UNIVERSITY OF BIRMINGHAM

Introduction to diagnostic accuracy meta-analysis

Yemisi Takwoingi
October 2015

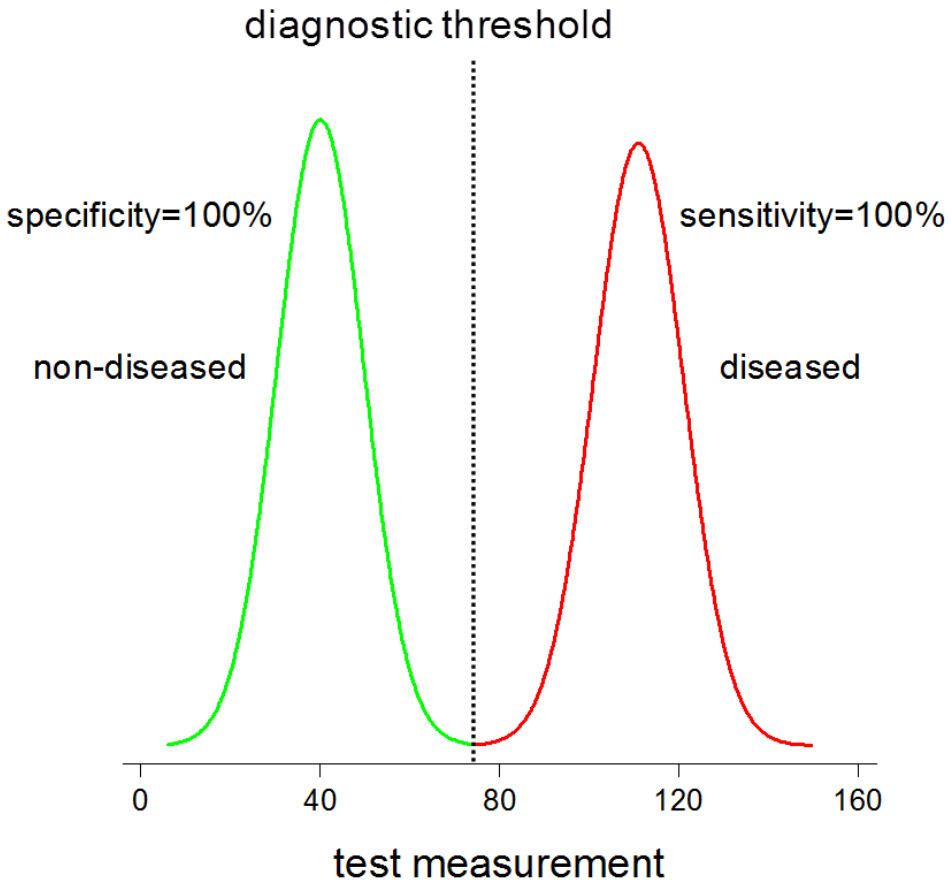
Learning objectives

- To appreciate the concept underlying DTA meta-analytic approaches
- To know the Moses-Littenberg SROC method and its limitations
- To understand the need for hierarchical models
- To know the recommended approaches
- To understand the fundamentals of the HSROC and bivariate models
- To be aware of approaches for investigation of heterogeneity and test comparisons

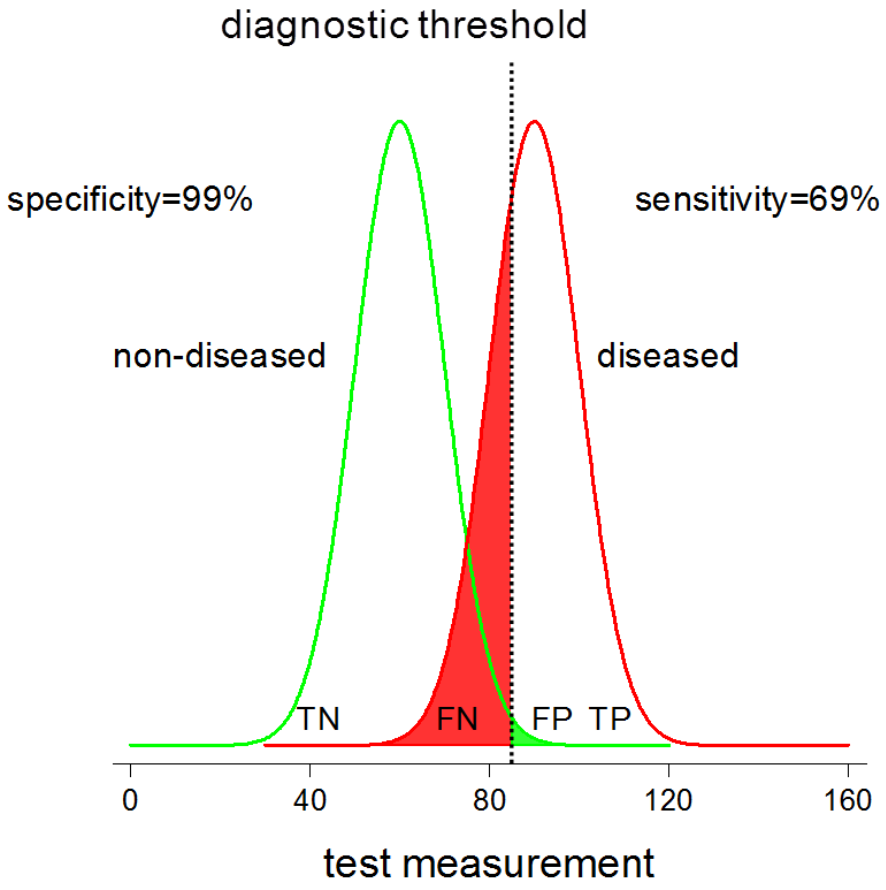
Outline

- Analysis of a single study – ROC curves
- Basic method for meta-analysis
- Hierarchical models
- Choice of method
- Data analysis in RevMan and external software

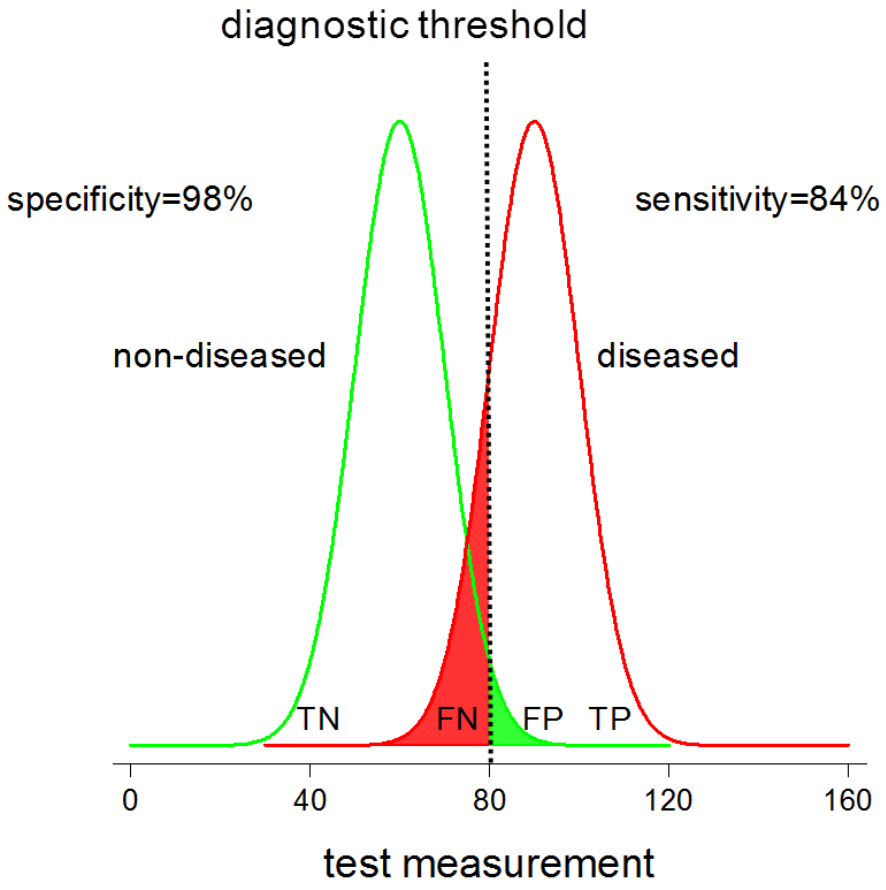
Heterogeneity in threshold within a study



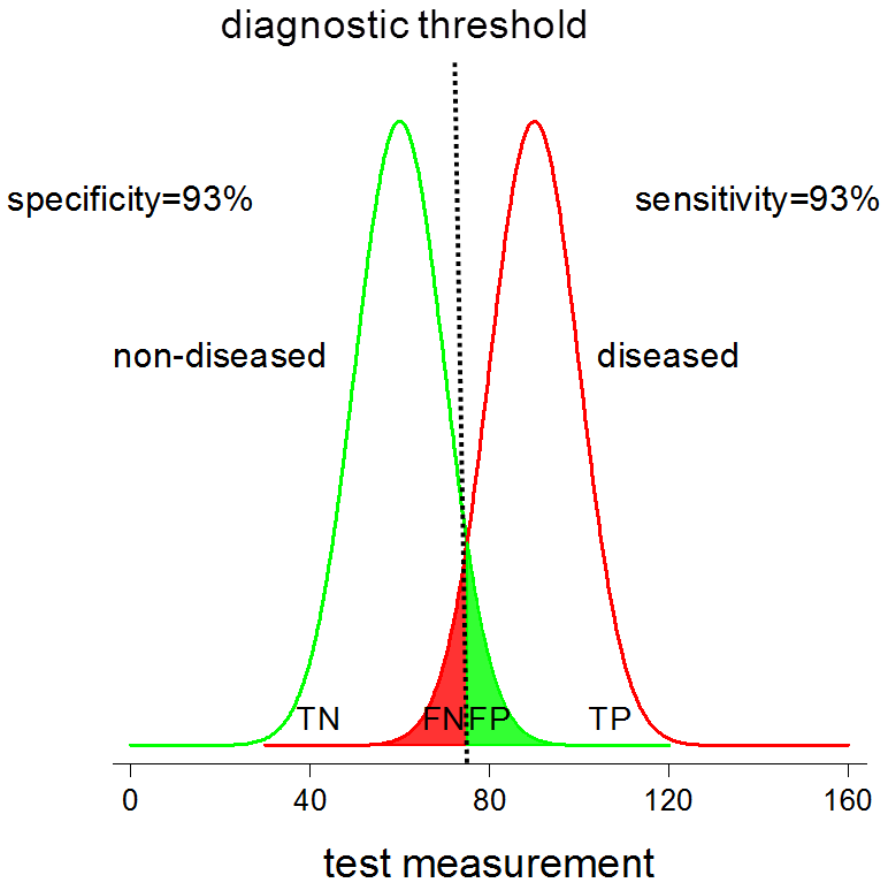
Heterogeneity in threshold within a study



