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

Meta-analysis of diagnostic accuracy studies

Mariska Leeflang
(with thanks to Yemisi Takwoingi, Jon Deeks and Hans Reitsma)

Diagnostic Test Accuracy Reviews

1. Framing the question
2. Identification and selection of studies
3. Quality assessment
4. Data extraction
5. Data analysis
6. Interpretation of the results

Ultimate goal of meta-analysis

Robust conclusions with respect to the research question(s)

Meta-Analysis

1. Calculation of an overall summary (average) of high precision, coherent with all observed data
2. Typically a “weighted average” is used where more informative (larger) studies have more say
3. Assess the degree to which the study results deviate from the overall summary
4. Investigate possible explanations for the deviations

The (meta-)analytic process

1. What analyses did you plan?
   a. Primary objective
   b. Subgroups, sensitivity analyses, etc.

2. What are the data at hand?
   a. Forest plots
   b. Raw ROC plots
   c. Variation in predefined covariates?

3. Is meta-analysis appropriate?
   a. Sufficient clinical/methodological homogeneity
   b. Enough studies per review question

4. Meta-analysis

Summary of which values?

Sensitivity

Specificity

Positive Predictive Value

Negative Predictive Value

Positive Likelihood Ratio

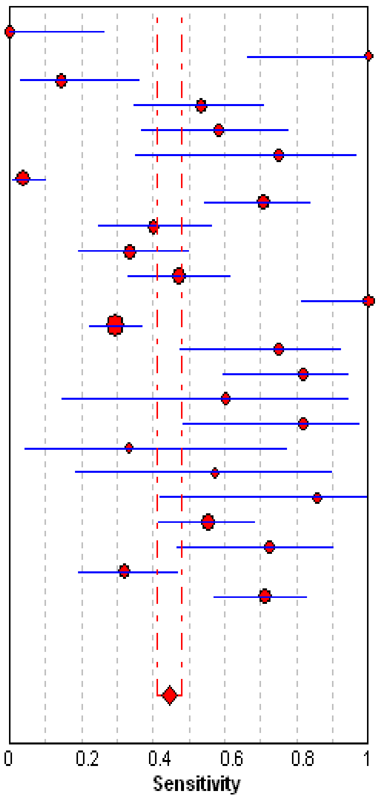
Negative Likelihood Ratio

Diagnostic Odds ratio

ROC curves

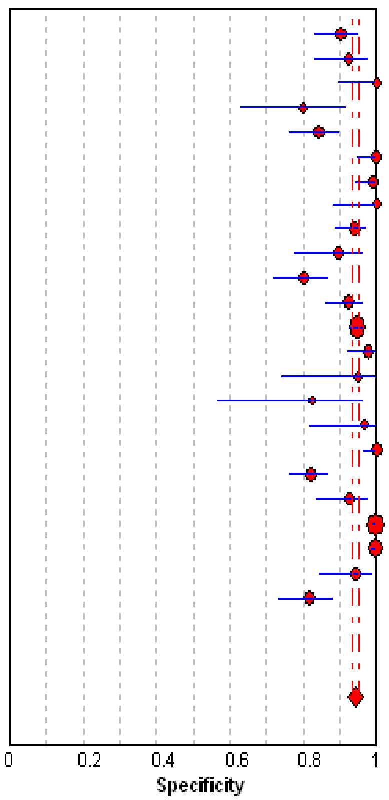
		Disease (Ref. test)		
		Pres.	Abs.	
Index - Test	+	TP	FP	
	-	FN	TN	

Pooling sensitivity and specificity?



	Sensitivity (95% CI)	
Allan	0.00	(0.00 - 0.26)
Busca	1.00	(0.66 - 1.00)
Moragues	0.14	(0.03 - 0.36)
Marr	0.53	(0.35 - 0.71)
Kawazu	0.58	(0.37 - 0.78)
Rovira	0.75	(0.35 - 0.97)
Buchheidt	0.04	(0.01 0.10)
Challier	0.71	(0.54 - 0.84)
Lai	0.40	(0.25 - 0.57)
Becker	0.33	(0.19 - 0.50)
Florent	0.47	(0.33 - 0.62)
Ulusakarya	1.00	(0.81 - 1.00)
Herbrecht	0.29	(0.22 - 0.37)
Williamson	0.75	(0.48 - 0.93)
Bretagne	0.82	(0.60 - 0.95)
Machetti	0.60	(0.15 - 0.95)
Pereira	0.82	(0.48 - 0.98)
White	0.33	(0.04 - 0.78)
Adam	0.57	(0.18 - 0.90)
Kallel	0.86	(0.42 - 1.00)
Pinel	0.55	(0.41 - 0.69)
Doermann	0.72	(0.47 - 0.90)
Maertens	0.32	(0.19 - 0.47)
Weisser	0.71	(0.57 - 0.83)

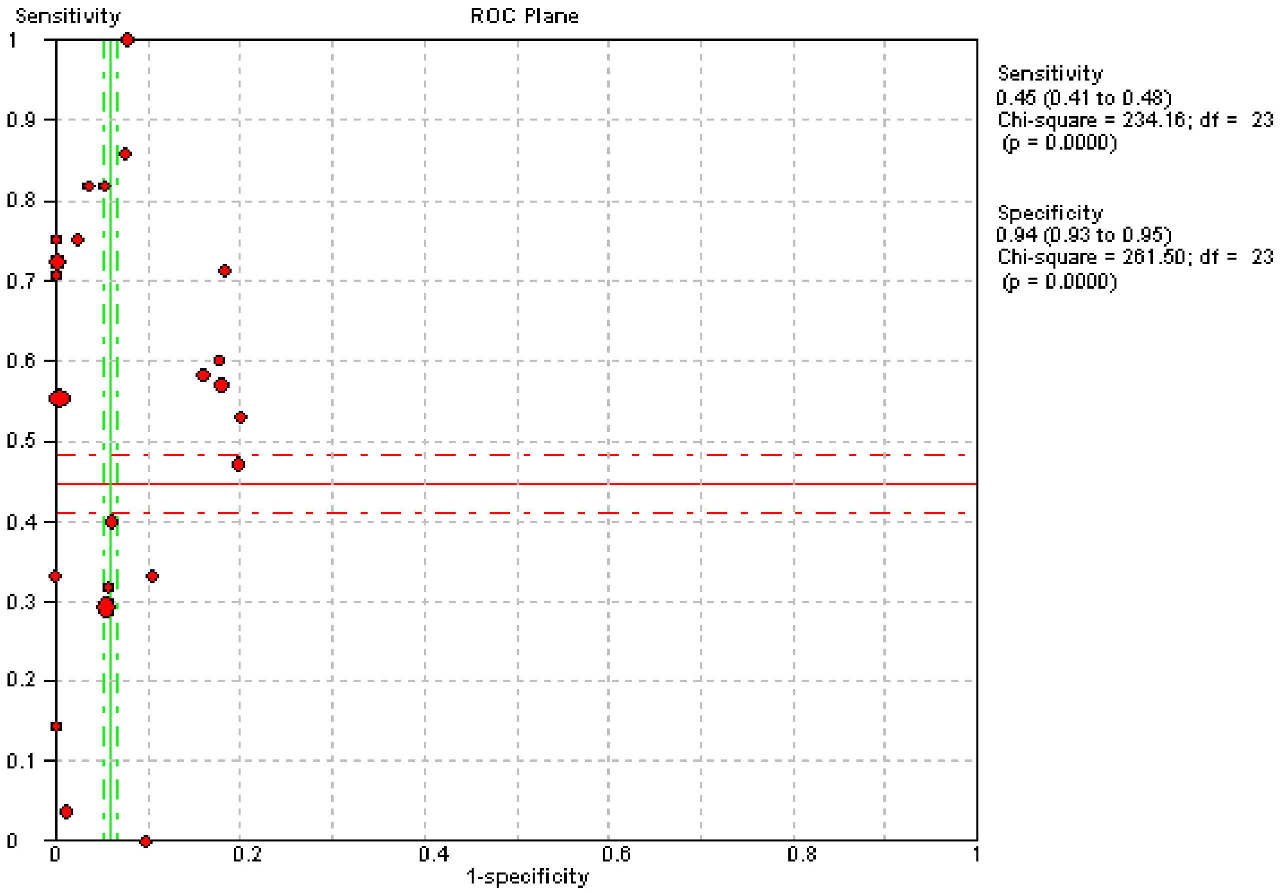
Pooled Sensitivity = 0.45 (0.41 to 0.48)
Chi-square = 234.16; df = 23 (p = 0.0000)
Inconsistency (I-square) = 90.2 %



	Specificity (95% CI)	
Allan	0.90	(0.83 - 0.95)
Busca	0.92	(0.83 - 0.97)
Moragues	1.00	(0.89 - 1.00)
Marr	0.80	(0.63 - 0.92)
Kawazu	0.84	(0.76 - 0.90)
Rovira	1.00	(0.95 - 1.00)
Buchheidt	0.99	(0.94 - 1.00)
Challier	1.00	(0.88 - 1.00)
Lai	0.94	(0.89 - 0.97)
Becker	0.90	(0.77 - 0.97)
Florent	0.80	(0.72 - 0.87)
Ulusakarya	0.92	(0.86 - 0.96)
Herbrecht	0.95	(0.93 - 0.96)
Williamson	0.98	(0.92 - 1.00)
Bretagne	0.95	(0.74 - 1.00)
Machetti	0.82	(0.57 - 0.96)
Pereira	0.96	(0.82 - 1.00)
White	1.00	(0.96 - 1.00)
Adam	0.82	(0.76 - 0.87)
Kallel	0.93	(0.83 - 0.98)
Pinel	1.00	(0.99 - 1.00)
Doermann	1.00	(0.99 - 1.00)
Maertens	0.94	(0.84 - 0.99)
Weisser	0.82	(0.73 - 0.88)

Pooled Specificity = 0.94 (0.93 to 0.95)
Chi-square = 261.50; df = 23 (p = 0.0000)
Inconsistency (I-square) = 91.2 %

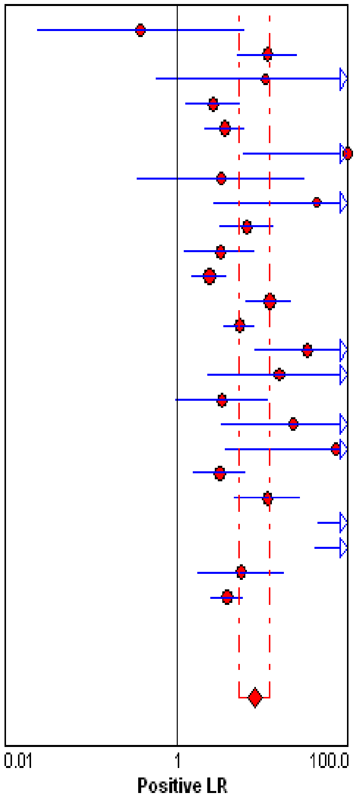
Pooling sensitivity and specificity?



Sensitivity
0.45 (0.41 to 0.48)
Chi-square = 234.16; df = 23
(p = 0.0000)

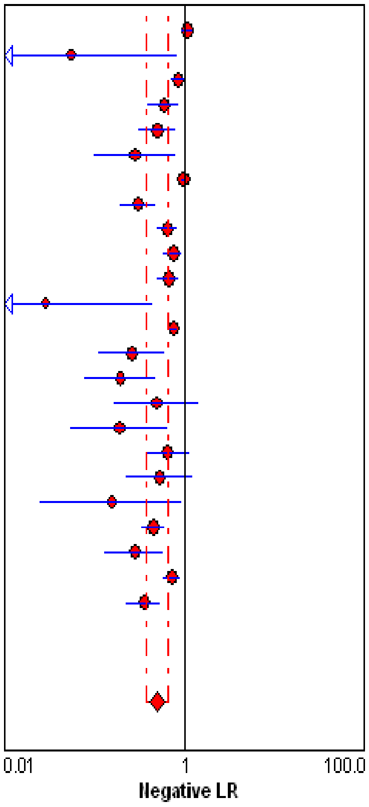
Specificity
0.94 (0.93 to 0.95)
Chi-square = 261.50; df = 23
(p = 0.0000)

Pooling Likelihood Ratios?



	Positive LR (95% CI)
Allan	0.38 (0.02 - 6.10)
Busca	11.40 (5.06 - 25.70)
Moragues	10.82 (0.59 - 199.44)
Marr	2.66 (1.27 - 5.56)
Kawazu	3.65 (2.16 - 6.16)
Rovira	96.78 (5.94 - 1,577.25)
Buchheidt	3.32 (0.35 - 31.32)
Challier	42.14 (2.68 - 662.97)
Lai	6.62 (3.16 - 13.86)
Becker	3.20 (1.25 - 8.20)
Florent	2.37 (1.49 - 3.79)
Ulusakarya	12.09 (6.54 - 22.35)
Herbrecht	5.41 (3.61 - 8.11)
Williamson	33.00 (8.15 - 133.68)
Bretagne	15.55 (2.28 - 105.81)
Machetti	3.40 (0.97 - 11.89)
Pereira	22.91 (3.28 - 160.17)
White	70.71 (3.75 - 1,334.98)
Adam	3.17 (1.57 - 6.41)
Kallel	11.49 (4.69 - 28.13)
Pinel	138.58 (43.72 - 439.21)
Doermann	292.50 (40.45 - 2,115.12)
Maertens	5.64 (1.74 - 18.27)
Weisser	3.88 (2.52 - 5.97)

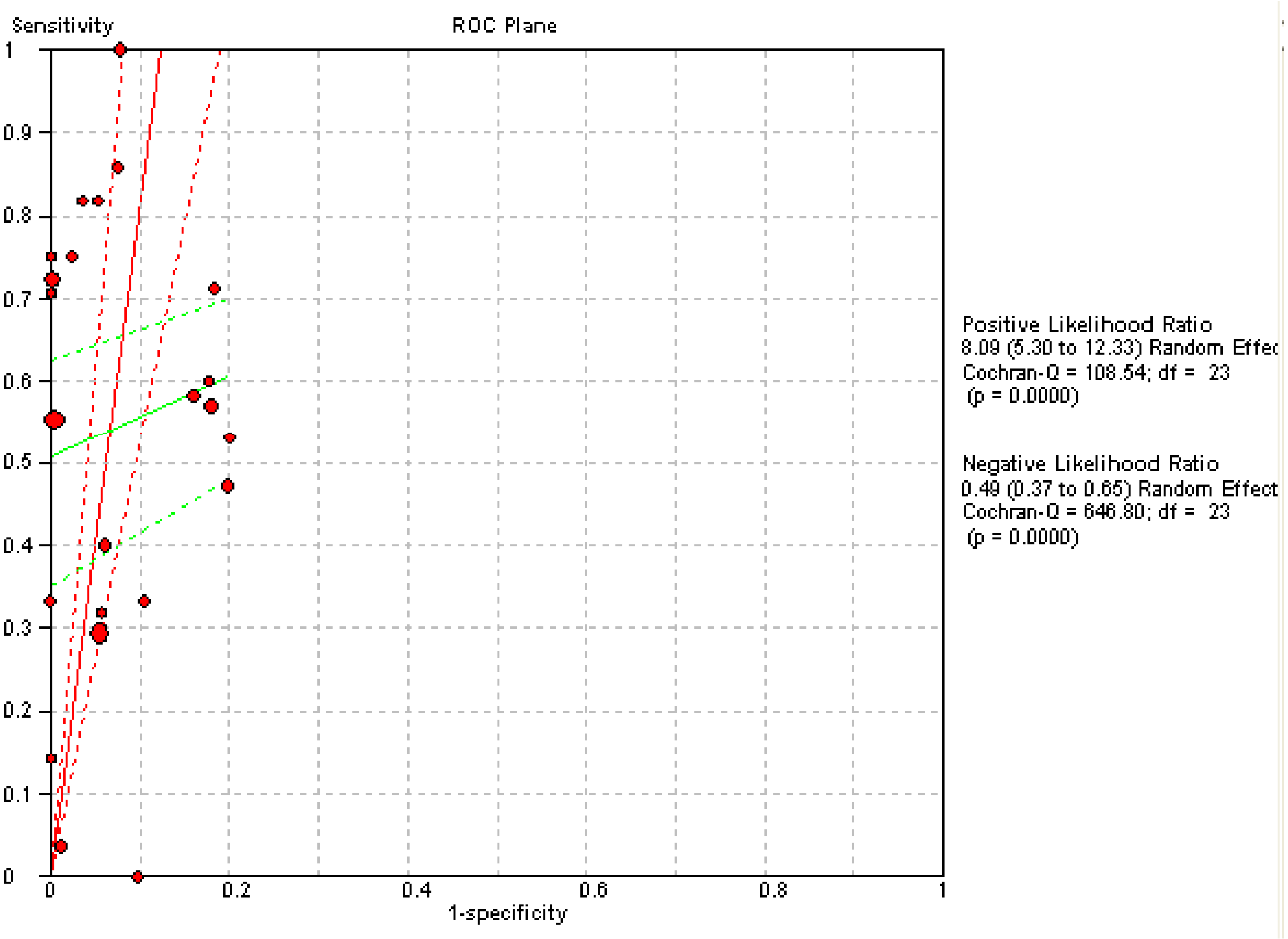
Random Effects Model
Pooled Positive LR = 8.09 (5.30 to 12.33)
Cochran-Q = 108.54; df = 23 (p = 0.0000)
Inconsistency (I-square) = 78.8 %
Tau-squared = 0.6818



	Negative LR (95% CI)
Allan	1.07 (0.94 - 1.21)
Busca	0.05 (0.00 - 0.81)
Moragues	0.85 (0.71 - 1.03)
Marr	0.59 (0.39 - 0.88)
Kawazu	0.50 (0.31 - 0.80)
Rovira	0.28 (0.10 - 0.80)
Buchheidt	0.97 (0.93 - 1.02)
Challier	0.30 (0.19 - 0.48)
Lai	0.64 (0.49 - 0.83)
Becker	0.74 (0.58 - 0.95)
Florent	0.66 (0.50 - 0.87)
Ulusakarya	0.03 (0.00 - 0.44)
Herbrecht	0.75 (0.67 - 0.83)
Williamson	0.26 (0.11 - 0.60)
Bretagne	0.19 (0.08 - 0.47)
Machetti	0.49 (0.16 - 1.45)
Pereira	0.19 (0.05 - 0.66)
White	0.65 (0.37 - 1.12)
Adam	0.52 (0.22 - 1.23)
Kallel	0.15 (0.03 - 0.95)
Pinel	0.45 (0.33 - 0.60)
Doermann	0.28 (0.13 - 0.59)
Maertens	0.72 (0.59 - 0.89)
Weisser	0.35 (0.23 - 0.55)

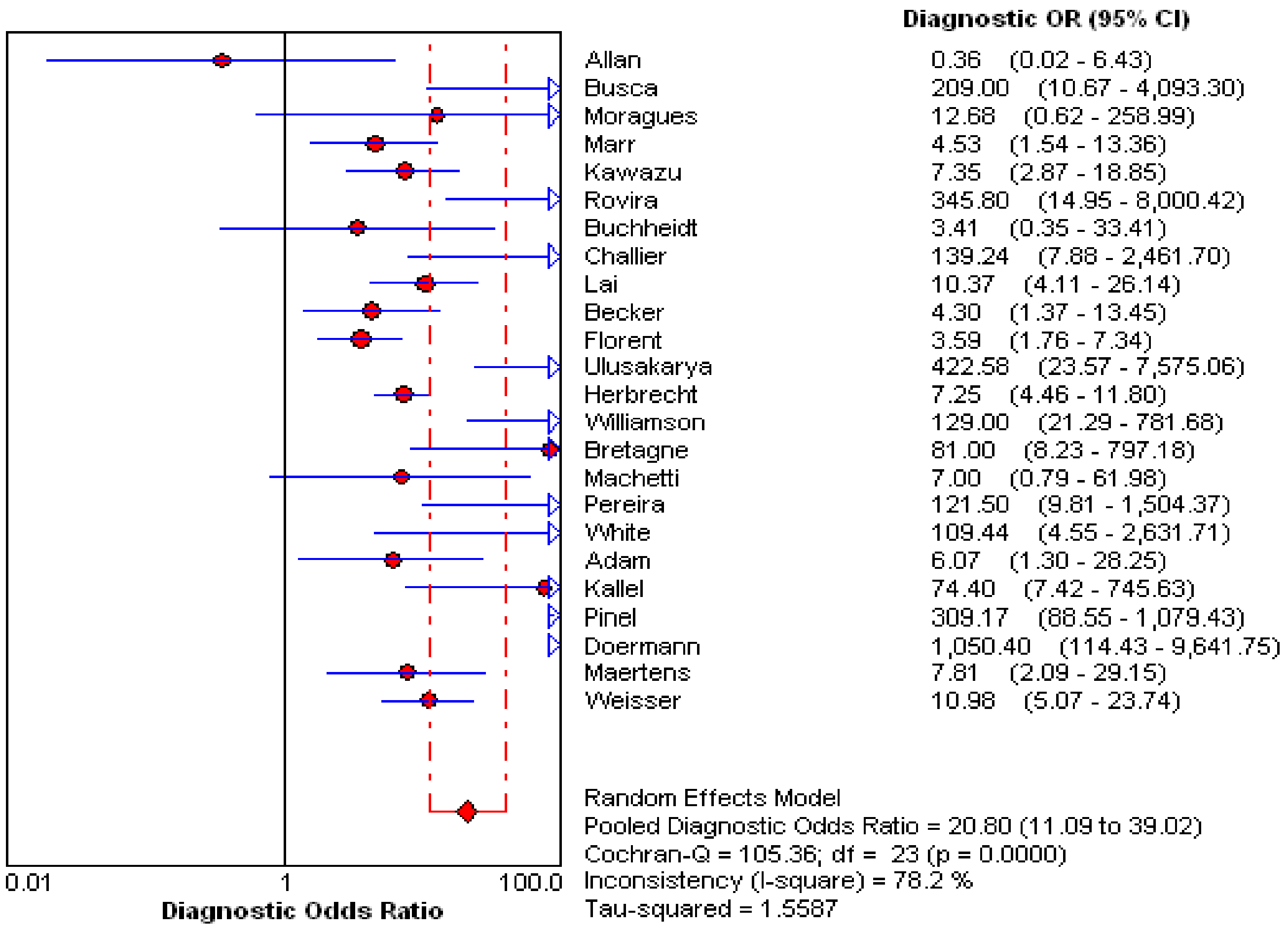
Random Effects Model
Pooled Negative LR = 0.49 (0.37 to 0.65)
Cochran-Q = 646.80; df = 23 (p = 0.0000)
Inconsistency (I-square) = 96.4 %
Tau-squared = 0.3522

Pooling LRs?



```
Positive Likelihood Ratio
8.09 (5.30 to 12.33) Random Effec
Cochran-Q = 108.54; df =  23
 (p = 0.0000)

Negative Likelihood Ratio
0.49 (0.37 to 0.65) Random Effect
Cochran-Q = 646.80; df =  23
 (p = 0.0000)
```

Pooling odds ratios?



	Diagnostic OR (95% CI)
Allan	0.36 (0.02 - 6.43)
Busca	209.00 (10.67 - 4,093.30)
Moragues	12.68 (0.62 - 258.99)
Marr	4.53 (1.54 - 13.36)
Kawazu	7.35 (2.87 - 18.85)
Rovira	345.80 (14.95 - 8,000.42)
Buchheidt	3.41 (0.35 - 33.41)
Challier	139.24 (7.88 - 2,461.70)
Lai	10.37 (4.11 - 26.14)
Becker	4.30 (1.37 - 13.45)
Florent	3.59 (1.76 - 7.34)
Ulusakarya	422.58 (23.57 - 7,575.06)
Herbrecht	7.25 (4.46 - 11.80)
Williamson	129.00 (21.29 - 781.68)
Bretagne	81.00 (8.23 - 797.18)
Machetti	7.00 (0.79 - 61.98)
Pereira	121.50 (9.81 - 1,504.37)
White	109.44 (4.55 - 2,631.71)
Adam	6.07 (1.30 - 28.25)
Kallel	74.40 (7.42 - 745.63)
Pinel	309.17 (88.55 - 1,079.43)
Doermann	1,050.40 (114.43 - 9,641.75)
Maertens	7.81 (2.09 - 29.15)
Weisser	10.98 (5.07 - 23.74)

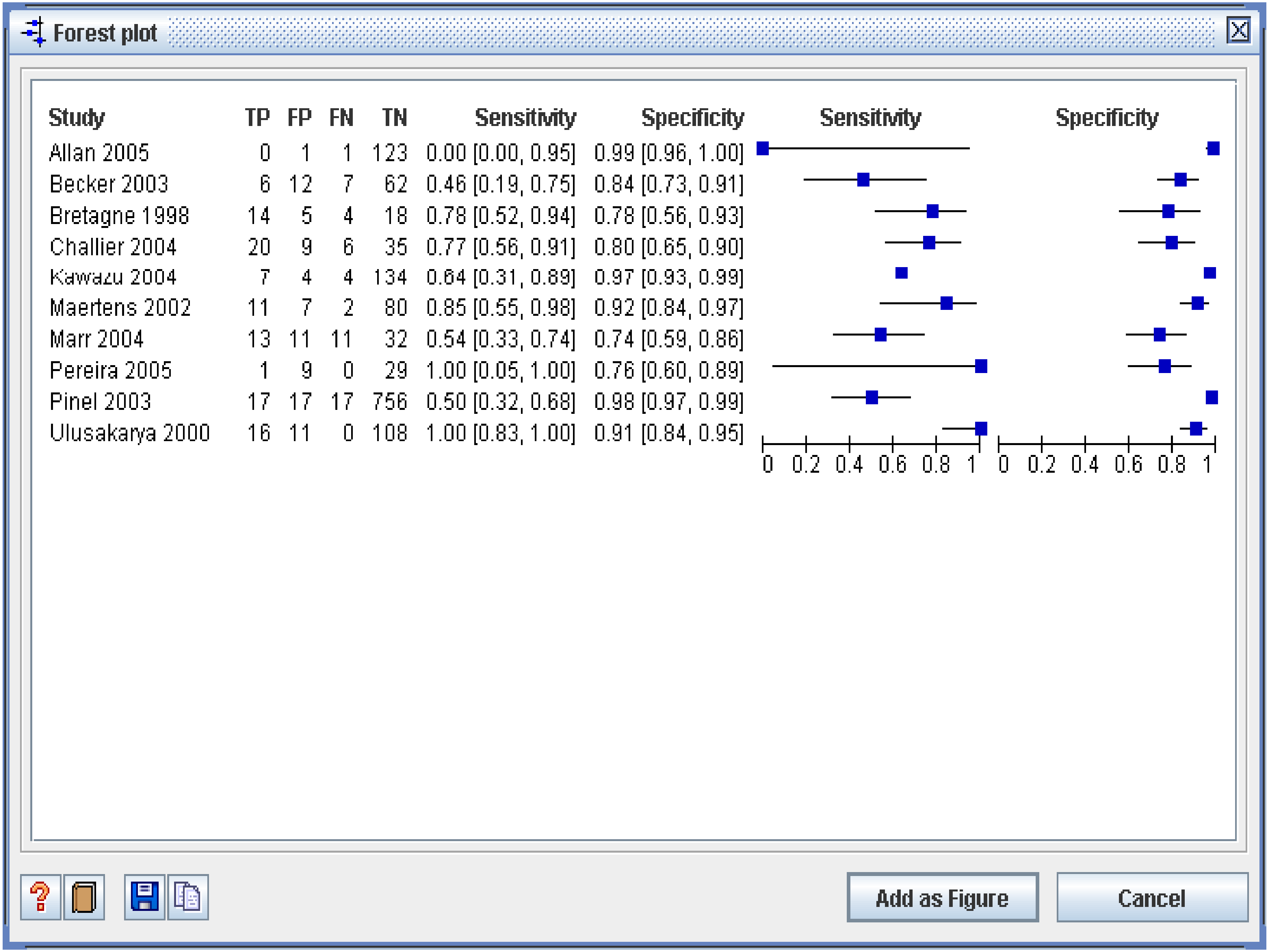
Random Effects Model
Pooled Diagnostic Odds Ratio = 20.80 (11.09 to 39.02)
Cochran-Q = 105.36; df = 23 (p = 0.0000)
Inconsistency (I-square) = 78.2 %
Tau-squared = 1.5587

Let’s focus on sensitivity and specificity

- Predictive values are directly depending on prevalence
- Pooling likelihood ratios may lead to misleading / impossible results
- Pooling odds ratios may be okay, but are difficult to interpret.
- From the pooled sensitivity and specificity, it is still possible to calculate LRs and PVs.

