Two key properties of any test

- Accuracy (also called ‘validity’)
- Precision (also called ‘reliability’ or ‘reproducibility’)
Precision and Accuracy
Precision and Accuracy

Quantifying precision/reliability

Observer Variation

• Intraobserver agreement
  Does the same clinician get the same result when repeating a symptom or sign on a patient who is clinically unchanged?

• Interobserver agreement
  Do 2 or more observers agree on the presence or absence of a finding in a patient who experienced no change in condition?

• Kappa (κ)
  Agreement beyond chance and can be used to describe both intra- and interobserver agreement

Note: Other measures are used for continuous measurements (e.g. correlation coefficient, limits of agreement, etc)
The development of a quality appraisal tool for studies of diagnostic reliability (QAREL)

Nicholas P. Lucas\textsuperscript{a,b,\*}, Petra Macaskill\textsuperscript{b}, Les Irwig\textsuperscript{b}, Nikolai Bogduk\textsuperscript{c}

\textsuperscript{a}School of Biomedical and Health Sciences, University of Western Sydney, Narre Warren, Campbelltown, Sydney, Australia
\textsuperscript{b}Screening and Test Evaluation Program, Sydney School of Public Health, University of Sydney, Edward Ford Building, Main Campus, Sydney, Australia
\textsuperscript{c}Department of Clinical Research, Royal Newcastle Centre, University of Newcastle, Newcastle, Australia

Accepted 6 October 2009

Quantifying accuracy

- Sensitivity and Specificity
- Likelihood ratios
- Positive and Negative Predictive Value
- Diagnostic Odds Ratio
Tests with dichotomous results

A standard Phase II/III diagnostic design for accuracy estimation

- Define gold standard
- Recruit consecutive patients in whom the test is indicated (in whom the disease is suspected)
- Perform gold standard and separate diseased and disease free groups
- Perform test on all and classify them as test positives or negatives
- Set up 2 x 2 table and compute:
  - Sensitivity
  - Specificity
  - Predictive values
  - Likelihood ratios
  - Diagnostic odds ratio
Evaluating a diagnostic test

• Diagnostic 2 X 2 table*:

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Test -</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

*When test results are not dichotomous, then can use ROC curves [see later]

Sensitivity [true positive rate]

The proportion of patients with disease who test positive = \( P(T+|D+) = \frac{TP}{(TP+FN)} \)
Specificity
[true negative rate]

- **Disease present**
  - **Test positive**: True positives
  - **Test negative**: False negatives

- **Disease absent**
  - **Test positive**: False positives
  - **Test negative**: True negatives

The proportion of patients without disease who test negative: \( P(T^-|D^-) = \frac{TN}{TN + FP} \).

Predictive value of a positive test

- **Disease present**
  - **Test positive**: True positives
  - **Test negative**: False negatives

- **Disease absent**
  - **Test positive**: False positives
  - **Test negative**: True negatives

Proportion of patients with positive tests who have disease: \( P(D^+|T^+) = \frac{TP}{TP + FP} \).
Predictive value of a negative test

<table>
<thead>
<tr>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>True positives</td>
</tr>
<tr>
<td>Test negative</td>
<td>False negative</td>
</tr>
</tbody>
</table>

Proportion of patients with negative tests who do not have disease = \( P(D^-|T^-) = \frac{TN}{TN+FN} \)

Example: Serological test for TB

<table>
<thead>
<tr>
<th>Culture (gold standard)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serological Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Negative</td>
<td>54</td>
<td>28</td>
</tr>
</tbody>
</table>

68 31 99

Sensitivity = 21%
Specificity = 90%

Clin Vacc Immunol 2006;13:702-03
All accuracy measures must be reported with confidence intervals!!

Sensitivity 20.6% (95% CI 12.7, 31.6)

Specificity 90.3% (75.1, 96.7)

Positive Predictive Value 82.4% (58.9, 93.8)

Negative Predictive Value 34.2% (24.8, 44.9)

For a given test, predictive values will depend on prevalence

Effect of Prevalence on Predictive Value: Positive Predictive Value of Prostatic Acid Phosphatase for Prostatic Cancer (Sensitivity = 70%, Specificity = 90%) in Various Clinical Settings*

