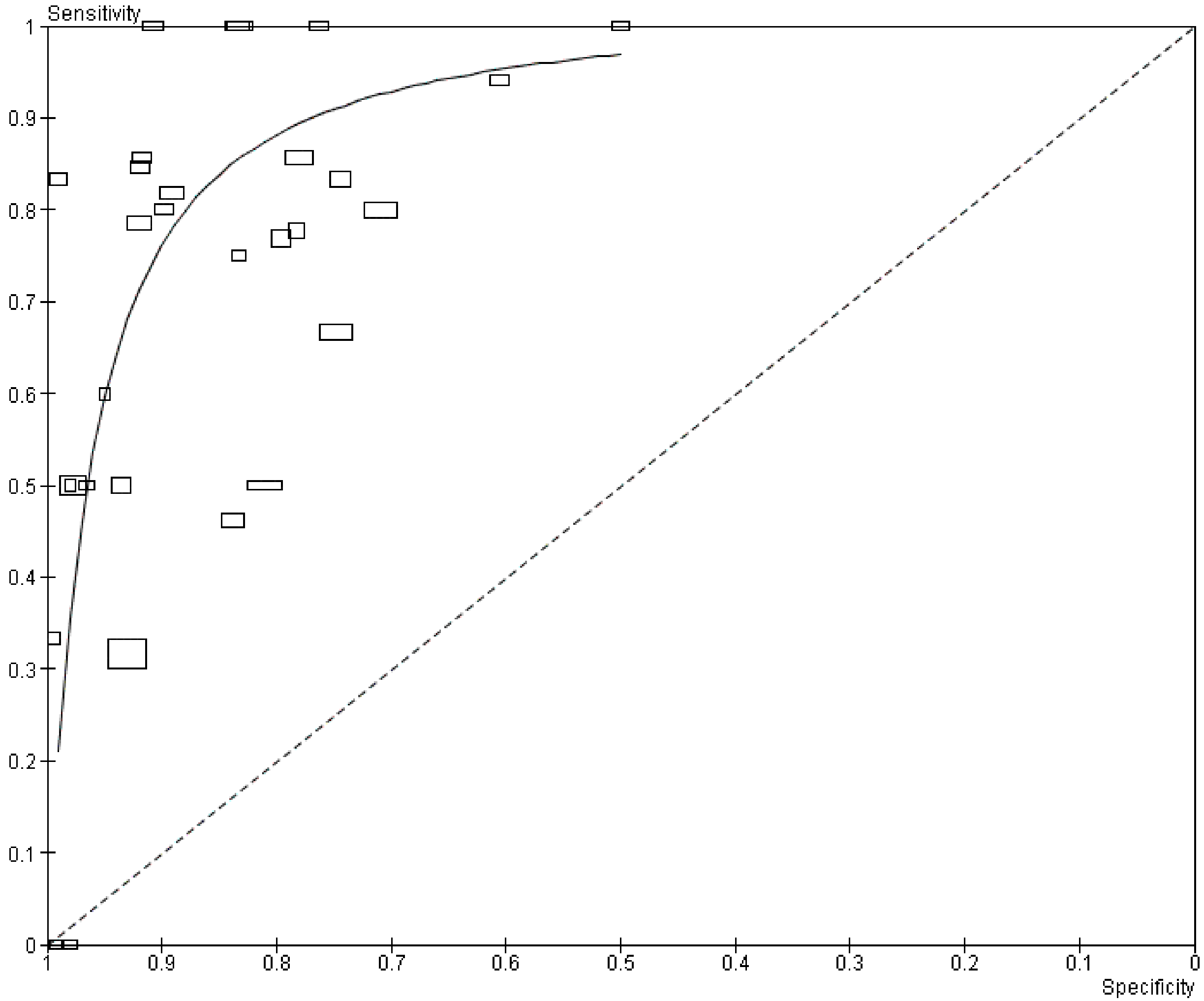
Study	TP	FP	FN	TN	Sensitivity	Specificity
Adam 2004	1	41	1	175	0.50 [0.01, 0.99]	0.81 [0.75, 0.86]
Allan 2005	0	1	1	123	0.00 [0.00, 0.97]	0.99 [0.96, 1.00]
Bialek 2002	1	8	0	8	1.00 [0.03, 1.00]	0.50 [0.25, 0.75]
Doermann 2002	10	4	2	407	0.83 [0.52, 0.98]	0.99 [0.98, 1.00]
Herbrecht 2002	31	49	67	650	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

- Threshold effect
  - Explicit (multiple cutoffs)
  - Implicit

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

- ○ Types of results of a DTA review
- ○ Interpretation of results
- ○ Presentation of results / Summary of Results (SoR) Table(s)

