Evaluation of Diagnostic Tests
July 18, 2011
Introduction to Clinical Research:
A Two-week Intensive Course

Milo A. Puhan, MD, PhD
Department of Epidemiology
Johns Hopkins
Bloomberg School of Public Health
Today’s learning objectives

- To describe the use of diagnostic tests in practice
- To know about the phases of diagnostic test evaluation
- To know about biases that affect diagnostic test accuracy studies
- To know an approach for designing randomized trials for diagnostic test evaluation
Today’s key messages

- Diagnostic tests are used for screening, as add-on tests, for triage, as replacement tests or for monitoring.

-Diagnostic test evaluation includes (at least) test accuracy studies, health outcomes studies and cost effectiveness studies.

- Biases related to the spectrum of patients and to the reference standard affect estimates of diagnostic test accuracy most.

- To ensure that informative RCTs, identify the critical comparisons between the old and new test-treatment.
Recommended books

Evidence-Based Diagnosis
Thomas B. Newman and Michael A. Kohn
Publisher: Cambridge University Press
Pub. Date: 2009
ISBN: 978-0-521-71402-0

Evidence Base of Clinical Diagnosis
Theory and Methods of Diagnostic Research by Andre Knottnerus and Frank Buntinx (Editor),
Publisher: Wiley, John & Sons,
Pub. Date: November 2008
Part I: Stages of diagnostic test evaluation
Brain natriuretic peptide (BNP) for diagnosing heart failure

Cut-offs between 15 and 100 pg/ml used

Toma et al Cardiovascular Medicine 2007;10:27–33
Diagnostic test should reduce uncertainty

Example 1: BNP for diagnosing heart failure in patients with dyspnea in ER

Pre-test probability:
- 20%

Diagnostic test:
- BNP

Post-test probability:
- 2%
Diagnostic test (may) have an indirect and direct impact on health outcomes
Use of a test (BNP) in practice

Stage of clinical management

- Early diagnostic work-up
  - Setting: Primary care
  - Available information: Patient history, Physical exam
  - Purpose of test: Screening, Triage

- Diagnosis not yet established
  - Setting: ER
  - Available information: Chest X-ray, ECG
  - Purpose of test: Add-on, Replacement

- Diagnosis established
  - Setting: ER, Specialized care
  - Available information: Echocardiography
  - Purpose of test: Prognostic information

- Under treatment
  - Setting: Primary or specialized care
  - Available information: Diagnostic work-up + treatments
  - Purpose of test: Monitor disease process
BNP used for triage to avoid unnecessary work-up

Heart failure among differential diagnoses after taking patient history and physical exam

- Positive test result
  - Heart failure work-up ± treatment
  - Health outcomes

- Negative test result
  - Consider other diagnoses
  - Health outcomes
BNP used as add-on test

Heart failure suggested after patient history, physical exam, ECG and Chest x-ray

Positive test result
Heart failure work-up ± treatment
Health outcomes

Negative test result
Consider other diagnoses
Health outcomes
BNP used as replacement test for echocardiography

Heart failure suggested after patient history, physical exam, ECG and Chest x-ray

Positive test result
- Treatment
- Health outcomes

Negative test result
- Consider other diagnoses
BNP used as a prognostic marker

Risk of 5-year mortality

- 0-10%
- >10-20%
- >20-30%
- >30%

Improved prediction by adding BNP?
BNP used for monitoring

Heart failure diagnosis and treatment established

Not in therapeutic range

- Adapt treatment

In therapeutic range

- Treatment unchanged
Test phases for diagnostic tests

**Phase I**
Investigates whether test results are different for patients ± disease

**Phase II**
Investigates whether patients with disease are more likely to have positive test results compared to patients without disease

**Phase III**
Investigates how well the test distinguishes between patients ± disease in patients suspected of having the disease

**Phase IV**
Investigates how informative a test is considering additional information available at the moment of testing.