Descriptive Analysis

- Forest plots
  - point estimate with 95% CI
  - paired: sensitivity and specificity side-by side



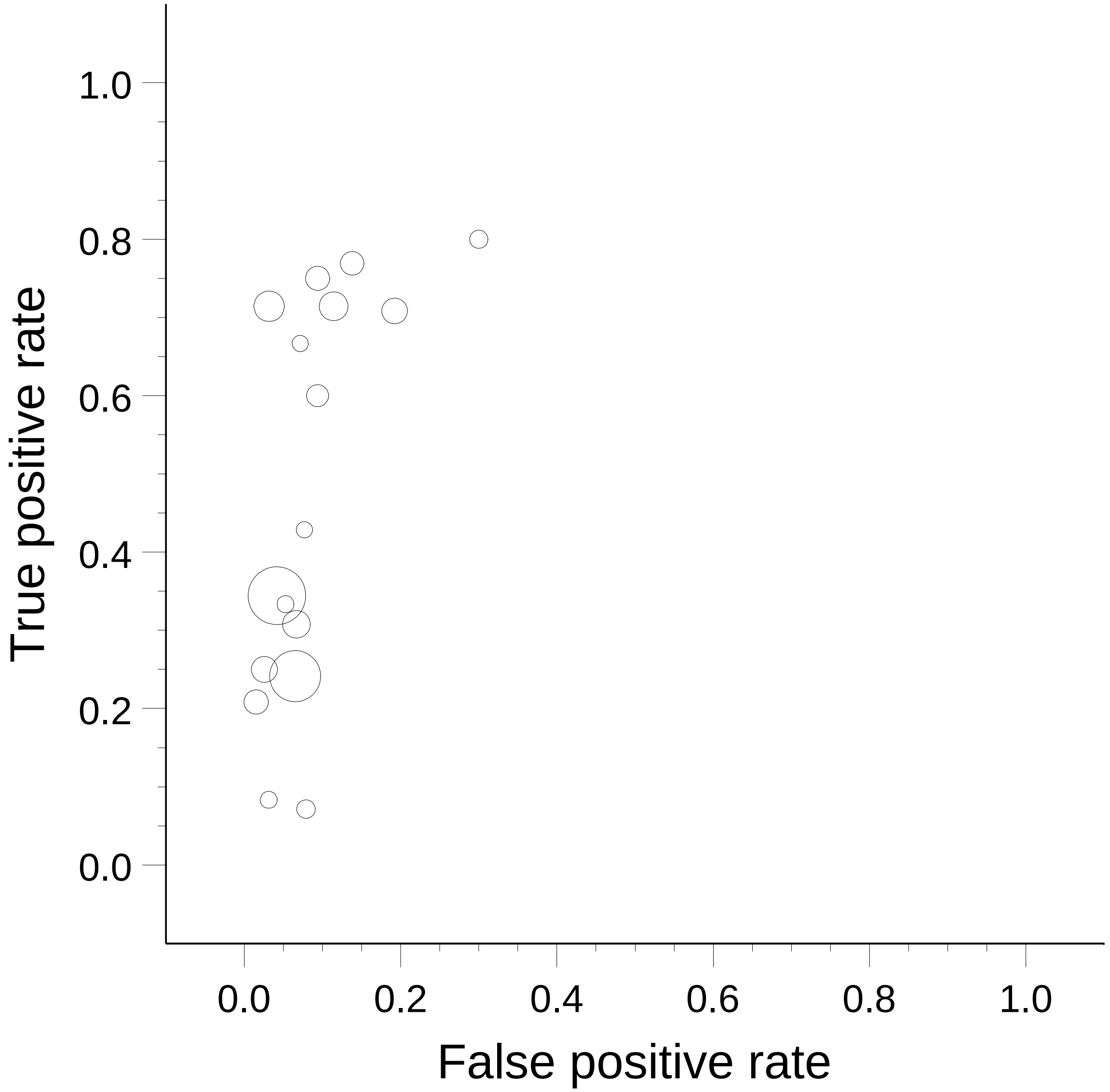
Study	TP	FP	FN	TN	Sensitivity	Specificity
Allan 2005	0	1	1	123	0.00 [0.00, 0.95]	0.99 [0.96, 1.00]
Becker 2003	6	12	7	62	0.46 [0.19, 0.75]	0.84 [0.73, 0.91]
Bretagne 1998	14	5	4	18	0.78 [0.52, 0.94]	0.78 [0.56, 0.93]
Challier 2004	20	9	6	35	0.77 [0.56, 0.91]	0.80 [0.65, 0.90]
Kawazu 2004	7	4	4	134	0.64 [0.31, 0.89]	0.97 [0.93, 0.99]
Maertens 2002	11	7	2	80	0.85 [0.55, 0.98]	0.92 [0.84, 0.97]
Marr 2004	13	11	11	32	0.54 [0.33, 0.74]	0.74 [0.59, 0.86]
Pereira 2005	1	9	0	29	1.00 [0.05, 1.00]	0.76 [0.60, 0.89]
Pinel 2003	17	17	17	756	0.50 [0.32, 0.68]	0.98 [0.97, 0.99]
Ulusakarya 2000	16	11	0	108	1.00 [0.83, 1.00]	0.91 [0.84, 0.95]

Descriptive Analysis

- Forest plots
  - point estimate with 95% CI
  - paired: sensitivity and specificity side-by side

- ROC plot
  - pairs of sensitivity & specificity in ROC space
  - bubble plot to show differences in precision

Plot in ROC Space

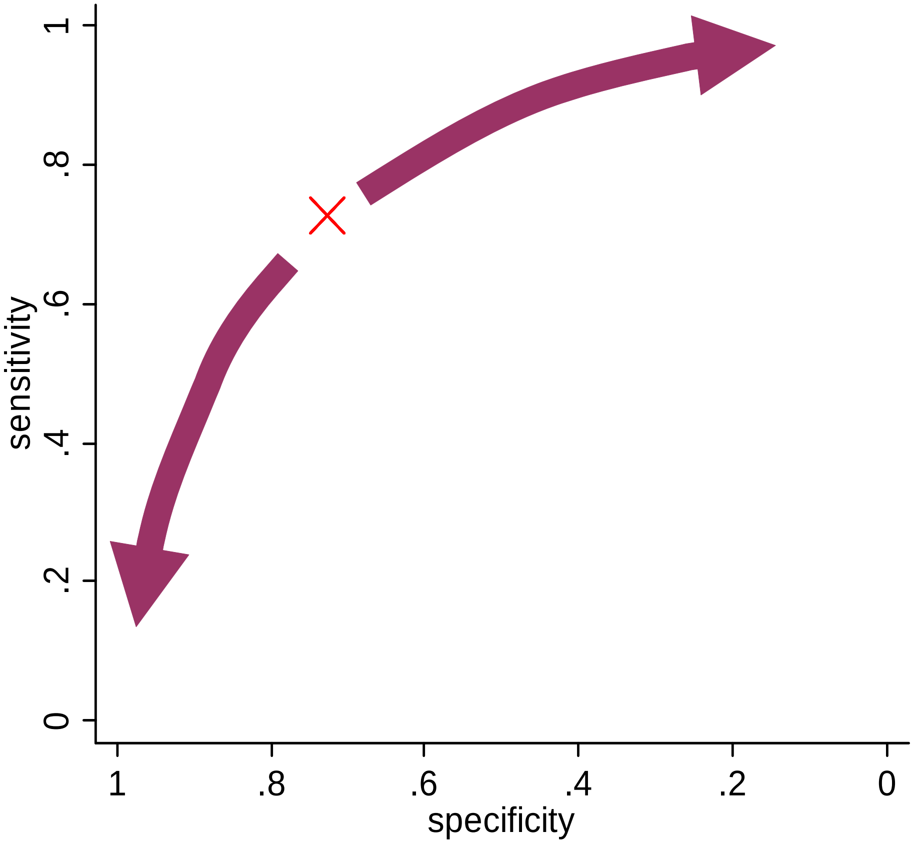


Different Approaches

- Pooling separate estimates
  - Not recommended
- Summary ROC model
  - Traditional approach, relative simple
- More complex models
  - Bivariate random approach
  - Hierarchical summary ROC approach

Threshold effects

Increasing threshold increases specificity but decreases sensitivity



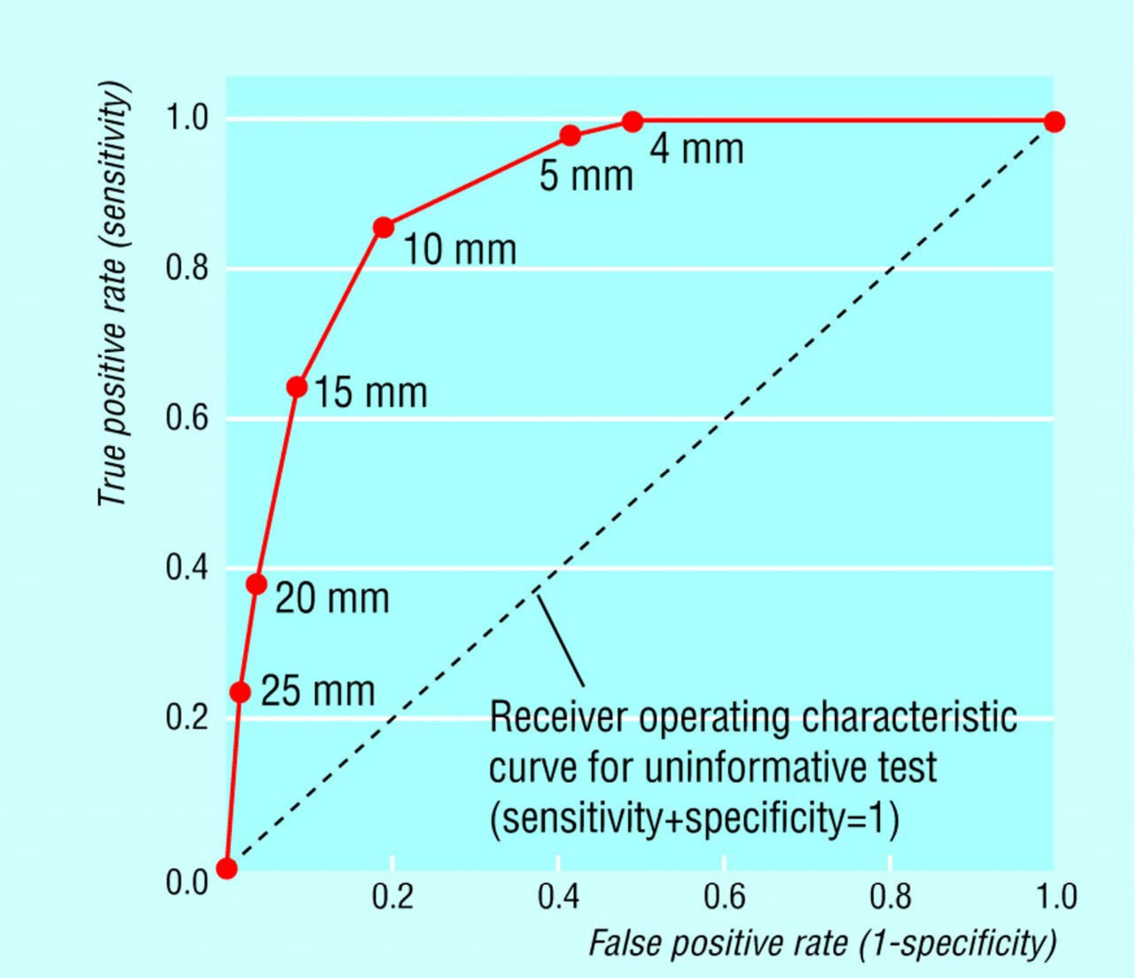
Decreasing threshold increases sensitivity but decreases specificity

Implicit and explicit threshold effects

- **Explicit threshold**: different thresholds are used for test positivity

- **Implicit threshold**: there is no or only one threshold, but in some cases tests are earlier regarded as positive than in other cases

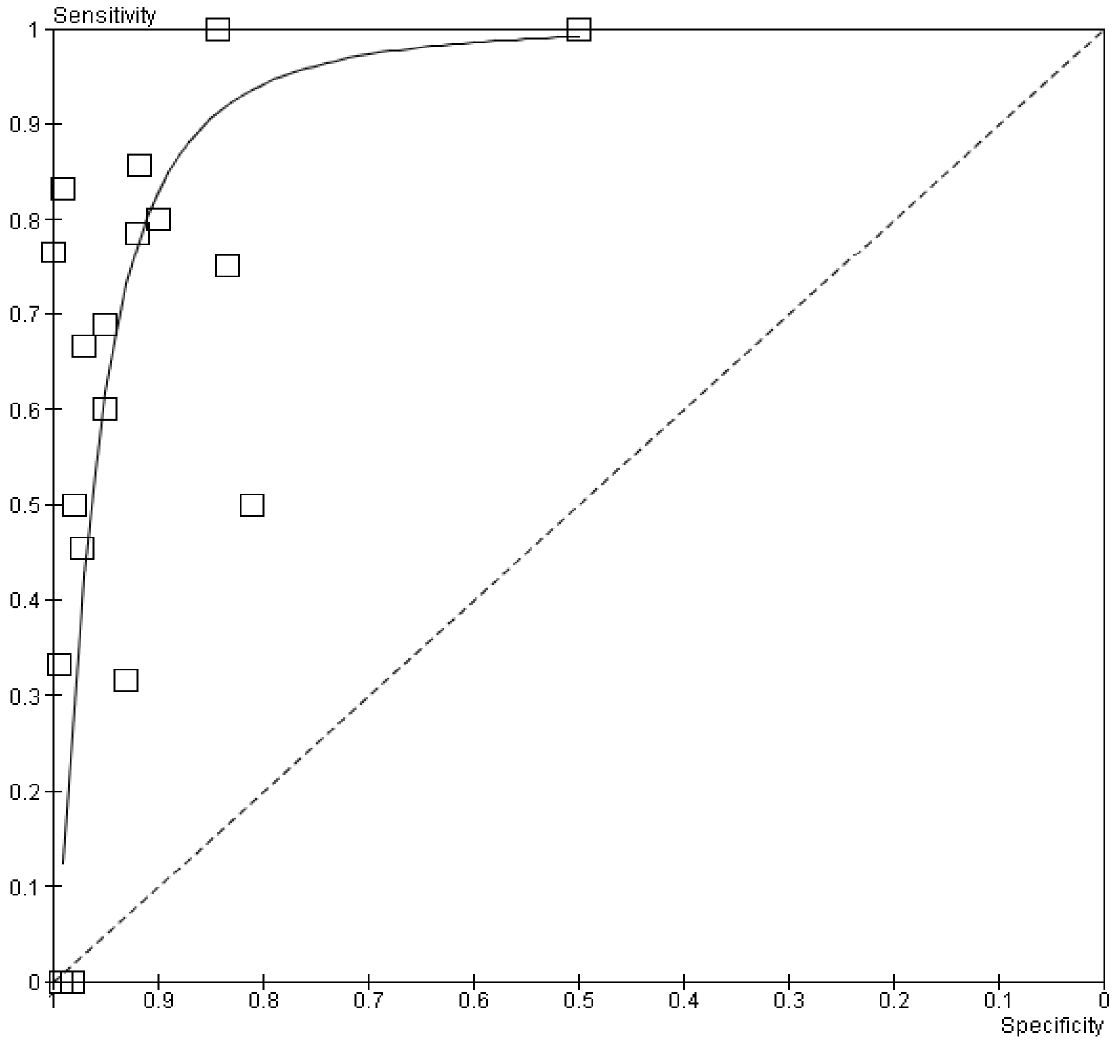
Explicit threshold: (*ROC*) curve



The ROC curve represents the relationship between the true positive rate (TPR) and the false positive rate (FPR) of the test at various thresholds used to distinguish disease cases from non-cases.

Deeks, J. J BMJ 2001;323:157-162

Implicit threshold



ELISA for invasive aspergillosis; cut-off value 1.5 ODI.

Diagnostic odds ratios

Ratio of the odds of positivity in the diseased to the odds of positivity in the non-diseased

$$Diagnostic\,OR = \frac{TP \times TN}{FP \times FN}$$

$$DOR = \frac{\left(\dfrac{sensitivity}{1 - sensitivity}\right)}{\left(\dfrac{1 - specificity}{specificity}\right)} = \frac{LR + ve}{LR - ve}$$

Diagnostic odds ratios

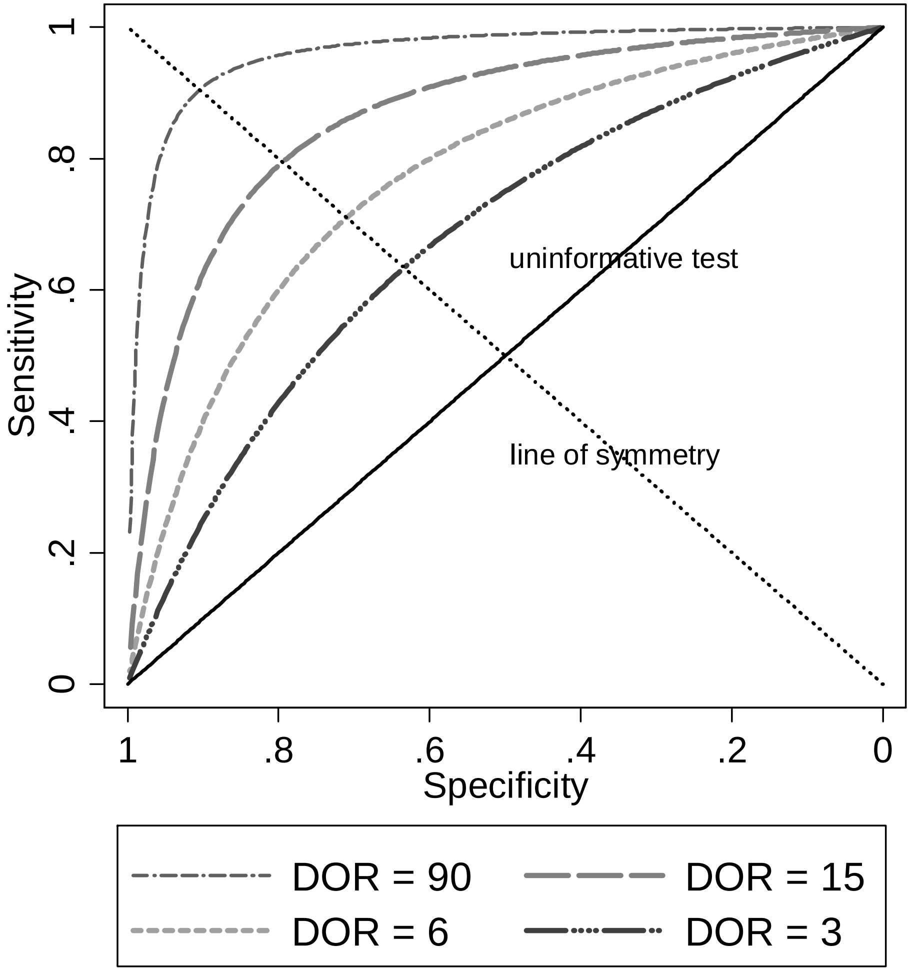
		Cervical Cancer (Biopsy)		
		Present	Absent	
HPV Test	+	65	93	158
	-	7	161	198
		72	254	356

$$\text{DOR} = \frac{65 \times 161}{93 \times 7} = 16$$

Diagnostic odds ratios

				Sensitivity			
Specificity	*50%*	*60%*	*70%*	*80%*	*90%*	*95%*	*99%*
50%	1	2	2	4	9	19	99
60%	2	2	4	6	14	29	149
70%	2	4	5	9	21	44	231
80%	4	6	9	16	36	76	396
90%	9	14	21	36	81	171	891
95%	19	29	44	76	171	361	1881
99%	99	149	231	396	891	1881	9801

Symmetrical *ROC* curves and diagnostic odds ratios



As DOR increases, the ROC curve moves closer to its ideal position near the upper-left corner.

ROC curve is asymmetric when test accuracy varies with threshold

Statistical modelling of ROC curves

- statisticians like straight lines with axes that are independent variables
- first calculate the logits of TPR and FPR
- and then graph the difference against their sum

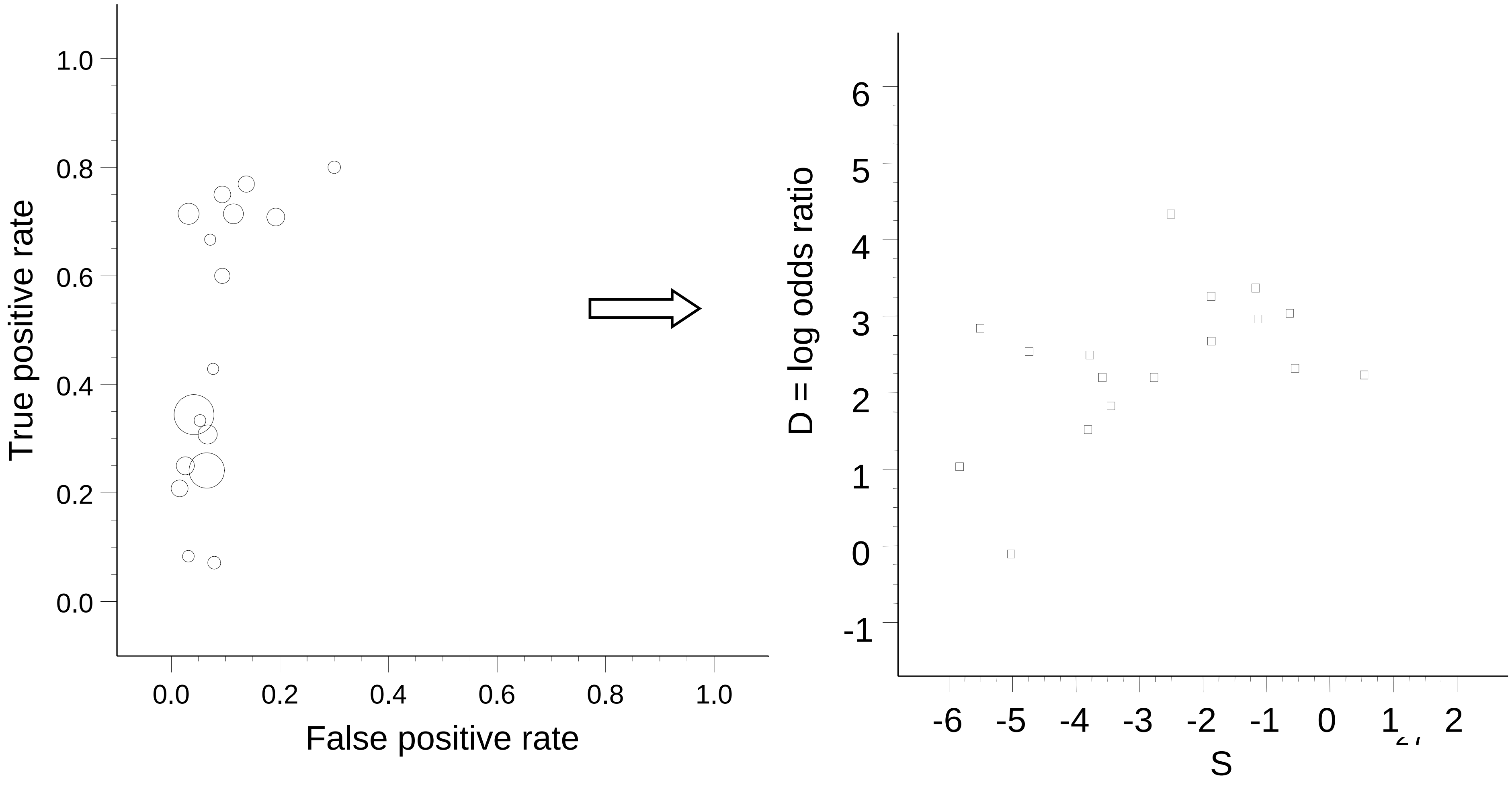
$$\text{logit}(TPR) = \ln\left(\frac{TPR}{1-TPR}\right)$$

$$\text{logit}(FPR) = \ln\left(\frac{FPR}{1-FPR}\right)$$

$$S = \text{logit}(TPR) + \text{logit}(FPR)$$

$$D = \text{logit}(TPR) - \text{logit}(FPR)$$

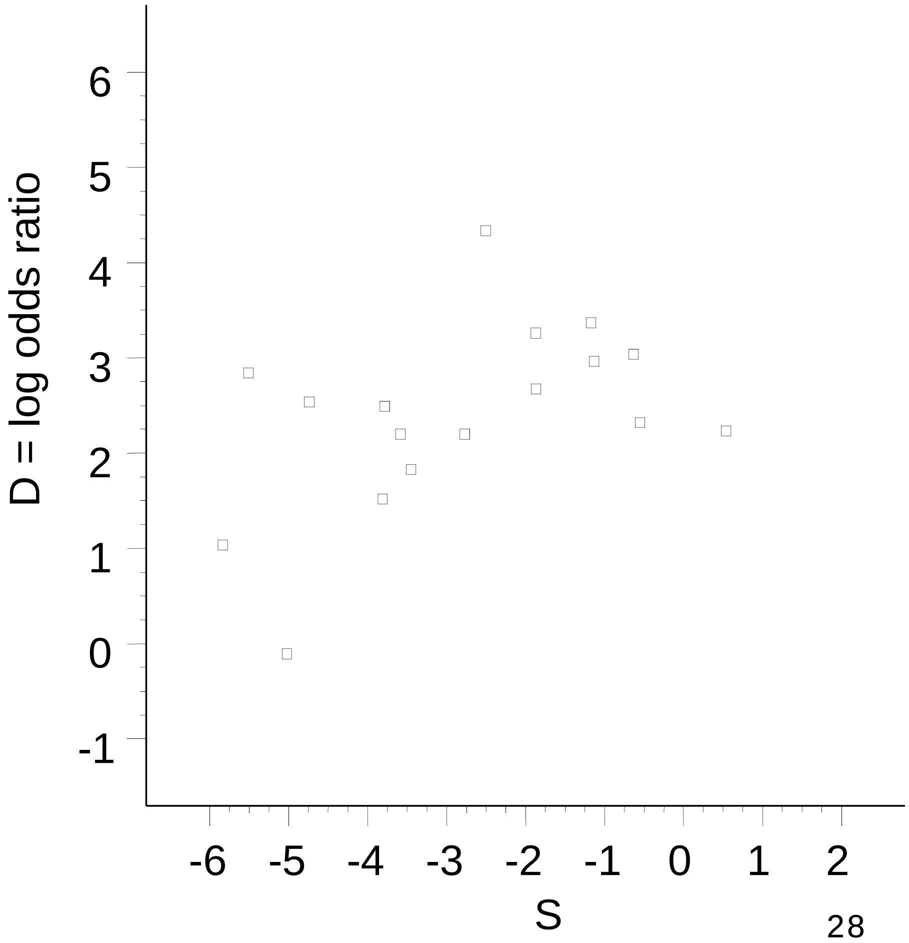
Translating ROC space to D versus S



Moses-Littenberg SROC method

What do the axes mean?