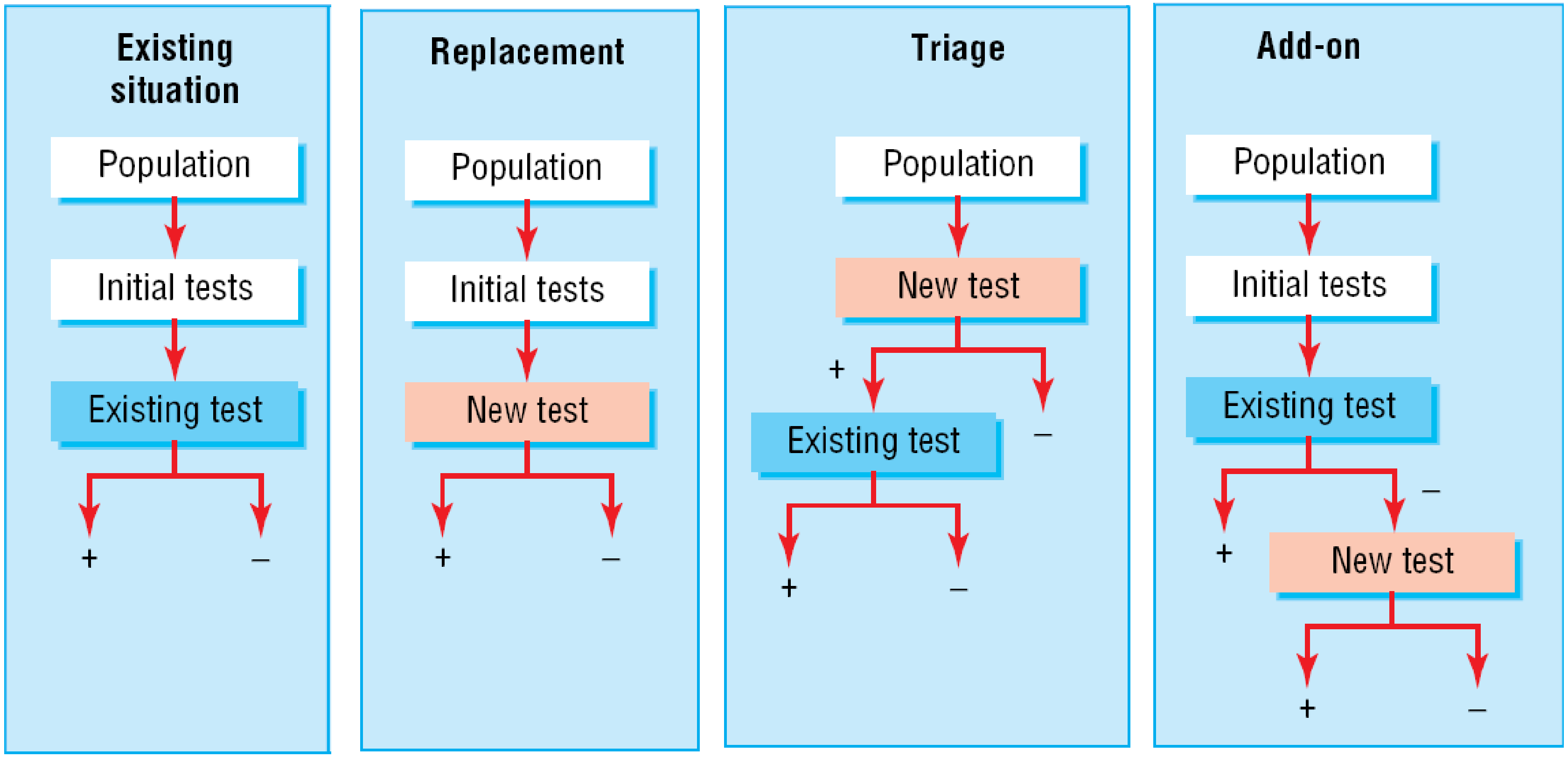
Purpose of test and test features

- ○ 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

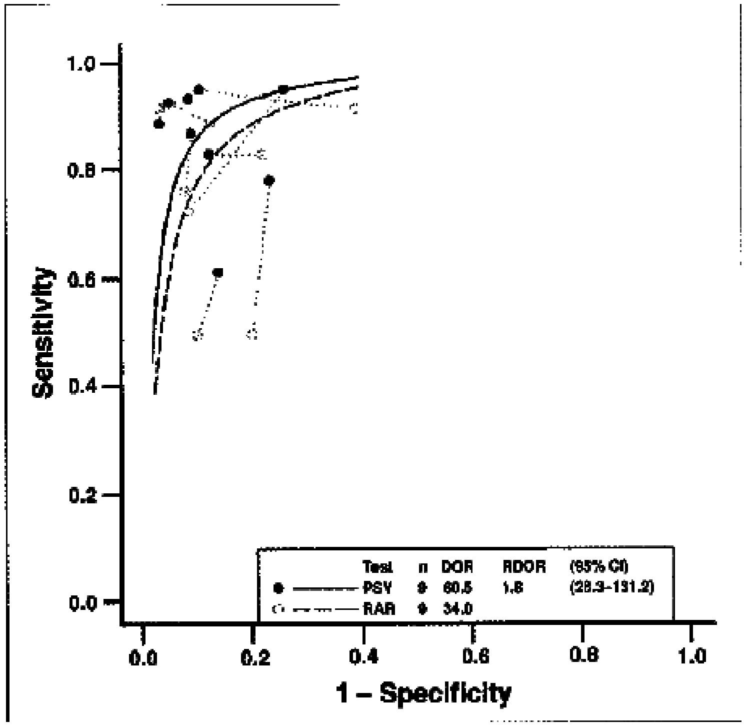
- more accurate
- less invasive, easier to do, less risky
- less uncomfortable for patients
- quicker to yield results
- technically less challenging
- more easily interpreted
- etc.