<table>
<thead>
<tr>
<th>Setting</th>
<th>Prevalence (Cases/100,000)</th>
<th>Positive Predictive Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General population</td>
<td>35</td>
<td>0.4</td>
</tr>
<tr>
<td>Men, age 75 or greater</td>
<td>500</td>
<td>5.6</td>
</tr>
<tr>
<td>Clinically suspicious prostatic</td>
<td>50,000</td>
<td>93.0</td>
</tr>
</tbody>
</table>

For a given test, predictive values will depend on prevalence

Likelihood Ratios (also called ‘Bayes Factor’)

- Likelihood ratio of a positive test: is the test more likely to be positive in diseased than non-diseased persons?
  - \( LR^+ = \frac{TPR}{FPR} \)
  - \( LR^+ = \frac{Pr(T+ | D+)}{Pr(T+ | D-)} \)
  - High \( LR^+ \) values help in RULING IN the disease
  - Values close to 1 indicate poor accuracy
  - E.g. \( LR^+ \) of 10 means a diseased person is 10 times more likely to have a positive test than a non-diseased person
Likelihood Ratio of a Positive Test

How more often a positive test result occurs in persons with compared to those without the target condition

\[ LR^+ = \frac{\Pr(T^+ | D^+)}{\Pr(T^+ | D^-)} \]

Likelihood Ratios

• Likelihood ratio of a negative test: is the test less likely to be negative in the diseased than non-diseased persons?

• \[ LR^- = \frac{\Pr(T^- | D^+)}{\Pr(T^- | D^-)} \]

• Low LR- values help in RULING OUT the disease
• Values close to 1 indicate poor accuracy
• E.g. LR- of 0.5 means a diseased person is half as likely to have a negative test than a non-diseased person
Likelihood Ratio of a Negative Test

How less likely a negative test result is in persons with the target condition compared to those without the target condition

\[ LR^- = \frac{Pr(T^- | D+)}{Pr(T^- | D-)} \]

LR: Impact on Likelihood of Disease

LR = 0.01  LR = 0.1  LR = 0.2  LR = 0.3  LR = 0.5  LR = 1.0  LR = 10  LR = 100

Less Likely  Less Likely  Less Likely  Less Likely  More Likely  More Likely  More Likely  More Likely

Increasing impact  Increasing impact

LR = 1

No Impact on Likelihood of Disease
LR: Impact on Likelihood of Disease

Quick review of odds vs. probability

- odds = probability / (1 - probability)
  
  $$\text{Odds}(D+) = \frac{\Pr(D+)}{1 - \Pr(D+)}$$

- probability = odds / (1 + odds)
  
  $$\Pr(D+) = \frac{\text{Odds}(D+)}{1 + \text{Odds}(D+)}$$
**Diagnostic Odds Ratio (DOR)**

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>True positives (a)</td>
<td>False positives (b)</td>
</tr>
<tr>
<td>Test negative</td>
<td>False negative (c)</td>
<td>True negatives (d)</td>
</tr>
</tbody>
</table>

\[
\text{DOR} = \frac{(a/c)}{(b/d)} = \frac{ad}{bc} = \frac{\text{Odds of } T+|D+}{\text{Odds of } T+|D-}
\]

**Example: Serological test for TB**

<table>
<thead>
<tr>
<th>Culture (gold standard)</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serological Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>14</td>
<td>3</td>
</tr>
<tr>
<td>Negative</td>
<td>54</td>
<td>28</td>
</tr>
</tbody>
</table>

\[
\text{LR+} = 2 \\
\text{LR-} = 0.9 \\
\text{DOR} = 2.4
\]

*Clin Vacc Immunol 2006;13:702-03*
Using LRs in practice

Scenario:
- Mr. A, a 27-year old male factory worker
- Fever and productive cough for the past 3 weeks
- Lost weight

Assess the patient and estimate the baseline risk (pre-test probability)

Based on initial history, how likely is it that Mr. A has pulmonary tuberculosis?

