Comparing the accuracy of diagnostic tests:

direct versus indirect comparisons

Mariska Leeflang
Birmingham, 19 July 2016
Comparing test accuracy

Consecutive series of patients

Randomization

Test A
Reference standard

Test B
Reference standard

Comparison of test accuracy
Comparing test accuracy

Consecutive series of patients

Test A

Test B

Reference standard

Comparison of test accuracy
Comparing test accuracy

Study A
- Consecutive series of patients
  - Test A
    - Reference standard
      - Test accuracy A

Study B
- Consecutive series of patients
  - Test B
    - Reference standard
      - Test accuracy B

Comparison of test accuracy
Direct versus Indirect comparisons

- Terminology:
  - Comparative study versus non-comparative study
  - Direct versus indirect comparisons
  - Controlled versus non-controlled comparisons

- Indirect comparisons more prone to bias than direct comparisons

- Even more true for diagnostic accuracy studies:
  - More heterogeneity: included patients, thresholds, settings, reference standards etc.
<table>
<thead>
<tr>
<th>Study</th>
<th>Ratio of relative diagnostic odds ratio (95% CI)</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CT vs US: unequivocal appendicitis</td>
<td>1.22 (0.39, 3.87)</td>
<td></td>
</tr>
<tr>
<td>SCT vs US: hepatocellular carcinoma</td>
<td>0.99 (0.10, 10.03)</td>
<td></td>
</tr>
<tr>
<td>MRI vs SCT: hepatocellular carcinoma</td>
<td>3.53 (0.46, 26.97)</td>
<td></td>
</tr>
<tr>
<td>MRI vs US: hepatocellular carcinoma</td>
<td>17.65 (1.74, 179.02)</td>
<td></td>
</tr>
<tr>
<td>CT vs US: head &amp; neck cancer</td>
<td>0.46 (0.09, 2.39)</td>
<td></td>
</tr>
<tr>
<td>NMP22 vs cytology: bladder cancer</td>
<td>1.83 (0.37, 9.17)</td>
<td></td>
</tr>
<tr>
<td>CYFRA-21-1 vs CEA: malignant pleural effusion</td>
<td>0.47 (0.16, 1.39)</td>
<td></td>
</tr>
<tr>
<td>Automated rapid ELFA vs membrane ELISAs: DVT</td>
<td>4.07 (0.81, 20.43)</td>
<td></td>
</tr>
<tr>
<td>MRI vs CT: staging pelvic lymph nodes</td>
<td>0.70 (0.26, 1.93)</td>
<td></td>
</tr>
<tr>
<td>Anti-CCP vs RF: rheumatoid arthritis</td>
<td>2.29 (1.02, 5.15)</td>
<td></td>
</tr>
<tr>
<td>Cornell voltage vs Romhilt-Estes score: left ventricular hypertrophy</td>
<td>2.27 (0.53, 9.81)</td>
<td></td>
</tr>
<tr>
<td>CT vs US: appendicitis in children</td>
<td>0.43 (0.10, 1.93)</td>
<td></td>
</tr>
<tr>
<td>MRI vs CT: cervical cancer</td>
<td>2.46 (0.86, 7.01)</td>
<td></td>
</tr>
<tr>
<td>McMurray test vs JLT: meniscal lesions of the knee</td>
<td>0.03 (0.00, 1.06)</td>
<td></td>
</tr>
<tr>
<td>MRI vs TEE: thoracic aortic dissection</td>
<td>2.11 (0.01, 594.17)</td>
<td></td>
</tr>
<tr>
<td>Semiquantitative vs qualitative catheter culture: bloodstream infection</td>
<td>0.98 (0.95, 1.02)</td>
<td></td>
</tr>
<tr>
<td>Triple vs double screen: Down's syndrome</td>
<td>1.61 (0.72, 3.60)</td>
<td></td>
</tr>
<tr>
<td>CT vs US: acute appendicitis in adults</td>
<td>1.22 (0.39, 3.87)</td>
<td></td>
</tr>
<tr>
<td>PET vs CT: lung cancer</td>
<td>1.22 (0.39, 3.87)</td>
<td></td>
</tr>
<tr>
<td>CT vs EUS: oesophageal cancer</td>
<td>2.46 (0.89, 6.80)</td>
<td></td>
</tr>
<tr>
<td>MRA vs DUS: carotid stenosis</td>
<td>0.90 (0.40, 2.04)</td>
<td></td>
</tr>
<tr>
<td>MRI vs SCT: carotid stenosis</td>
<td>1.08 (0.46, 2.53)</td>
<td></td>
</tr>
<tr>
<td>LDDE vs Thallium-201</td>
<td>0.47 (0.20, 1.10)</td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>1.18 (0.89, 1.57)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Weights are from random effects analysis

By: Yemisi Takwoingi, Birmingham
Aims of this study

- In a large review about one topic: do we see the same effects?
- Can we use covariates to correct for these effects?
Study base

- 217 D-dimer test evaluations for deep venous thrombosis and 111 for pulmonary embolism were analyzed.

- Two reviewers independently extracted study characteristics, including study quality.

- More than 9 different tests.

- Seventeen comparisons could be analyzed; 14 DVT and 3 PE.

- In all studies except one, all patients received both tests in the comparison.

Example: DVT-test1 versus DVT-test4

41 test-evaluations
21 test1
20 test2
6 studies evaluated both tests
Example: DVT-test1 versus DVT-test4
SAS proc nlmixed

$$\text{logitp} = (_\text{sens} + _\text{usens} + b*\text{assay} + c*\text{comp} + d*\text{assay*comp})*_\text{dis} + (_\text{spec} + _\text{uspec} + b2*\text{assay} + c2*\text{comp} + d2*\text{assay*comp})*_\text{nondis};$$

transformation to probability:
$$p = \frac{\exp(\text{logitp})}{1+\exp(\text{logitp})};$$

no. of correct classifications within a study is assumed to be binomially distributed: model _pos ~ binomial(_n,p);

equal variance assumed for both assays and design-types
Indirect versus Direct  (DVT 1 vs 4)
What if they find a difference?

- **Comparative >> Non-comparative**
  - Both above 1: A is better than B; comp bigger difference
  - Both below 1: B is better than A; comp smaller difference
  - Non-comp below 1 and Comp above 1

- **Comparative << Non-comparative**
  - Both above 1: A is better than B; comp smaller difference
  - Both below 1: B is better than A; comp bigger difference
  - Non-comp above 1 and Comp below 1
Can we adjust for these differences?

Other factors?
E.g. differences in:
- population
- referral routes
- selection of patients
- verification of results
- flow and timing
- test conduct
- etc

Directness of comparison

Difference between test A and test B
In two out of five analyses

1. Interaction term significant for sensitivity: $P=0.006$
   1. ‘time-interval’: $P_{\text{interaction}}=0.086$
   2. ‘year of publication’: $P_{\text{interaction}}=0.096$

2. Interaction term significant for specificity: $P=0.039$
   1. ‘all results verified’: $P_{\text{interaction}}=0.160$
   2. ‘only one reference standard’: $P_{\text{interaction}}=0.083$
Illustration: test 1 versus test 4
Conclusions

- Direct comparisons may give different answers than indirect comparisons
- This is mainly at the level of significance
- Adjusting for directness seems to be possible in some instances, but no systematic effects were found
- Next step: how to deal with this information when doing a comparative accuracy meta-analysis?
- Adding too many covariates is not wise (for 1 vs 4 already 4 covariates on 41 studies)