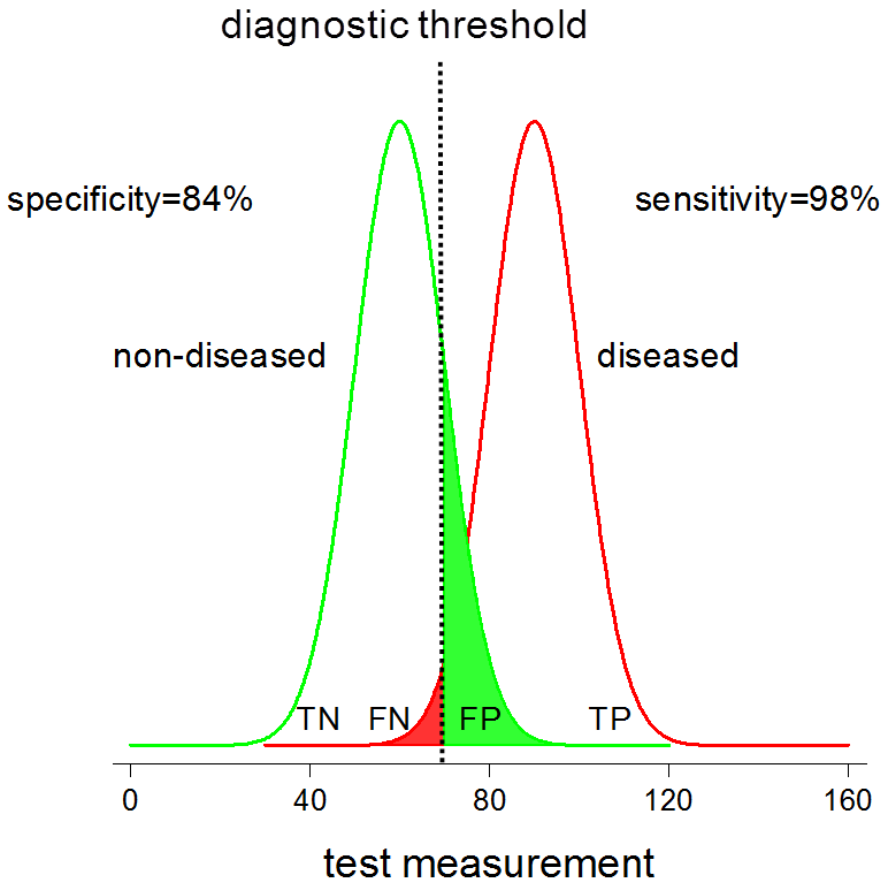
Heterogeneity in threshold within a study



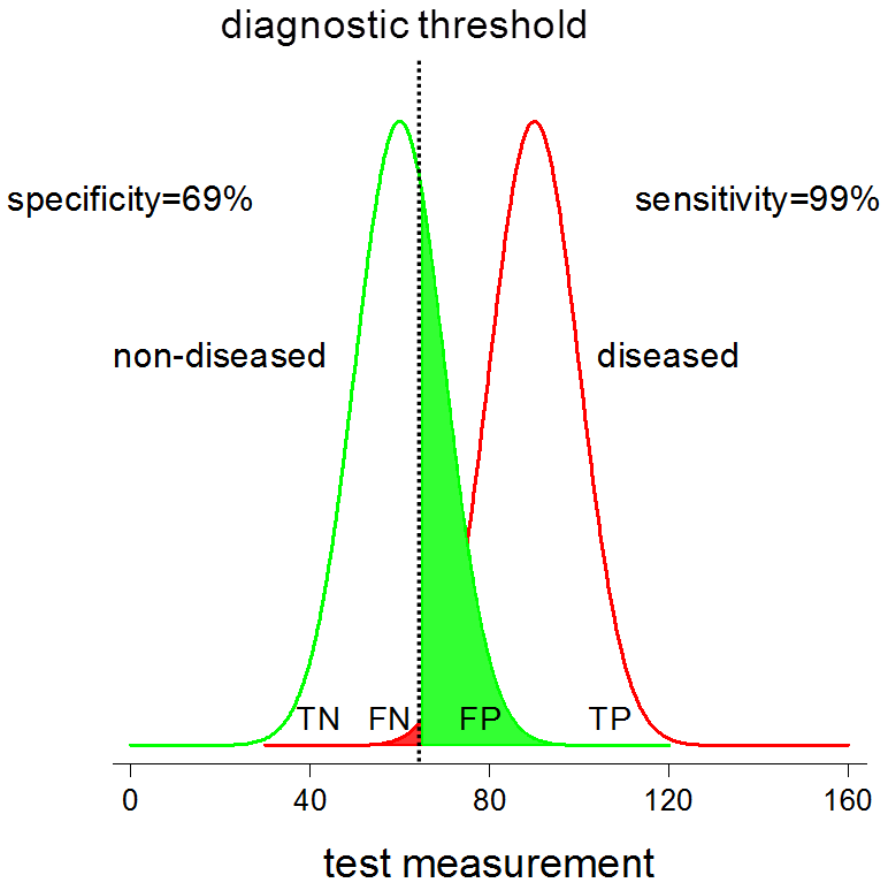
Heterogeneity in threshold within a study



Heterogeneity in threshold within a study

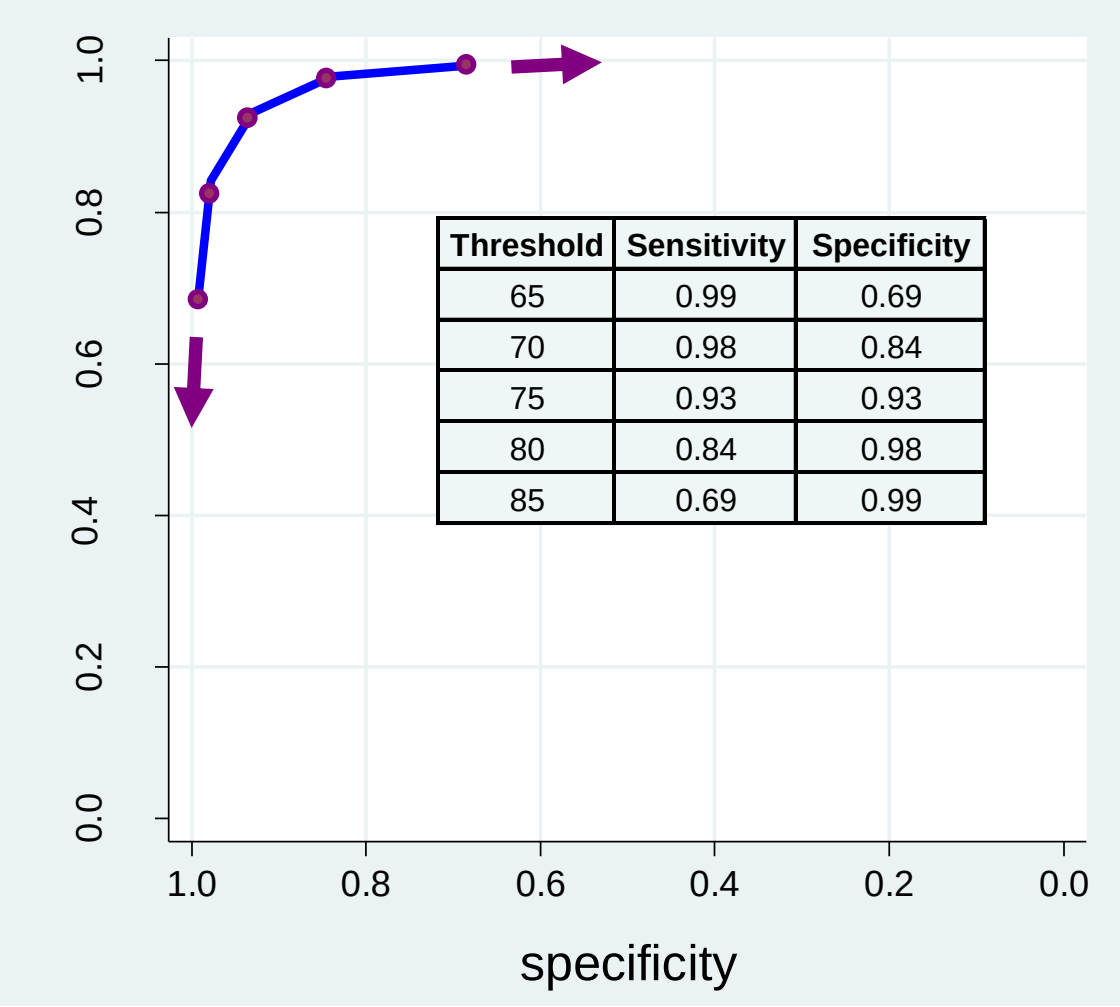


Heterogeneity in threshold within a study



Threshold effect

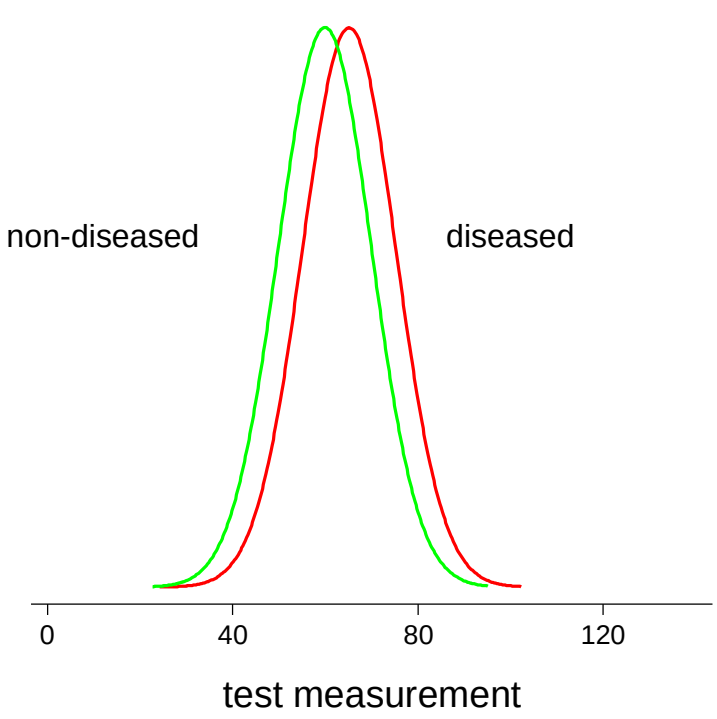
Increasing threshold decreases sensitivity but increases specificity