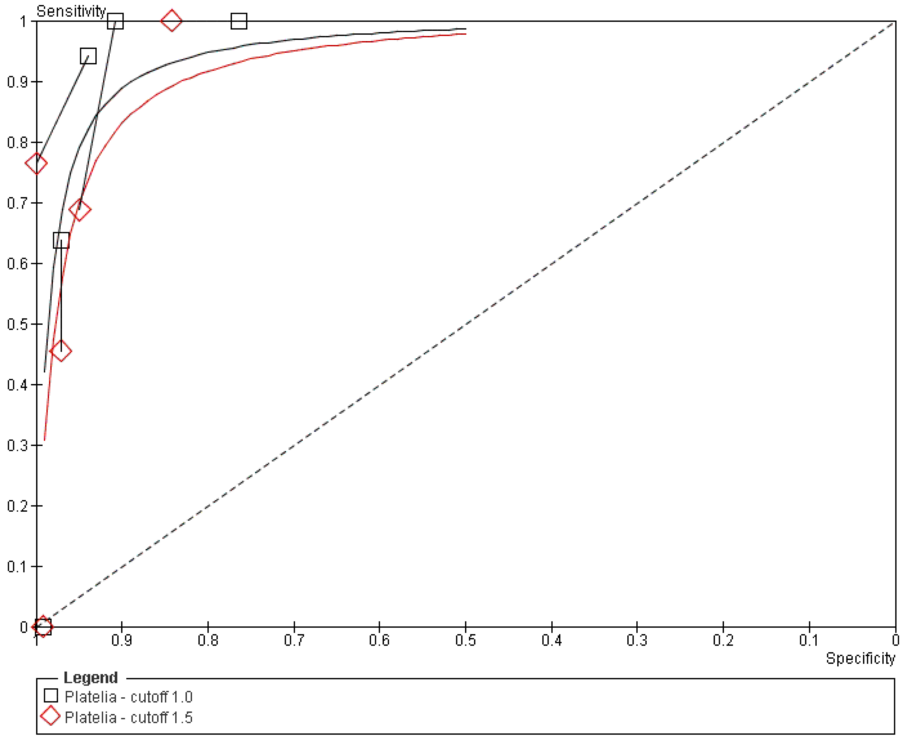
Replacement: preferred design

- 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

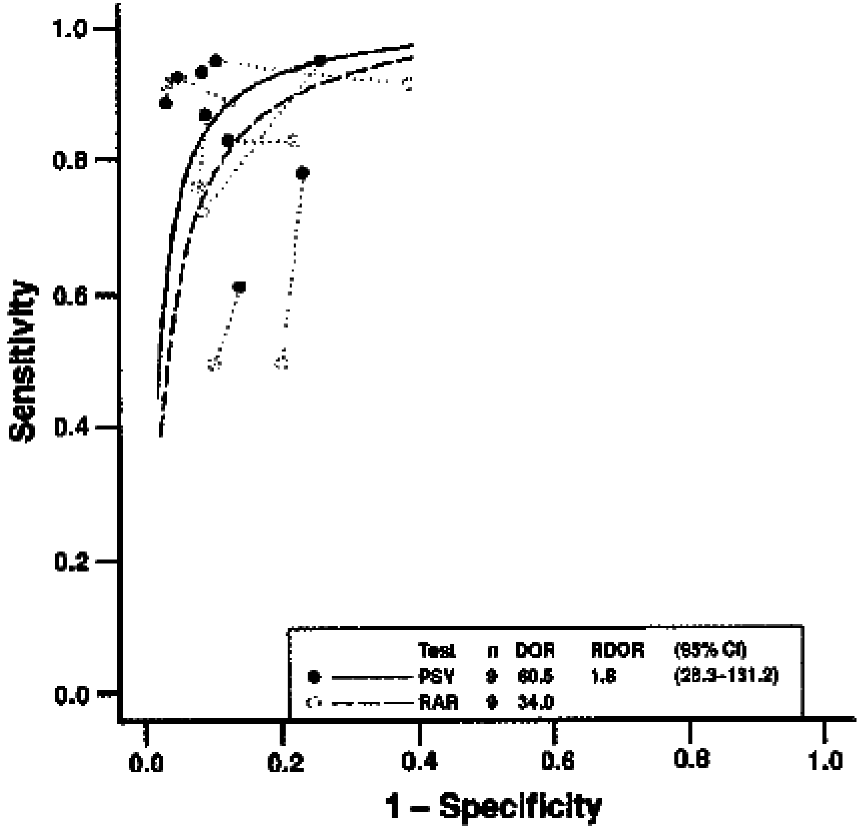
a. Curve B “Northwest” of curve A