How might the result of a serological test change the likelihood of TB in this patient?
Likelihood Ratios

**Pre-Test Probability**

Mr. A
Pre-Test Prob. 50%

Serological test
LR+ = 2

Post-Test Probability

Post-Test Prob. 70%

Likelihood Ratios

**Pre-Test Probability**

Mr. A
Pre-Test Prob. 50%

Serological test
LR- = 0.9

Post-Test Probability

Post-Test Prob. 45%
Using LRs in practice

Scenario:
- Ms. B, a 18 year old engineering student
- Fever and non-productive cough for the past 4 days
- Nobody in the household has had TB

Likelihood Ratios

<table>
<thead>
<tr>
<th>Pre-Test Probability</th>
<th>Post-Test Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms. B Pre-Test Prob. 10%</td>
<td>Post-Test Prob. 20%</td>
</tr>
</tbody>
</table>

Serological test
LR+ = 2
**Likelihood Ratios**

**Pre-Test Probability**

- Ms. B
- Pre-Test Prob. 10%

**Post-Test Probability**

- Post-Test Prob. 10%

---

**Serological test**

LR+ = 0.9

---

**Where do we get LRs from?**

The Rational Clinical Examination: Evidence-Based Clinical Decision Making

Copyright © American Medical Association. All rights reserved. | JAMA | The McGraw-Hill Companies, Inc.

Quick Reference

<table>
<thead>
<tr>
<th>Prior Probability</th>
<th>Test/Finding</th>
<th>LR+</th>
<th>LR-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1: biopsy on precision and accuracy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 2: Abdominal Abdominal Impression</td>
<td>Occurs in 4% to 8% of older men. The prevalence in older women is less than 2%.</td>
<td>15 (9.6 - 29)</td>
<td>0.51 (0.38 - 0.67)</td>
</tr>
<tr>
<td>Physical examination for anorexia</td>
<td></td>
<td>12 (7.4 - 29)</td>
<td>0.62 (0.45 - 0.81)</td>
</tr>
<tr>
<td>Physical examination for anorexia &gt; 2.0 cm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 3: Systolic-Diastolic Blood Pressure</td>
<td></td>
<td>39 (18 - 145)</td>
<td>0.62 (0.45 - 0.81)</td>
</tr>
</tbody>
</table>

The Rational Clinical Examination

Copyright © American Medical Association. All rights reserved. | JAMA | The McGraw-Hill Companies, Inc.
Are sens/spec and LRs inherent properties of a test?

- Most textbooks will say that sens and spec do not depend on disease prevalence
- This is not true
- In reality, sens/spec and LRs vary across populations because of differences in disease spectra (case-mix) and several other factors
- This is equivalent to “effect modification” in epidemiology
Example

Sens and Spec across populations

Ex:
Sensitivity + specificity of serum CEA for detection of colorectal cancer, across stages

Variation in performance in high vs low endemic countries: example

T-cell interferon-γ release assays for the rapid immunodiagnosis of tuberculosis: clinical utility in high-burden vs. low-burden settings

Keerati Dheeran°, Richard van Zyl Smit°, Motasim Badri° and Madhukar Pai°

High incidence countries

Low incidence countries
Tests with continuous or multi-level results

Example: WBC count in bacteremia

Figure 4.4 Histogram showing distributions of the nonbacteremic and bacteremic populations across the WBC count intervals.

Table 4.3. Sensitivity and specificity of the WBC count as a predictor of bacteremia at different cut-offs for considering the test “positive” (data from Lee and Harper 1998)

<table>
<thead>
<tr>
<th>WBC count interval (× 1,000/µL)</th>
<th>Percent of bacteremia patients in interval</th>
<th>Percent of no bacteremia patients in interval</th>
<th>Sensitivity (using bottom of interval as cut-off)</th>
<th>1 – Specificity (using bottom of interval as cut-off)</th>
</tr>
</thead>
<tbody>
<tr>
<td>≥30</td>
<td>11.8%</td>
<td>0.8%</td>
<td>11.8%</td>
<td>0.8%</td>
</tr>
<tr>
<td>25 to &lt;30</td>
<td>9.4%</td>
<td>1.8%</td>
<td>21.3%</td>
<td>2.6%</td>
</tr>
<tr>
<td>20 to &lt;25</td>
<td>26.8%</td>
<td>5.4%</td>
<td>48.0%</td>
<td>8.0%</td>
</tr>
<tr>
<td>15 to &lt;20</td>
<td>37.8%</td>
<td>15.5%</td>
<td>85.8%</td>
<td>23.5%</td>
</tr>
<tr>
<td>10 to &lt;15</td>
<td>11.8%</td>
<td>32.1%</td>
<td>97.6%</td>
<td>55.6%</td>
</tr>
<tr>
<td>5 to &lt;10</td>
<td>2.4%</td>
<td>38.1%</td>
<td>100%</td>
<td>93.7%</td>
</tr>
<tr>
<td>0 to &lt;5</td>
<td>0.0%</td>
<td>6.3%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 4.5 ROC curve corresponding to the distributions in Figure 4.4.
Multi-level likelihood ratios