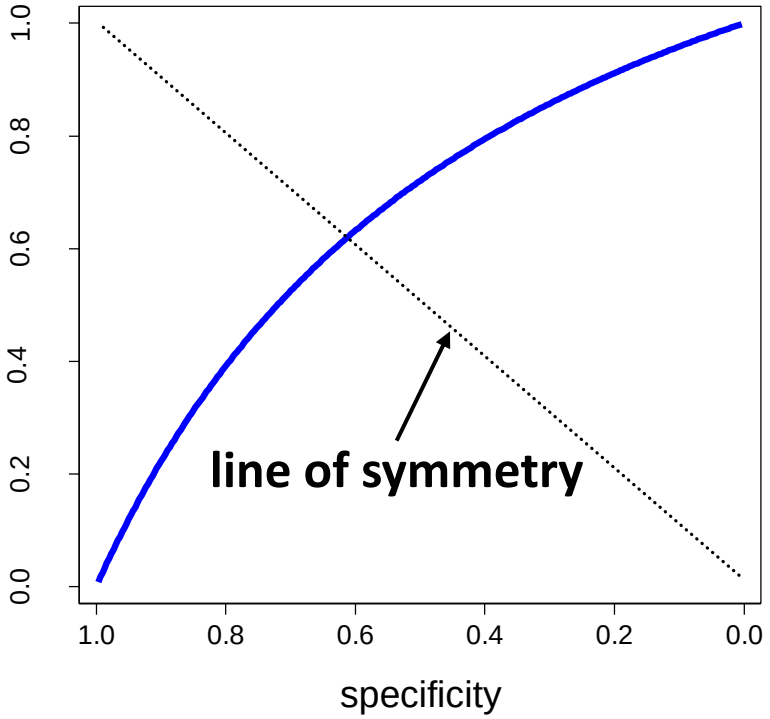


Threshold	Sensitivity	Specificity
65	0.99	0.69
70	0.98	0.84
75	0.93	0.93
80	0.84	0.98
85	0.69	0.99

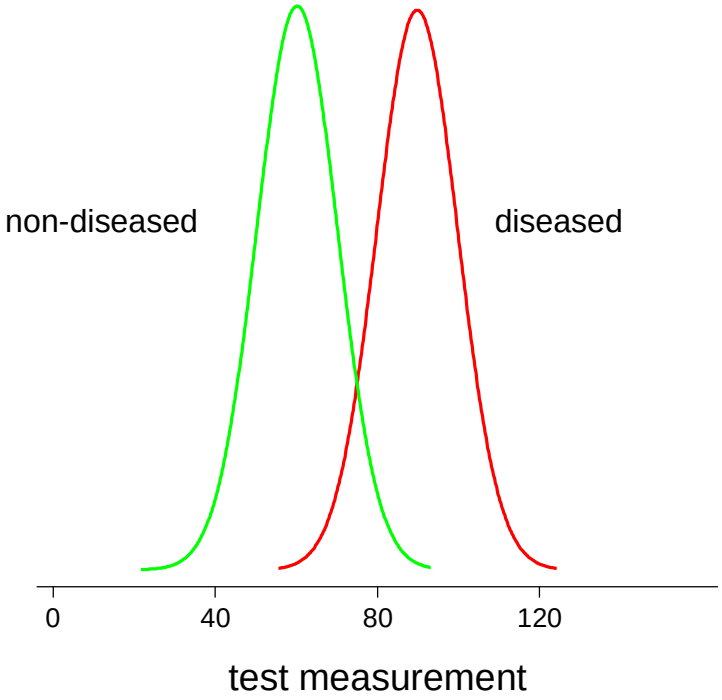
Decreasing threshold decreases specificity but increases sensitivity

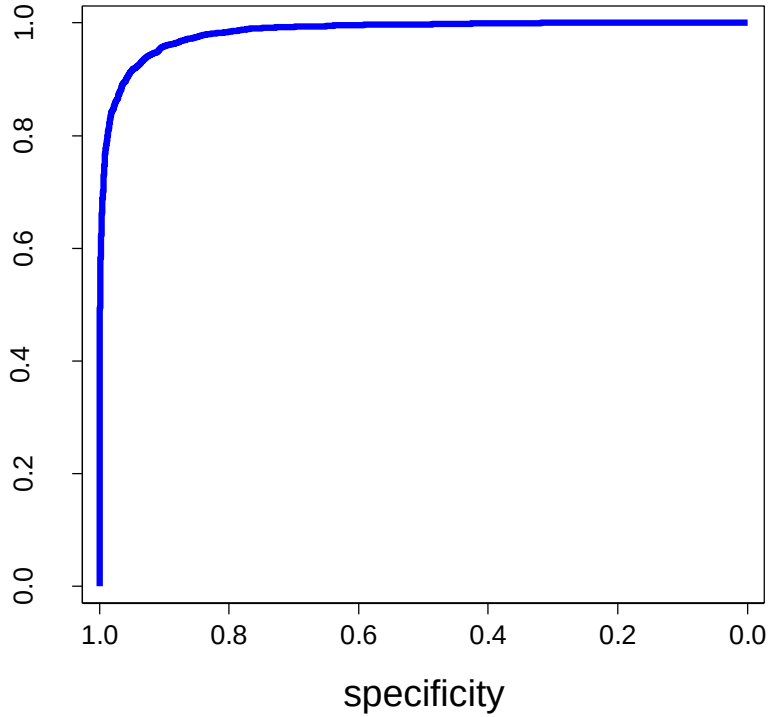
Distributions and ROC plot (small difference, same spread)





Distributions and ROC plot (large difference, same spread)





Diagnostic odds ratios

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

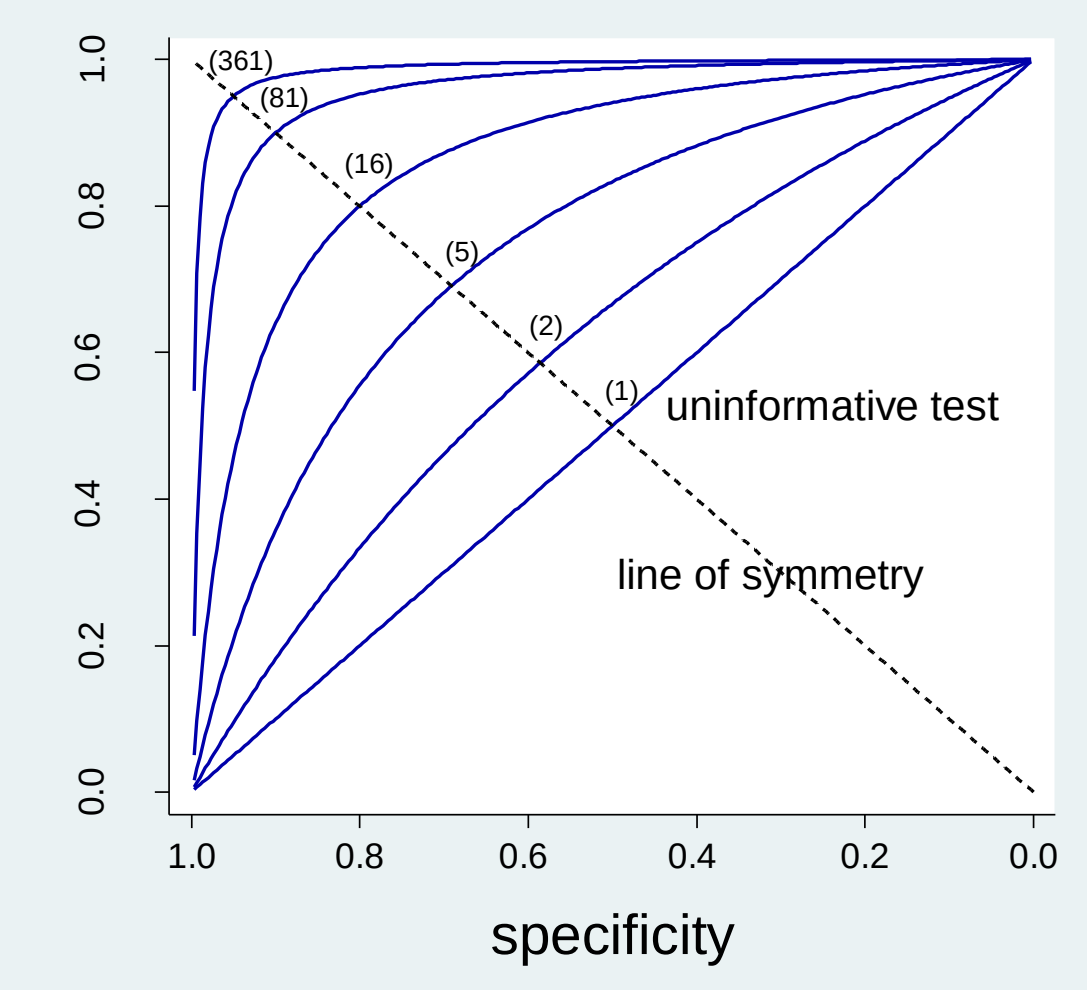
$$Diagnostic\ OR = \frac{TP \times TN}{FP \times FN}$$