**Phase V**
Investigates whether using the test leads to better health outcomes

**Phase VI**
Investigates whether using the test leads to better health outcomes at acceptable costs
Phases of diagnostic test evaluation

**Phase I**
- Healthy subjects
- Patient with heart failure

**Phase II**
- Patient with heart failure
- Healthy subjects

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>90</td>
<td>20</td>
<td>90%</td>
<td>80%</td>
<td>82%</td>
<td>89%</td>
<td>36</td>
<td>4.5</td>
<td>0.13</td>
</tr>
<tr>
<td>Test negative</td>
<td>10</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phases of diagnostic test evaluation

Phase III

Patients suspected of having disease

<table>
<thead>
<tr>
<th></th>
<th>heart failure</th>
<th>no heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test positive</td>
<td>85</td>
<td>250</td>
</tr>
<tr>
<td>Test negative</td>
<td>15</td>
<td>650</td>
</tr>
</tbody>
</table>

Sens: 85%  Spec: 72%

PPV: 25%  DOR: 15
NPV: 96%  +LR: 3.0
-LR: 0.21

Phase IV

Patients suspected of having disease

Age, gender, smoking and coronary heart disease status known

Probability of heart failure?
Phase V

**Randomized trial**

Patients suspected of having heart failure

![Decision tree diagram]

*Health outcome*

- Treat and follow-up
- ± Treat and follow-up

or

**Before-after study**

*Until to 2000*

Patients suspected of having heart failure

± Treat and follow-up

*Health outcome*

*From 2001*

Patients suspected of having heart failure

![Decision tree diagram]

- Treat and follow-up
Phases of diagnostic test evaluation

Phase VI

Randomized trial

Patients suspected of having heart failure

\[ R \]

\[ \pm \text{Treat and follow-up} \]

\[ \begin{array}{c}
\circ \text{Treat and follow-up} \\
- \quad \\
\end{array} \]

\[ \text{Health outcome} \]

\[ + \text{ costs} \]
Outcomes for phase V and VI studies
Study designs for the evaluation of diagnostic tests

Phase I  Cross-sectional case-control study (pro- or retrospective)

Phase II Cross-sectional case-control study (pro- or retrospective)

Phase III Cross-sectional study of patients suspected of having disease (pro- or retrospective)

Phase IV Cross-sectional study of patients suspected of having disease

Phase IV Randomized trial or before-after study

Phase V Cost effectiveness study

Systematic review
Test phases are not well established for diagnostic studies

Clinical Trials

- Phase 1
  - Safety (maximum tolerated dose)
  - Pharmacokinetics

- Phase 2
  - Prelim. Efficacy
  - Dosage response

- Phase 3
  - Efficacy
  - Clinically relevant effects

- Phase 4
  - Safety surveillance

Diagnostic studies

19 models have been proposed...

Lijmer et al. Med Decis Making 2009; 29; E13
Synthesis of models of diagnostic test evaluation phases

Table 1  Summary of Proposals for the Phased Evaluation of Medical Tests

<table>
<thead>
<tr>
<th>Levels/Phases</th>
<th>van der Loop</th>
<th>Zweig</th>
<th>Guyatt</th>
<th>Freedman Memorandum</th>
<th>Fryback</th>
<th>Kent</th>
<th>Taylor</th>
<th>Silverstein</th>
<th>Schouw</th>
<th>Mackenzie</th>
<th>Pearl</th>
<th>Houn</th>
<th>Gatsonis</th>
<th>Sackett</th>
<th>Haddow</th>
<th>Pepe</th>
<th>Taube</th>
<th>Brueel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical efficacy</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1–3</td>
<td>1</td>
<td>1</td>
<td>1–3</td>
<td>1–2</td>
<td>1</td>
<td>1–3</td>
<td>1–3</td>
<td>1</td>
<td>1–3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Intended use</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnostic accuracy</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4–6</td>
<td>3</td>
<td>1–2</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Usual range</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subgroups</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clinical population</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diagnostic thinking efficacy</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Therapeutic efficacy</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3–4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient outcome efficacy</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4–3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Societal efficacy</td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Technical requirements → Test accuracy → Effects on decisions → Effects on patient outcomes → Effects on health care system

Lijmer et al. Med Decis Making 2009; 29; E13
Part II: Biases in diagnostic test accuracy studies
The diagnostic test accuracy study

**Aim:** To obtain (unbiased) estimates of diagnostic test accuracy such as sensitivity, specificity, likelihood ratios, etc.

