Bias in diagnostic research and sources of variation

Karen R Steingart, MD, MPH
karenst@uw.edu
Chennai, 13 December 2010

Disclosure and acknowledgements

• I serve as co-chair of the Evidence Synthesis subgroup of Stop TB Partnership’s New Diagnostics Working Group

• Slides used by permission of Madhu Pai

• Description of QUADAS-2, used by permission of Penny Whiting
The medical literature can be compared to a jungle. It is fast growing, full of deadwood, sprinkled with hidden treasure and infested with spiders and snakes. Morgan. Can Med Assoc J, 134, Jan 15, 1986

Overview

• Discuss major forms of bias and sources of variation in diagnostic studies

• Describe assessment of methodological quality of diagnostic accuracy studies
Diagnostic studies lack methodological rigor

Diagnostic studies in four prominent general medical journals


**Lack of rigor: example from TB literature**

- 12 meta-analyses; > 500 diagnostic studies
  - 65% used prospective design
  - 33% used consecutive or random sampling
  - 72% used a cross-sectional design; 1/3 used case-control
  - Blinding reported in 34%

"Bias is any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth." *

Biases
- can arise through problems in design, execution, analysis, and interpretation
- can lead to over or underestimates of test accuracy

Any factor that influences the assessment of disease status or test results can produce bias

More definitions

• Variability arises from differences among studies, such as population demographics, disease prevalence, choice of cut-off value

• Assessment of methodological quality is the process of appraising the design and conduct of the studies included in a systematic review of diagnostic studies - addresses both bias and variation

In a perfect world, the ideal study design…

• All consecutive (or random) patients with the suspected disease enrolled
• Criteria for enrollment should be clearly stated
• Blind comparison of the index test and the reference test
• The group of patients enrolled should cover the spectrum of disease that is likely to be encountered in practice
Can you explain all of these biases reported from diagnostic studies?

centripetal
clinical review
co-intervention
comparator review
diagnostic access
diagnostic review
diagnostic safety
diagnostic suspicion
differential verification
disease progression
extrinsic interobserver variability
inappropriate reference standard
Incorporation
indeterminate results
intraobserver variability
intrinsic interobserver variability
loss to follow-up
observer variability
partial verification
patient cohort
patient filtering
popularity
population
referral
sampling
spectrum
temporal effects
test review
withdrawal
work-up bias
yet-another-bias

“Everything should be made as simple as possible but not simpler.”
### Sources of bias in diagnostic studies

- Bias due to an inappropriate/imperfect reference standard
- Spectrum bias
- Verification (work-up) bias
  - Partial verification bias
  - Differential verification bias
- Lack of blinding
- Incorporation bias
- Bias due to withdrawals, indeterminates, etc

### An ideal reference standard...

- provides error-free classification of all participants
- verifies all test results
- both study test and reference standard can be performed within a short interval to avoid changes in target disease status
Bias due to inappropriate or imperfect reference standard

• The “gold standard” is the best performing test available, but it is rarely perfect
• Imperfect reference standards are commonly used in diagnostic studies
• May lead to over or underestimation of test accuracy

Misclassification of disease status

• How accurately can you measure the following?
  – Depression
  – Tuberculosis in children
  – Latent TB infection
  – Dementia
  – Migraine
  – Attention deficit disorder
  – Cause of death
  – Irritable bowel syndrome
  – Chronic fatigue syndrome
Rarely, you get tests that are nearly perfect

The OraQuick test on oral fluid specimens had better performance with a sensitivity of 100% (95% CI 98, 100) and a specificity of 100% (95% CI 99, 100), as compared to the OraQuick test on finger stick specimens with a sensitivity of 100% (95% CI 98, 100), and a specificity of 99.7% (95% CI 98.4, 99.9).

But even ‘nearly perfect’ tests run into problems!
What if the reference standard is imperfect or missing?