b. Curves cross

c. Curves in different areas

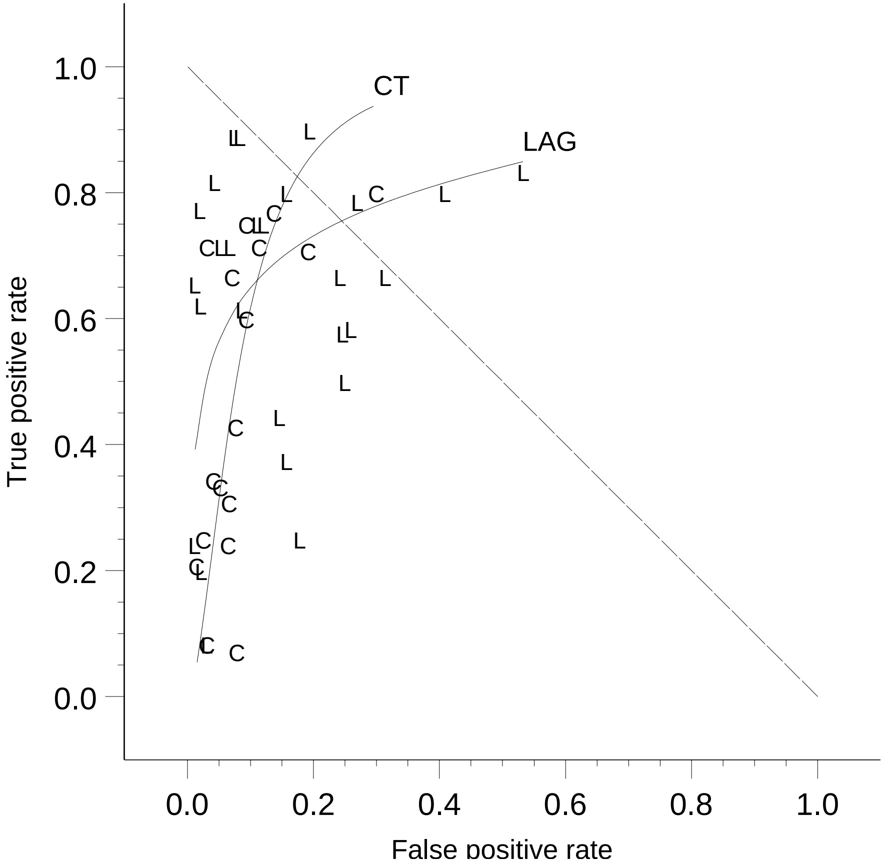
a. B more accurate than A

- ○ 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?

