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?

<table>
<thead>
<tr>
<th>Disease (Ref. test)</th>
<th>Pres.</th>
<th>Abs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index - Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>-</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

- Sensitivity
- Specificity
- Positive Predictive Value
- Negative Predictive Value
- Positive Likelihood Ratio
- Negative Likelihood Ratio
- Diagnostic Odds ratio
- ROC curves
Pooling sensitivity and specificity?

- Sensitivity (95% CI):
  - Allen: 0.60 (0.37 - 0.83)
  - Busca: 1.00 (0.86 - 1.00)
  - Monques: 0.88 (0.83 - 0.93)
  - Mann: 0.93 (0.89 - 0.97)
  - Kawazu: 0.56 (0.47 - 0.66)
  - Reina: 0.75 (0.55 - 0.97)
  - Queen-Harville: 0.67 (0.56 - 0.78)
  - Chatter: 0.44 (0.35 - 0.54)
  - Lee: 0.71 (0.55 - 0.87)
  - Becker: 0.33 (0.19 - 0.50)
  - Florent: 0.47 (0.33 - 0.62)
  - Uusariya: 1.00 (0.81 - 1.00)
  - Herbrecht: 0.29 (0.22 - 0.37)
  - Williamson: 0.75 (0.48 - 0.96)
  - Bresnahan: 0.62 (0.30 - 0.85)
  - Mezetti: 0.86 (0.65 - 0.95)
  - Parada: 0.82 (0.68 - 0.98)
  - White: 0.33 (0.04 - 0.72)
  - Adan: 0.57 (0.48 - 0.68)
  - Kalil: 0.57 (0.48 - 0.68)
  - Pfeil: 0.55 (0.41 - 0.69)
  - Doerrman: 0.72 (0.47 - 0.98)
  - Martin: 0.32 (0.19 - 0.47)
  - Wasser: 0.11 (0.05 - 0.21)

- Pooled Sensitivity: 0.46 (0.41 to 0.50)
  - Chi-square = 234.46, df = 23 (p = 0.0000)
  - Inconsistency (I-square) = 90.2%

- Specificity (95% CI):
  - Allen: 0.90 (0.83 - 0.95)
  - Busca: 0.94 (0.83 - 0.97)
  - Monques: 1.00 (0.99 - 1.00)
  - Mann: 0.80 (0.63 - 0.92)
  - Kawazu: 0.84 (0.47 - 0.89)
  - Reina: 1.00 (0.95 - 1.00)
  - Queen-Harville: 0.99 (0.94 - 1.00)
  - Chatter: 0.88 (0.66 - 0.97)
  - Lee: 0.94 (0.89 - 0.97)
  - Becker: 0.90 (0.77 - 0.97)
  - Florent: 0.89 (0.72 - 0.97)
  - Uusariya: 0.92 (0.86 - 0.95)
  - Herbrecht: 0.95 (0.93 - 0.98)
  - Williamson: 0.95 (0.92 - 0.98)
  - Bresnahan: 0.95 (0.74 - 0.99)
  - Mezetti: 0.86 (0.75 - 0.98)
  - Parada: 0.98 (0.82 - 1.00)
  - White: 1.00 (0.96 - 1.00)
  - Adan: 0.98 (0.76 - 0.98)
  - Kalil: 0.99 (0.83 - 0.98)
  - Pfeil: 1.00 (0.96 - 1.00)
  - Doerrman: 0.94 (0.64 - 0.95)
  - Martin: 0.82 (0.73 - 0.88)

- Pooleed Specificity: 0.94 (0.86 to 0.95)
  - Chi-square = 261.50, df = 23 (p = 0.0000)
  - Inconsistency (I-square) = 91.2%
Pooling sensitivity and specificity?
Pooling Likelihood Ratios?
Pooling LRs?
Pooling odds ratios?