- Difference in logits is the log of the DOR
- Sum of the logits is a marker of diagnostic threshold



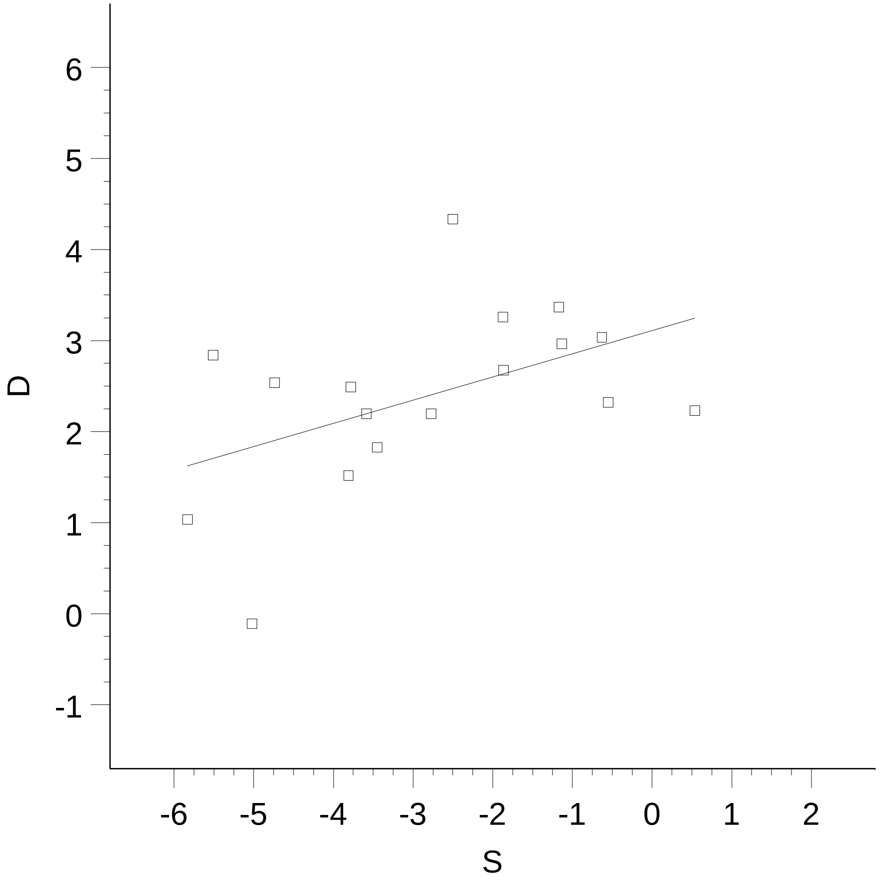
Moses-Littenberg SROC method

- Regression models can be used to fit the straight lines to model relationship between test accuracy and test threshold

$$D = a + bS$$

  - Outcome variable D is the difference in the logits
  - Explanatory variable S is the sum of the logits
  - Ordinary or weighted regression – weighted by sample size or by inverse variance of the log of the DOR

Linear Regression



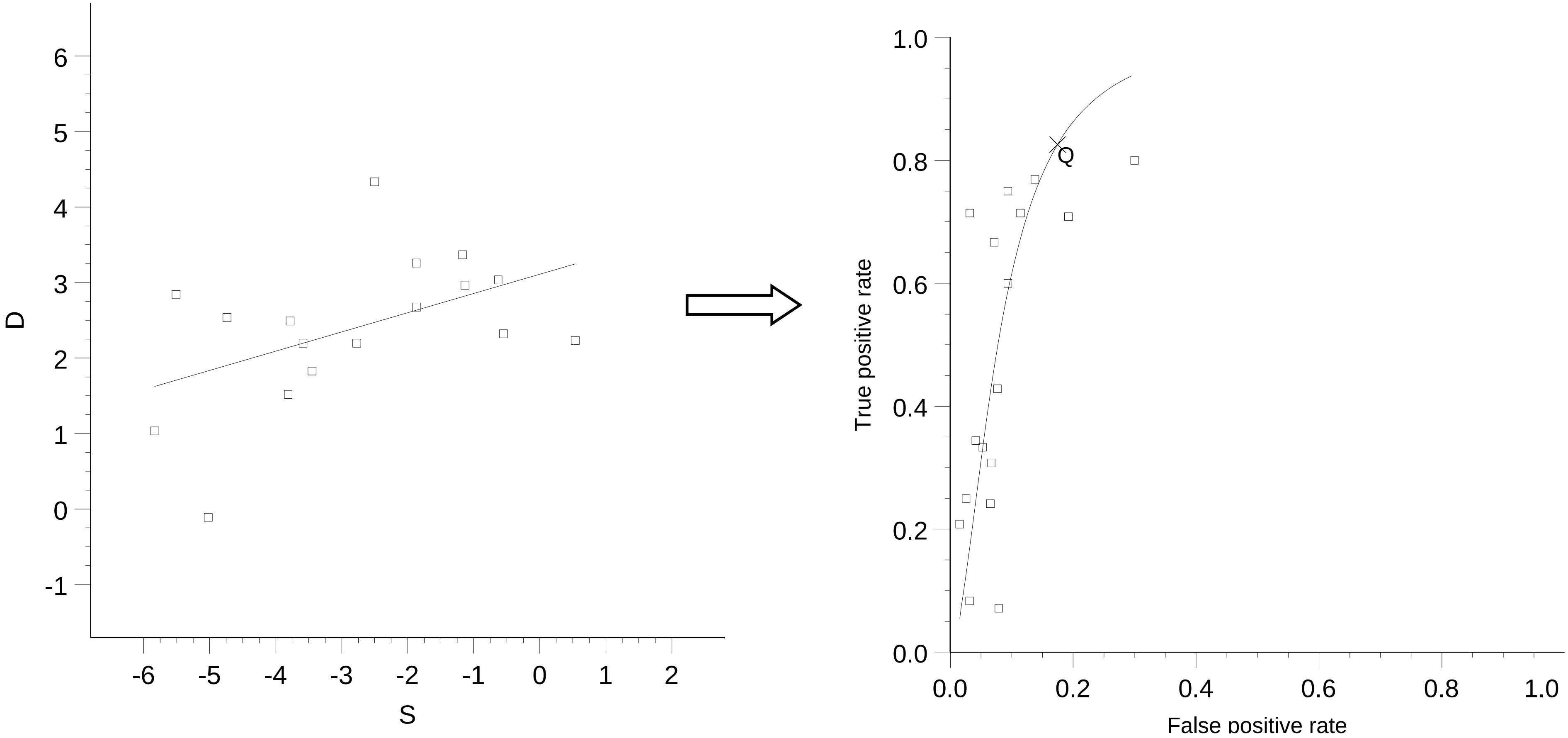
Producing summary ROC curves

- Transform back to the ROC dimensions

$$TPR = \frac{1}{1 + \frac{1}{e^{a/(1-b)}} \times \left(\frac{FPR}{1 - FPR} \right)^{\frac{1+b}{1-b}}}$$

- where ‘a’ is the intercept, ‘b’ is the slope
  - when the ROC curve is symmetrical, b=0 and the equation is simpler

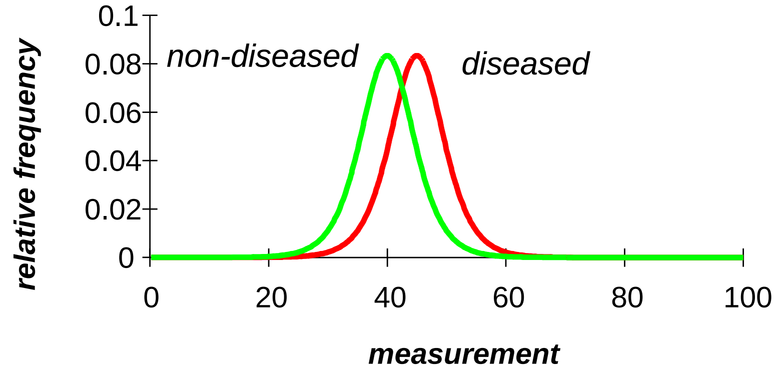
Linear Regression & Back Transformation

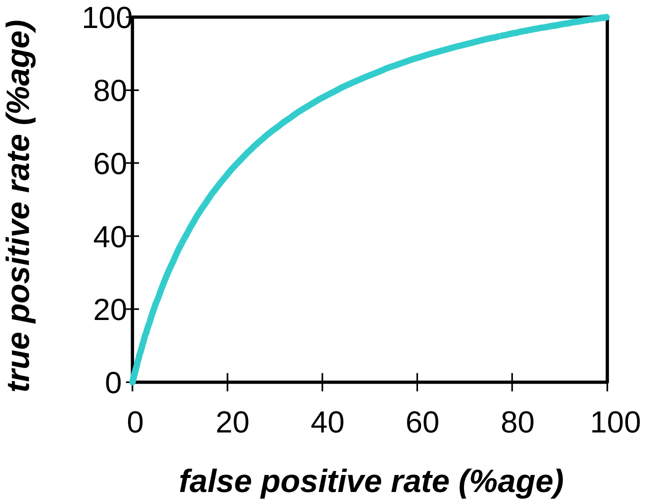


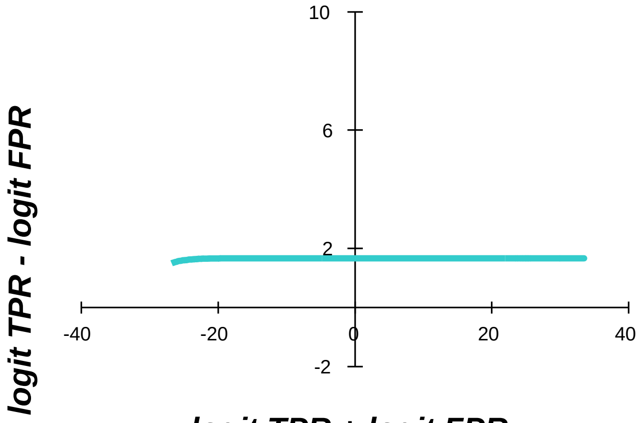
Different situations

- What is the relationship between the underlying distribution and the ROC curve and the D versus S line?
- Let’s have a look at different situations.

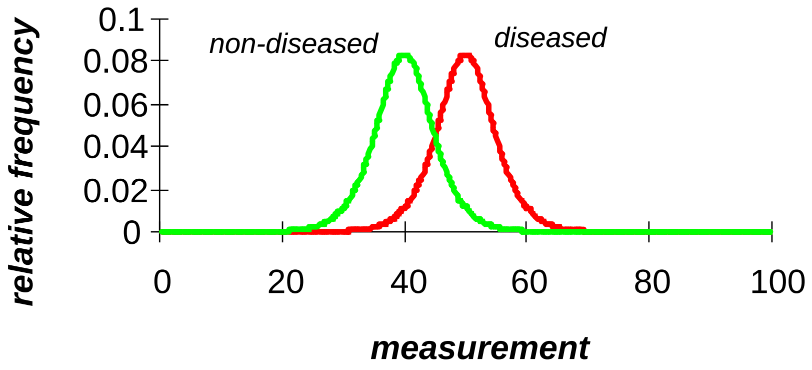
ROC curve and logit difference and sum plot: small difference, same spread

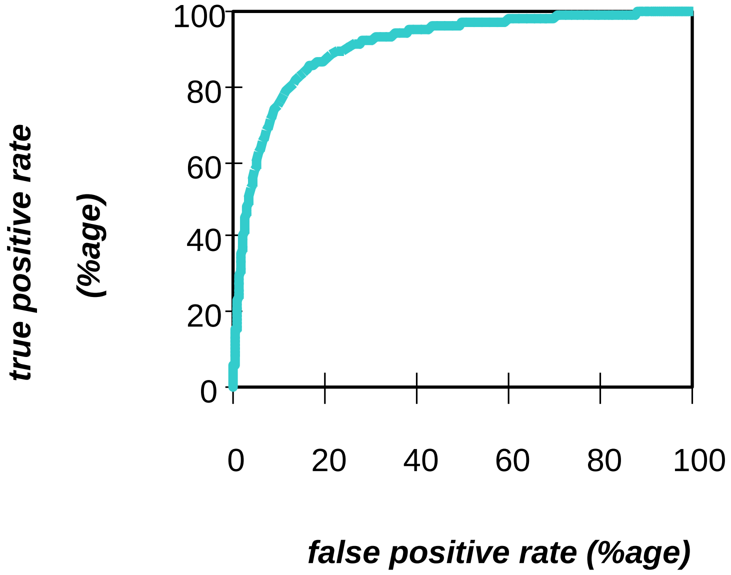


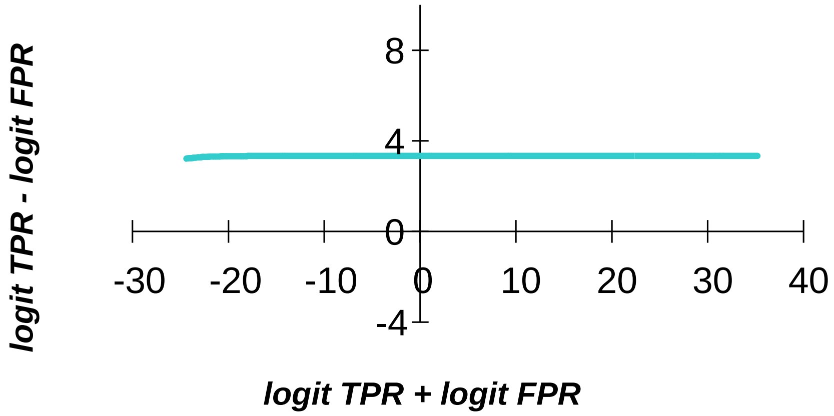




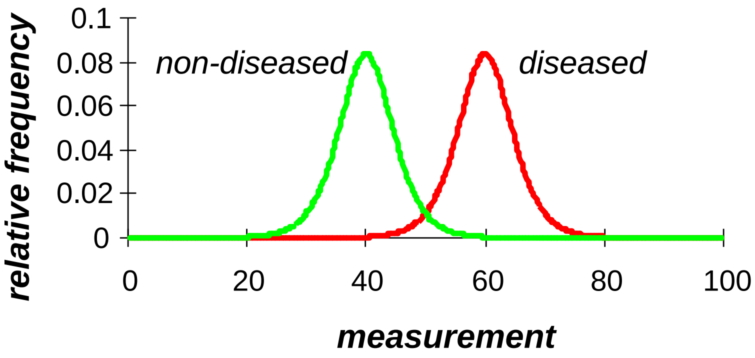
ROC curve and logit difference and sum plot: moderate difference, same spread

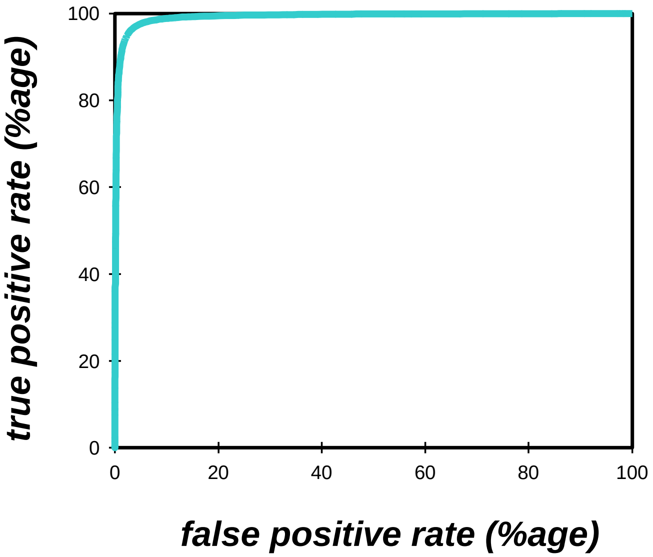


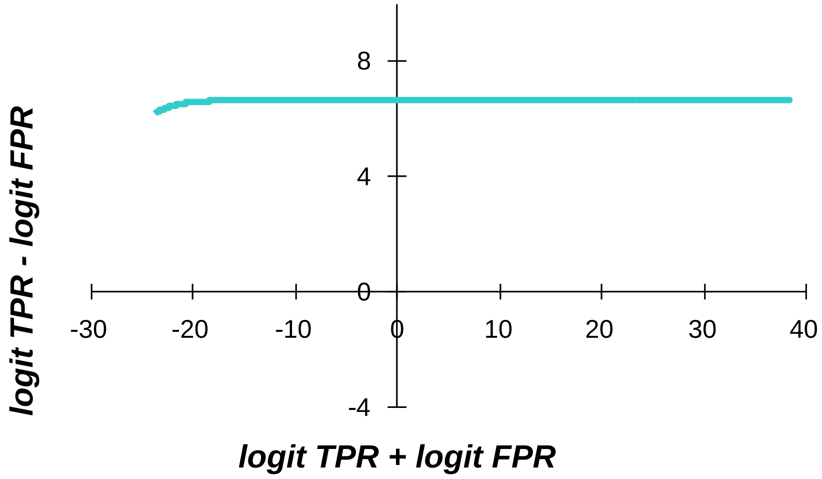




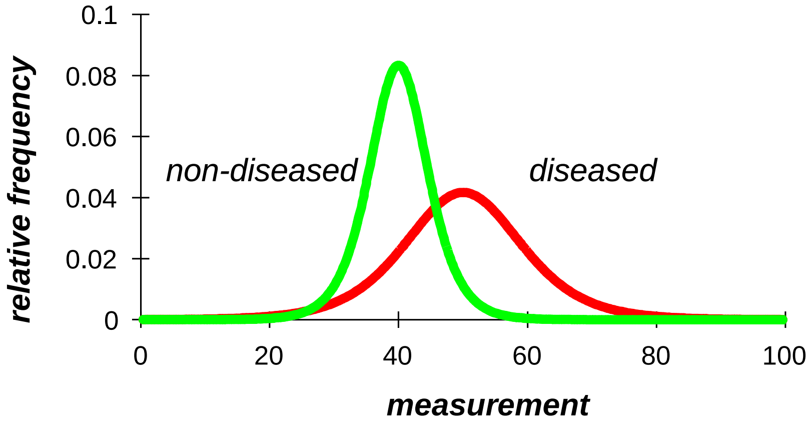
ROC curve and logit difference and sum plot: large difference, same spread

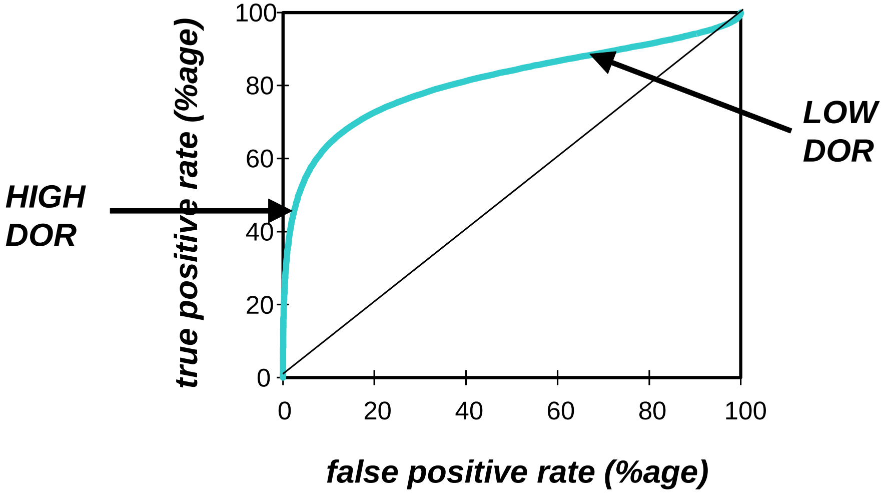


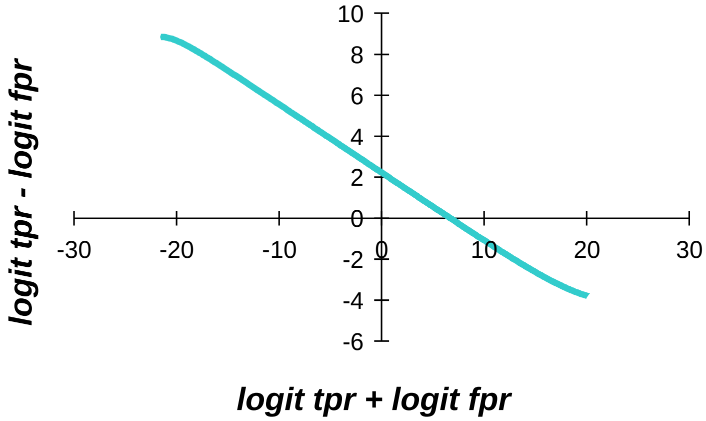




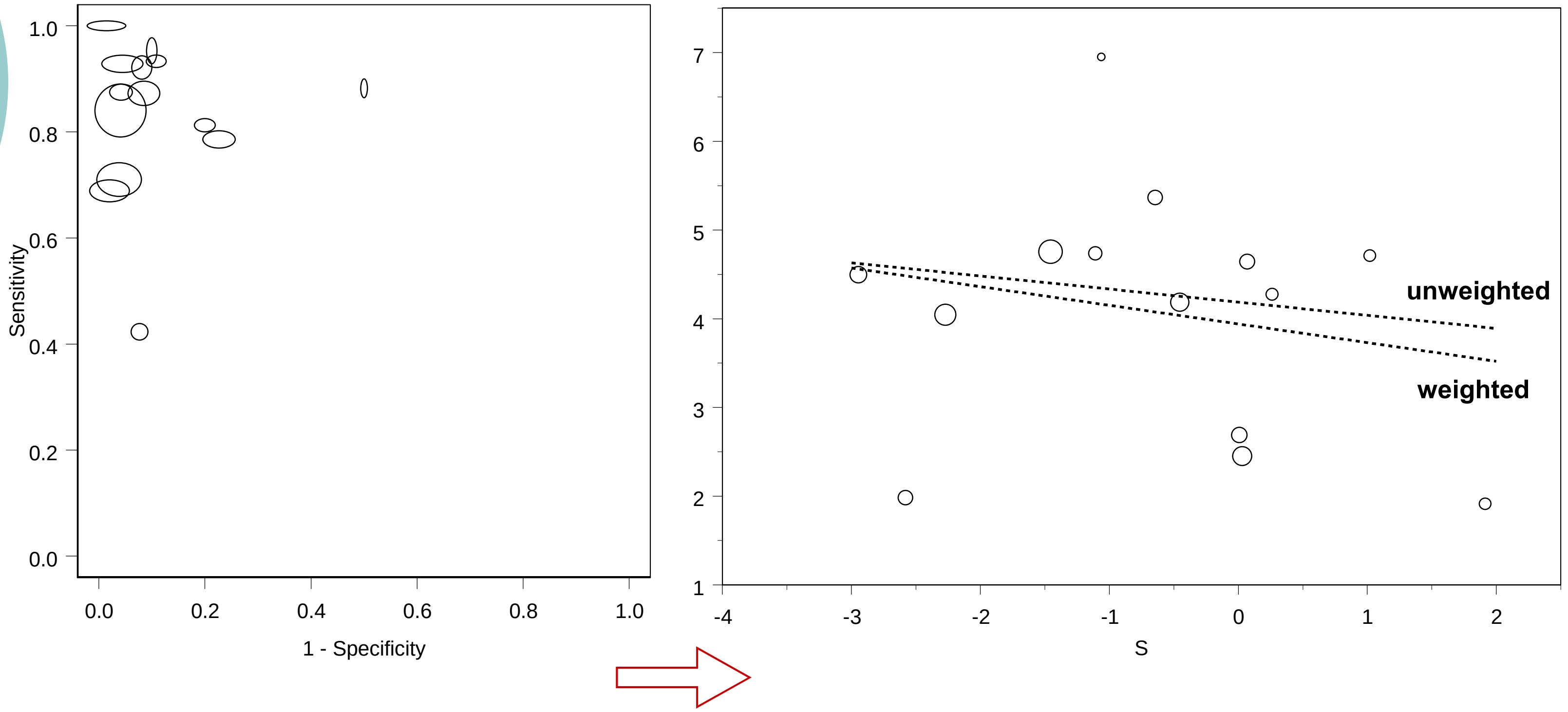
ROC curve and logit difference and sum plot: moderate difference, unequal spread