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

Diagnostic odds ratios

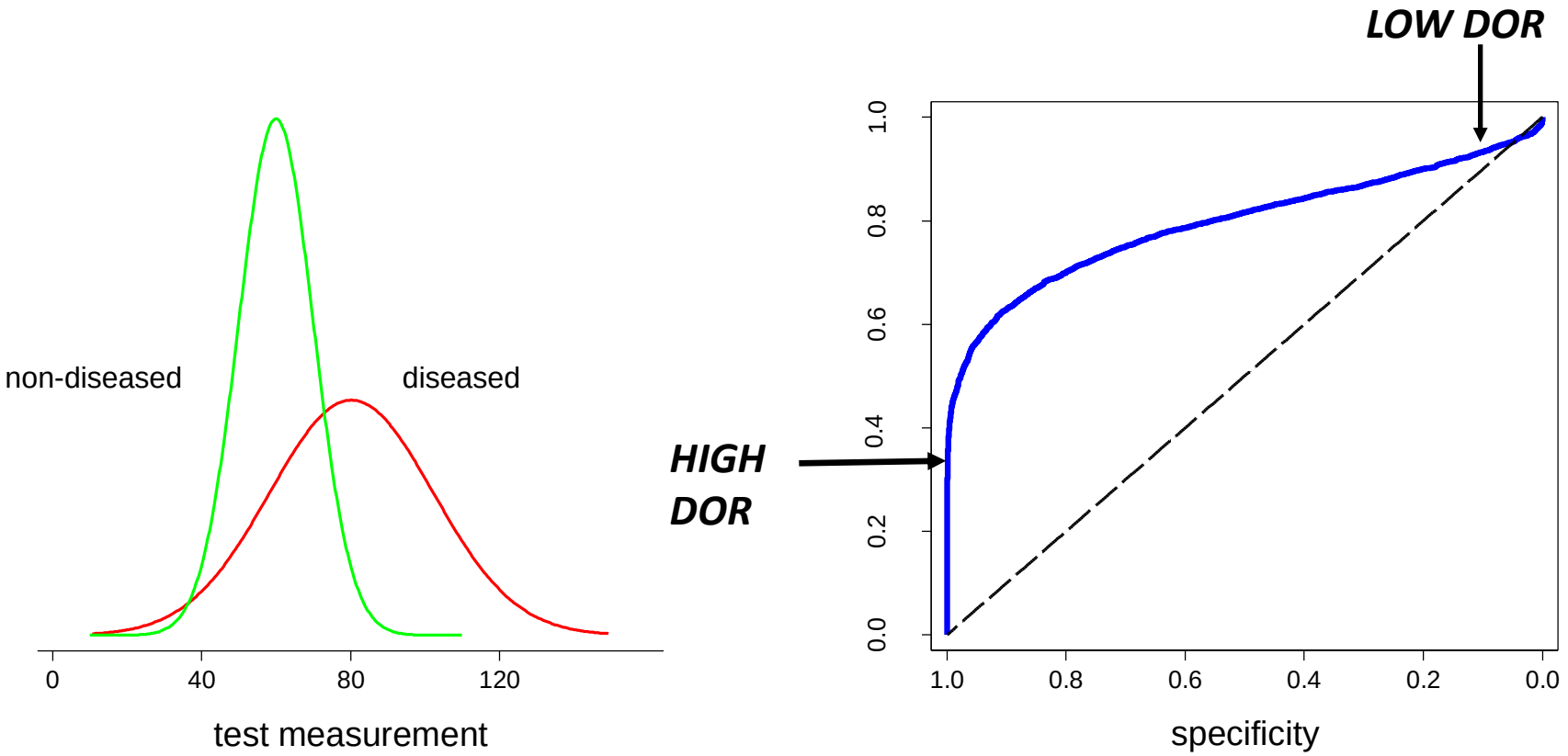
	Sensitivity						
Specificity	50%	60%	70%	80%	90%	95%	99%
50%	1	2	2	4	9	19	99
60%	2	2	4	6	14	29	149
70%	2	4	5	9	21	44	231
80%	4	6	9	16	36	76	396
90%	9	14	21	36	81	171	891
95%	19	29	44	76	171	361	1881
99%	99	149	231	396	891	1881	9801

Symmetrical ROC curves and diagnostic odds ratios



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

Asymmetrical ROC curves and diagnostic odds ratios



ROC curve is asymmetric when test accuracy varies with threshold

Scope of a DTA review

- Multiple objectives are possible
- 3 main types of analyses based on review question and objectives
    1. What is the diagnostic accuracy of a particular test?
    2. How does the accuracy of two or more tests compare?
    3. How does test accuracy vary with clinical and methodological characteristics?

The meta-analysis process

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

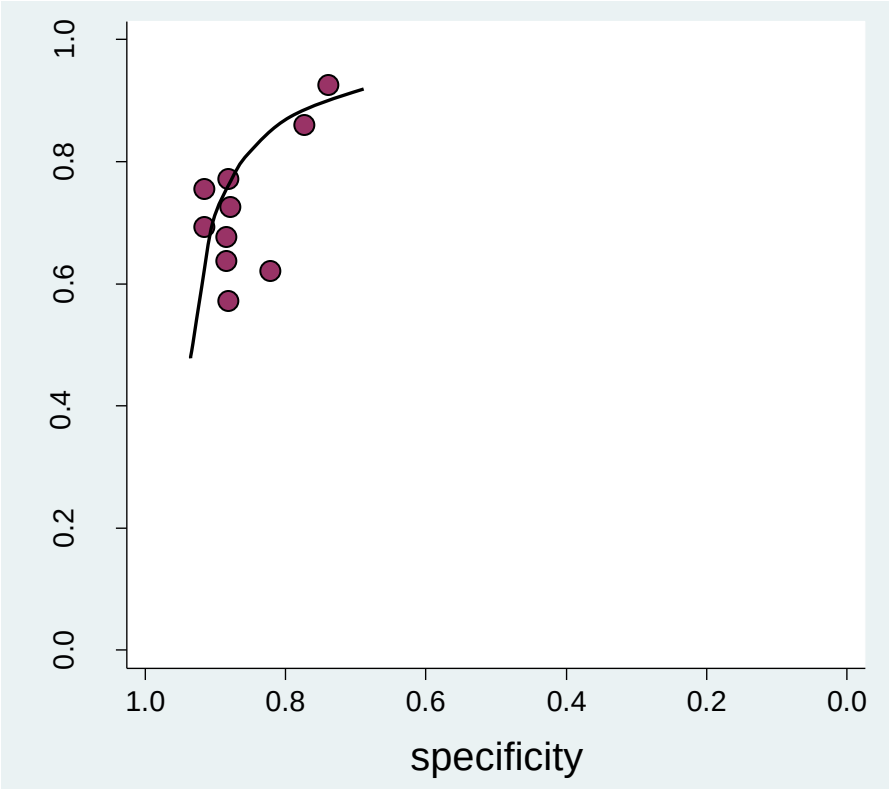
Challenges

- There are two summary statistics for each study
  - sensitivity and specificity and each have different implications

- Threshold effects induce correlations between sensitivity and specificity and often seem to be present
  - thresholds can vary between studies
  - the same threshold can imply different sensitivities and specificities in different groups

- Heterogeneity is the norm
  - substantial variation in sensitivity and specificity are noted in most reviews

Approach for meta-analysis



- Current statistical methods use a single estimate of sensitivity and specificity for each study
- Estimate the underlying ROC curve based on studies analysing different thresholds
- Analyses at specified threshold
  - Estimate summary sensitivity and summary specificity
- Compare ROC curves between tests
  - Allows comparison unrestricted to a particular threshold

Moses-Littenberg modelling of ROC curves

ROC curve transformation to linear plot

- Calculate the logits of TPR and FPR
- Plot their difference against their sum

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

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

Moses-Littenberg SROC method

- Regression models used to fit 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

- What do the axes mean?
  - Difference in logits is the log of the DOR
  - Sum of the logits is a marker of diagnostic threshold