**Study design:** Cross-sectional study with patients suspected of having disease (phase III)

<table>
<thead>
<tr>
<th>Patients suspected of having disease</th>
<th>Index test</th>
<th>Reference test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient with disease</td>
<td>Test positive</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Test negative</td>
<td>15</td>
</tr>
<tr>
<td>Patients without disease</td>
<td></td>
<td>Patients without disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
</tr>
<tr>
<td></td>
<td></td>
<td>650</td>
</tr>
</tbody>
</table>

---

25
The “perfect” diagnostic test accuracy study

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described (setting)
- Prior information available (comprehensive ascertainment of patient history, exam, tests)

- Aim of test (triage, replacement, add-on)
- Well defined protocol
- Well defined threshold
- Performed for all patients
- Maximized reliability (intra- and inter-rater)
- Blinded towards reference test

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test
Sources of bias in diagnostic test accuracy studies

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described (setting)
- Prior information available (comprehensive ascertainment patient history, exam, tests)
- Aim of test (triage, replacement, add-on)
- Well defined protocol
- Well defined threshold
- Performed for all patients
- Maximized reliability (intra- and inter-rater)
- Blinded towards reference test
- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test
### Sources of variability in diagnostic test accuracy studies

<table>
<thead>
<tr>
<th>Patients suspected of having disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index test</td>
</tr>
<tr>
<td>Reference test</td>
</tr>
</tbody>
</table>

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described (setting)
- Prior information available (comprehensive ascertainment patient history, exam, tests)

- Aim of test (triage, replacement, add-on)
- Well defined protocol
- Well defined threshold
- Performed for all patients
- Maximized reliability (intra- and inter-rater)
- Blinded towards reference test

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test
Bias from definition or recruitment of population - Spectrum bias

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described
- Prior information available (comprehensive ascertainment patient history, exam, tests)

Setting: Outpatient cardiology clinic

Patients: referred from primary care with suspected new heart failure

Number of patients

Clinical manifestations

- None
- Very severe
Bias from definition or recruitment of population - Spectrum bias

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described
- Prior information available (comprehensive ascertainment patient history, exam, tests)

Setting: Outpatient cardiology clinic

Patients: referred from primary care with suspected new heart failure

BNP levels

Clinical manifestations

None

Very severe
Patients suspected of having disease

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described
- Prior information available (comprehensive ascertainment patient history, exam, tests)

Setting: Outpatient cardiology clinic

Patients: referred from primary care with suspected new heart failure

BNP levels vs Clinical manifestations

Bias from definition or recruitment of population - Spectrum bias
Bias from definition or recruitment of population - Spectrum bias

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described
- Prior information available (comprehensive ascertainment patient history, exam, tests)

Setting: Outpatient cardiology clinic

Patients: referred from primary care with suspected new heart failure

![Biplot showing BNP levels vs. clinical manifestations]

<table>
<thead>
<tr>
<th>Heart failure</th>
<th>No heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>9</td>
</tr>
<tr>
<td>Test -</td>
<td>2</td>
</tr>
</tbody>
</table>

Sens: 82% Spec: 74%
Bias from definition or recruitment of population - Spectrum bias

Patients suspected of having disease

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described
- Prior information available (comprehensive ascertainment patient history, exam, tests)

Setting: Outpatient cardiology clinic

Patients: referred from primary care with suspected new heart failure

Clinical manifestations

<table>
<thead>
<tr>
<th>BNP levels</th>
<th>None</th>
<th>Very severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test -</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test +
- Heart failure: 5
- No heart failure: 2

Test -
- Heart failure: 1
- No heart failure: 10

Sens: 83%  Spec: 83%
Healthy controls frequently included in DTA studies

**Purpose:** Urinary tumor markers that help in the early detection of bladder cancer promise a significant improvement in sensitivity, specificity and convenience over conventional, invasive diagnostic tests. We assessed the diagnostic efficacy of hyaluronidase (HYAL1) and survivin for early bladder cancer detection.