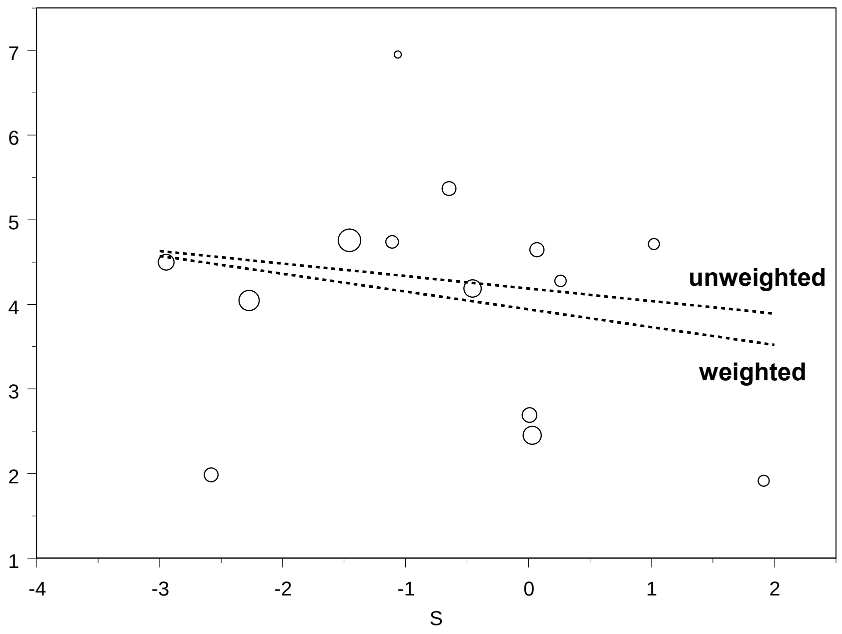


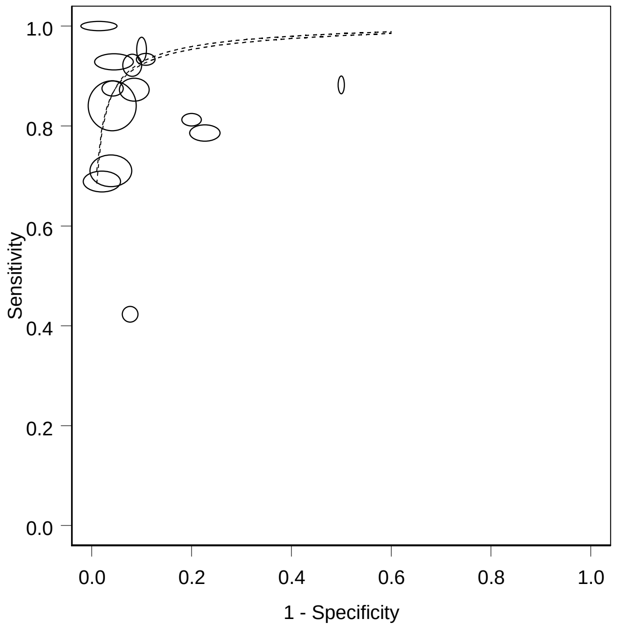
SROC regression: another example



Transformation linearizes relationship between accuracy and threshold so that linear regression can be used

PSV example *cont*.





inverse transformation

The SROC curve is produced by using the estimates of a and b to compute the expected sensitivity (*tpr)* across a range of values for 1-specificity (*fpr)*

Problems with the Moses-Littenberg SROC method

- Poor estimation
  - Tends to underestimate test accuracy due to zero-cell corrections and bias in weights
- Validity of significance tests
  - Sampling variability in individual studies not properly taken into account
  - P-values and confidence intervals erroneous
- Operating points
  - knowing average sensitivity/specificity is important but cannot be obtained
  - Sensitivity for a given specificity can be estimated

Advanced models –
HSROC and Bivariate methods

- Hierarchical / multi-level
  - allows for both within and between study variability, and within study correlations between diseased and non-diseased groups
- Logistic
  - correctly models sampling uncertainty in the true positive proportion and the false positive proportion
  - no zero cell adjustments needed
- Random effects
  - allows for heterogeneity between studies
- Regression models
  - used to investigate sources of heterogeneity

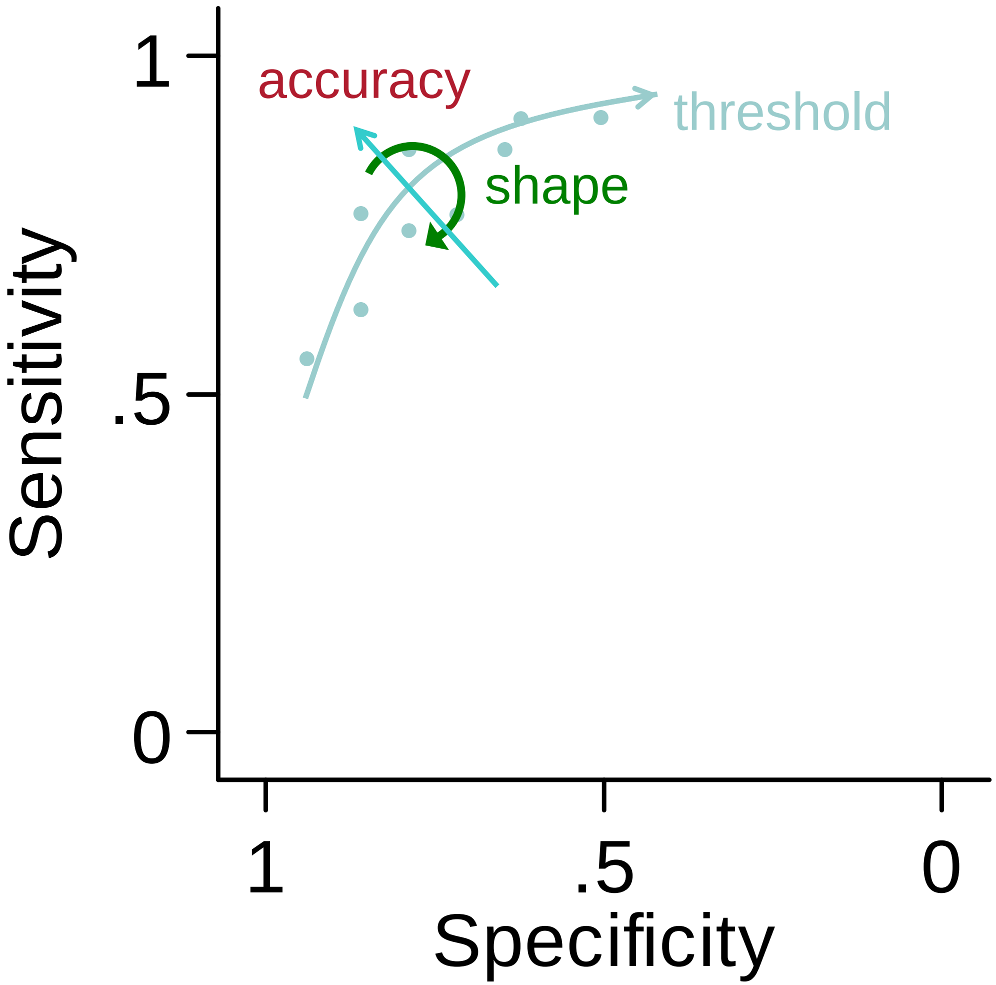
Parameterizations

- HSROC
  - Mean lnDOR
  - Variance lnDOR
  - Mean threshold
  - Variance threshold
  - Shape of ROC

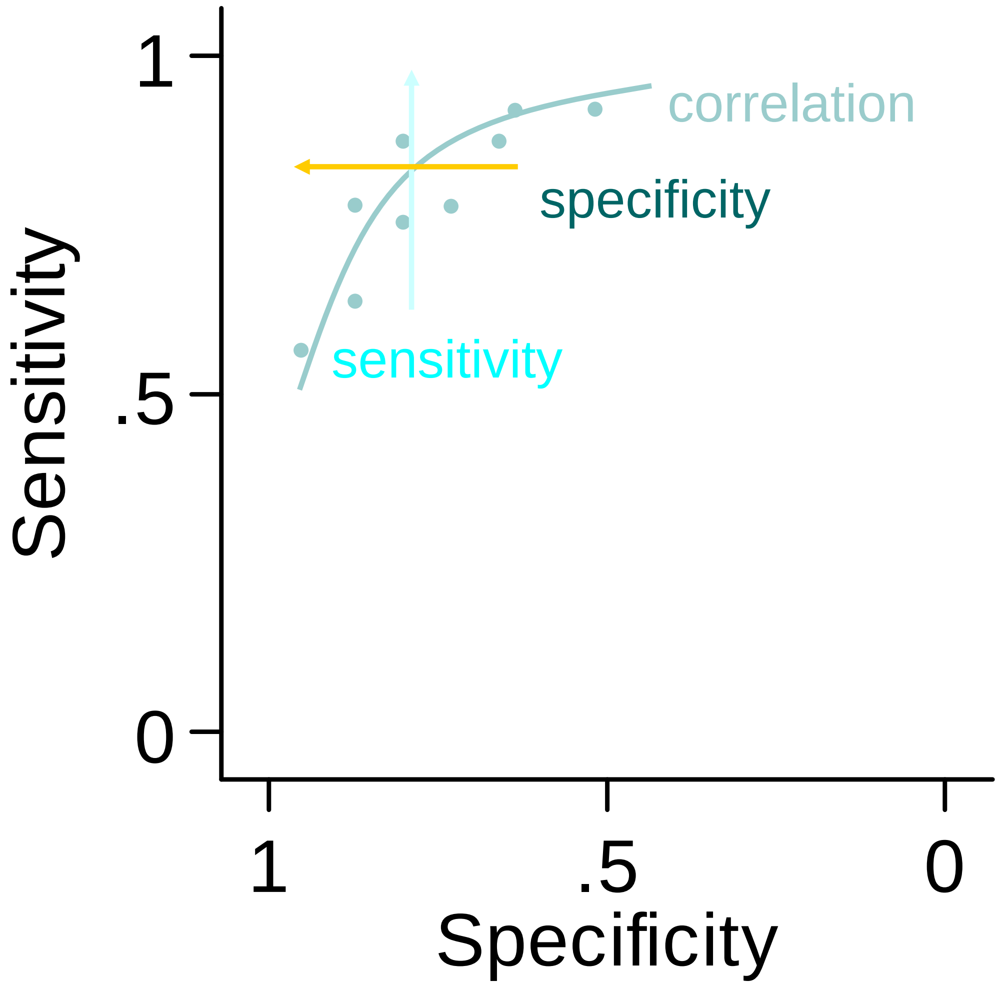
- Bivariate
  - Mean logit sens
  - Variance logit sens
  - Mean logit spec
  - Variance logit spec
  - Correlation between sensitivity and specificity

Other than the parameterization, the models are mathematically equivalent, see Harbord R, Deeks J *et al.* A unification of models for meta-analysis of diagnostic accuracy studies. *Biostatistics* 2006;1:1-21.

Hierarchical SROC model



Bivariate model



Outputs from the models

HSROC

- Estimates underlying SROC curve, and the average operating point on the curve (mean DOR and mean threshold)
- Possible to estimate mean sensitivity, specificity and mean likelihood ratios, with standard errors obtained using the delta method
- Confidence and prediction ellipses estimable

Bivariate

- Estimates the average operating point (mean sensitivity and specificity), confidence and prediction ellipses
- Possible to estimate mean likelihood ratios, with standard errors obtained using the delta method
- Underlying SROC curve estimable

Fitting the models

HSROC

- Hierarchical model with non-linear regression, random effects and binomial error
- Original code in winBUGs
- Easy to fit in PROC NLMIXED in SAS

Bivariate

- Hierarchical model with linear regression, random effects and binomial error
- Easy to fit in PROC NLMIXED in SAS, can be fitted in PROC MIXED
- Also in GLLAMM in STATA, MLWin

Syntax Proc NLMIXED - HSROC

```
proc nlmixed data=diag ;
   parms alpha=4 theta=0 beta=0
    s2ua=1 s2ut=1;
   logitp = (theta + ut + (alpha + ua) * dis) *
              exp(-(beta)*dis);
   p = exp(logitp)/(1+exp(logitp));
   model  pos ~ binomial(n,p);
   random ua ut ~ normal([0 , 0],
       [s2ua,0,s2ut]) subject=study;
```

Mean accuracy

Mean threshold

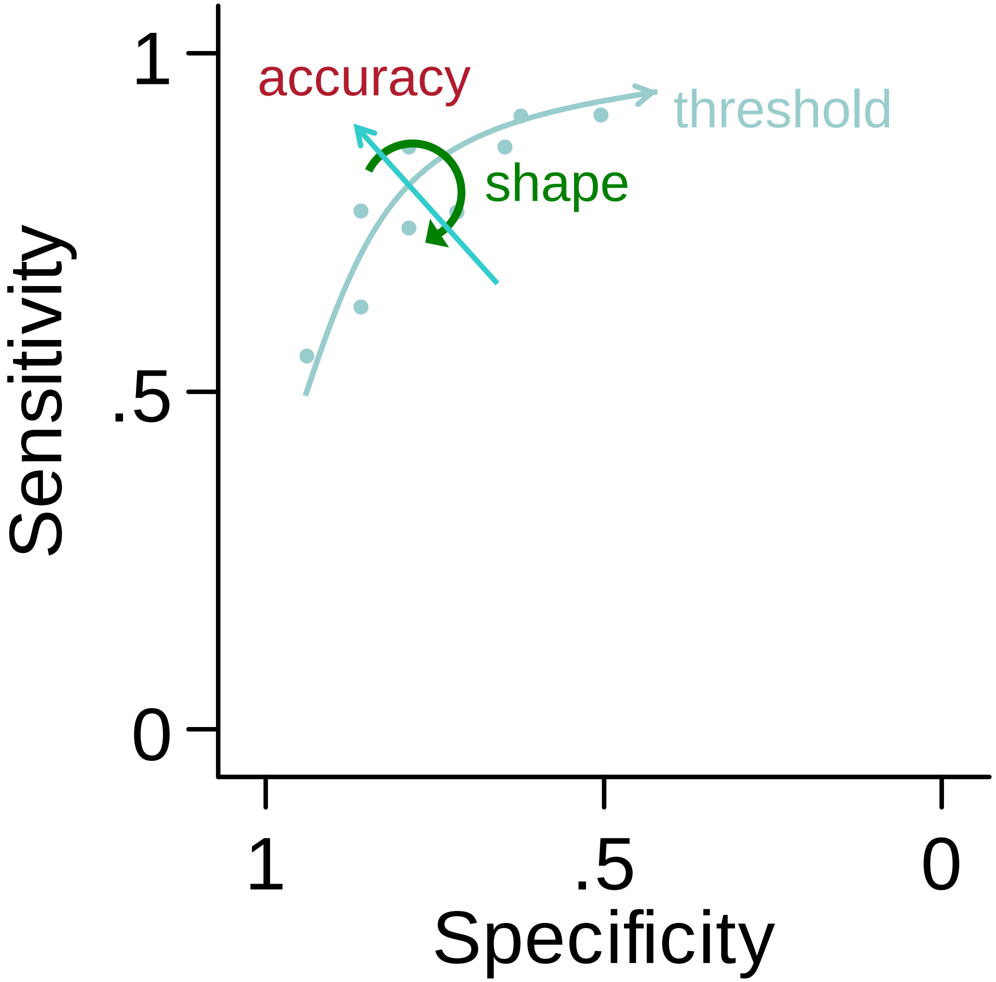
Random effect threshold

Random effect threshold

shape

Disease indicator

Hierarchical SROC model



Syntax Proc NLMIXED - Bivariate

```
proc nlmixed data=diag ;
   parms msens=1 mspec=2
    s2usens=0.2 s2uspec=0.6 cov=0;
   logitp = (msens + usens)*dis +
                (mspec + uspec)*nondis;
   p = exp(logitp)/(1+exp(logitp));
   model  pos ~ binomial(n,p);
   random usens uspec ~ normal([0 , 0],
         [s2usens,cov,s2uspec]) subject=study;
```

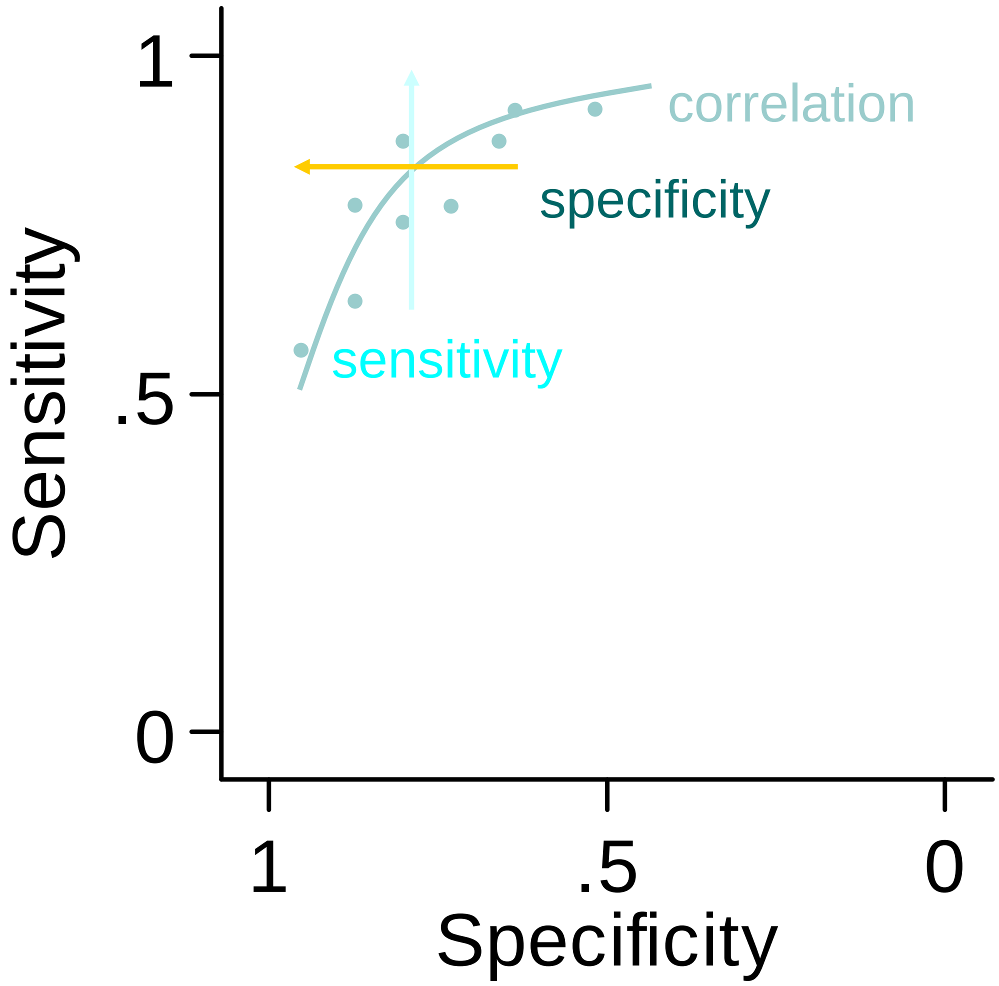
Random effect sensitivity

Mean sensitivity

Random effect specificity

Mean specificity

Bivariate model



METADAS

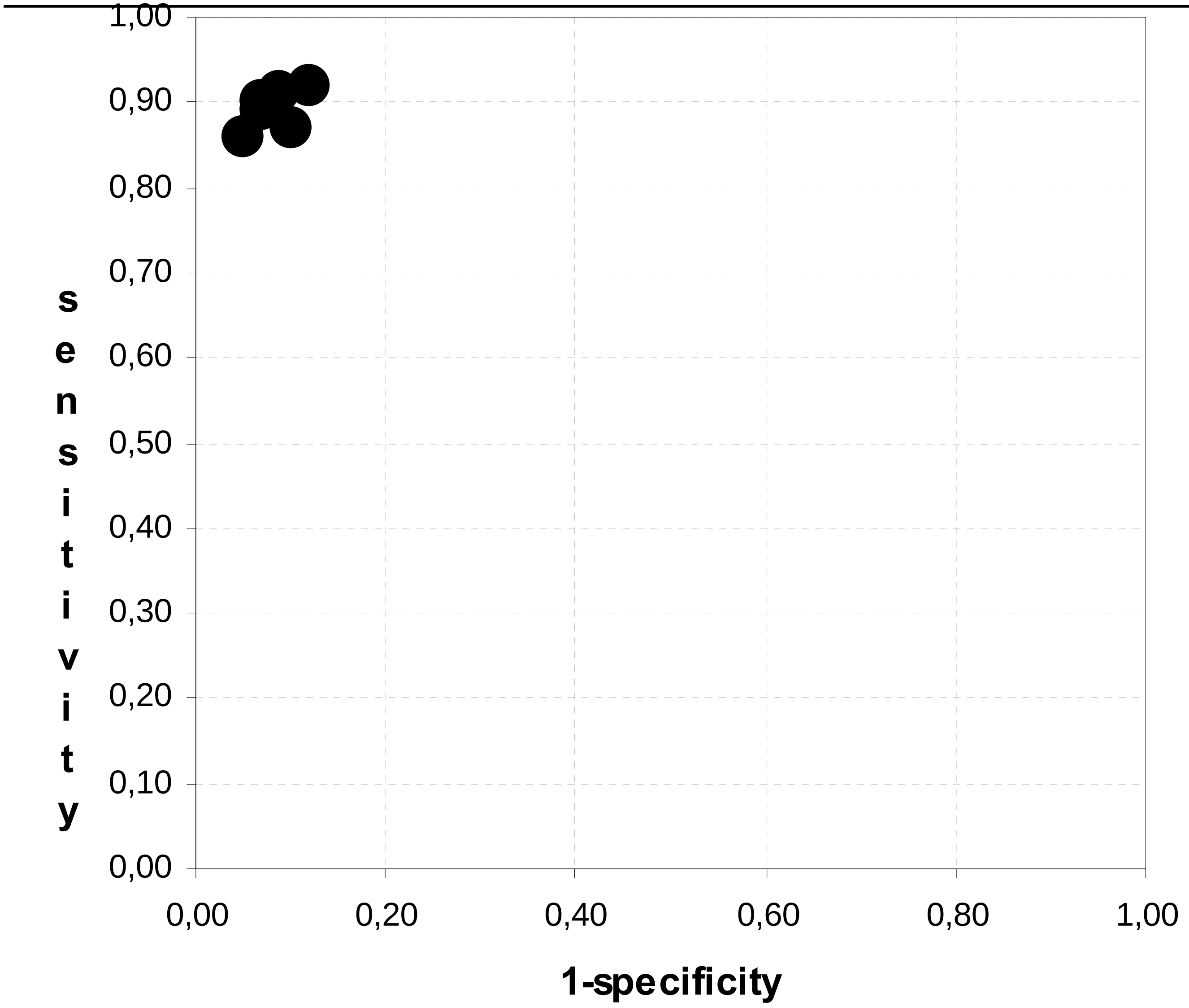
- SAS macro developed to automate HSROC/bivariate analysis using PROC NLMIXED

- Can be used together with Review Manager 5 (Cochrane review Software):
  - Plot summary curve(s)
  - Display summary point(s)
  - Display 95% confidence and/or prediction regions for summary point(s)