Producing summary ROC curves

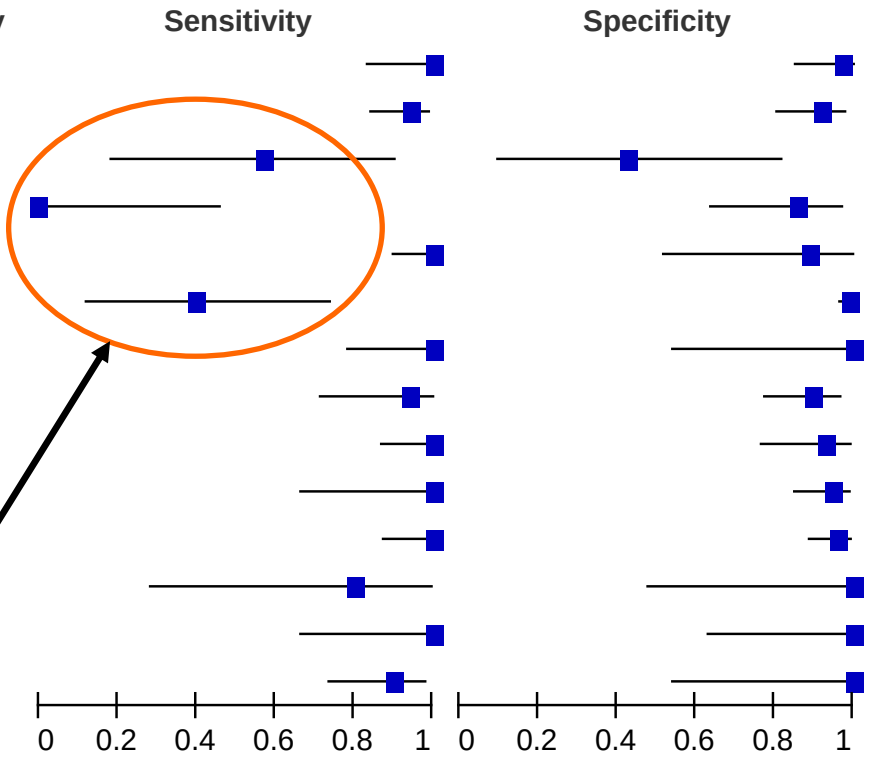
- Transform back to the ROC dimensions

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

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

Example: MRI for suspected deep vein thrombosis

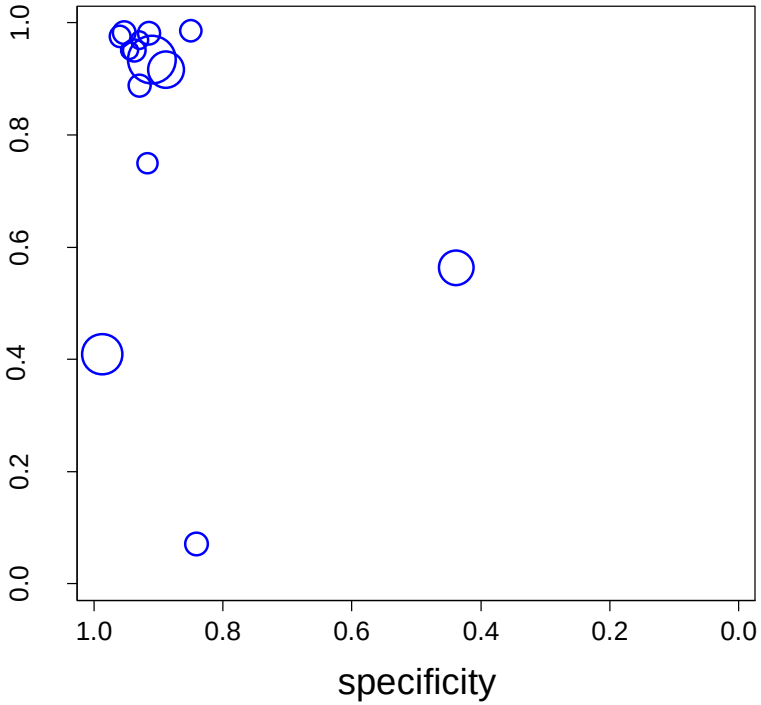
Study	TP	FP	FN	TN	Sensitivity	Specificity
Fraser 2003	20	1	0	34	1.00 [0.83, 1.00]	0.97 [0.85, 1.00]
Fraser 2002	49	4	3	45	0.94 [0.84, 0.99]	0.92 [0.80, 0.98]
Sica 2001	4	4	3	3	0.57 [0.18, 0.90]	0.43 [0.10, 0.82]
Jensen 2001	0	3	6	18	0.00 [0.00, 0.46]	0.86 [0.64, 0.97]
Catalano 1997	34	1	0	8	1.00 [0.90, 1.00]	0.89 [0.52, 1.00]
Larcom 1996	4	2	6	191	0.40 [0.12, 0.74]	0.99 [0.96, 1.00]
Laissy 1996	15	0	0	6	1.00 [0.78, 1.00]	1.00 [0.54, 1.00]
Evans 1996	16	5	1	43	0.94 [0.71, 1.00]	0.90 [0.77, 0.97]
Spritzer 1993	26	2	0	26	1.00 [0.87, 1.00]	0.93 [0.76, 0.99]
Evans 1993	9	3	0	52	1.00 [0.66, 1.00]	0.95 [0.85, 0.99]
Carpenter 1993	27	3	0	71	1.00 [0.87, 1.00]	0.96 [0.89, 0.99]
Vukov 1991	4	0	1	5	0.80 [0.28, 0.99]	1.00 [0.48, 1.00]
Pope 1991	9	0	0	8	1.00 [0.66, 1.00]	1.00 [0.63, 1.00]
Erdman 1990	27	0	3	6	0.90 [0.73, 0.98]	1.00 [0.54, 1.00]

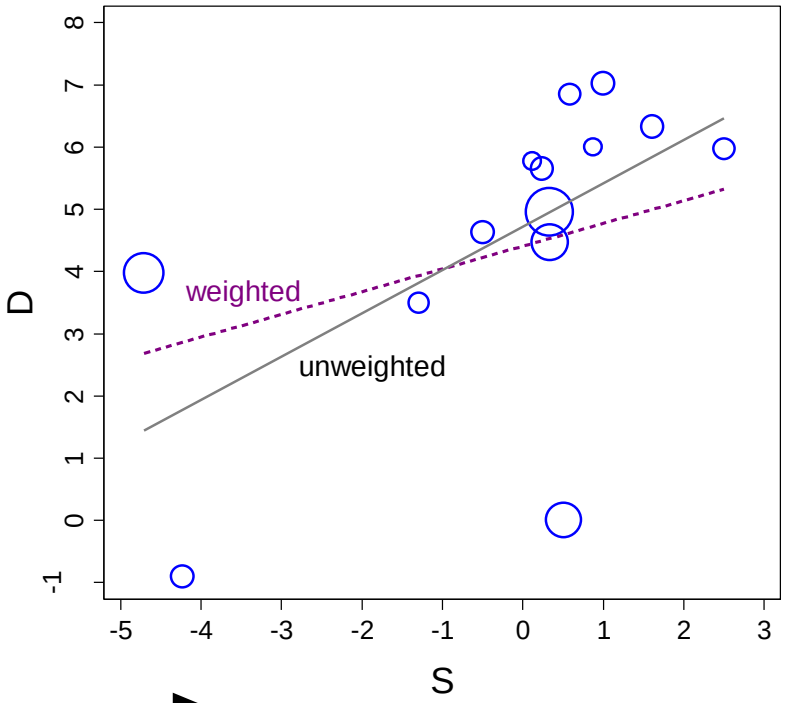


Low sensitivity probably due to failure of MRI to detect distal thrombi

Sampson et al. Eur Radiol (2007) 17: 175–181

SROC regression: MRI for suspected deep vein thrombosis

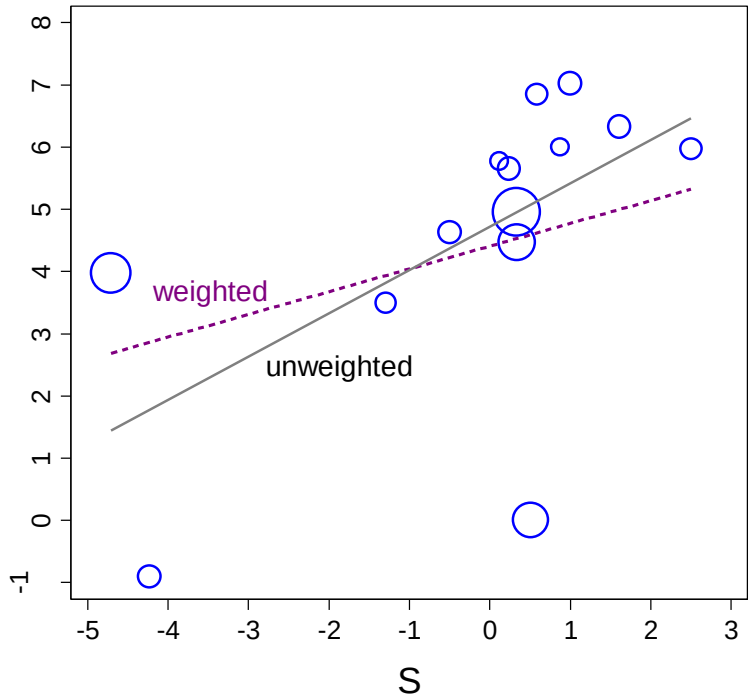




Linear transformation

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

SROC regression: MRI for suspected deep vein thrombosis



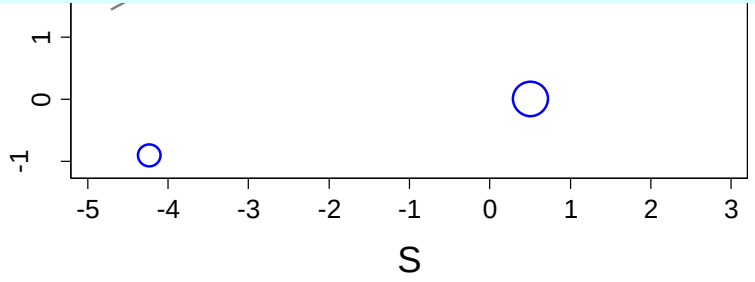
Inverse transformation

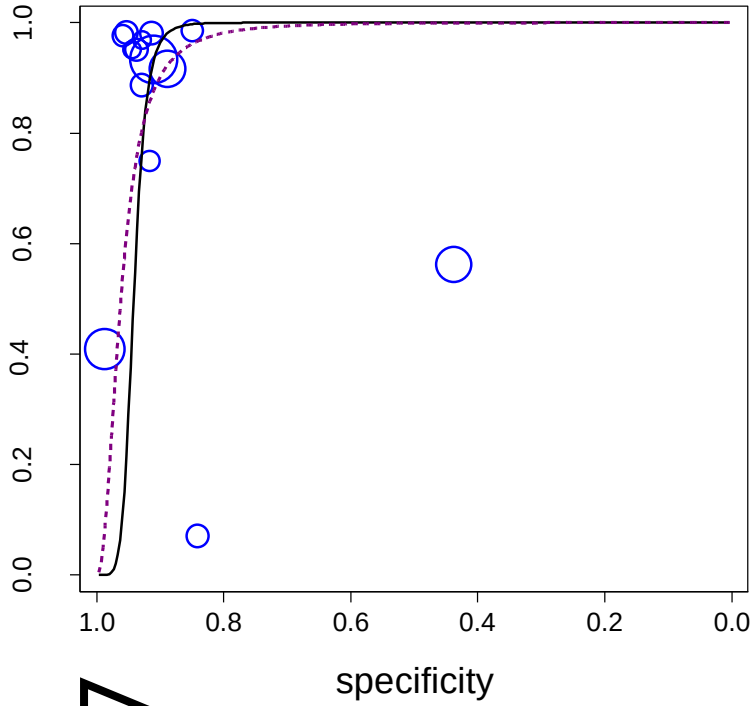
The SROC curve is produced by using the estimates of *a* and *b* to compute the expected sensitivity (*tpr)* across a range of values for 1-specificity (*fpr)*