**Materials and Methods:** The study included 166 patients diagnosed with bladder carcinoma, 112 with benign bladder lesions and 100 healthy volunteers who served as controls. All underwent serological assessment of schistosomiasis antibody, urine cytology, and hyaluronidase (HYAL1) and survivin RNA estimation by qualitative and semiquantitative reverse transcriptase-polymerase chain reaction in urothelial cells from voided urine.

**Results:** Positivity rates of HYAL1 RNA and survivin RNA on qualitative reverse transcriptase-polymerase chain reaction were significantly different among the 3 groups. Mean rank using semiquantitative method was increased in the malignant vs the other groups. The best cutoff for HYAL1 and survivin RNA was 0.25 each. Using these cutoffs HYAL1 and survivin RNA sensitivity was 91% and 75%, respectively, with absolute specificity. HYAL1 RNA detected all patients with stages 0 and I bladder cancer ($p < 0.037$). Urine cytology sensitivity improved when combined with hyaluronidase or survivin RNA on semiquantitative reverse transcriptase-polymerase chain reaction.

**Conclusions:** The detection of urinary HYAL1 and survivin RNA is a promising noninvasive test for bladder cancer early detection. HYAL1 RNA was more sensitive and specific than urine cytology. Semiquantitative reverse transcriptase-polymerase chain reaction is favored for its high sensitivity and specificity.
Often strong conclusions despite being phase II studies

unknown. The aim of this study was to determine the value of serum GP73 in the diagnosis of HCC.

**Methods** Serum GP73 and alpha-fetoprotein (AFP) were compared in a total of 4217 human subjects in this multicentre study, including 1690 healthy adults, 337 hepatitis B virus (HBV) carriers, 512 patients with cirrhosis, 789 patients with HCC, 61 patients with other malignant liver lesions, 206 patients with benign liver lesions and 622 patients with 14 different kinds of non-liver cancers. The main outcome measures were the specificity and sensitivity of GP73 in patients at risk for the development of HCC.

**Results** Using 8.5 relative units as a cut-off value, the sensitivity and specificity of serum GP73 for HCC were 74.6% (95% CI 71.5% to 77.6%) and 97.4% (95% CI 96.8

**Conclusions** GP73 is an accurate serum marker for the detection of HCC and its recurrence after surgery, with higher sensitivity and specificity than AFP. Clinical implementation of serum GP73 measurement as a standard test for HCC is recommended.

Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

Meta-analyses of at least 5 test accuracy studies

18 meta-analyses found including 193 studies

Quality assessment + diagnostic odds ratio (DOR)

Association of quality of studies with diagnostic odds ratio

PubMed, EMBASE, DARE, Cochrane

\[
\frac{80 \times 450}{20 \times 50} = 36
\]

DOR low quality

DOR high quality

= relative DOR

Relative Diagnostic Odds Ratios of the 9 Study Characteristics

- **Case-Control**: $3.0 \ (2.0-4.5)$
- Different Reference Tests: $2.2 \ (1.5-3.3)$
- Partial Verification: $1.0 \ (0.8-1.3)$
- Not Blinded: $1.3 \ (1.0-1.9)$
- Nonconsecutive: $0.9 \ (0.7-1.1)$
- Retrospective: $1.0 \ (0.7-1.4)$
- No Description Test: $1.7 \ (1.1-2.5)$
- No Description Population: $1.4 \ (1.1-1.7)$
- No Description Reference: $0.7 \ (0.6-0.9)$

Bias from definition or recruitment of population – Prospective vs retrospective

- Spectrum of patients adequate for setting and intention-to-diagnose
- Prospective data collection
- Consecutive recruitment
- Referral filter described
- Prior information available (comprehensive ascertainment patient history, exam, tests)

Patients suspected of having disease

• No verification of disease status

Validated BNP measurements?

Same reference standard for all?