![Graph showing pooled diagnostic odds ratios]

Random Effects Model
Pooled Diagnostic Odds Ratio = 20.00 (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
<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allan 2005</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>123</td>
<td>0.00 [0.00, 0.95]</td>
<td>0.99 [0.96, 1.00]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Becker 2003</td>
<td>6</td>
<td>12</td>
<td>7</td>
<td>62</td>
<td>0.46 [0.19, 0.75]</td>
<td>0.84 [0.73, 0.91]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bretagne 1998</td>
<td>14</td>
<td>5</td>
<td>4</td>
<td>18</td>
<td>0.78 [0.52, 0.94]</td>
<td>0.78 [0.56, 0.93]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Challier 2004</td>
<td>20</td>
<td>9</td>
<td>6</td>
<td>35</td>
<td>0.77 [0.56, 0.91]</td>
<td>0.80 [0.65, 0.90]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kawazu 2004</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>134</td>
<td>0.64 [0.31, 0.89]</td>
<td>0.97 [0.93, 0.99]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maertens 2002</td>
<td>11</td>
<td>7</td>
<td>2</td>
<td>80</td>
<td>0.85 [0.55, 0.98]</td>
<td>0.92 [0.84, 0.97]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marr 2004</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>32</td>
<td>0.54 [0.33, 0.74]</td>
<td>0.74 [0.59, 0.86]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pereira 2005</td>
<td>1</td>
<td>9</td>
<td>0</td>
<td>29</td>
<td>1.00 [0.05, 1.00]</td>
<td>0.76 [0.60, 0.89]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pinel 2003</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>756</td>
<td>0.50 [0.32, 0.68]</td>
<td>0.98 [0.97, 0.99]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ulusakarya 2000</td>
<td>16</td>
<td>11</td>
<td>0</td>
<td>108</td>
<td>1.00 [0.83, 1.00]</td>
<td>0.91 [0.84, 0.95]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

False positive rate

True positive rate

False positive rate
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

Decreasing threshold increases sensitivity but decreases specificity for predicting spontaneous birth.

Fetal fibronectin sensitivity decreases specificity.

Increasing threshold increases specificity but decreases sensitivity.
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

\[ \text{Diagnostic OR} = \frac{TP \times TN}{FP \times FN} \]

\[ DOR = \left( \frac{\text{sensitivity}}{1 - \text{sensitivity}} \right) \left( \frac{1 - \text{specificity}}{\text{specificity}} \right) = \frac{LR + ve}{LR - ve} \]
### Diagnostic odds ratios

<table>
<thead>
<tr>
<th>HPV Test</th>
<th>Cervical Cancer (Biopsy)</th>
<th>Present</th>
<th>Absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td></td>
<td>65</td>
<td>93</td>
<td>158</td>
</tr>
<tr>
<td>-</td>
<td></td>
<td>7</td>
<td>161</td>
<td>198</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td>254</td>
<td>356</td>
</tr>
</tbody>
</table>

\[
DOR = \frac{65 \times 161}{93 \times 7} = 16
\]
## 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.

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

\[
\text{logit}(TPR) = \ln \left( \frac{TPR}{1 - TPR} \right) \quad S = \text{logit}(TPR) + \text{logit}(FPR) \\
\text{logit}(FPR) = \ln \left( \frac{FPR}{1 - FPR} \right) \quad D = \text{logit}(TPR) - \text{logit}(FPR)
\]
Translating ROC space to D versus S

\[ D = \log \text{odds ratio} \]
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
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

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

- 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

![Graph showing ROC curve and logit difference and sum plot.]

- **Non-diseased** distribution
- **Diseased** distribution

**Axes:**
- **Measurement**
- **Relative frequency**

**Plots:**
- **True positive rate (%)** vs. **False positive rate (%)**
- **Logit tpr - logit fpr** vs. **Logit tpr + logit fpr**

**Legend:**
- **LOW DOR**
- **HIGH DOR**
SROC regression: another example

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

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

![Graph showing sensitivity vs. specificity with correlation line](image)

- Sensitivity
- Specificity
- Correlation line
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;

Hierarchical SROC model

Sensitivity

Specificity

accuracy

threshold

shape
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;
Bivariate model

Sensitivity

Specificity

correlation

specificity

sensitivity
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

Sensitivity

1-specificity
F/T PSA in the Detection of Prostate cancer
Dip-stick Testing for Urinary Tract Infection

![Graph showing sensitivity and 1-specificity]

- Sensitivity
- 1-specificity
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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Older Model*</th>
<th>Advanced models**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chance variability</td>
<td>+/-</td>
<td>+</td>
</tr>
<tr>
<td>Threshold differences</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Subgroup</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Unexplained variation</td>
<td>+/-</td>
<td>+</td>
</tr>
</tbody>
</table>

* 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 1:
- both sens & spec higher
Prospective vs. Retrospective studies

Data collection:

- **Prosp**: □ □
- **Retro**: × × ×
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.