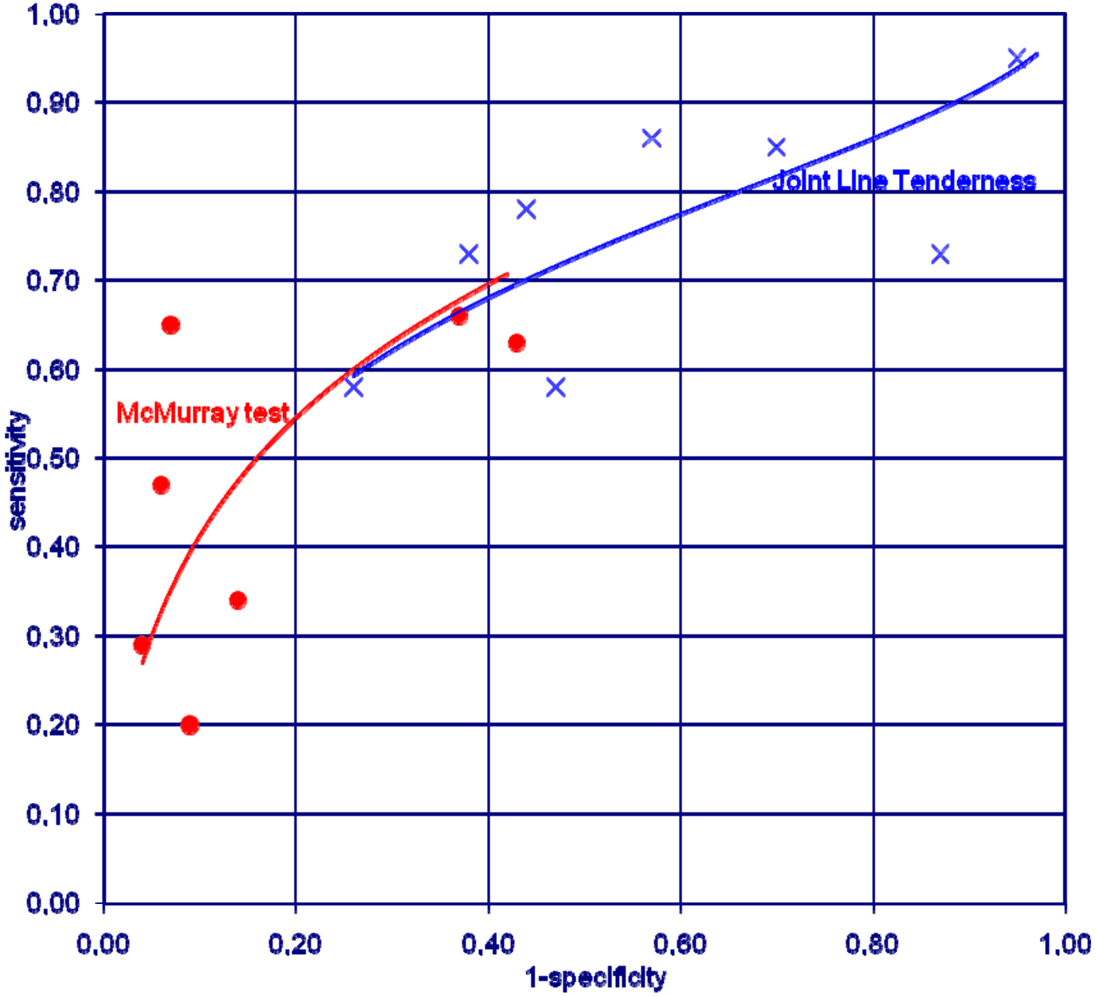


c. Curves in different areas

- 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

- ○ Types of results of a DTA review
- ○ Interpretation of results
- ○ Small groups
- ○ Presentation of results / Summary of Results (SoR) Table(s)

Small groups

1. Role of the index test
2. Requirements for the index test (e.g. high sens, small no. of false positives)?
3. What will happen with index test positives and negatives?
4. Consequences for TPs and TNs?
5. Consequences for FPs and FNs?
6. If sROC, where should the curve lie to meet the requirements of the index test?

Outline

- Types of results of a DTA review
- Interpretation of results
- Small groups
- Presentation of results / Summary of Results (SoR) Table(s)

Summary of Results Table

- Mandatory Table
- Analogous to Summary of Findings Table of Intervention reviews
- No standard format yet
- GRADE Working Group in process of developing SoR template
- Input from authors more than welcome!

SoR Table – Heading

- State review question (one Table for each main question)
- Report features of
    - Population
    - Prevalence
    - Setting
    - Index test (including cut-offs)
    - Reference test

SoR – Essential features (?)

- Summary sensitivity/specificity + 95% CI (and/or other accuracy metrics)
- Consistency of results between studies
- Number of studies/participants
- Average prevalence of target condition (range)
- Overall study quality
- Notes, including other limitations

SoR – GRADE Working Group

- Heading (like before)
- Overall quality rating (limitations)
- Directness
- Inconsistency
- Imprecision
- Summary Sens and Spec (+ 95%-CI)
- Consequences of TP, FP, TN, FN