Methods for diagnostic research where reference standard is imperfect or missing

1. Adjust for missing data on reference standard
2. Correct for imperfections in reference standard (based on previous research about the degree of imperfection)
3. Combine multiple pieces of information to construct a reference standard
4. Validate the index test results with other relevant clinical characteristics
Example: in the absence of a gold standard for latent TB infection…

a) use the tuberculin skin test as the gold standard
b) use both TST and IGRA
c) use active TB as surrogate for LTBI
d) use exposure gradient among contacts of active TB cases; examine if IGRA or TST correlates more closely with exposure
e) use future progression from latent infection to active disease

Interferon-gamma release assays for the diagnosis of latent tuberculosis infection in HIV-infected individuals - A systematic review and meta-analysis, Cattamanchi et al, accepted manuscript, JAIDS

• “Studies evaluating the performance of IGRAs are hampered by the lack of an adequate gold standard to distinguish the presence or absence of LTBI. …we developed a hierarchy of outcomes that could support a role for IGRAs in identifying HIV-infected individuals who could benefit from isoniazid preventive therapy….”
Spectrum bias (a form of selection bias)

Could the selection of patients have introduced bias?

• Extreme case, case-control design where study enrolls patients with definite disease and healthy controls, estimates of accuracy may be inflated

• However, the use of a case-control design does not always produce biased estimates of accuracy, for example enrolling diseased controls will reduce the potential for bias

Example: spectrum bias

<table>
<thead>
<tr>
<th>Antigen name</th>
<th>Specificity (%)</th>
<th>Patients with nontuberculous respiratory disease</th>
<th>Healthy subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recombinant 38 kDa</td>
<td>97 (90–99) (6)</td>
<td>90 (57–99) (6)</td>
<td></td>
</tr>
<tr>
<td>Recombinant malate synthase</td>
<td>97 (91–100) (4)</td>
<td>99 (81–100) (4)</td>
<td></td>
</tr>
<tr>
<td>Recombinant CFP-10</td>
<td>99 (92–100) (3)</td>
<td>90 (43–99) (3)</td>
<td></td>
</tr>
<tr>
<td>Native 38 kDa</td>
<td>96 (90–99) (6)</td>
<td>98 (92–100) (4)</td>
<td></td>
</tr>
<tr>
<td>DAT</td>
<td>55 (39–76) (4)</td>
<td>97 (88–100) (3)</td>
<td></td>
</tr>
</tbody>
</table>

*a The data represent the posterior means (95% credible intervals) (number of studies).*
Example: spectrum bias - NAAT for tuberculous meningitis

Case-control studies had a two-fold higher DOR than cross-sectional studies


Empirical evidence of sources of bias in diagnostic studies

**Figure.** Relative Diagnostic Odds Ratios and 95% Confidence Intervals (CI) of the 9 Study Characteristics Examined With a Multivariate Regression Analysis

<table>
<thead>
<tr>
<th>Study Characteristics</th>
<th>Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>8.0 (2.0-4.5)</td>
</tr>
<tr>
<td>Different Reference Tests</td>
<td></td>
</tr>
<tr>
<td>Partial Verification</td>
<td>1.0 (0.8-1.3)</td>
</tr>
<tr>
<td>Not Blinded</td>
<td>1.5 (1.0-2.0)</td>
</tr>
<tr>
<td>Nonconsecutive</td>
<td>0.9 (0.7-1.1)</td>
</tr>
<tr>
<td>Nonconsecutive Random sample</td>
<td>1.0 (0.7-1.6)</td>
</tr>
<tr>
<td>No Description Test</td>
<td>1.7 (1.1-2.5)</td>
</tr>
<tr>
<td>No Description Population</td>
<td>1.4 (1.1-1.7)</td>
</tr>
<tr>
<td>No Description Reference</td>
<td>0.7 (0.5-0.9)</td>
</tr>
</tbody>
</table>

Evidence of bias and variation in diagnostic accuracy studies. Rutjes. CMAJ.2006

**Figure.** Lower estimate of diagnostic accuracy of diagnostic accuracy vs. Higher estimate of diagnostic accuracy.
Verification bias (work up bias)

**Risk of bias if...**

- ...not all of the study group receive confirmation of diagnosis by the same reference standard
- ...if index test result influences decision to perform the reference standard or which reference standard to use
- Partial verification: reference standard is performed on test-positives, but not test-negatives
- Differential verification: reference standard used for test-positives differs from that used for test-negatives

