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

[Diagram showing distribution of test measurements with diagnostic threshold, specificity=99%, sensitivity=69%, and classification into non-diseased and diseased categories.]
Heterogeneity in threshold within a study

specificity=98%
sensitivity=84%

non-diseased
diseased

TN FN FP TP

test measurement
Heterogeneity in threshold within a study

- Diagnostic threshold
- Specificity = 93%
- Sensitivity = 93%
- Non-diseased
- Diseased
- TN, FNFP, TP

Test measurement
Heterogeneity in threshold within a study

specificity = 84%
sensitivity = 98%

non-diseased
diseased

TN  FN  FP  TP

0  40  80  120  160

diagnostic threshold

test measurement
Heterogeneity in threshold within a study
Threshold effect

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>0.99</td>
<td>0.69</td>
</tr>
<tr>
<td>70</td>
<td>0.98</td>
<td>0.84</td>
</tr>
<tr>
<td>75</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>80</td>
<td>0.84</td>
<td>0.98</td>
</tr>
<tr>
<td>85</td>
<td>0.69</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Increasing threshold decreases sensitivity but increases specificity.

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

<table>
<thead>
<tr>
<th>Specificity</th>
<th>50%</th>
<th>60%</th>
<th>70%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>19</td>
<td>99</td>
</tr>
<tr>
<td>60%</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>14</td>
<td>29</td>
<td>149</td>
</tr>
<tr>
<td>70%</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>21</td>
<td>44</td>
<td>231</td>
</tr>
<tr>
<td>80%</td>
<td>4</td>
<td>6</td>
<td>9</td>
<td>16</td>
<td>36</td>
<td>76</td>
<td>396</td>
</tr>
<tr>
<td>90%</td>
<td>9</td>
<td>14</td>
<td>21</td>
<td>36</td>
<td>81</td>
<td>171</td>
<td>891</td>
</tr>
<tr>
<td>95%</td>
<td>19</td>
<td>29</td>
<td>44</td>
<td>76</td>
<td>171</td>
<td>361</td>
<td>1881</td>
</tr>
<tr>
<td>99%</td>
<td>99</td>
<td>149</td>
<td>231</td>
<td>396</td>
<td>891</td>
<td>1881</td>
<td>9801</td>
</tr>
</tbody>
</table>
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

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

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

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

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

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

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

SROC regression: MRI for suspected deep vein thrombosis

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

Linear 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 ($fpr$) weighted/unweighted 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 ($fpr$).

The SROC regression: MRI for suspected deep vein thrombosis

$a = 4.721, b = 0.697$

$$TPR = \frac{1}{1 + \frac{1}{e^{4.721/(1+0.697)}} \times \left( \frac{FPR}{1-FPR} \right)^{1+0.697/1-0.697}}$$
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
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).
# Effect of continent on Type 1 RDTs

<table>
<thead>
<tr>
<th>Continent</th>
<th>Number of studies</th>
<th>Number of patients</th>
<th>Number of malaria cases</th>
<th>Pooled sensitivity (95% CI)</th>
<th>Pooled specificity (95% CI)</th>
<th>Test&lt;sup&gt;1&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Africa</td>
<td>39</td>
<td>21958</td>
<td>7445</td>
<td>94.0 (91.3, 95.9)</td>
<td>93.0 (89.8, 95.3)</td>
<td>p=0.01</td>
</tr>
<tr>
<td>Asia</td>
<td>24</td>
<td>15810</td>
<td>4060</td>
<td>96.7 (93.7, 97.8)</td>
<td>96.7 (94.4, 98.1)</td>
<td></td>
</tr>
<tr>
<td>South America</td>
<td>2</td>
<td>2294</td>
<td>461</td>
<td>88.7 (61.9, 97.4)</td>
<td>99.4 (96.4, 100.0)</td>
<td></td>
</tr>
</tbody>
</table>

<sup>1</sup>Statistical test for evidence of a difference between groups
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
SROC plot of HRP-2 and pLDH antibody based RDT types: paired data
Comparison between HRP-2 and pLDH antibody based RDT types – summary estimates

<table>
<thead>
<tr>
<th></th>
<th>Number of studies</th>
<th>Number of patients</th>
<th>Number of malaria cases</th>
<th>Pooled sensitivity (95% CI)</th>
<th>Pooled specificity (95% CI)</th>
<th>Test¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Indirect comparison (using all studies)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRP-2 based</td>
<td>75</td>
<td>43307</td>
<td>12857</td>
<td>95.8 (94.4, 96.8)</td>
<td>95.0 (93.2, 96.3)</td>
<td></td>
</tr>
<tr>
<td>pLDH based</td>
<td>19</td>
<td>14787</td>
<td>4674</td>
<td>87.8 (83.9, 90.9)</td>
<td>98.7 (98.2, 99.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Ratio</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.92 (0.89, 0.95), p&lt;0.001</td>
<td>1.04 (1.03, 1.05), p&lt;0.001</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>

¹Statistical 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

<table>
<thead>
<tr>
<th></th>
<th>Number of studies</th>
<th>Number of patients</th>
<th>Number of malaria cases</th>
<th>Pooled sensitivity (95% CI)</th>
<th>Pooled specificity (95% CI)</th>
<th>Test¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Indirect comparison (using all studies)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRP-2 based</td>
<td>75</td>
<td>43307</td>
<td>12857</td>
<td>95.8 (94.4, 96.8)</td>
<td>95.0 (93.2, 96.3)</td>
<td></td>
</tr>
<tr>
<td>pLDH based</td>
<td>19</td>
<td>14787</td>
<td>4674</td>
<td>87.8 (83.9, 90.9)</td>
<td>98.7 (98.2, 99.1)</td>
<td></td>
</tr>
<tr>
<td><strong>Ratio</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.92 (0.89, 0.95), p&lt;0.001</td>
<td>1.04 (1.03, 1.05), p&lt;0.001</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td><strong>Direct comparison (using only studies that directly compared the two)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HRP-2 based</td>
<td>9</td>
<td>10626</td>
<td>3672</td>
<td>97.8 (95.0, 98.9)</td>
<td>93.9 (86.5, 97.4)</td>
<td></td>
</tr>
<tr>
<td>pLDH based</td>
<td>9</td>
<td>10623</td>
<td>3672</td>
<td>92.4 (85.0, 96.3)</td>
<td>98.4 (96.3, 99.3)</td>
<td></td>
</tr>
<tr>
<td><strong>Ratio</strong></td>
<td></td>
<td></td>
<td></td>
<td>0.95 (0.91, 0.99), p=0.008</td>
<td>1.05 (1.01, 1.09), p=0.02</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>

¹Statistical test for evidence of a difference between groups
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


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