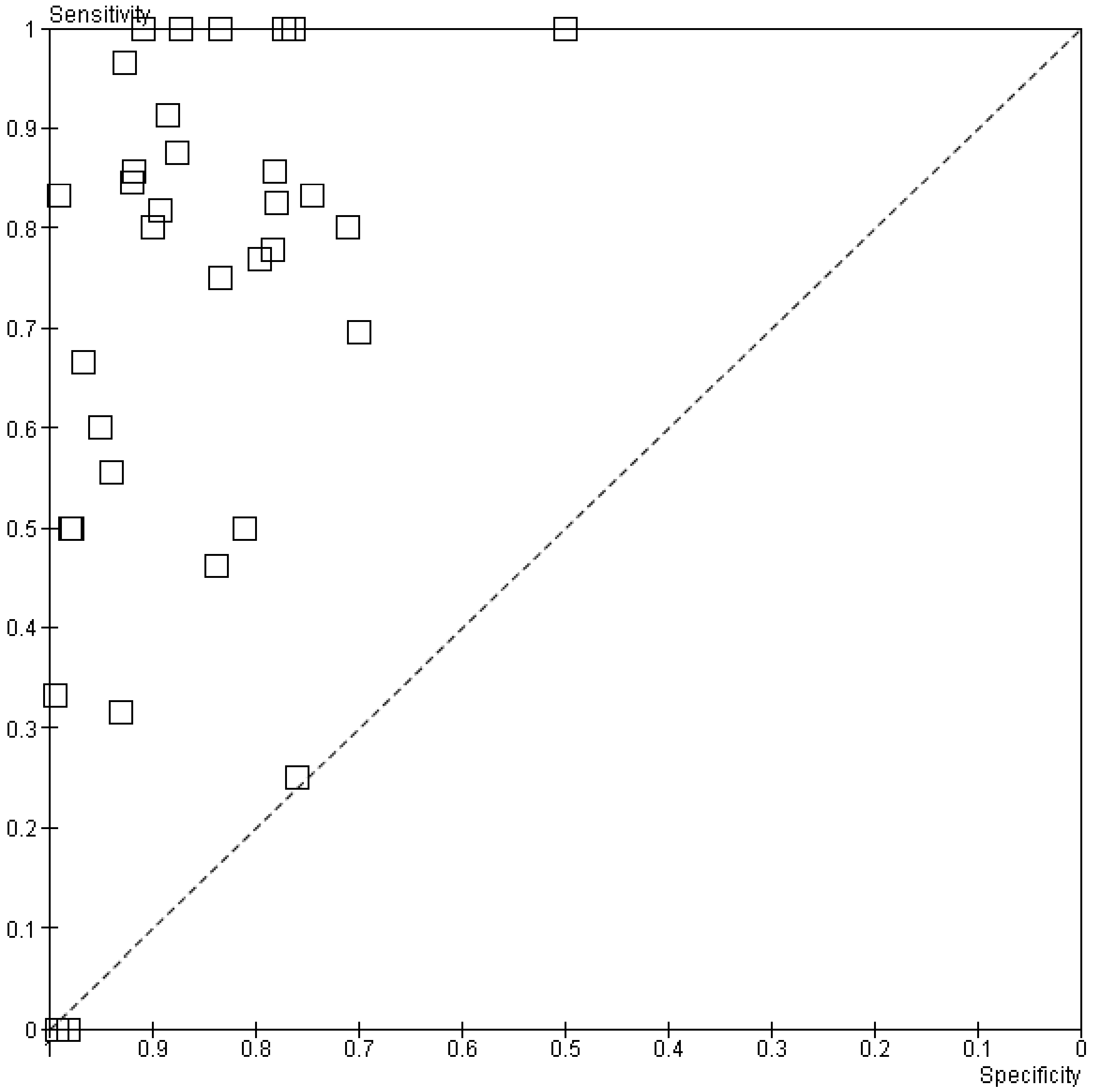
Part 2

dealing with heterogeneity

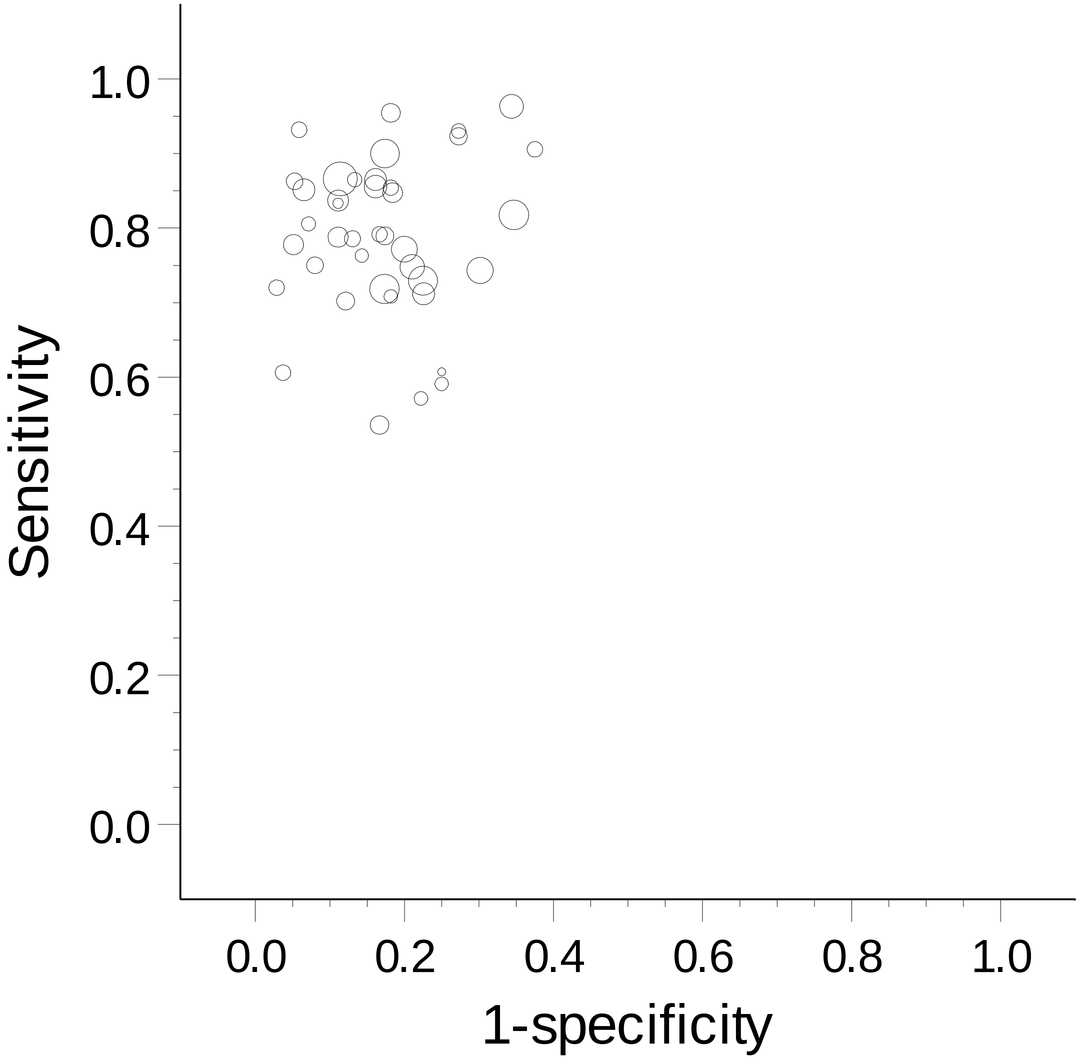
The meta-analyst's dream!



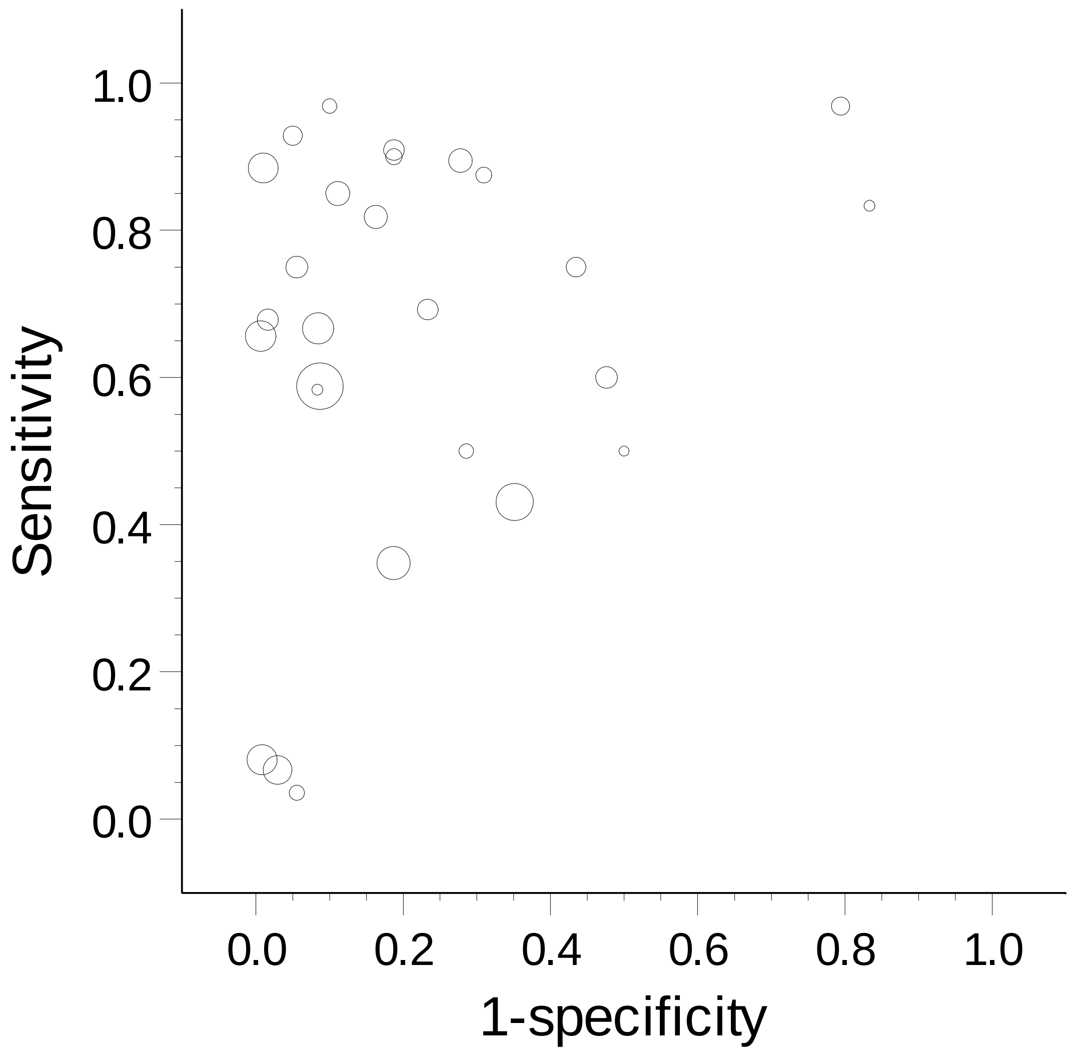
Realistic situation: vast heterogeneity



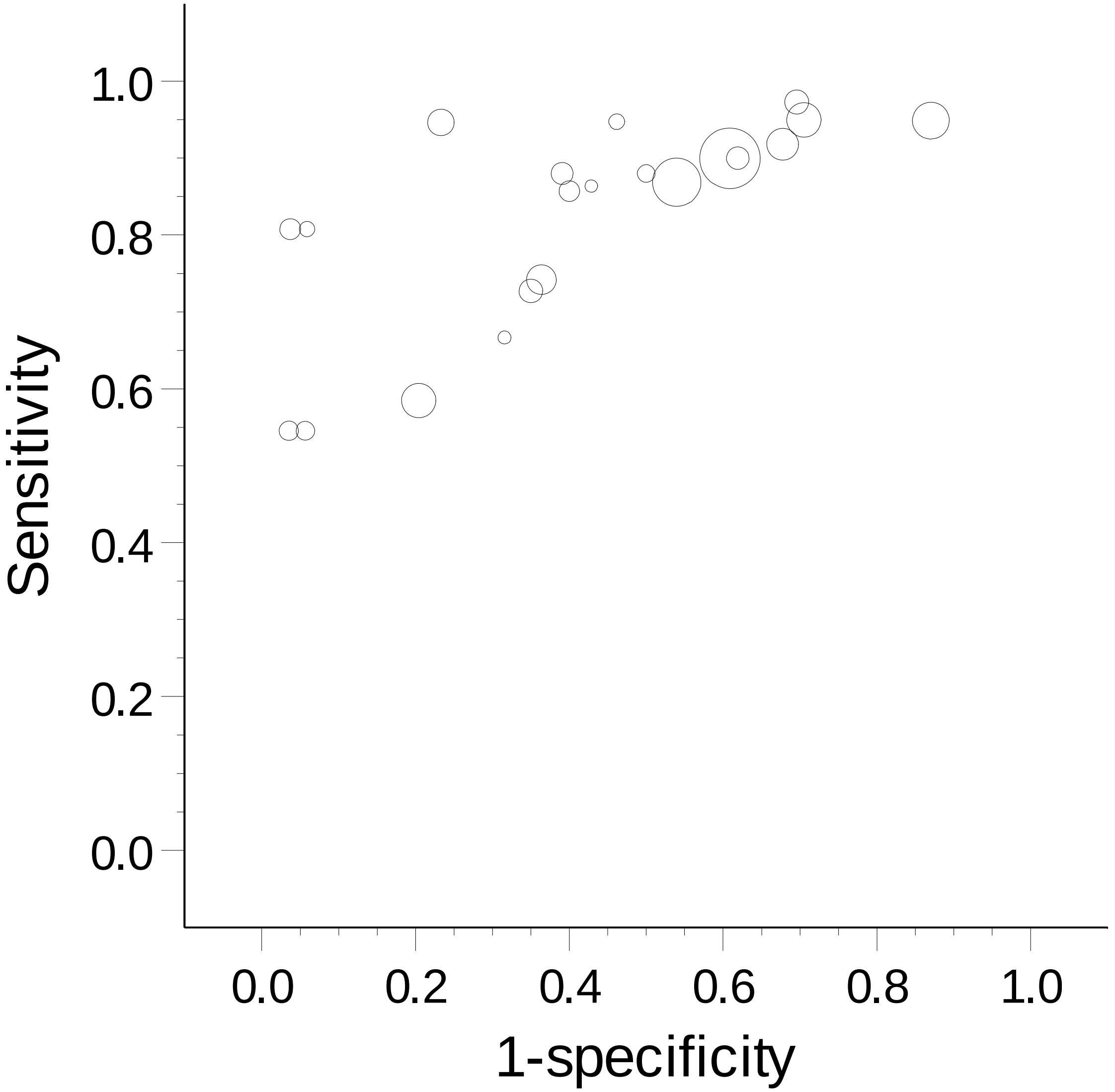
Echocardiography in Coronary Heart Disease