Same assessors for reference standard for all?
Relative Diagnostic Odds Ratios of the 9 Study Characteristics

- **Case-Control**: 3.0 (2.0-4.5)
- **Different Reference Tests**: 2.2 (1.5-3.3)
- **Partial Verification**: 1.0 (0.8-1.3)
- **Not Blinded**: 1.3 (1.0-1.9)
- **Nonconsecutive**: 0.9 (0.7-1.1)
- **Retrospective**: 1.0 (0.7-1.4)
- **No Description Test**: 1.7 (1.1-2.5)
- **No Description Population**: 1.4 (1.1-1.7)
- **No Description Reference**: 0.7 (0.6-0.9)

Index test

- Aim of test (triage, replacement, add-on)
- Well defined protocol
- Well defined threshold
- Performed for all patients
- Maximized reliability (intra- and inter-rater)
- Blinded towards reference test

Bias from index test - Test review bias (blinding)

Blinding

Heart sound level

No blinding
Relative Diagnostic Odds Ratios of the 9 Study Characteristics

- Case-Control: 3.0 (2.0-4.5)
- Different Reference Tests: 2.2 (1.5-3.3)
- Partial Verification: 1.0 (0.8-1.3)
- Not Blinded: 1.3 (1.0-1.9)
- Nonconsecutive: 0.9 (0.7-1.1)
- Retrospective: 1.0 (0.7-1.4)
- No Description Test: 1.7 (1.1-2.5)
- No Description Population: 1.4 (1.1-1.7)
- No Description Reference: 0.7 (0.6-0.9)

Biases from reference test

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

Challenges for verification of disease status by reference test

- Partial verification bias
- Differential verification bias
- Disease progression bias
- Incorporation bias
- Blinding – Diagnosis review bias
Reference test

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

How to define heart failure?

- Clinical criteria?
- Echocardiography?
- Response to treatment?
Disease progression bias

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

Patients suspected of having disease

Index test

Delay of reference test
- Disease status may have changed
- Particularly problematic for acute diseases (infections)
- Problematic for chronic diseases if prognostic criteria used as reference standard

Reference test
Partial verification bias

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

• No verification of disease status
  - Missed to perform reference test
    → randomly
  → in patients with lower disease probability
Partial verification bias

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

**Reference test**

<table>
<thead>
<tr>
<th>Heart failure</th>
<th>No heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test +</strong></td>
<td><strong>Test -</strong></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
</tr>
</tbody>
</table>

Sens: 82% Spec: 74%

<table>
<thead>
<tr>
<th>Heart failure</th>
<th>Unclear</th>
<th>No heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test +</strong></td>
<td><strong>Test -</strong></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>19</td>
<td></td>
</tr>
</tbody>
</table>

Complete case analysis

Sens: 78% Spec: 76%

If uncorrelated to Test +/- tends to underestimate accuracy

- prevalence low → sensitivity more affected
- prevalence high → specificity more affected
Partial verification bias

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

Reference test

<table>
<thead>
<tr>
<th></th>
<th>Heart failure</th>
<th>Unclear</th>
<th>No heart failure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>7</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Test -</td>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
</tbody>
</table>

Sens: 85% Spec: 76%
Methods to correct for partial and differential verification bias

<table>
<thead>
<tr>
<th>Main classification</th>
<th>Main characteristic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subdivision</strong></td>
<td></td>
</tr>
<tr>
<td>A. Impute or adjust for</td>
<td><strong>Impute the outcome of the reference standard in those patients who did not receive verification by the reference standard or adjust estimates of accuracy based on complete cases</strong></td>
</tr>
<tr>
<td>missing data on reference standard</td>
<td></td>
</tr>
<tr>
<td>B. Correct imperfect</td>
<td><strong>Correct estimates of accuracy or perform sensitivity analysis to examine the impact of using imperfect reference standard based on external data about the degree of imperfection</strong></td>
</tr>
<tr>
<td>reference standard</td>
<td></td>
</tr>
<tr>
<td>C. Construct reference standard</td>
<td><strong>Information from different tests is combined to construct the reference standard outcome. Groups of patients receive either different tests (differential verification and discrepant analysis) or the same set of tests after which these results are combined by:</strong></td>
</tr>
<tr>
<td>Differential verification</td>
<td>(a) deterministic predefined rule (composite reference standard)</td>
</tr>
<tr>
<td>Discrepant analysis</td>
<td>(b) consensus procedure among experts (panel diagnosis)</td>
</tr>
<tr>
<td>Composite reference standard</td>
<td>(c) a statistical model based on actual data (latent class analysis)</td>
</tr>
<tr>
<td>Panel or consensus diagnosis</td>
<td></td>
</tr>
<tr>
<td>Latent class analysis</td>
<td></td>
</tr>
<tr>
<td>D. Validate index test results</td>
<td><strong>Explore meaningful relations between index test results and other relevant clinical characteristics. An important way to validate is to use dedicated follow-up to capture clinical events of interest in relation to index test results, including randomised diagnostic studies</strong></td>
</tr>
</tbody>
</table>
Incorporation bias