SROC regression: MRI for suspected deep vein thrombosis

$$a = 4.721, b = 0.697$$

$$TPR = \frac{1}{1 + \frac{1}{e^{4.721/(1+0.697)}} \times \left(\frac{FPR}{1-FPR}\right)^{\frac{1+0.697}{1-0.697}}}$$

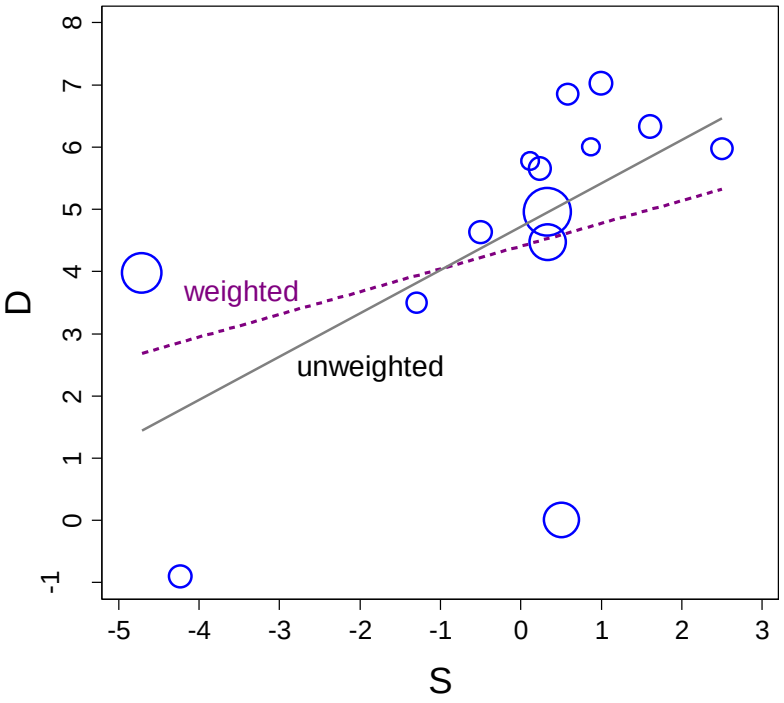


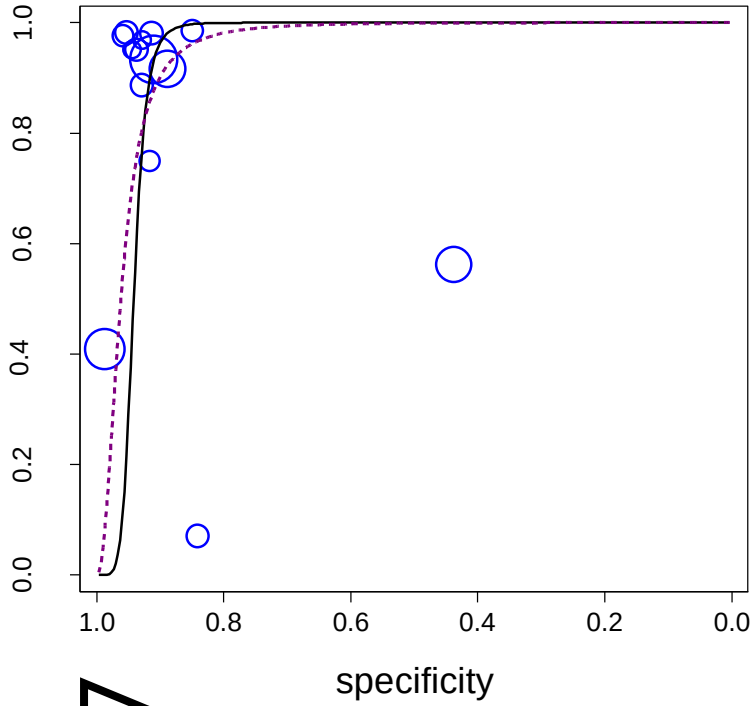


Inverse transformation

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)*

SROC regression: MRI for suspected deep vein thrombosis





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 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

Why we need hierarchical models

- Heterogeneity in reviews of diagnostic studies is common

- Valid methods for statistical inference are required
  - summary estimates and confidence intervals/ regions
  - investigating heterogeneity
  - test comparisons

- Moses-Littenberg method does not meet these requirements

Hierarchical models

- Hierarchical / multi-level
  - allows for both within (sampling error) and
  - between study variability (through inclusion of random effects)
- Logistic
  - correctly models sampling uncertainty in the true positive proportion and the false positive proportion
  - no zero cell adjustments needed
- Regression models
  - used to investigate sources of heterogeneity

Hierarchical models

Two models have been proposed for the meta-analysis of studies of diagnostic accuracy:

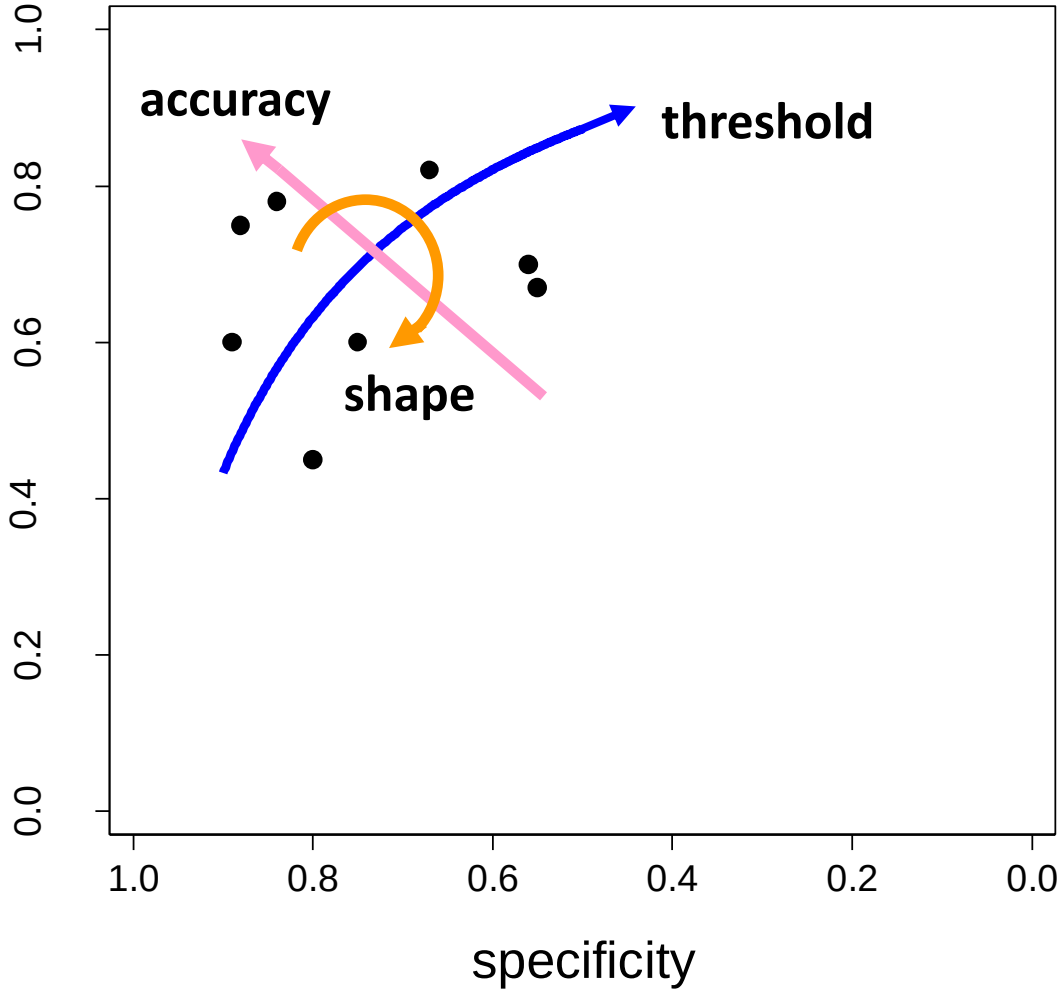
- **HSROC model**: primary objective is to fit a summary ROC curve

  and

- **Bivariate model**: primary objective is to obtain a summary estimate of 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



Models the summary ROC in terms of:

- "threshold"
- accuracy
- shape of the curve (dependence of accuracy on "threshold")