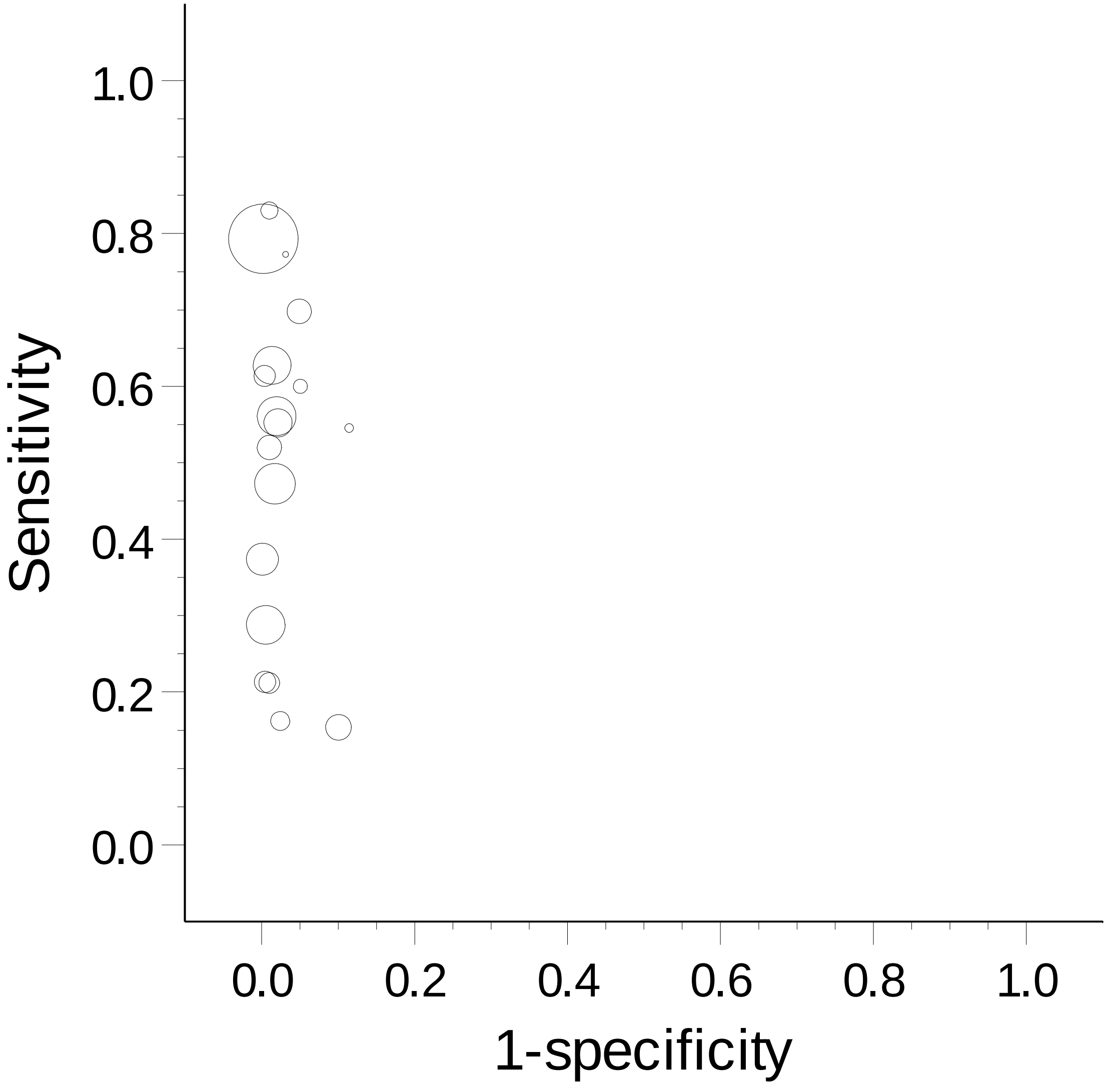
GLAL in Gram Negative Sepsis



F/T PSA in the Detection of Prostate cancer



Dip-stick Testing for Urinary Tract Infection



Sources of Variation

I. Chance variation

II. Differences in threshold

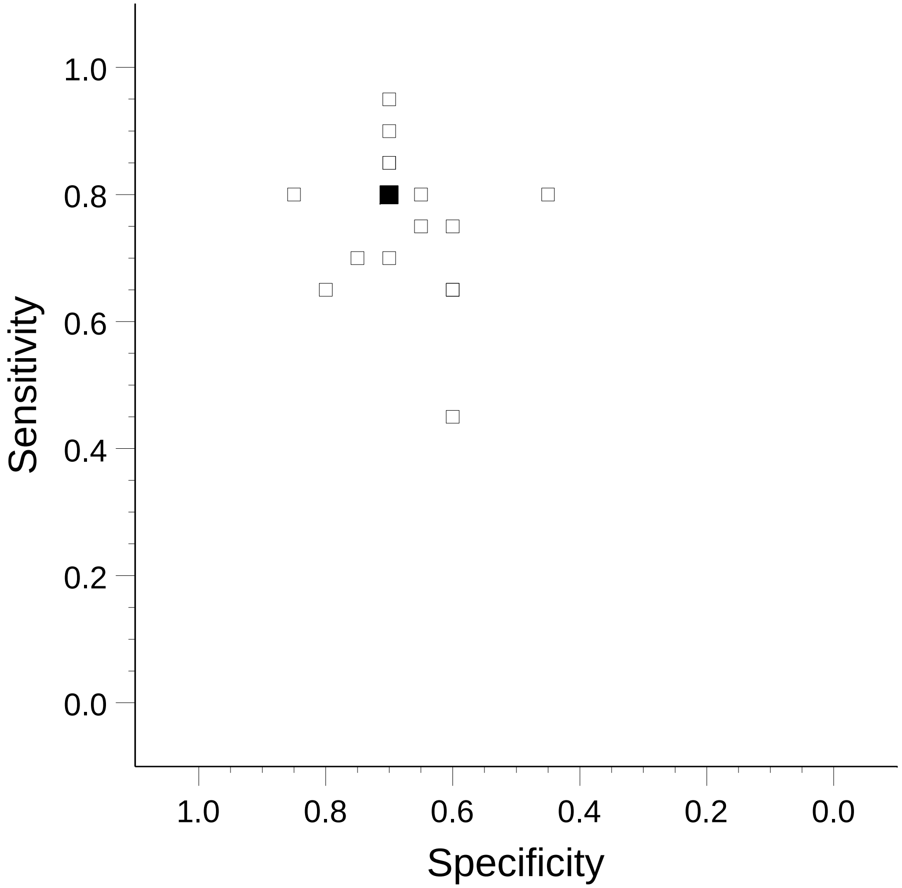
III. Bias

IV. Clinical subgroups

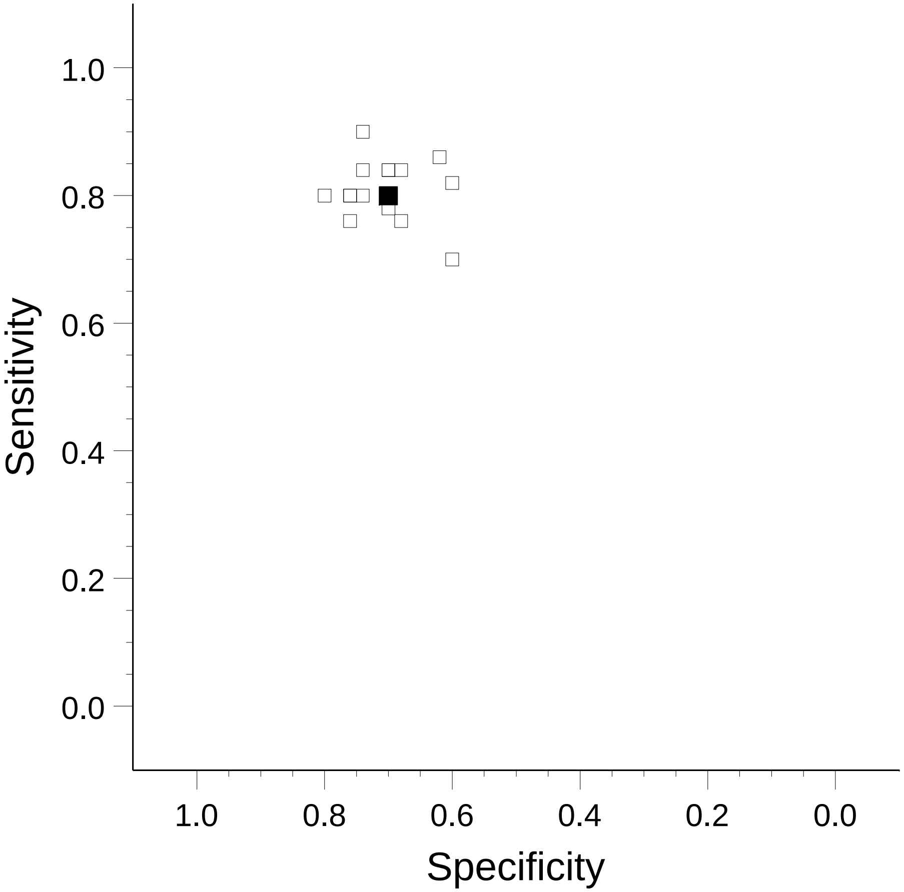
V. Unexplained variation

Sources of Variation: Chance

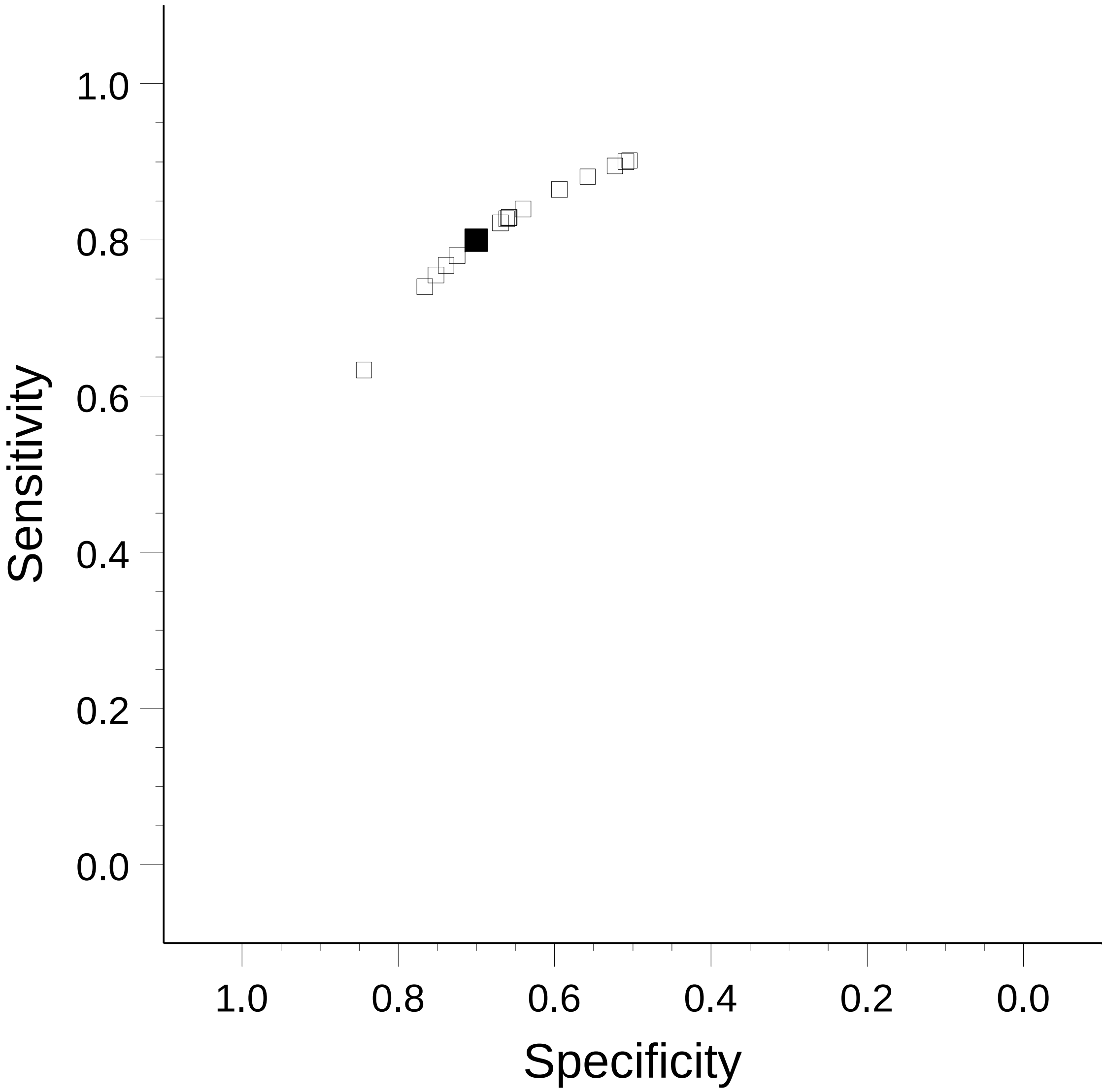
Chance variability: sample size=40



Chance variability: sample size=100



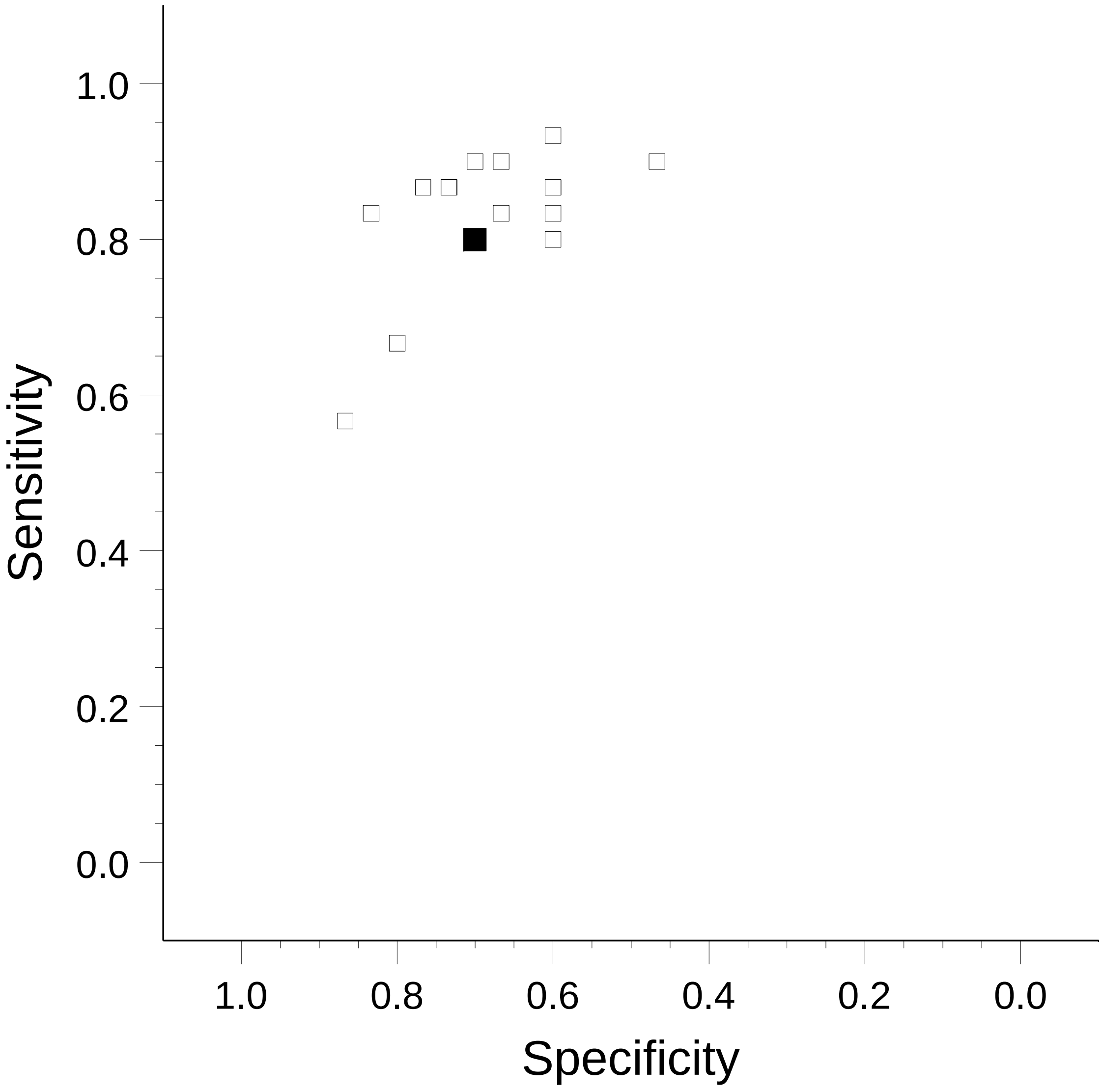
Sources of Variation: Threshold



Threshold:

- perfect negative correlation
- no chance variability

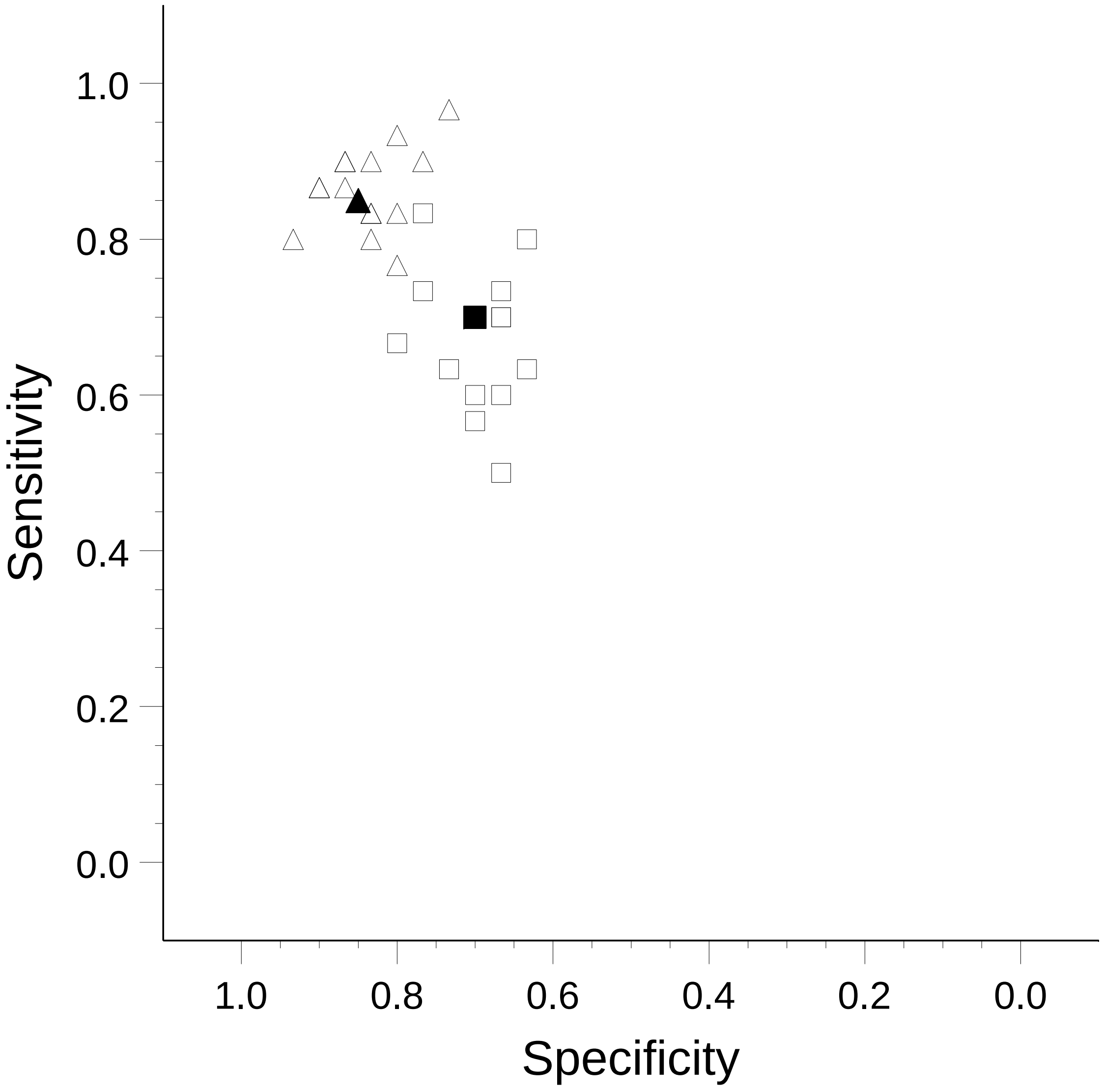
Sources of Variation: Threshold



Threshold:

- perfect negative correlation
- + chance variability ss=60

Sources of Variation: Bias & Subgroup



Bias & Subgroup:

- sens & spec higher
- ss=60
- no threshold

Sources of Variation

I. Chance variation

II. Differences in threshold

III. Bias

IV. Subgroups

V. Unexplained variation

Comparison

Feature	Older Model*	Advanced models**
Chance variability	+ / -	+
Threshold differences	+	+
Subgroup	+	+
Unexplained variation	+ / -	+

* Moses-Littenberg model

** Hierarchical and bivariate models

Exploring heterogeneity

Summarise data per subgroup

- Subgroup analyses
- Meta-regression analysis

Covariates

- Study characteristics (patients, index tests, reference standard, setting, disease stage, etc.)
- Methodological quality items (QUADAS items)

Subgroup analysis and meta-regression

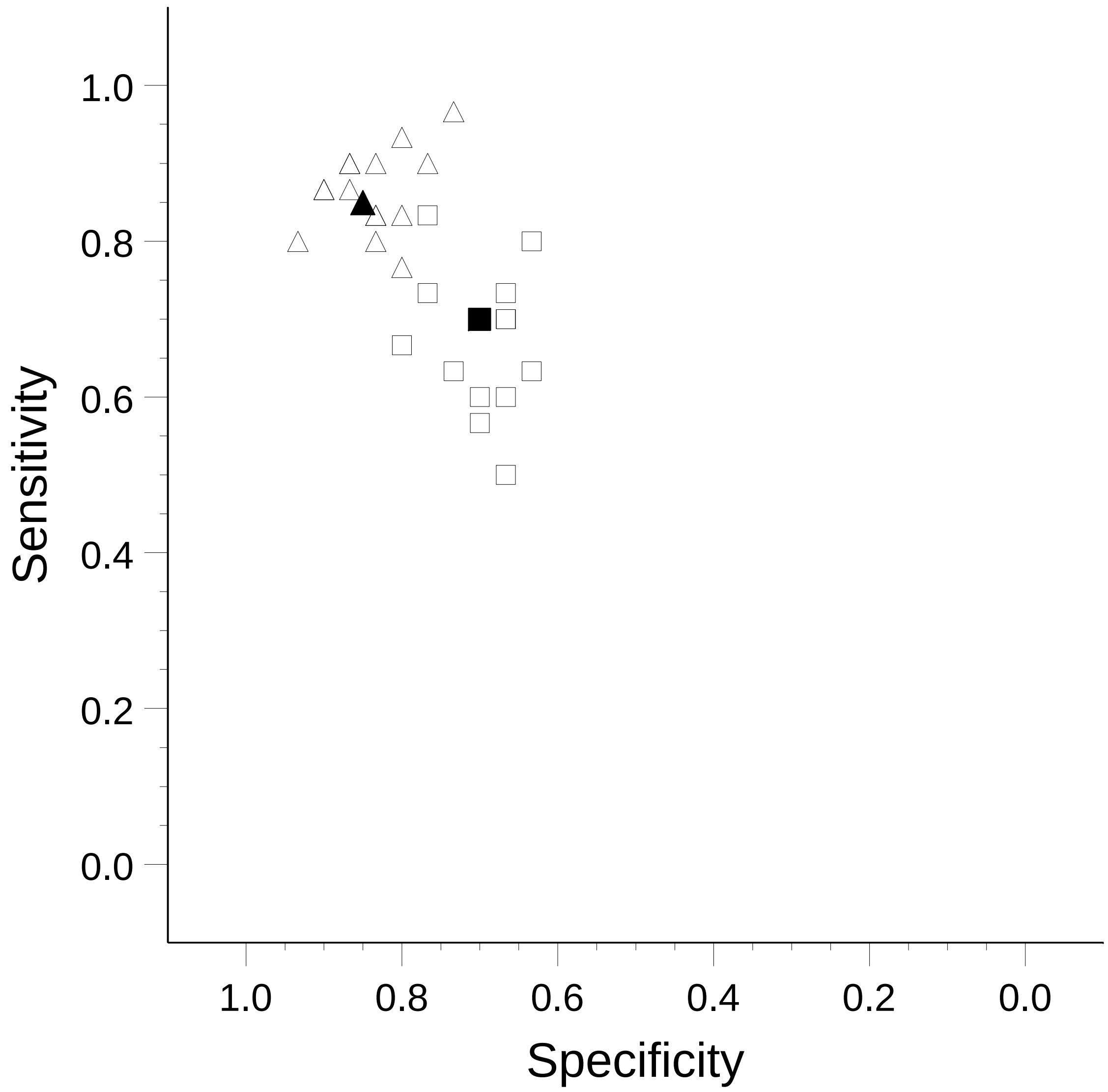
- Advanced models can easily incorporate study-level covariates

- Different questions can be addressed:
  - differences in summary points of sensitivity or specificity
  - differences in overall accuracy
  - differences in threshold
  - differences in shape of SROC curve

Limitations of meta-regression

- Validity of covariate information
  - poor reporting on design features
- Population characteristics
  - information missing or crudely available
- Lack of power
  - small number of contrasting studies

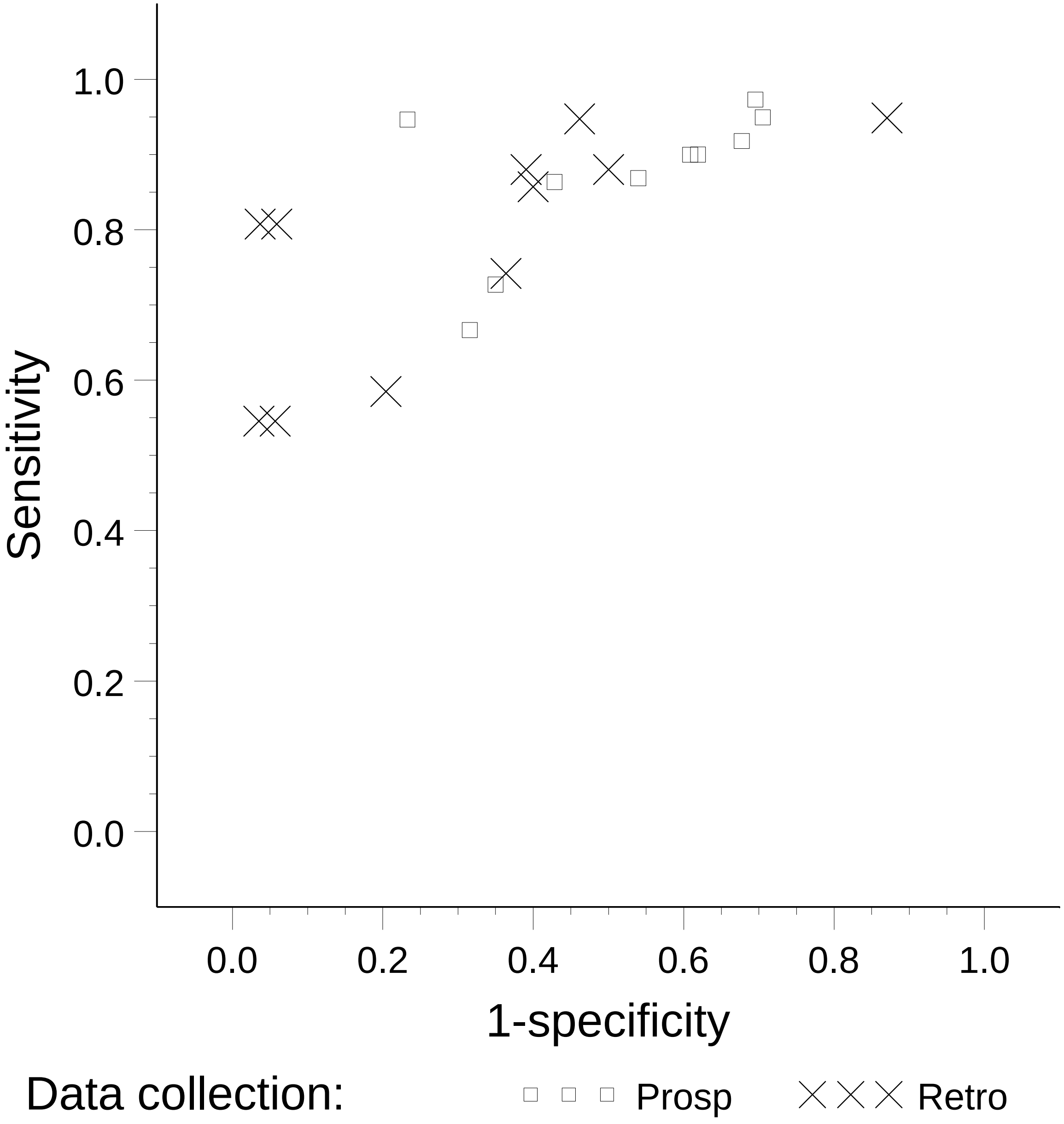
Subgroup analyses



Subgroup 1:

- both sens & spec higher

Prospective vs. Retrospective studies



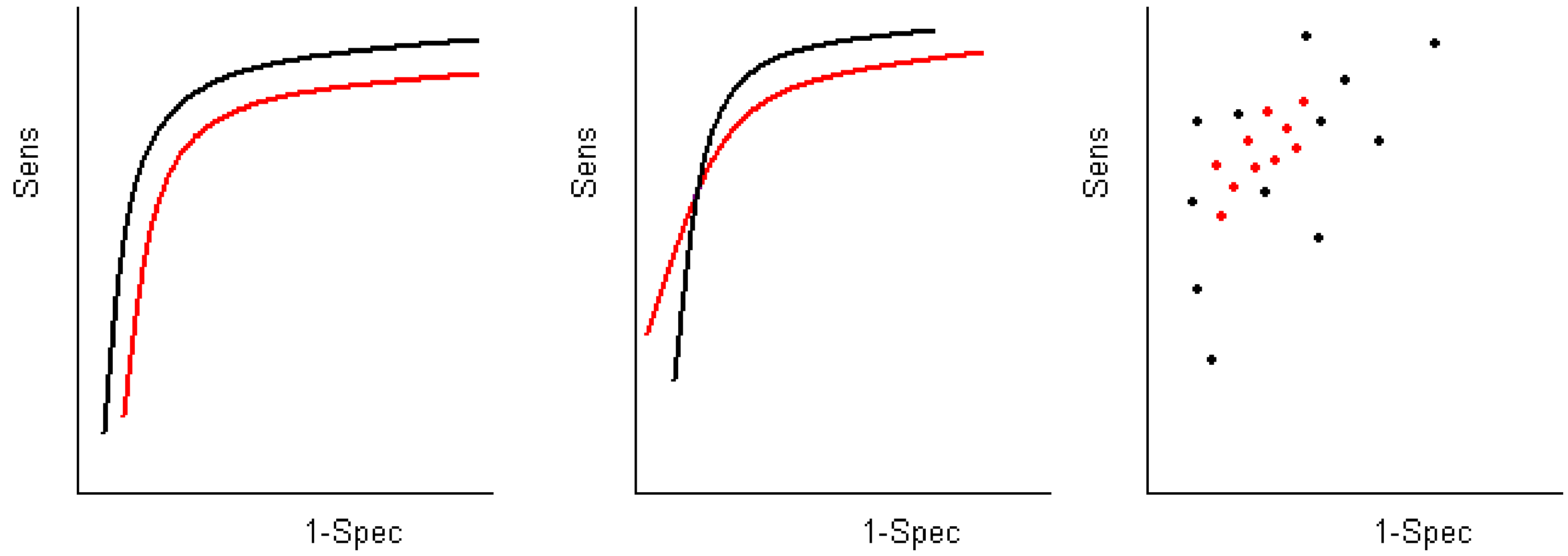
This may look easy, but…

- The following slides give the results of a study we did to incorporate the effects of quality into a meta-analysis.

Leeflang et al. Impact of adjustment for quality on results of metaanalyses of diagnostic accuracy. Clin Chem. 2007;53:164-72.

Effects of high/low Q?

1. Change in DOR
2. Change in consistency of DOR
3. Change in heterogeneity



Hypotheses

Deficiencies in study quality have been associated with inflated estimates and with heterogeneity.

Accounting for quality differences will therefore lead to ...

- ... less optimistic summary estimates.
- ... more homogenous results.

Incorporation Strategies

1. **Ignoring (sometimes graphs are shown)**
   pooling all studies, disregarding quality

2. **Subgroup Analysis**
   also: quality as criterion for inclusion
   also: stratification → more than one subgroup
   also: sensitivity analysis

3. **Regression analysis**
   Stepwise multivariable regression analysis and
   Multivariable regression analysis with a fixed set of covariates

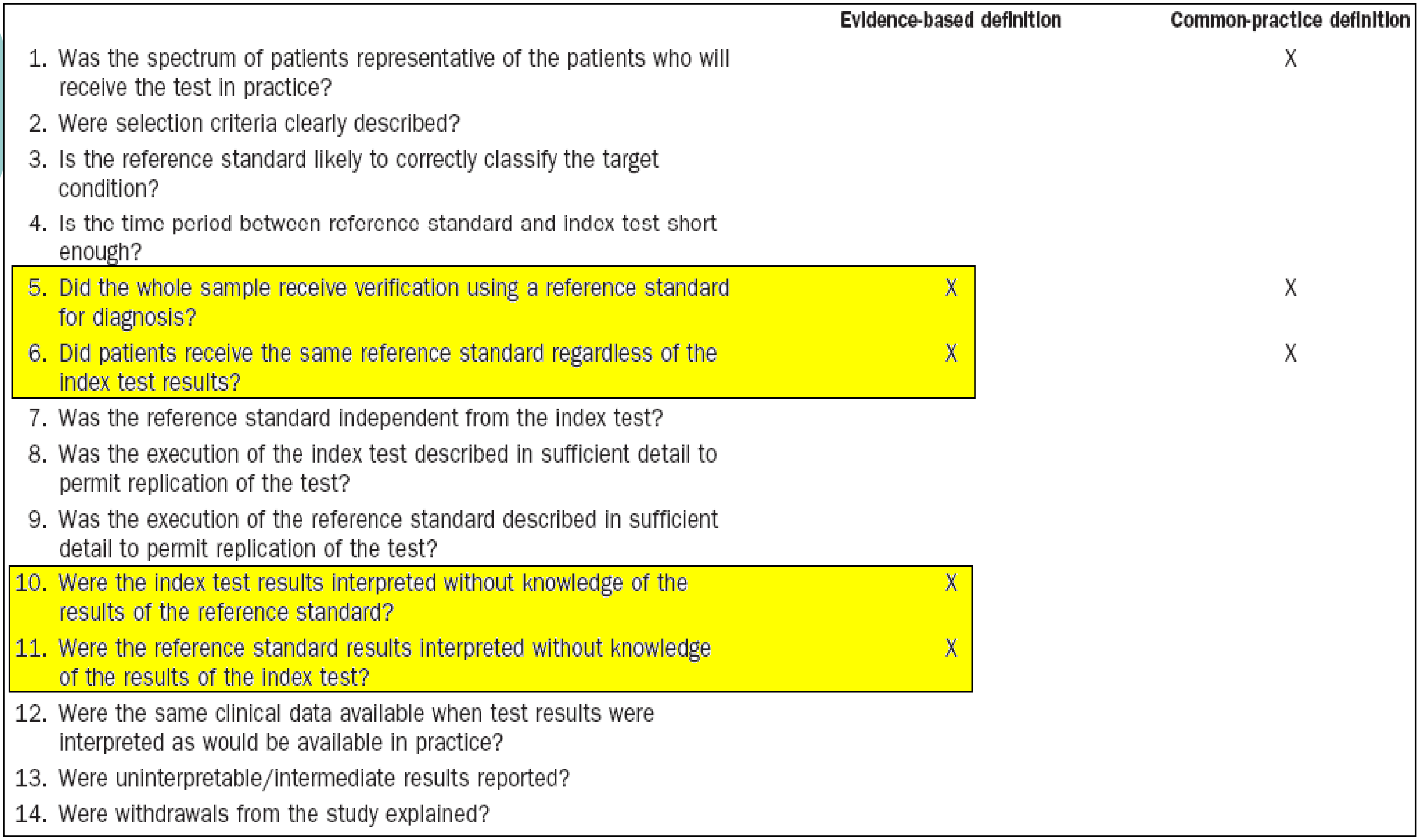
4. **Weighted pooling**
   'not done'

5. **Sequential analysis**
   highest quality → → → lowest quality
   cumulative meta-analysis

Methods

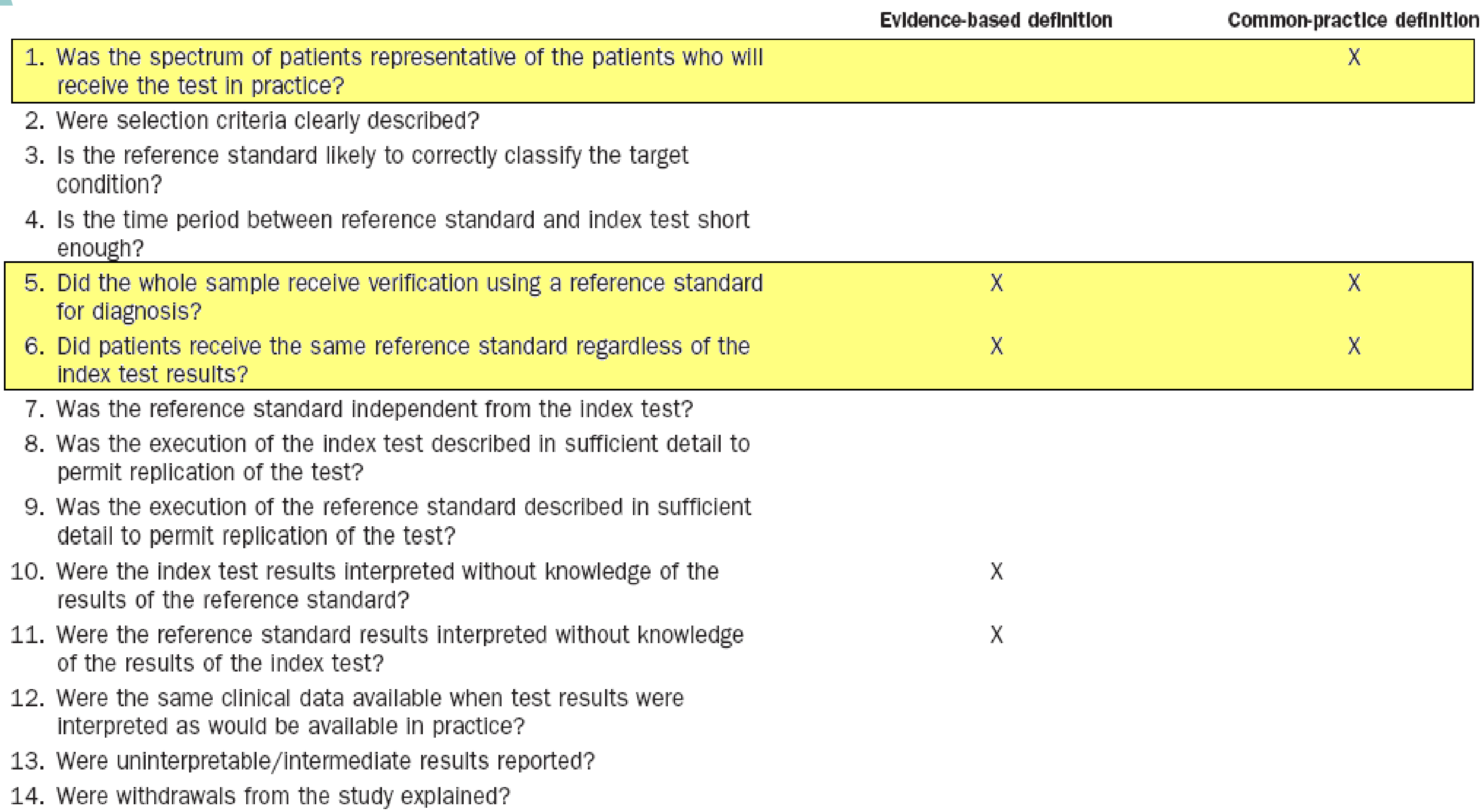
- Quality assessment in 487 studies included in 30 systematic reviews.
- QUADAS checklist used (Whiting et al. BMC Med Res Methodol, 2003)
- Two definitions for high-quality:
  1. Evidence-based definition
  2. Common practice definition
- Three methods for incorporation of quality:
  1. Exclusion of low quality studies
  2. Multivariable regression analysis with all items involved
  3. Stepwise multivariable regression analysis (p>0.2)
- Comparison of DORs, 95% CI of DORs, and changes in a hypothetical decision.