- Good measure of target condition
- Well defined protocol
- Performed for all patients
- Performed at same time as index test
- Maximized reliability (intra- and inter-rater)
- Blinded towards index test

Reference test

Diagnosis of multiple sclerosis

Index test

MRI

Reference test

Clinical follow-up, cerebrospinal fluid + MRI
Relative Diagnostic Odds Ratios of the 9 Study Characteristics

- **Case-Control**: 3.0 (2.0-4.5)
- **Different Reference Tests**: 2.2 (1.5-3.3)
- **Partial Verification**: 1.0 (0.8-1.3)
- **Not Blinded**: 1.3 (1.0-1.9)
- **Nonconsecutive**: 0.9 (0.7-1.1)
- **Retrospective**: 1.0 (0.7-1.4)
- **No Description Test**: 1.7 (1.1-2.5)
- **No Description Population**: 1.4 (1.1-1.7)
- **No Description Reference**: 0.7 (0.6-0.9)

# Solutions to minimize bias from reference standard

<table>
<thead>
<tr>
<th>Sources of bias</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Applies to all</td>
<td>- Learn from previous studies</td>
</tr>
<tr>
<td></td>
<td>- Do systematic review</td>
</tr>
<tr>
<td></td>
<td>- Write protocol</td>
</tr>
<tr>
<td></td>
<td>- Take time and find consensus</td>
</tr>
<tr>
<td>- Case definition</td>
<td>- Double/triple reading</td>
</tr>
<tr>
<td>- Intra- and inter rater reliability</td>
<td>- Adjudication committee (expert panels)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>- Partial verification bias</td>
<td>- Ensure quality control for data collection</td>
</tr>
<tr>
<td>- Differential verification bias</td>
<td>- Use “realistic” reference tests</td>
</tr>
<tr>
<td></td>
<td>- Foresee missings and consider prognostic criteria</td>
</tr>
<tr>
<td></td>
<td>- Statistical methods</td>
</tr>
<tr>
<td>- Disease progression bias</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- No relevant time delay between index and reference test</td>
</tr>
<tr>
<td>- Incorporation bias and diagnosis review bias</td>
<td>- Appropriate case definition</td>
</tr>
<tr>
<td></td>
<td>- Ensure blinding</td>
</tr>
</tbody>
</table>
Part III: Randomized trials for diagnostic test evaluation
## Test phases for diagnostic tests

<table>
<thead>
<tr>
<th>Phase</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I</td>
<td>Investigates whether test results are different for patients ± disease</td>
</tr>
<tr>
<td>Phase II</td>
<td>Investigates whether patients with disease are more likely to have positive test results compared to patients without disease</td>
</tr>
<tr>
<td>Phase III</td>
<td>Investigates how well the test distinguishes between patients ± disease in patients suspected of having the disease</td>
</tr>
<tr>
<td>Phase IV</td>
<td>Investigates how informative a test is considering additional information available at the moment of testing.</td>
</tr>
<tr>
<td>Phase V</td>
<td>Investigates whether using the test leads to better health outcomes</td>
</tr>
<tr>
<td>Phase VI</td>
<td>Investigates whether using the test leads to better health outcomes at acceptable costs</td>
</tr>
</tbody>
</table>
Phase V

Randomized trial

Patients suspected of having heart failure

R

+ Treat and follow-up

− ± Treat and follow-up

Health outcome
The critical questions when assessing patient outcomes

- What is the intended incremental value of the test on outcomes (short- and long-term patient outcomes and costs)?