Accuracy and threshold specified as random study effects

HSROC model

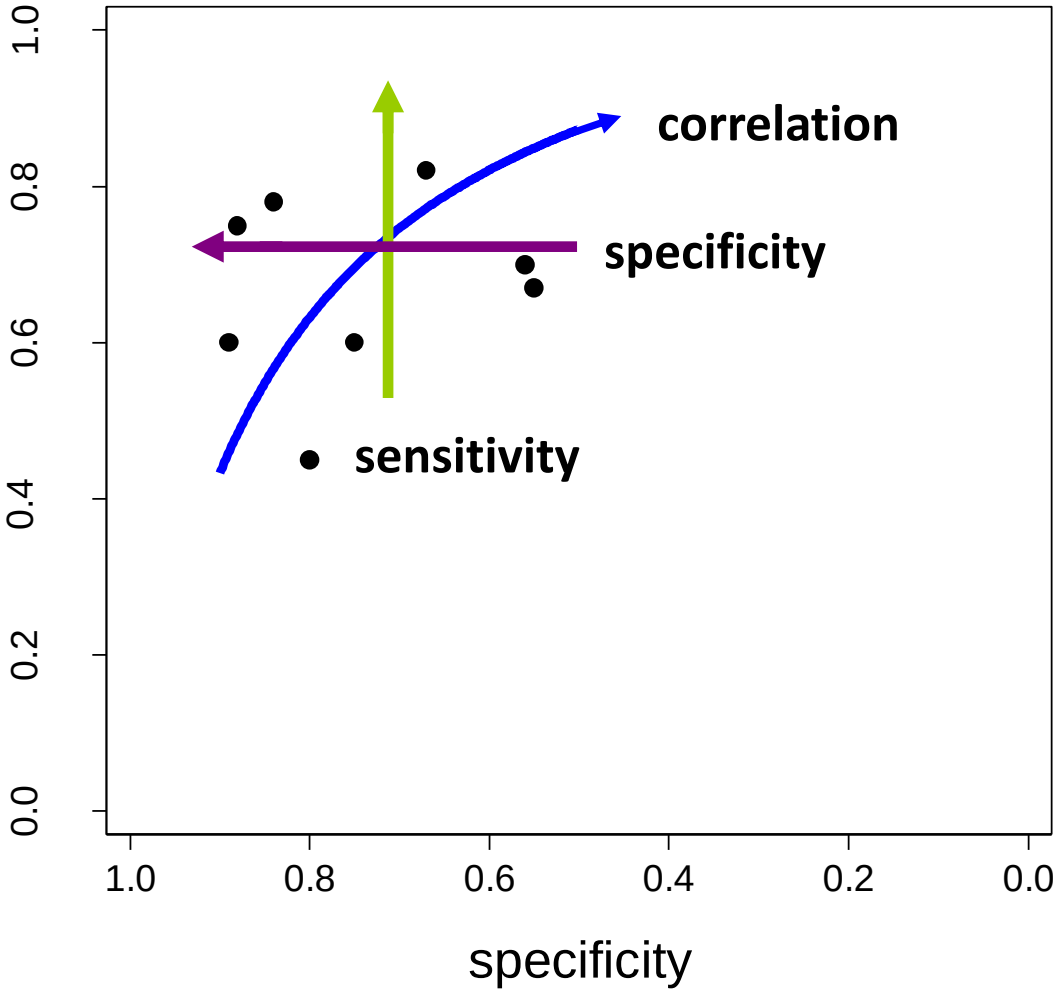
A properly formulated model for estimating summary ROC curve

Models ROC curves using estimates of log DOR, a pseudo threshold parameter, and a shape parameter

Random effects included for variation between studies in accuracy and threshold

Both within and between study variability are taken into account

Bivariate model



Combines two random effects meta-analysis of sensitivity and specificity in a single model

Models both logit(sensitivity) and logit(specificity) and the correlation between them

logit(sensitivity) and logit(specificity) are specified as random study effects

Bivariate model

Combines two random effects meta-analysis of sensitivity and specificity in a single model

Models both logit(sensitivity) and logit(specificity) and the correlation between them

logit(sensitivity) and logit(specificity) are specified as random study effects

Both within and between study variability are taken into account

Fitting the models

HSROC model

Hierarchical model with non-linear regression, random effects and binomial error

Easy to fit in PROC NLMIXED in SAS

Original code in winBUGs

Bivariate model

Hierarchical model with linear regression, random effects and binomial error

Easy to fit in PROC NLMIXED in SAS

Also in STATA, MLWin and R

Outputs from the models

- Underlying SROC curve, and the average operating point on the curve
- Confidence and prediction ellipses estimable
- Possible to derive other summary estimates

Which method to use?

HSROC and bivariate models are mathematically equivalent when there are no covariates in the models. Harbord R, Deeks J et al. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics 2006;1:1-21.

A unification of the models

- **HSROC model**: primary objective is to fit a summary ROC curve
- **Bivariate model**: primary objective is to obtain a summary estimate of sensitivity and specificity

If there are no covariates included in the model, the methods are equivalent :

- Parameter estimates from the HSROC model can be used to derive the summary point and corresponding confidence region
- Parameter estimates from the bivariate model can be used to obtain the HSROC curve

Which model to choose?

If covariates are included in the model to explore reasons for heterogeneity in test performance, the choice will be guided by both:

- ***The research question:*** Whether we want to make inferences about
  - the curve or
  - the summary point
- ***The available data:*** how this affects meaningful interpretation of the results

Should I be estimating a summary point or a summary curve ?

Reasons to prefer points ...

- Estimate the mean sensitivity and mean specificity together with a confidence region – **understandable quantities**

- Possible to estimate differences in sensitivity and specificity between tests (giving absolute numbers of extra TP and TN) and their statistical significance – **the quantities required to judge the consequences of the difference**

Limitation of points ...

- Points should only be calculated when studies **share a common threshold** value
  - Pooling studies that do **not** use a common threshold leads to a summary point that is uninterpretable
  - Restricting to a **common threshold** reduces data available
  - Choice of **common threshold** is often arbitrary
  - The **common threshold** may not be the threshold a reader wants to know about
  - A **common threshold** for non-numeric tests may be hard to define
  - The ranking of two tests may not be consistent across different **common threshold** choices

Reasons to prefer curves ...

- Estimation **unrestricted by threshold** – data from **all relevant studies** can be included
- Greater power to make comparisons between tests or investigate potential reasons for heterogeneity
- Expected trade-off in sensitivity and specificity is observed, and sensitivity estimated for a fixed specificity
- **However**, *if studies use a common threshold there will be limited information to inform the shape of the curve*

DTA meta-analysis: Investigating heterogeneity

Sources of heterogeneity

I. Chance variation

II. Differences in (implicit) threshold

III. Bias

IV. Clinical subgroups

V. Unexplained variation

Quantifying heterogeneity

- I^2 statistic is a univariate measure
  - does not account for heterogeneity due to threshold effects

- Not generally recommended for DTA reviews
- Prediction regions provide a visual indication of between study heterogeneity for summary points

Approaches to investigations of heterogeneity

Covariates to be investigated should be specified

- Graphical presentation
  - Forest plots
  - SROC plots
- Subgroup analyses
  - Limited to categorical variables
  - Low power
- Meta-regression
  - relationship of test accuracy with categorical or continuous covariate can be explored

Meta-regression

- Hierarchical models can incorporate a study-level covariate to
  - investigate the relative accuracy of 2 or more tests
  - investigate heterogeneity
- 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

Investigation of heterogeneity - example

Rapid diagnostic tests for diagnosing uncomplicated *P. falciparum* malaria in endemic countries (Review)

Abba K, Deeks JJ, Olliaro P, Naing CM, Jackson SM, Takwoingi Y, Donegan S, Garner P

Investigation of sources of heterogeneity

We planned to investigate heterogeneity in relation to the index test (by commercial test, test type and grouped by HRP-2/pLDH) and reference tests (microscopy vs PCR), as well as the study participants' age, endemicity of malaria, and geographic area (by continent).

Cochrane Database of Systematic Reviews
6 JUL 2011 DOI: 10.1002/14651858.CD008122.pub2

Effect of continent on Type 1 RDTs

	Number of studies	Number of patients	Number of malaria cases	Pooled sensitivity (95% CI)	Pooled specificity (95% CI)	Test[1]
			Continent			
Africa	39	21958	7445	94.0 (91.3, 95.9)	93.0 (89.8, 95.3)	p=0.01
Asia	24	15810	4060	96.7 (93.7, 97.8)	96.7 (94.4, 98.1)	
South America	2	2294	461	88.7 (61.9, 97.4)	99.4 (96.4, 100.0)	

[1]Statistical test for evidence of a difference between groups

Cochrane Database of Systematic Reviews
6 JUL 2011 DOI: 10.1002/14651858.CD008122.pub2

Limitations of meta-regression

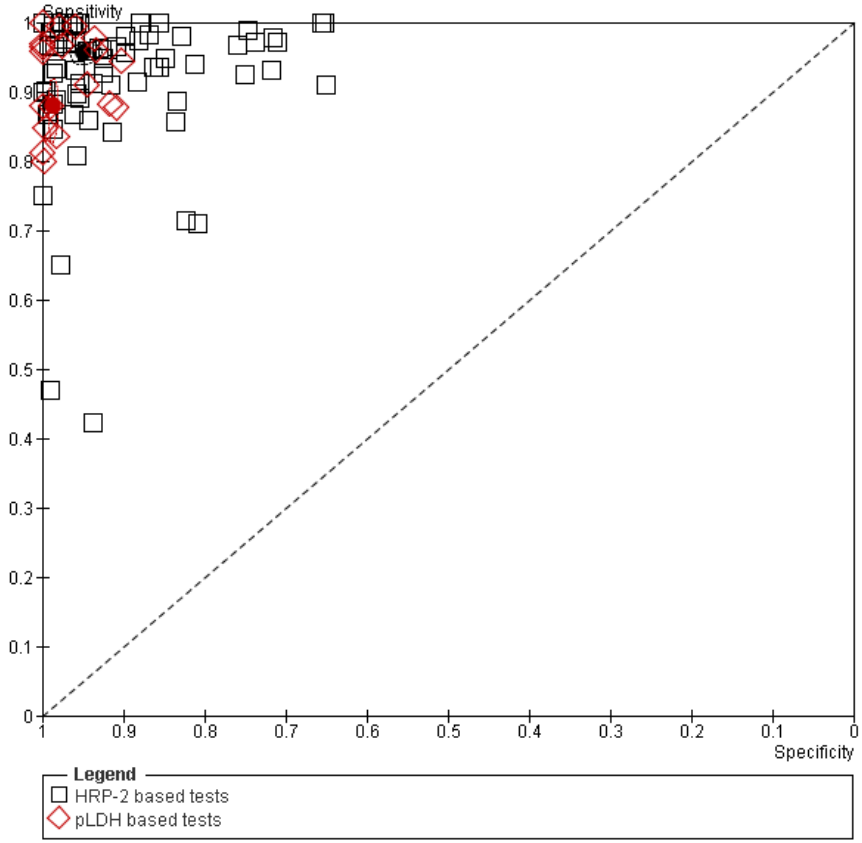
- Validity of covariate information
  - poor reporting on design features
- Population characteristics
  - information missing or crudely available
- Lack of power
  - small number of contrasting studies