Evidence-based definition



	Evidence-based definition	Common-practice definition
1. Was the spectrum of patients representative of the patients who will receive the test in practice?		X
2. Were selection criteria clearly described?		
3. Is the reference standard likely to correctly classify the target condition?		
4. Is the time period between reference standard and index test short enough?		
5. Did the whole sample receive verification using a reference standard for diagnosis?	X	X
6. Did patients receive the same reference standard regardless of the index test results?	X	X
7. Was the reference standard independent from the index test?		
8. Was the execution of the index test described in sufficient detail to permit replication of the test?		
9. Was the execution of the reference standard described in sufficient detail to permit replication of the test?		
10. Were the index test results interpreted without knowledge of the results of the reference standard?	X	
11. Were the reference standard results interpreted without knowledge of the results of the index test?	X	
12. Were the same clinical data available when test results were interpreted as would be available in practice?		
13. Were uninterpretable/intermediate results reported?		
14. Were withdrawals from the study explained?		

Common practice definition



	Evidence-based definition	Common-practice definition
1. Was the spectrum of patients representative of the patients who will receive the test in practice?		X
2. Were selection criteria clearly described?		
3. Is the reference standard likely to correctly classify the target condition?		
4. Is the time period between reference standard and index test short enough?		
5. Did the whole sample receive verification using a reference standard for diagnosis?	X	X
6. Did patients receive the same reference standard regardless of the index test results?	X	X
7. Was the reference standard independent from the index test?		
8. Was the execution of the index test described in sufficient detail to permit replication of the test?		
9. Was the execution of the reference standard described in sufficient detail to permit replication of the test?		
10. Were the index test results interpreted without knowledge of the results of the reference standard?	X	
11. Were the reference standard results interpreted without knowledge of the results of the index test?	X	
12. Were the same clinical data available when test results were interpreted as would be available in practice?		
13. Were uninterpretable/intermediate results reported?		
14. Were withdrawals from the study explained?		

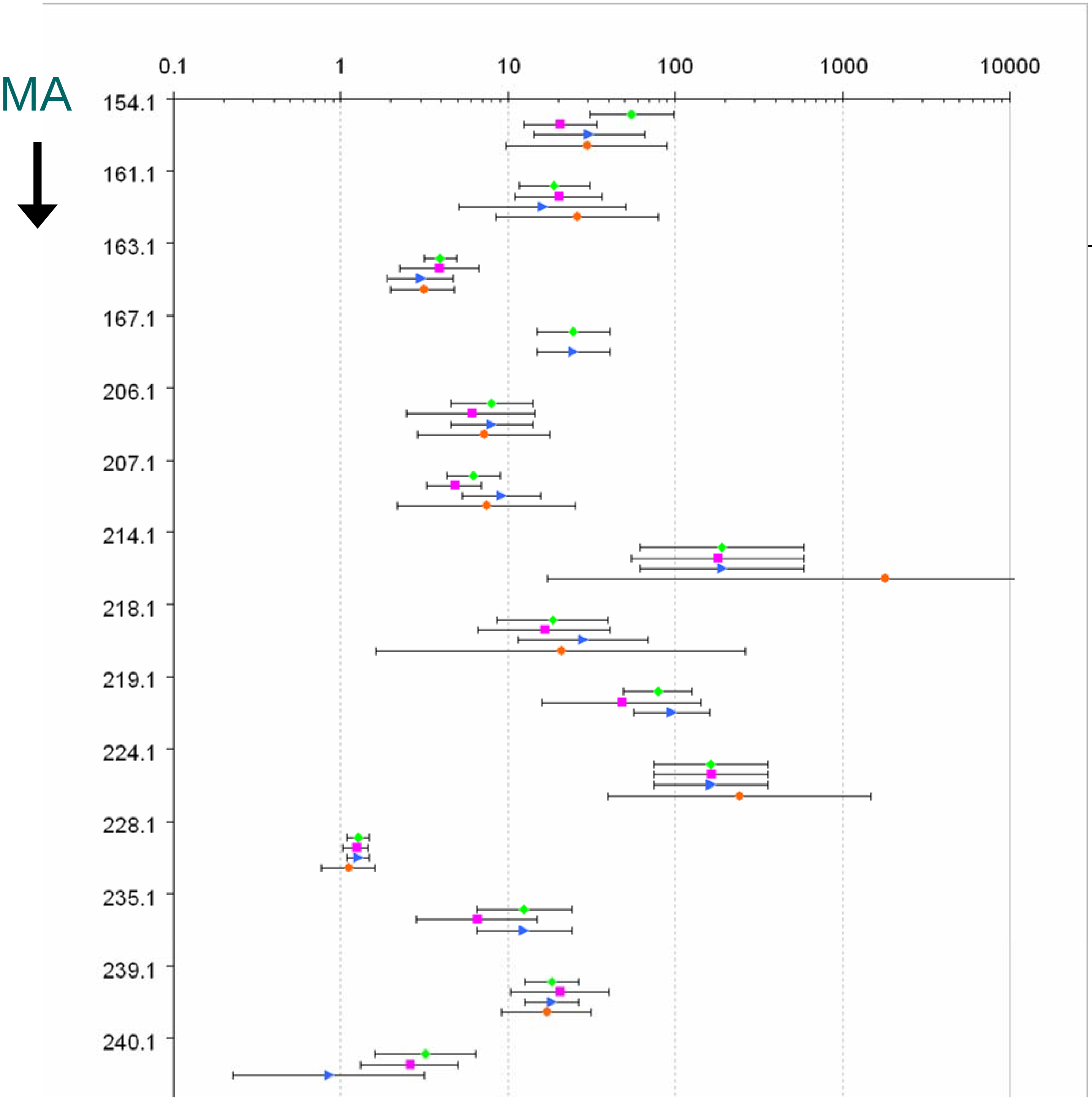
Results

- Nonreporting of items was common, especially for blinding of index or reference test; time-interval between index test and reference test; and about inclusion of patients.
- Evidence-based definition: 72 high quality studies (15%); 12 reviews contained no high-quality studies.
- Common-practice definition: 70 high quality studies (14%); 9 reviews contained no high-quality studies.
- Fulfilling all 8 criteria: only 10 out of 487 studies were of high quality and only 1 meta-analysis out of 31 contained more than 3 high-quality studies…

The Strategies

◆	*Ignoring quality:*	Pooling all studies
■	*Analyzing subgroups:*	Only pooling high-quality studies; high quality defined as fulfilling a certain subset of criteria.
▲	*Stepwise multivariable regression analysis:*	QUADAS-items with a p-value <0.2 univariate are entered in a multivariable regression model
●	*Multivariable regression analysis with a set of covariates:*	A standard set of three QUADAS-items was used as covariates in each meta-analysis.

ID MA


0.1
1
10
100
1000
10000
154.1
161.1
163.1
167.1
206.1
207.1
214.1
218.1
219.1
224.1
228.1
235.1
239.1
240.1

DOR

Conclusions?

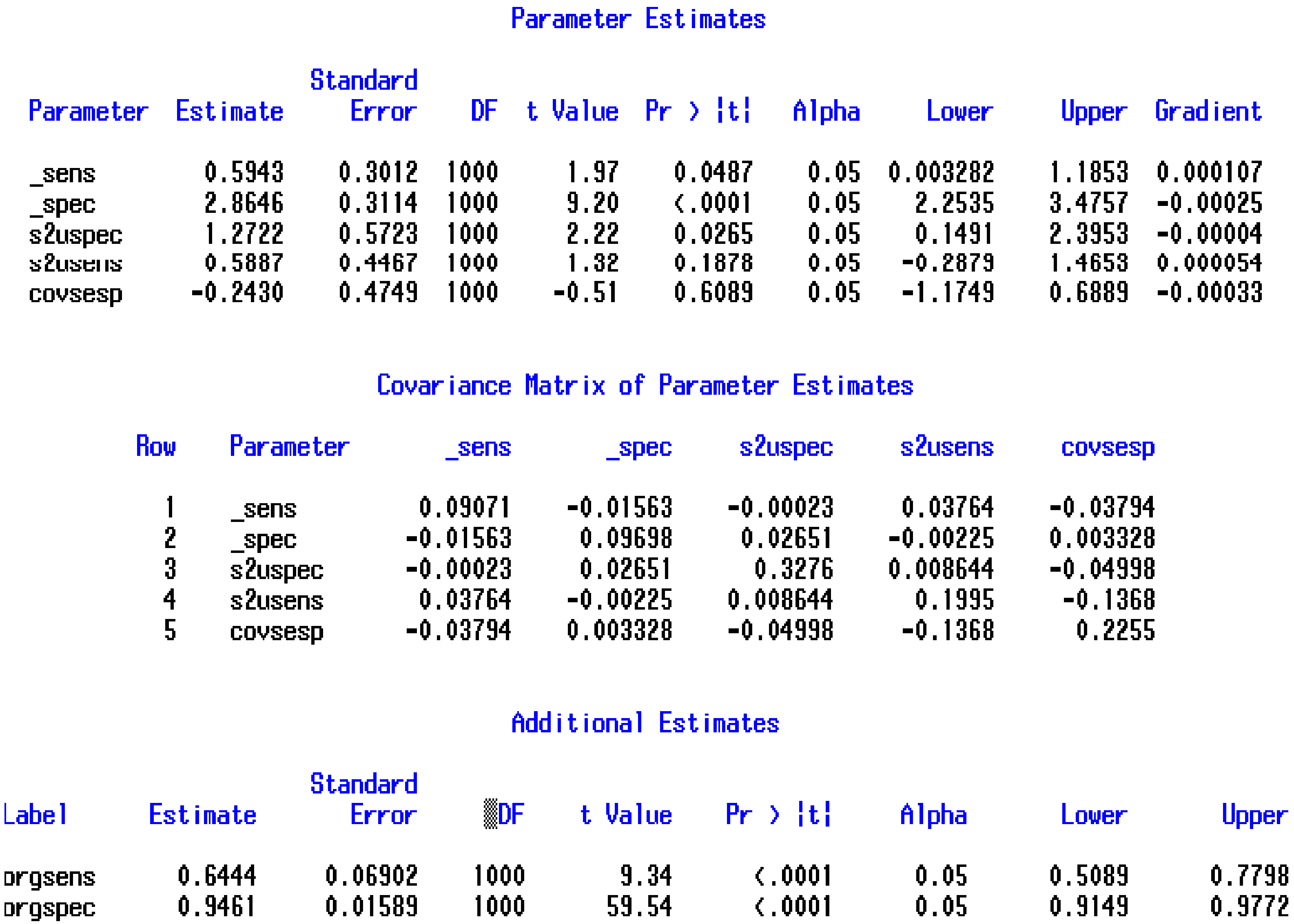
We found no evidence for our hypothesis that adjusting for quality leads to less optimistic and more homogenous results.

Explanations:	Poor reporting
	Small sample size (30 SRs, small studies)
	Opposite effects of quality items
	DOR in stead of sensitivity and specificity
	Relation quality – estimates not straightforward

Still, poor quality will affect the trustworthiness. Therefore, report quality of individual studies and overall quality.

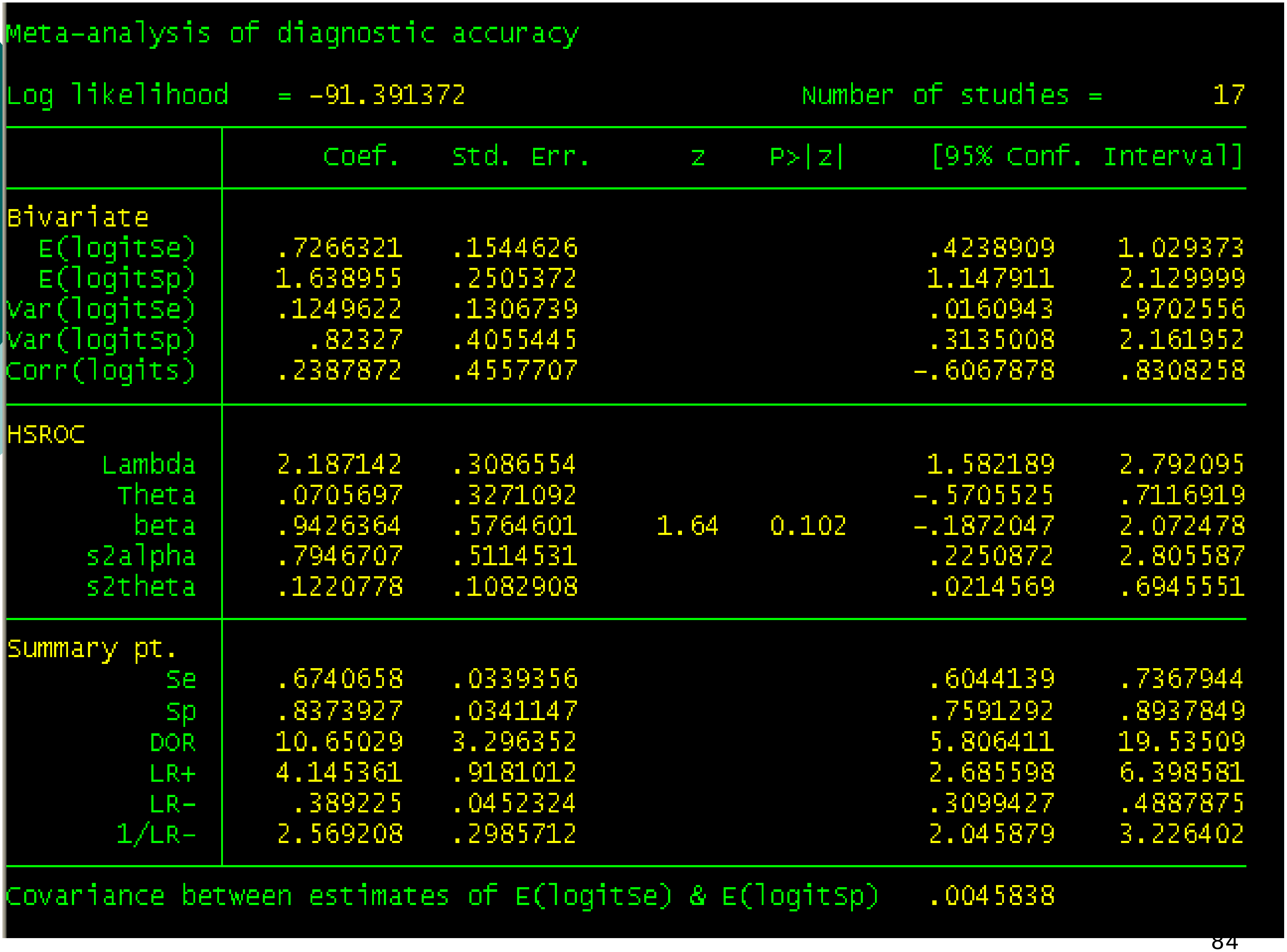
Exercise

- What do the results of a meta-analysis mean...?
- I have some Output from SAS and STATA and would like to invite you to have a look at them.



```
                              Parameter Estimates

                         Standard
Parameter    Estimate       Error     DF   t Value   Pr > |t|   Alpha      Lower      Upper   Gradient

_sens          0.5943      0.3012   1000      1.97     0.0487    0.05   0.003282     1.1853   0.000107
_spec          2.8646      0.3114   1000      9.20     <.0001    0.05     2.2535     3.4757   -0.00025
s2uspec        1.2722      0.5723   1000      2.22     0.0265    0.05     0.1491     2.3953   -0.00004
s2usens        0.5887      0.4467   1000      1.32     0.1878    0.05    -0.2879     1.4653   0.000054
covsesp       -0.2430      0.4749   1000     -0.51     0.6089    0.05    -1.1749     0.6889   -0.00033


                     Covariance Matrix of Parameter Estimates

       Row    Parameter        _sens        _spec      s2uspec      s2usens      covsesp

         1    _sens          0.09071     -0.01563     -0.00023      0.03764     -0.03794
         2    _spec         -0.01563      0.09698      0.02651     -0.00225     0.003328
         3    s2uspec       -0.00023      0.02651       0.3276     0.008644     -0.04998
         4    s2usens        0.03764     -0.00225     0.008644       0.1995      -0.1368
         5    covsesp       -0.03794     0.003328     -0.04998      -0.1368       0.2255


                              Additional Estimates

                          Standard
Label        Estimate        Error      DF    t Value    Pr > |t|     Alpha       Lower       Upper

orgsens        0.6444      0.06902    1000       9.34      <.0001      0.05      0.5089      0.7798
orgspec        0.9461      0.01589    1000      59.54      <.0001      0.05      0.9149      0.9772
```

Bivariate or HSROC?   What do the parameters mean?



```
Meta-analysis of diagnostic accuracy

Log likelihood   = -91.391372                Number of studies =       17
-------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
Bivariate    |
  E(logitSe) |   .7266321   .1544626                       .4238909   1.029373
  E(logitSp) |   1.638955   .2505372                       1.147911   2.129999
Var(logitSe) |   .1249622   .1306739                       .0160943   .9702556
Var(logitSp) |     .82327   .4055445                       .3135008   2.161952
Corr(logits) |   .2387872   .4557707                      -.6067878   .8308258
-------------------------------------------------------------------------------
HSROC        |
      Lambda |   2.187142   .3086554                       1.582189   2.792095
       Theta |   .0705697   .3271092                      -.5705525   .7116919
        beta |   .9426364   .5764601     1.64   0.102     -.1872047   2.072478
     s2alpha |   .7946707   .5114531                       .2250872   2.805587
     s2theta |   .1220778   .1082908                       .0214569   .6945551
-------------------------------------------------------------------------------
Summary pt.  |
          Se |   .6740658   .0339356                       .6044139   .7367944
          Sp |   .8373927   .0341147                       .7591292   .8937849
         DOR |   10.65029   3.296352                       5.806411   19.53509
         LR+ |   4.145361   .9181012                       2.685598   6.398581
         LR- |    .389225   .0452324                       .3099427   .4887875
       1/LR- |   2.569208   .2985712                       2.045879   3.226402
-------------------------------------------------------------------------------
Covariance between estimates of E(logitSe) & E(logitSp)    .0045838
```

Part 3

Test Comparisons

Differences between tests

- Diagnosis of lymph node metastasis in women with cervical cancer

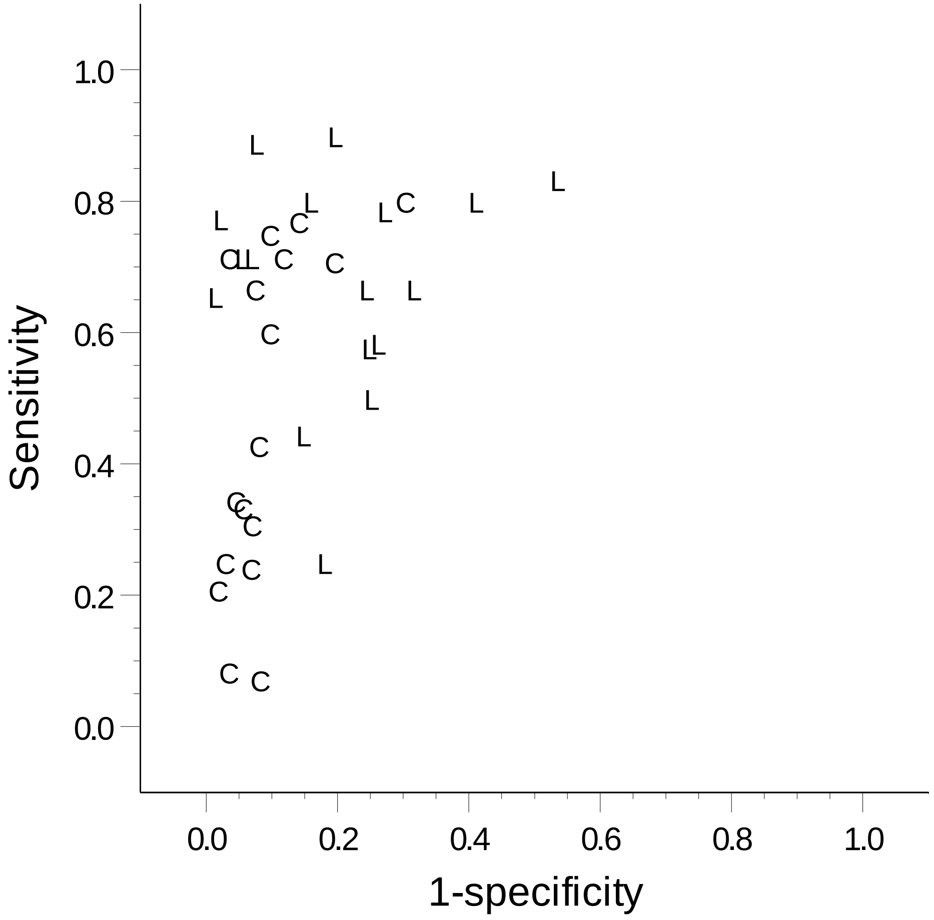
- 2 imaging modalities:
  - lymphangiography (LAG, n=17)
  - CT (n=17)

- Published meta-analysis JAMA 1997;278:1096-1101

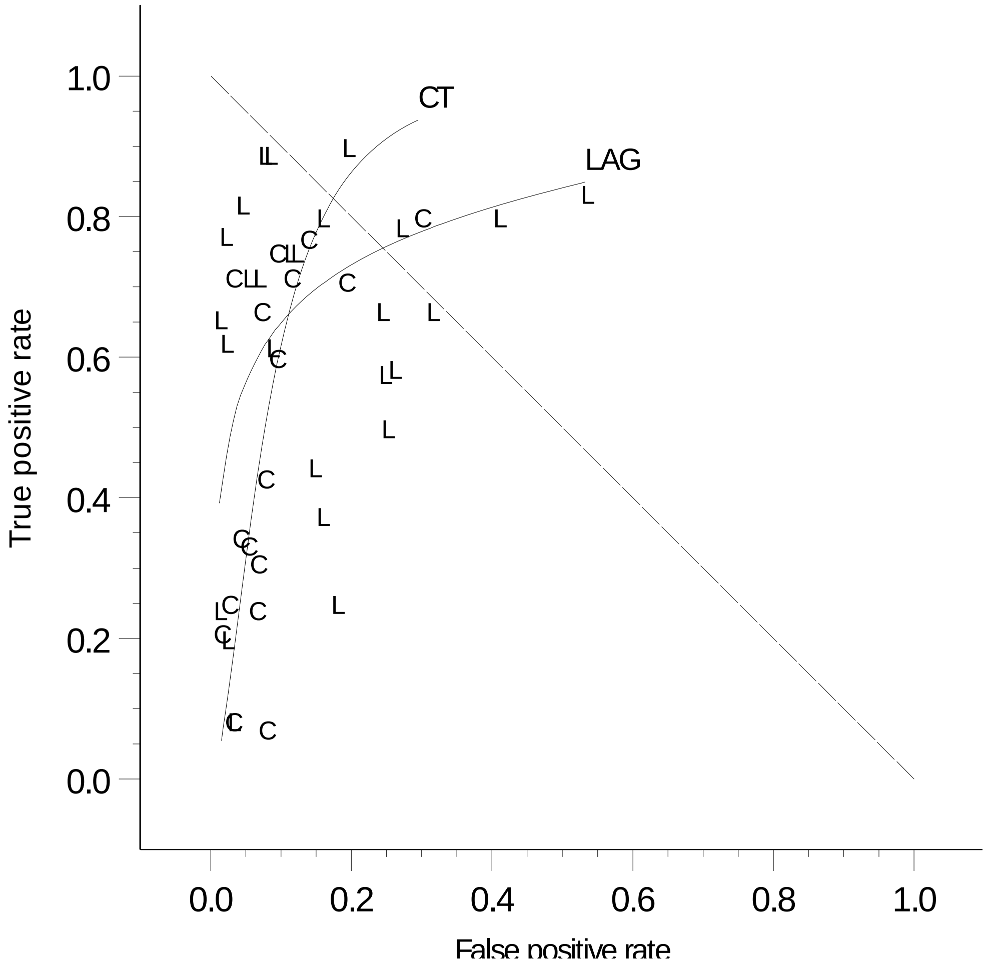
- Modelled by adding covariate for test into the model statement, and parameter estimates for differences in:

