Reference standards: problems in evaluating diagnostics in the pediatric TB setting

Lori E. Dodd
Biostatistics Research Branch
NIAID
### Diagnostic accuracy: definitions

<table>
<thead>
<tr>
<th></th>
<th>TB+</th>
<th>TB-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xpert+</td>
<td>80</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Xpert-</td>
<td>20</td>
<td>90</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

**Sensitivity:** Proportion of test positives amongst those with true disease.

\[
\text{(80/100)}
\]

**Specificity:** Proportion of test negatives amongst those without disease.

\[
\text{(90/100)}
\]

- Inherent discriminatory ability
- Not of direct clinical reference
Diagnostic accuracy: definitions

Positive predictive value:
Likelihood of disease given a positive test.
(80/90)

Negative predictive value:
Likelihood of no-disease given negative test.
(90/110)

- Clinically meaningful
- DEPEND on prevalence

<table>
<thead>
<tr>
<th></th>
<th>TB+</th>
<th>TB-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xpert +</td>
<td>80</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Xpert-</td>
<td>20</td>
<td>90</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>
These assume we know the gold standard without error
Imperfect reference standard:

• Disease state classified imperfectly
  – Disease missed
  – Disease over-called
• May result in test accuracy appearing better or worse than it truly is
## Imperfect reference: Example 1

Reference standard (say, culture) misses true disease 60% of the time

<table>
<thead>
<tr>
<th></th>
<th>TB+</th>
<th>TB-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xpert+</td>
<td>80</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Xpert-</td>
<td>20</td>
<td>90</td>
<td>110</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TB+</th>
<th>TB-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xpert+</td>
<td>32</td>
<td>58</td>
<td>90</td>
</tr>
<tr>
<td>Xpert-</td>
<td>8</td>
<td>102</td>
<td>110</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>40</td>
<td>160</td>
<td>200</td>
</tr>
</tbody>
</table>

Sensitivity=80%  
Specificity=90%  

Sensitivity=80%  
Specificity=64%
# Imperfect reference: Example 2

Reference standard (say, culture) misses only Xpert- 60% of the time

<table>
<thead>
<tr>
<th></th>
<th>TB+</th>
<th>TB-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TRUTH</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xpert+</td>
<td>80</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Xpert-</td>
<td>20</td>
<td>90</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TB+</th>
<th>TB-</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Imperfect Reference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xpert+</td>
<td>80</td>
<td>10</td>
<td>90</td>
</tr>
<tr>
<td>Xpert-</td>
<td>8</td>
<td>102</td>
<td>110</td>
</tr>
<tr>
<td>Total</td>
<td>88</td>
<td>112</td>
<td>200</td>
</tr>
</tbody>
</table>

Sensitivity = 80%
Specificity = 90%

Sensitivity = 91%
Specificity = 91%
Additional information is needed to obtain reasonable estimates of diagnostic accuracy
Latent Class Analysis

• Multiple (likely imperfect) tests
• LCA is a statistical method to estimate diagnostic test accuracy when gold standard not available.
• Calculates test accuracy and prevalence without defining what disease is
• Methods make assumptions about relationship between multiple tests (one of which is the index test)
Latent Class Analysis

Culture → ? → Smear → ? → Xpert

Unobserved disease state
Latent Class Analysis

• Various LCA methods exist
  – Different methods often give different estimates of accuracy and prevalence

• Methods rely on assumptions that cannot be evaluated
  – If assumptions not met, estimates of accuracy and prevalence not meaningful.

• No formal def’n of disease
  – Interpretation of prevalence and accuracy unclear

• Related Bayesian methods make assumptions that are difficult to verify too
Latent Class Analysis: Hearing study example

Three experimental tests for hearing impairment in infants

Gold standard known on all subjects

Ignore gold standard and apply LCA

<table>
<thead>
<tr>
<th>Estimates of sensitivity with gold standard</th>
<th>0.66</th>
<th>0.63</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCA estimates of sensitivity</td>
<td>0.84</td>
<td>0.76</td>
<td>0.89</td>
</tr>
</tbody>
</table>

| Estimates of specificity with gold standard | 0.60 | 0.64 | 0.54 |
| LCA estimates of specificity               | 0.87 | 0.87 | 0.69 |

Pepe and Janes, *Biostatistics*, 2007
Gold standard verification on subset

• If gold standard exists, but can only be collected on a random subset:
  – Expensive or time-consuming to obtain
  – Statistical methods become more practical and robust.

• If gold standard exists, but can only be collected on test positives
  – Unethical to obtain on test negatives
  – True positive and false positive rates cannot be estimated.
Gold standard verification on external sample

• Sensitivity and specificity (with respect to gold) standard known from previous study
  – e.g., culture thought to have sensitivity of 30-40% and high specificity

• Must be confident these estimates are reasonable and applicable to current study population
Clinical outcomes

For example, TB defined as:
Culture positive or clinical presentation of disease within 2 months
May not be reasonable if subjects are treated prior to development of clinical disease
Time interval will depend on rate of disease pathogenesis and risk of infection after index test sample collection
Advantage: has practical interpretation
Composite reference

Pre-define criteria based on multiple tests:
  e.g., Culture positive or smear positive or antigen detection

Expert consensus review

Experimental test must not be included in reference standard

Advantage: May better reflect clinical reality and treatment decisions
Reporting

- Multiple reference standards are possible:
  - e.g., culture, clinical outcomes, clinical scoring
- Reporting accuracy with respect to each is informative
  - Gives information about experimental test’s strengths and weaknesses
Standardized reporting criteria

Common reporting criteria critical for comparing studies and facilitating test development

Multiple sources exist:

- CONSORT
- STARD
- QUADAS