**Table 4.4. Likelihood ratios for WBC and bacteremia (from Lee and Harper 1998)**

<table>
<thead>
<tr>
<th>WBC Count (×1,000/µL)</th>
<th>Bacteremia</th>
<th>No bacteremia</th>
<th>LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>30–35</td>
<td>11.8%</td>
<td>0.8%</td>
<td>15.2</td>
</tr>
<tr>
<td>25–30</td>
<td>9.4%</td>
<td>1.8%</td>
<td>5.3</td>
</tr>
<tr>
<td>20–25</td>
<td>26.8%</td>
<td>5.4%</td>
<td>4.9</td>
</tr>
<tr>
<td>15–20</td>
<td>37.8%</td>
<td>15.5%</td>
<td>2.4</td>
</tr>
<tr>
<td>10–15</td>
<td>11.8%</td>
<td>32.1%</td>
<td>0.37</td>
</tr>
<tr>
<td>5–10</td>
<td>2.4%</td>
<td>38.1%</td>
<td>0.06</td>
</tr>
<tr>
<td>0–5</td>
<td>0.0%</td>
<td>6.3%</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Using ROCs to compare tests

Figure 2. Whole blood was stimulated with Mycobacterium tuberculosis specific antigens or saline. The diagnostic potential of interferon (IFN)-γ, IFN-γ inducible protein (IP)-10 and monocyte chemotactic protein (MCP)-2 was determined by receiver operating characteristic curve analysis using antigen-dependent values. Students were used as gold standard for noninfected tuberculosis patients were used as gold standard for infected. ● maximum Youden’s index (Yi) for IFN-γ (4 pg·mL(-1)), □ maximum Yi for MCP-2 (97 pg·mL(-1)). ■ cut-off applied in the QuantiFERON T-Tube test (Collins, Camague, Australia) 17.5 pg·mL(-1), △ maximum Yi for IP-10 test, used as cut-off for the IP-10 test (227 pg·mL(-1)), ○ cut-off for the IP-10 test (573 pg·mL(-1)), ▲ selected pragmatic cut-off for the IP-10 test (665 pg·mL(-1)) ——— IFN-γ, - - - - IP-10, --- MCP-2.

Ruhwald, ERJ 2008

Figure 4.2 Test discriminates poorly between patients with disease (D+) and patients without disease (D-). (A) The distribution of test results in D+ patients is very similar to the distribution in D- patients. (B) This "bad" ROC curve approaches a 45 degree diagonal line.

Figure 4.3 Test discriminates well between patients with the disease (D+) and patients without the disease (D-). (A) The distribution of test results in D+ patients differs substantially from the distribution in D- patients. (B) This "good" ROC curve nears the upper left corner of the grid.
After understanding ROC curves, it should be obvious that:

- the case of a dichotomous test accuracy (i.e. the usual 2 x 2 table) is merely a single point on some underlying ROC curve.
- in other words, all tests have some underlying ROC curve
- we can easily change the sens/spec by shifting the point on the ROC curve.

ROC: pros and cons

Pros:
- Provides a wholistic picture (a global assessment of a test’s accuracy)
- Not dependent on disease prevalence
- Does not force us to pick a single cut-off point
- Shows the trade off between sens and spec
- Great for comparing accuracy of competing tests
- Can be applied to any diagnostic system: weather forecasting, lie detectors, medical imaging, to detection of cracks in metals!
ROC: pros and cons

Cons:

- Not very intuitive for clinicians; the ROC and AUC cannot be directly used for any given patient
- Clinicians prefer simple yes/no test results
- You can have the same AUC, but different shapes
- Does not fit into the EBM framework of working with LRs and probabilities
- Very hard to meta-analyze

Articles

Measuring the Accuracy of Diagnostic Systems

JOHN A. SWEET

Diagnostic centers of several kinds are used to obtain information about the Diagnostic methods of a decision or event. For example, a diagnostic test may yield a "YES" or "NO" result. The results obtained from sets of diagnostic tests can be used to evaluate the accuracy of the diagnostic test. The accuracy of the diagnostic test can be evaluated by comparing the results obtained from the diagnostic test with the results obtained from a gold standard or reference test.

Two classic papers on ROC

The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve

James A. Hanley, Ph.D.
Barbara J. McNeil, M.D., Ph.D.