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 \( \rightarrow \) 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 \( \rightarrow \) \( \rightarrow \) lowest quality
   cumulative meta-analysis
Methods

- Quality assessment in 487 studies included in 30 systematic reviews.
- 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

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

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

<table>
<thead>
<tr>
<th>Ignoring quality:</th>
<th>Pooling all studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Analyzing subgroups:</strong></td>
<td>Only pooling high-quality studies; high quality defined as fulfilling a certain subset of criteria.</td>
</tr>
<tr>
<td><strong>Stepwise multivariable regression analysis:</strong></td>
<td>QUADAS-items with a p-value &lt;0.2 univariate are entered in a multivariable regression model</td>
</tr>
<tr>
<td><strong>Multivariable regression analysis with a set of covariates:</strong></td>
<td>A standard set of three QUADAS-items was used as covariates in each meta-analysis.</td>
</tr>
</tbody>
</table>
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.
Bivariate or HSROC? What do the parameters mean?

### Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>_sens</td>
<td>0.5943</td>
<td>0.3012</td>
<td>1000</td>
<td>1.97</td>
<td>0.0487</td>
<td>0.05</td>
<td>0.003282</td>
<td>1.1853</td>
<td>0.000107</td>
<td></td>
<td></td>
</tr>
<tr>
<td>_spec</td>
<td>2.8646</td>
<td>0.3114</td>
<td>1000</td>
<td>9.20</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>2.2535</td>
<td>3.4757</td>
<td>-0.00025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2uspec</td>
<td>1.2722</td>
<td>0.5723</td>
<td>1000</td>
<td>2.22</td>
<td>0.0265</td>
<td>0.05</td>
<td>0.1491</td>
<td>2.3953</td>
<td>-0.00004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s2usens</td>
<td>0.5687</td>
<td>0.4487</td>
<td>1000</td>
<td>1.32</td>
<td>0.1870</td>
<td>0.05</td>
<td>-0.2873</td>
<td>1.4853</td>
<td>0.000054</td>
<td></td>
<td></td>
</tr>
<tr>
<td>covsesp</td>
<td>-0.2430</td>
<td>0.4749</td>
<td>1000</td>
<td>-0.51</td>
<td>0.6089</td>
<td>0.05</td>
<td>-1.1749</td>
<td>0.6889</td>
<td>-0.00033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Covariance Matrix of Parameter Estimates

<table>
<thead>
<tr>
<th>Row</th>
<th>Parameter</th>
<th>_sens</th>
<th>_spec</th>
<th>s2uspec</th>
<th>s2usens</th>
<th>covsesp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>_sens</td>
<td>0.09071</td>
<td>-0.01563</td>
<td>-0.00023</td>
<td>0.03764</td>
<td>-0.03794</td>
</tr>
<tr>
<td>2</td>
<td>_spec</td>
<td>-0.01563</td>
<td>0.09698</td>
<td>0.02651</td>
<td>-0.00225</td>
<td>0.003328</td>
</tr>
<tr>
<td>3</td>
<td>s2uspec</td>
<td>-0.00023</td>
<td>0.02651</td>
<td>0.3276</td>
<td>0.008644</td>
<td>-0.04998</td>
</tr>
<tr>
<td>4</td>
<td>s2usens</td>
<td>0.03764</td>
<td>-0.00225</td>
<td>0.008644</td>
<td>0.1995</td>
<td>-0.1368</td>
</tr>
<tr>
<td>5</td>
<td>covsesp</td>
<td>-0.03794</td>
<td>0.003328</td>
<td>-0.04998</td>
<td>-0.1368</td>
<td>0.2255</td>
</tr>
</tbody>
</table>

### Additional Estimates

<table>
<thead>
<tr>
<th>Label</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>DF</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
<th>Alpha</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>orgsens</td>
<td>0.6444</td>
<td>0.06902</td>
<td>1000</td>
<td>9.34</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.5089</td>
<td>0.7798</td>
<td></td>
<td></td>
</tr>
<tr>
<td>orgspec</td>
<td>0.9461</td>
<td>0.01589</td>
<td>1000</td>
<td>59.54</td>
<td>&lt;.0001</td>
<td>0.05</td>
<td>0.9149</td>
<td>0.9772</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Meta-analysis of diagnostic accuracy