  - Sensitivity and specificity for bivariate
  - Log DOR, threshold and shape for HSROC

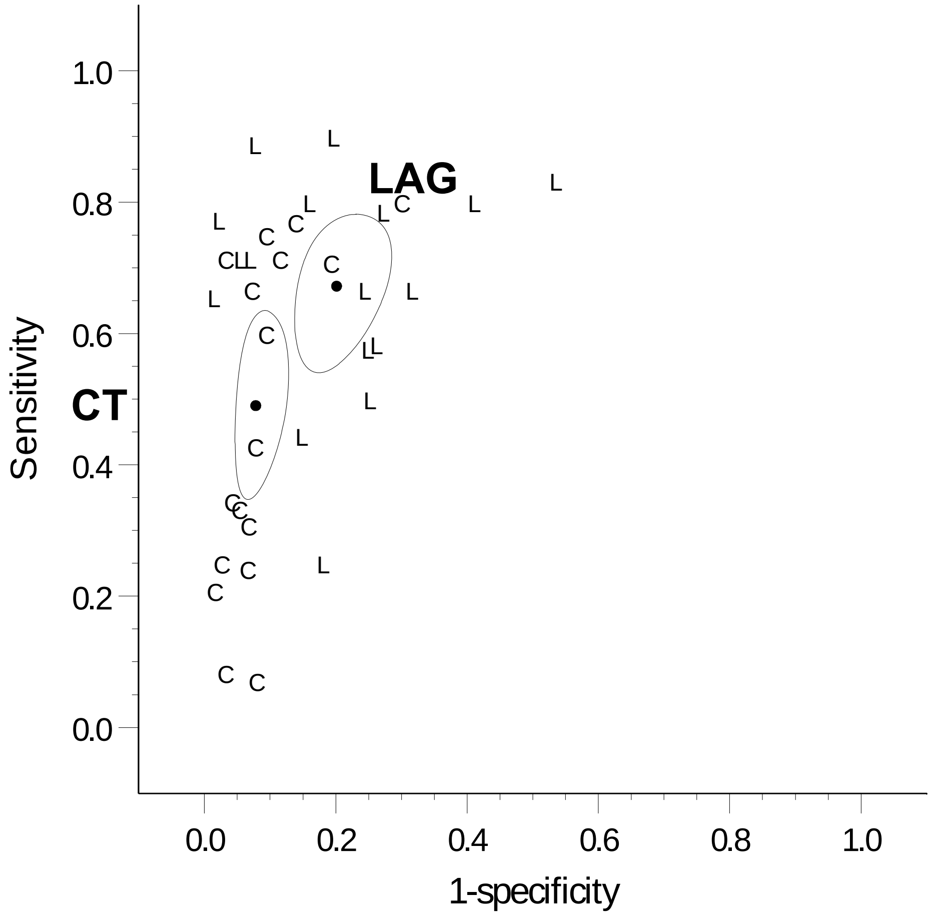
ROC plot of individual study results (L=lymphangiography C=CT)



Summary ROC estimates



Average operating points and confidence ellipses



Difference between average operating points

Imaging modality	Sensitivity (95% CI)	Specificity (95% CI)
LAG	0.67 (0.57 to 0.76)	0.80 (0.73 to 0.85)
CT	0.49 (0.37 to 0.61)	0.92 (0.88 to 0.95)
P-value Lag vs. CT	0.023	0.0002

Summary points or SROC curves?

- Clinical interpretation
  - Need to estimate performance at a threshold, using sensitivity, specificity or/and likelihood ratios

- Single threshold or mixed thresholds?
  - Summary curve describes how test performance varies across thresholds. Studies do not need to report a common threshold to contribute.
  - Summary point must relate to a particular threshold. Only studies reporting a common threshold can be combined.

Summary points or SROC curves?

- Comparing tests and subgroups
  - Often wish to use as much data as possible –
    - if this means mixing thresholds SROC curves are needed
    - if still a common threshold either method appropriate
  - Possible to assess impact of threshold as a covariate
  - SROC curves allow identification of crossing lines

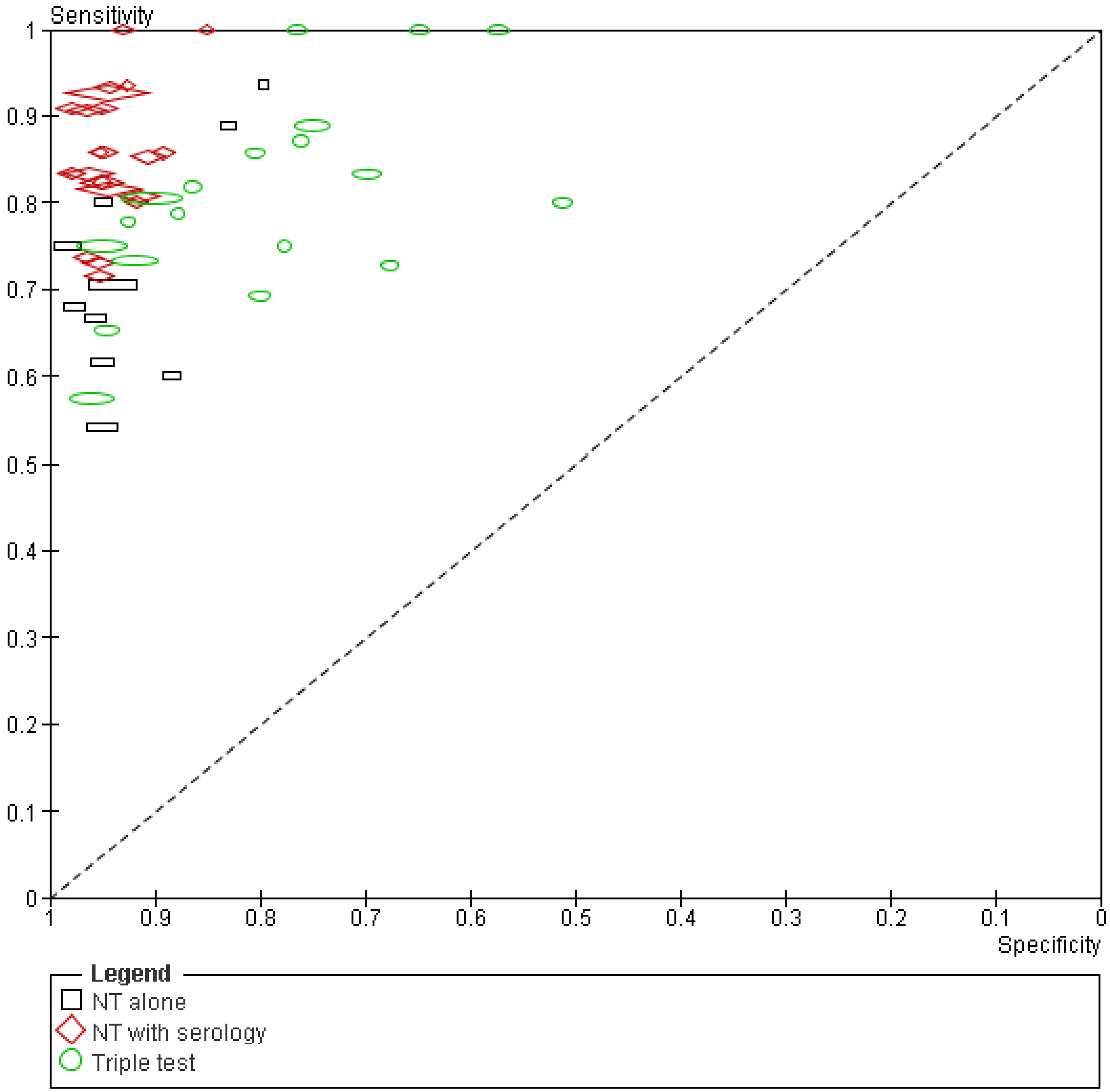
- A Cochrane review may include both an analysis of the SROC curves, and estimation of average threshold specific operating points

Comparative analyses

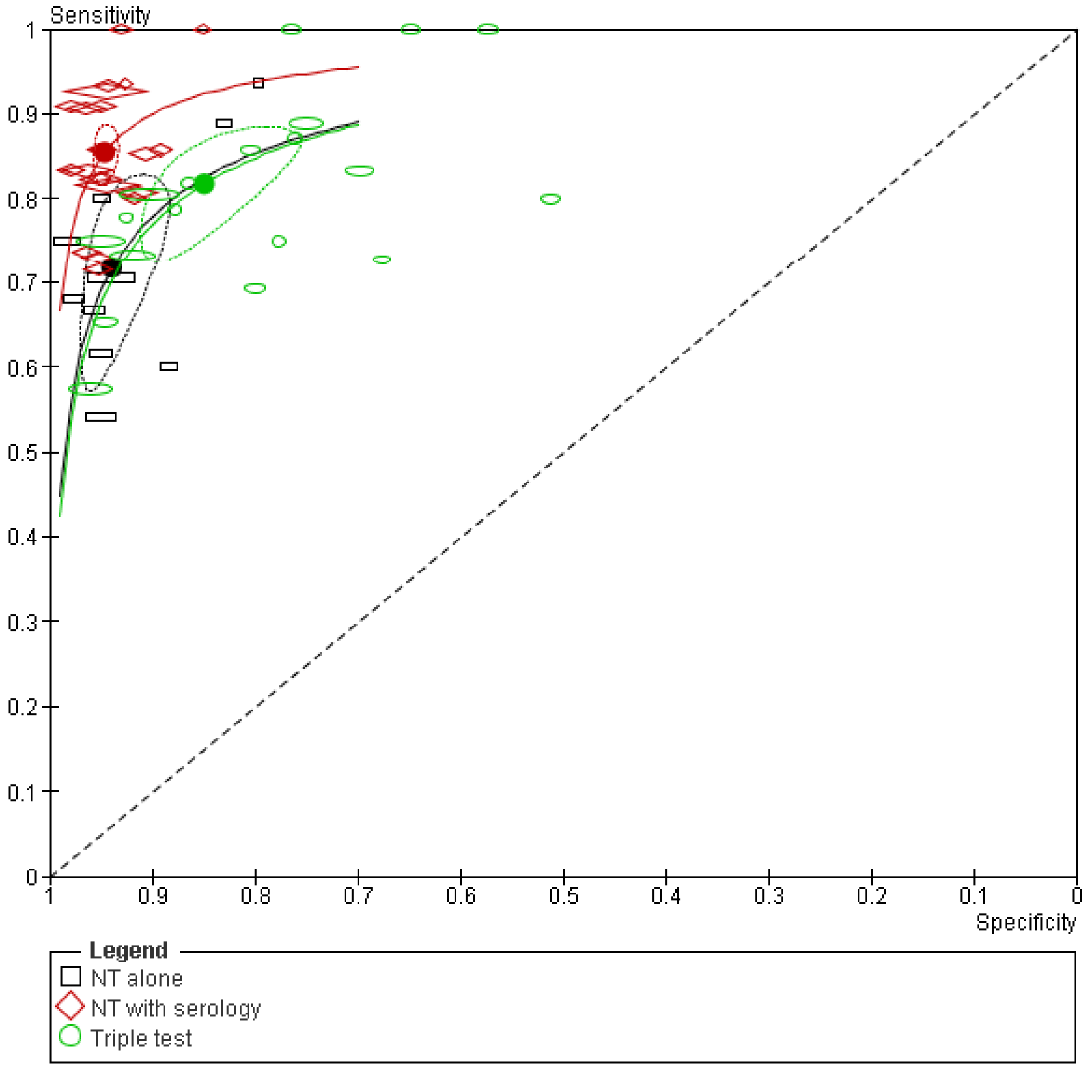
- Indirect comparisons
  - Different tests used in different studies
  - Potentially confounded by other differences between the studies
- Direct comparisons
  - Patients receive both tests or randomized to tests
  - Differences in accuracy more attributable to the tests
  - Few studies may be available and may not be representative

Example of pilot Cochrane Review Down' Syndrome screening review

	Studies	Participants
1st trimester - NT alone	10	79,412
1st trimester - NT and serology	22	222,171
2nd trimester - triple test (serology)	19	72,797


Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Specificity
Legend
□ NT alone
◇ NT with serology
○ Triple test

Indirect comparison



NT alone

Sensitivity: 72% (63%-79%)

Specificity: 94% (91% -96%)

DOR: 39 (26-60)

NT with serology

Sensitivity: 86% (82%-90%)

Specificity: 95% (93%-96%)

DOR: 110 (84-143)

RDOR: 2.8 (1.7-4.6),
p <0.0001

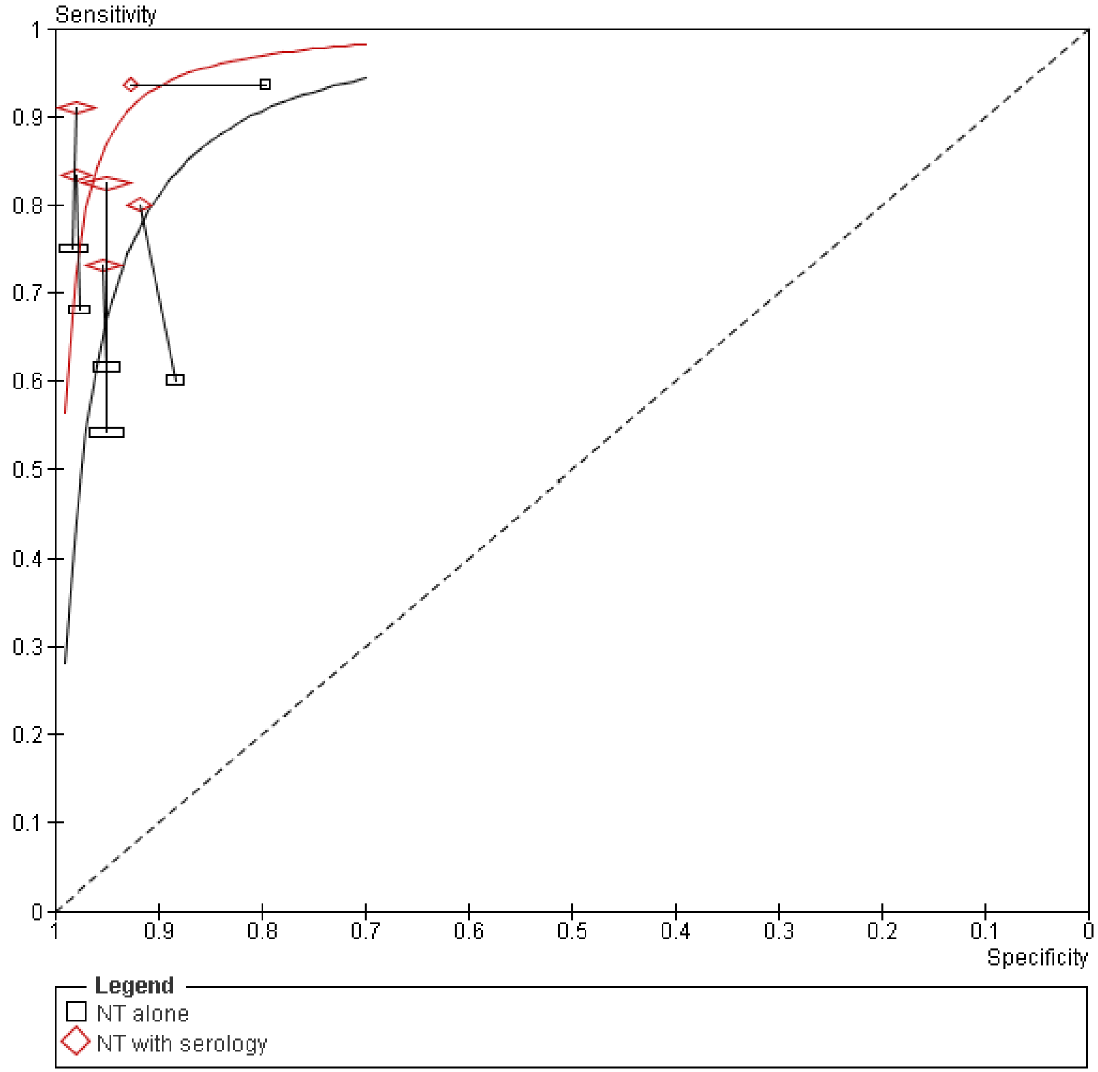
Triple test

Sensitivity: 82% (76%-86%)

Specificity: 83% (77%-87%)

DOR: 21 (15-30)

RDOR: 0.5 (0.3-0.9),
p = 0.03



DIRECT COMPARISONS

NT alone

Sensitivity: 71% (59%-82%)

Specificity: 95% (91%-98%)

DOR: 41 (16-67)

NT with serology

Sensitivity: 85% (77%-93%)

Specificity: 96% (93%-98%)

DOR: 123 (40-206)

Triple test

No paired studies available

Indirect versus Direct comparisons

NT alone

Sensitivity: 72% (63%-79%)

Specificity: 94% (91% -96%)

DOR: 39 (26-60)

NT with serology

Sensitivity: 86% (82%-90%)

Specificity: 95% (93%-96%)

DOR: 110 (84-143)

RDOR: 2.8 (1.7-4.6),
p <0.0001

NT alone

Sensitivity: 71% (59%-82%)

Specificity: 95% (91%-98%)

DOR: 41 (16-67)

NT with serology

Sensitivity: 85% (77%-93%)

Specificity: 96% (93%-98%)

DOR: 123 (40-206)

Part 4

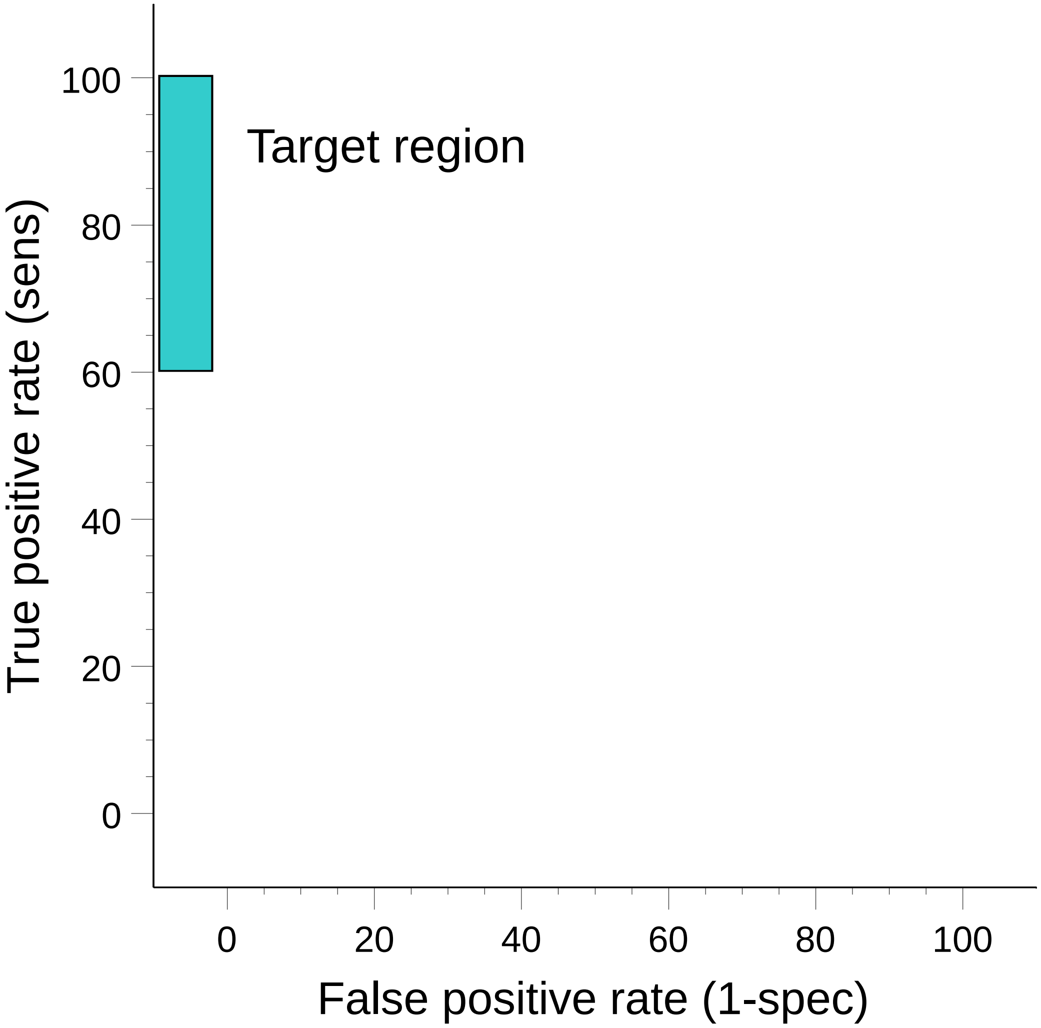
Some other issues

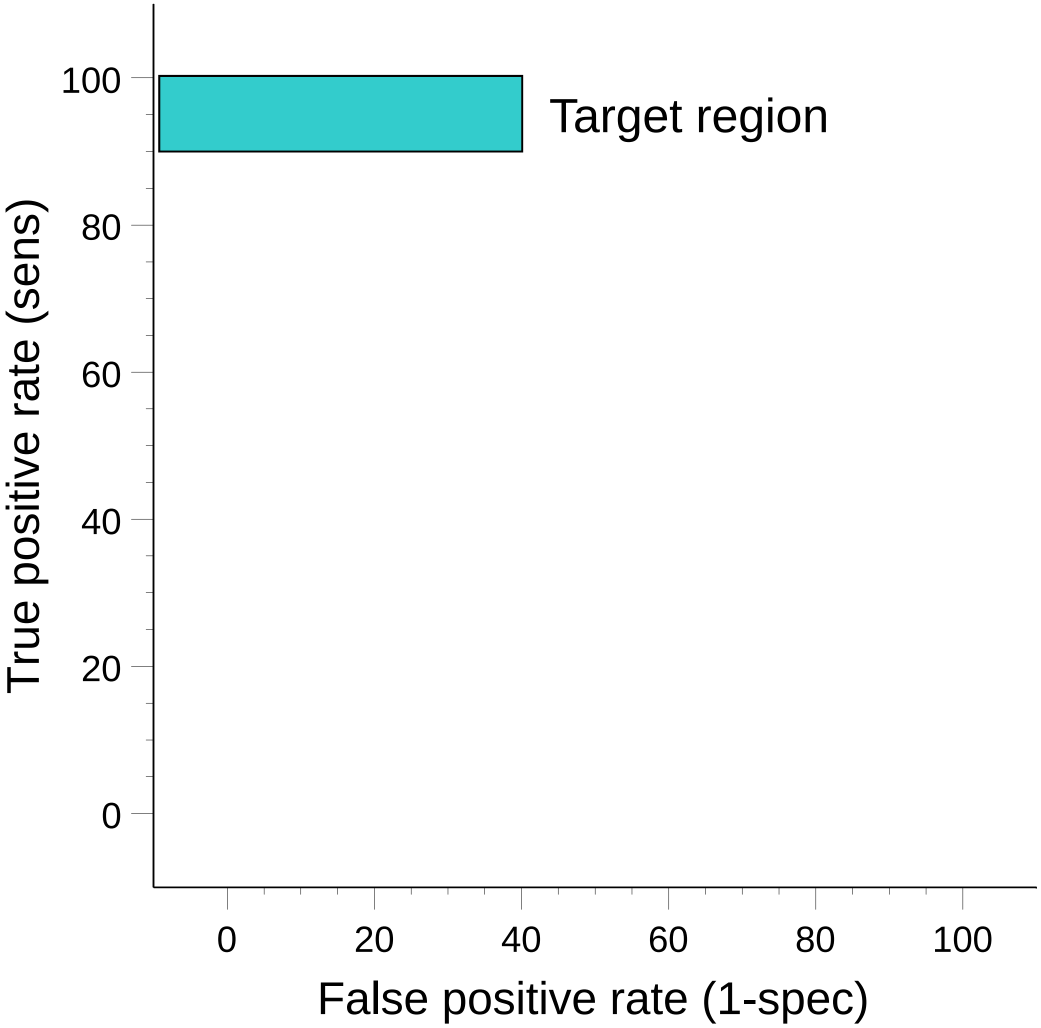
Another approach…

- Hypothesis testing is not common in diagnostic test accuracy research or in diagnostic meta-analyses.

- But you could test whether the studies you found or whether the summary estimate falls within a certain target region.

Target region




100
80
60
40
20
0
True positive rate (sens)
Target region
0
20
40
60
80
100
False positive rate (1-spec)

Publication bias

- In systematic reviews of intervention studies, publication bias is an important form of bias
- To investigate publication bias in reviews, funnel plots are used.
- In diagnostic reviews, funnel plots are seriously misleading and alternatives have poor power.

Publication bias - background

- many studies are done without ethical review or study registration → prospective registration is therefore not available

- diagnostic test accuracy studies do not test hypotheses, so there is no ‘significance’ involved

- we have no clue whether publication bias exists for diagnostic accuracy studies and how the mechanisms behind it may work

Summary

- Part 1: meta-analysis introduction
- Part 2: heterogeneity
- Part 3: test comparisons
- Part 4: some other issues