- What type of evidence is needed to assess this incremental value?

Recommended approach

- Define the purpose of the test
- Display the existing test-treatment strategy
- Display the new test-treatment strategy
- Identify the critical comparison to assess the incremental value
- Assess whether existing evidence suffices or if RCTs are required
Test-treatment strategy for replacement tests

Target population
- Prior tests
  - Existing test
    - Test result
      - Management
        - Test pos pathway: TF, FP
        - Test neg pathway: TN, FN
  - Test safety & other attributes?
    - Sensitivity and specificity?
      - Change in management?
        - Test pos pathway: TF, FP
        - Test neg pathway: TN, FN
      - Treatment effects?
        - Management

Patient outcomes
Example: Liquid-based cytology to replace Pap smear for cervical cancer screening in order to reduce repeated testing (poor Pap smear quality)

Target population

Pap smear → Test result → Management

Test procedure identical for women
Reference standard for both: Biopsy SR show: Sensitivity and specificity very similar

LBC → Test result → Management

No change in management

Test pos pathway
Test neg pathway

Test pos pathway
Test neg pathway

Treatment effects not different

No long-term RCT needed

RCT to compare short-term effects from testing

Patient outcomes
**Test-treatment strategy for add-on tests**

1. **Target population**
   - Prior tests

2. **Existing test**
   - Test result
     - Management
       - Test pos pathway A
       - Test neg pathway B

3. **Test safety & other attributes?**
   - Sensitivity & specificity
     - Treated populations?
       - Treatment effects?

4. **Test result**
   - Management
     - Test pos pathway A*
     - Test neg pathway B*

5. **Patient outcomes**
Example: MRI as add-on test to mammography and ultrasound in breast cancer screening to detect extra cases and inform decision on type of surgery

Target population
Prior tests

Mammography + ultrasound
Test result
Management

Test pos
BCS or mastectomy
Test neg
continue screening

MRI more sensitive
→ more cases and detects multifocal disease

Test pos BCS or mastectomy
Treatment effects unclear

Test neg
MRI
Test result
Management

Test pos BCS or mastectomy*

Test neg
MRI
Test result
Management

Test neg
continue screening

Patient outcomes

RCT to compare short-term effects from testing

RCT to compare long-term effects of different treatments (different surgery and populations)
Test-treatment strategy for triage tests

Target population
Prior tests

Existing test

Test result

Test safety & other attributes?

Triage test

Test result

Management

Treated populations?

Test pos
Add existing test

Test pos
Management

Test neg pathway B*

Test neg
pathway B

Test pos
pathway A

Test neg
pathway B

Treatment effects?

Management

Test pos
pathway A

Test neg
pathway B

Patient outcomes
Example: Triage D-Dimer test to reduce the number of ultrasounds in patients at low risk for DVT

Target population
Prior tests

Ultrasound
Test result

Patient convenience
D-Dimer
Test result
Management

Same treated populations
D-Dimer sensitivity >98%

Patient outcomes

Management

Test pos
pathway A
Test neg
pathway B

Test pos
Add ultrasound
Test neg
pathway B*

Test result
Management

Test pos
pathway A
Test neg
pathway B

RCT to compare short-term effects from testing

RCT to compare long-term effects may not be necessary
Use of randomized trials for test evaluation

- Define the purpose of the test
- Display the existing test-treatment strategy
- Display the new test-treatment strategy
- Identify the critical comparison to assess the incremental value
- Assess whether existing evidence suffices or if RCTs are required
Today’s key messages

- Diagnostic tests are used for screening, as add-on tests, for triage, as replacement tests or for monitoring.

- Diagnostic test evaluation includes (at least) test accuracy studies, health outcomes studies and cost effectiveness studies.

- Biases related to the spectrum of patients and to the reference standard affect estimates of diagnostic test accuracy most.

- To ensure that informative RCTs, identify the critical comparisons between the old and new test-treatment.