Log likelihood = -91.391372

|                      | Coef.      | Std. Err. |    z  | P>|z| | [95% Conf. Interval] |
|----------------------|------------|-----------|-------|------|----------------------|
| Bivariate            |            |           |       |      |                      |
| E(logitSe)           | 0.7266321  | 0.1544626 | 0.4238909 | 1.029373 |
| E(logitSp)           | 1.638955   | 0.2505372 | 1.147911 | 2.129999 |
| var(logitSe)         | 0.1249622  | 0.1306739 | 0.0160943 | 0.9702556 |
| var(logitSp)         | 0.82327    | 0.4055445 | 0.3135008 | 2.161952 |
| Corr(logits)         | 0.2387872  | 0.4557707 | -0.6067878 | 0.8308258 |
| HSROC                |            |           |       |      |                      |
| Lambda               | 2.187142   | 0.3086554 | 1.582189 | 2.792095 |
| Theta                | 0.0705697  | 0.3271092 | -0.5705525 | 0.7116919 |
| beta                 | 0.9426364  | 0.5764601 | -1.872047 | 2.072478 |
| s2alpha              | 0.7946707  | 0.5114531 | 2.250872 | 2.805587 |
| s2theta              | 0.1220778  | 0.1082908 | 0.0214569 | 0.6945551 |
| Summary pt.          |            |           |       |      |                      |
| Se                   | 0.6740658  | 0.0339356 | 0.6044139 | 0.7367944 |
| Sp                   | 0.8373927  | 0.0341147 | 0.7591292 | 0.8937849 |
| DOR                  | 10.65029   | 3.296352  | 5.806411 | 19.53509 |
| LR+                  | 4.145361   | 1.9181012 | 2.685598 | 6.398581 |
| LR-                  | 0.389225   | 0.0452324 | 0.3099427 | 0.4887875 |
| 1/LR-                | 2.569208   | 0.2985712 | 2.045879 | 3.226402 |

Covariance between estimates of E(logitSe) & E(logitSp) = 0.0045838
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

<table>
<thead>
<tr>
<th>Imaging modality</th>
<th>Sensitivity (95% CI)</th>
<th>Specificity (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAG</td>
<td>0.67 (0.57 to 0.76)</td>
<td>0.80 (0.73 to 0.85)</td>
</tr>
<tr>
<td>CT</td>
<td>0.49 (0.37 to 0.61)</td>
<td>0.92 (0.88 to 0.95)</td>
</tr>
<tr>
<td>P-value Lag vs. CT</td>
<td>0.023</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Studies</th>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st trimester - NT alone</td>
<td>10</td>
</tr>
<tr>
<td>1st trimester - NT and serology</td>
<td>22</td>
</tr>
<tr>
<td>2nd trimester - triple test (serology)</td>
<td>19</td>
</tr>
</tbody>
</table>
**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

<table>
<thead>
<tr>
<th></th>
<th>NT alone</th>
<th>NT with serology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NT alone</strong></td>
<td>Sensitivity: 72% (63%-79%)</td>
<td>Sensitivity: 86% (82%-90%)</td>
</tr>
<tr>
<td></td>
<td>Specificity: 94% (91% -96%)</td>
<td>Specificity: 95% (93%-96%)</td>
</tr>
<tr>
<td></td>
<td>DOR: 39 (26-60)</td>
<td>DOR: 110 (84-143)</td>
</tr>
<tr>
<td><strong>NT with serology</strong></td>
<td>Sensitivity: 86% (82%-90%)</td>
<td>Specificity: 95% (93%-96%)</td>
</tr>
<tr>
<td></td>
<td>Specificity: 95% (93%-96%)</td>
<td>DOR: 110 (84-143)</td>
</tr>
<tr>
<td></td>
<td>DOR: 123 (40-206)</td>
<td>RDOR: 2.8 (1.7-4.6), p &lt;0.0001</td>
</tr>
<tr>
<td></td>
<td>RDOR: 123 (40-206)</td>
<td></td>
</tr>
</tbody>
</table>
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

![Graph showing true positive rate against false positive rate (1-spec)]
Target region
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