**Example: verification bias - performance of prostate-specific antigen (PSA)**

- In the past, men were only recommended for biopsy (the gold standard for assessment of prostate cancer) if PSA > 4 ng/ml
- If the true disease state is known for only a subset of participants, and that subset is determined by the PSA result, data are subject to "verification bias"
- More recently, in one large study, 15% of men with a PSA level at or below 4.0 ng/mL had prostate cancer*

*Thompson et al. NEJM. 2004; 350(22):2239–2246
Empirical evidence of verification bias reported in 3 systematic reviews of diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Question?</th>
<th>Lijmer</th>
<th>Whiting</th>
<th>Rutjes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did investigators perform the same gold standard on all patients regardless of the study test results?</td>
<td>Different gold standard used for some patients RDOR 2.2 (95% CI 1.5, 3.3)</td>
<td>Inappropriate gold standard (some empirical support)</td>
<td>Different gold standard used for some patients RDOR 1.6 (95% CI 0.9, 2.9)</td>
</tr>
<tr>
<td>Gold standard not used for some patients RDOR 1.0 (95% CI 0.8, 1.3)</td>
<td>Gold standard not used for some patients (strong empirical support)</td>
<td>Gold standard not used for some patients RDOR 1.1 (95% CI 0.7, 1.7)</td>
<td></td>
</tr>
</tbody>
</table>

Adapted from Furukawa and Guyatt. CMAJ 2006; 174(4):481-2

Lack of blinding (also called review bias)

- Diagnostic studies may be:
  - Unblinded
  - Single blind (study test or ref. standard result is blinded)
  - Double blind (study test and ref. std results are blinded)
- Lack of blinding can lead to overestimation of test accuracy
Lack of blinding

- Blinding is more important when the interpretation of test results is subjective (e.g., pain)
- Blinding is less important when study test and gold standard are produced by an automated system with little or no ambiguity in the reading of results (e.g., CD4 count)
- Lab tests can be easily blinded by coding specimens

Example: blinding

Blinded evaluation of commercial urinary lipoarabinomannan for active tuberculosis: a pilot study

P. Daley, J. S. Michael, P. Himai, A. Lathe, P. Chodla, D. Mathai, K. R. John, M. Patel

Blinding
Urine specimens were labelled with a four-digit random number by the laboratory investigator. The technician was not aware of the identity of each specimen. A table connecting random numbers with study numbers was kept by the laboratory investigator in a locked file.

Analysis
Two hundred pulmonary and extra-pulmonary TB suspects were recruited as part of a diagnostic evaluation project, in which the sample size had been calculated to provide a test positivity of 32.6%, with a (95% CI 14.5-50.7) and a (NPV) of 78.6% (95% CI 71.0-85.2).
Empirical evidence of lack of blinding reported in 3 systematic reviews of diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Question?</th>
<th>Lijmer</th>
<th>Whiting</th>
<th>Rutjes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did investigators interpret the results of the study test and the gold standard independently and blindly from each other?</td>
<td>Nonblinded reading of results RDOR 1.3 (95% CI 1.0, 1.9)</td>
<td>Review bias (some empirical support)</td>
<td>Nonblinded reading of results RDOR 1.1 (95% CI 0.8, 1.6)</td>
</tr>
</tbody>
</table>

Adapted from Furukawa and Guyatt. CMAJ 2006; 174(4):481-2

Incorporation bias

- If the study test is included in reference standard (i.e., used to establish diagnosis)
- Example: Tuberculin skin test for TB in children. What is the most appropriate reference standard for pediatric TB?
- Empirical evidence is lacking, but incomplete reporting makes it difficult to evaluate potential sources of bias - use common sense
Evidence of bias and variation in diagnostic accuracy studies. Rutjes. CMAJ.2006