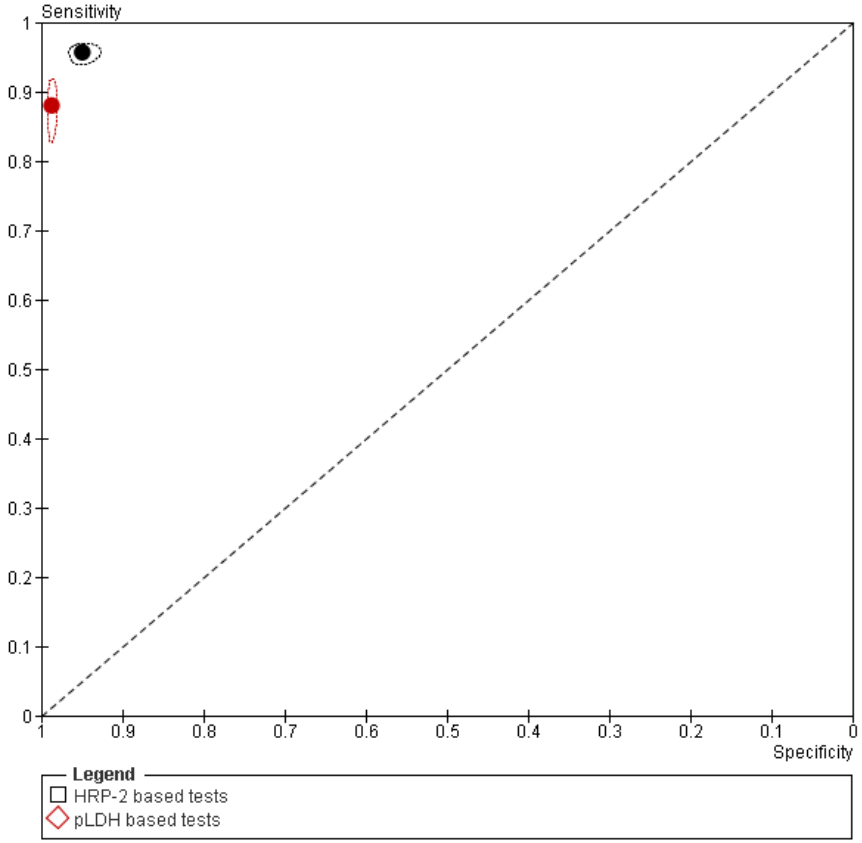
DTA meta-analysis: Test comparisons

Issues in test comparisons

- Pool all available studies assessing the performance of one or more tests
  - Can lead to bias due to confounding
- Adjusting for potential confounders is often not feasible
- Restrict analysis to studies that evaluated both tests in the same patients, or randomized patients to receive each test
  - removes the need to adjust for confounders

Comparison between HRP-2 and pLDH antibody based RDT types: all studies

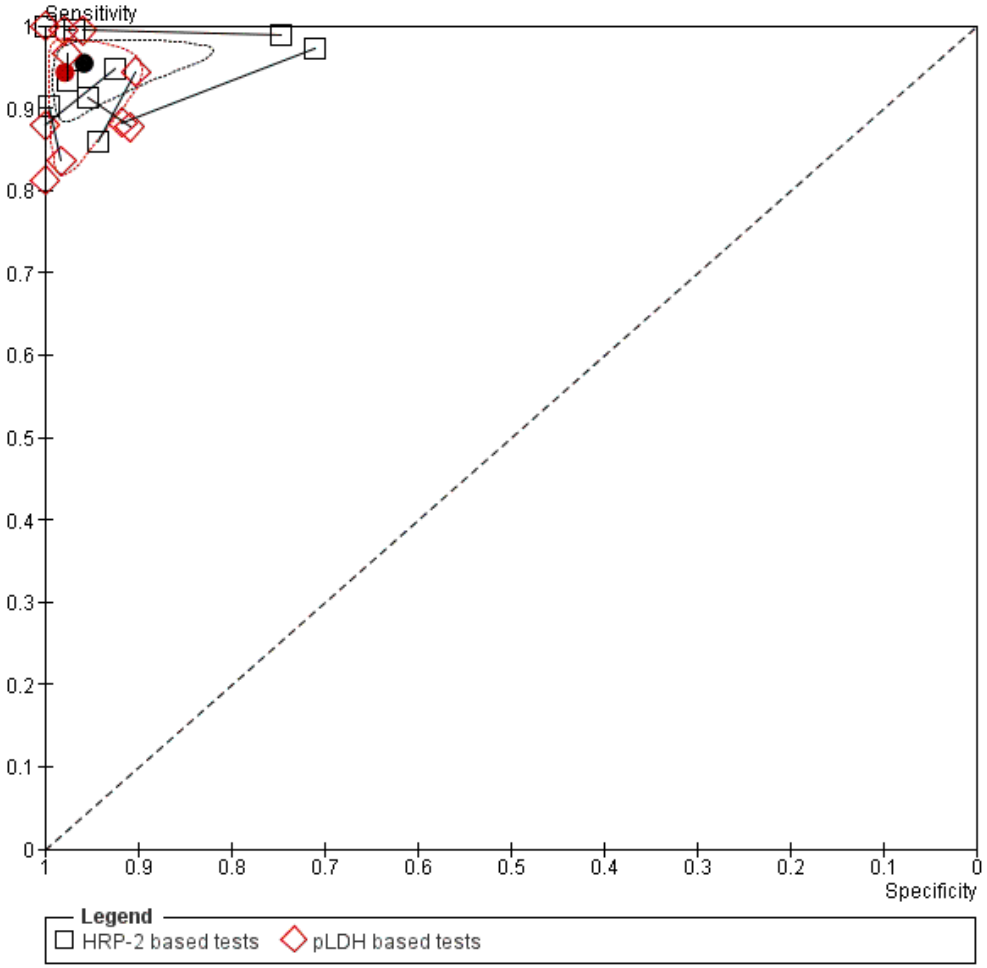




75 HRP-2 studies and 19 pLDH studies

Cochrane Database of Systematic Reviews
6 JUL 2011 DOI: 10.1002/14651858.CD008122.pub2

SROC plot of HRP-2 and pLDH antibody based RDT types: paired data



Cochrane Database of Systematic Reviews
6 JUL 2011 DOI: 10.1002/14651858.CD008122.pub2
http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008122.pub2/full#CD008122-fig-0015

Comparison between HRP-2 and pLDH antibody based RDT types – summary estimates

	Number of studies	Number of patients	Number of malaria cases	Pooled sensitivity (95% CI)	Pooled specificity (95% CI)	Test[1]
Indirect comparison (using all studies)						
HRP-2 based	75	43307	12857	95.8 (94.4, 96.8)	95.0 (93.2, 96.3)	
pLDH based	19	14787	4674	87.8 (83.9, 90.9)	98.7 (98.2, 99.1)	
Ratio				*0.92 (0.89, 0.95), p<0.001*	*1.04 (1.03, 1.05), p<0.001*	*p<0.001*

[1]Statistical test for evidence of a difference between groups

Cochrane Database of Systematic Reviews
6 JUL 2011 DOI: 10.1002/14651858.CD008122.pub2

Comparison between HRP-2 and pLDH antibody based RDT types – summary estimates

	Number of studies	Number of patients	Number of malaria cases	Pooled sensitivity (95% CI)	Pooled specificity (95% CI)	Test[1]
Indirect comparison (using all studies)						
HRP-2 based	75	43307	12857	95.8 (94.4, 96.8)	95.0 (93.2, 96.3)	
pLDH based	19	14787	4674	87.8 (83.9, 90.9)	98.7 (98.2, 99.1)	
Ratio				*0.92 (0.89, 0.95), p<0.001*	*1.04 (1.03, 1.05), p<0.001*	*p<0.001*
Direct comparison (using only studies that directly compared the two)						
HRP-2 based	9	10626	3672	97.8 (95.0, 98.9)	93.9 (86.5, 97.4)	
pLDH based	9	10623	3672	92.4 (85.0, 96.3)	98.4 (96.3, 99.3)	
Ratio				*0.95 (0.91, 0.99) p=0.008*	*1.05 (1.01, 1.09), p=0.02*	*p<0.001*

[1]Statistical test for evidence of a difference between groups

Cochrane Database of Systematic Reviews
6 JUL 2011 DOI: 10.1002/14651858.CD008122.pub2

Take home message (1)

- Moses & Littenberg method useful for exploratory analysis but should not be used for statistical inference
- Hierarchical (mixed) models appropriately model the data, allow for threshold effects, and estimate unexplained heterogeneity.
- The HSROC and bivariate models are mathematically identical when there are no covariates.
  - Differences occur when investigating heterogeneity and comparing tests.
- Models estimate summary ROC curves, average operating points, confidence and prediction regions.
- Most important to ascertain whether summary curves or summary points are appropriate.

Take home message (2)

- Heterogeneity is expected in test accuracy meta-analysis
- Heterogeneity should be investigated whenever possible
- Ideally, investigations should be pre-planned
- Available evidence for test comparisons
  - All studies (comparative and non-comparative studies)
  - Restricted to studies that have directly compared the tests (comparative studies)
  - Within study test comparisons are desirable as they are less subject to bias

References

- Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. J Clin Epidemiol. 2005;58(10):982-90.

- Chu H, Cole SR. Bivariate meta-analysis for sensitivity and specificity with sparse data: a generalized linear mixed model approach (letter to the Editor). J Clin Epidemiol. 2006;59(12):1331-2.

- Rutter CM, Gatsonis C. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. Stat Med. 2001;20(19)2865-84.

- Harbord RM, Deeks JJ, Egger M, Whiting P, Sterne JA. A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics. 2007;8(2):239-51.

- Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration; 2010. Available at http://srdta.cochrane.org/.

- Takwoingi Y, Riley R, Deeks J. Meta-analyses of diagnostic accuracy studies in mental health. Evid Based Ment Health. Epub ahead of print October 7 2015. DOI: 10.1136/eb-2015-102228

ACKNOWLEDGEMENTS

Materials for this presentation are based in part on material adapted from members of the Cochrane Screening and Diagnostic Test Methods Group

See http://dta.cochrane.org/dta-author-training-online-learning for additional training materials