Bias due to withdrawals, indeterminates, missing data

• Example: “High sensitivity of IGRA in HIV+ TB patients”
  – Sensitivity of IGRA ~90%
    • But nearly 30% of all patients had indeterminate IGRA results!
    • These results were excluded for computation of sensitivity
Sensitivity of QuantiFERON-TB Gold In-Tube and T-SPOT.TB in HIV-infected persons with confirmed active tuberculosis (low/middle-income countries)

<table>
<thead>
<tr>
<th>Study</th>
<th>Country</th>
<th>Sensitivity (95% CI)</th>
<th>% Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>QFT-GIT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aabye 2009</td>
<td>Tanzania</td>
<td>65 (52, 76)</td>
<td>16</td>
</tr>
<tr>
<td>Kabeer 2009</td>
<td>India</td>
<td>66 (50, 80)</td>
<td>15</td>
</tr>
<tr>
<td>Leidl 2009</td>
<td>Uganda</td>
<td>74 (49, 91)</td>
<td>12</td>
</tr>
<tr>
<td>Markova 2009</td>
<td>Bulgaria</td>
<td>92 (64, 100)</td>
<td>14</td>
</tr>
<tr>
<td>Raby 2008</td>
<td>Zambia</td>
<td>63 (49, 75)</td>
<td>16</td>
</tr>
<tr>
<td>Tsiouris 2006</td>
<td>South Africa</td>
<td>65 (44, 83)</td>
<td>13</td>
</tr>
<tr>
<td>Veldman 2009</td>
<td>South Africa</td>
<td>30 (15, 49)</td>
<td>14</td>
</tr>
<tr>
<td>Subtotal</td>
<td></td>
<td>65 (52, 77)</td>
<td>100</td>
</tr>
<tr>
<td>TSPOT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cattamanchi 2010</td>
<td>Uganda</td>
<td>54 (45, 64)</td>
<td>25</td>
</tr>
<tr>
<td>Jiang 2009</td>
<td>China</td>
<td>66 (47, 81)</td>
<td>19</td>
</tr>
<tr>
<td>Leidl 2009</td>
<td>Uganda</td>
<td>89 (67, 99)</td>
<td>20</td>
</tr>
<tr>
<td>Markova 2009</td>
<td>Bulgaria</td>
<td>62 (32, 86)</td>
<td>12</td>
</tr>
<tr>
<td>Oni 2010</td>
<td>South Africa</td>
<td>68 (57, 78)</td>
<td>25</td>
</tr>
<tr>
<td>Subtotal</td>
<td></td>
<td>68 (56, 80)</td>
<td>100</td>
</tr>
</tbody>
</table>

Metcalfe et al.- Methods

- We used the following definitions for primary outcomes

- (1) Sensitivity - the proportion of individuals with a positive IGRA result among those with culture-positive TB (we included indeterminate IGRA results in the denominator if they occurred in individuals with culture positive TB)
Assessment of methodological quality of diagnostic accuracy studies
Users’ guide for a diagnostic study

Users’ Guide for an Article About Interpreting Diagnostic Test Results

Are the results valid?
- Did participating patients present a diagnostic dilemma?
- Did investigators compare the test to an appropriate, independent reference standard?
- Were those interpreting the test and reference standard blind to the other results?
- Did investigators perform the same reference standard to all patients regardless of the results of the test under investigation?

What are the results?
- What likelihood ratios were associated with the range of possible test results?

How can I apply the results to patient care?
- Will the reproducibility of the test result and its interpretation be satisfactory in my clinical setting?
- Are the study results applicable to the patients in my practice?
- Will the test results change my management strategy?
- Will patients be better off as a result of the test?

QUality Assessment of Diagnostic Accuracy Studies (QUADAS)

- Systematically developed based on empirical evidence and a formal consensus method (modified Delphi)

- Recommended tool by Cochrane Collaboration

QUADAS-2, currently being piloted

- Four core domains: Patient selection; Index test; Reference standard; Flow and timing
  - Assessed for Risk of Bias (ROB) and Applicability
  - ‘Signalling’ questions which are scored as ‘Yes’, ‘No’, ‘Unclear’
  - ROB and Applicability are scored as ‘Low’, ‘High’, ‘Unclear’

QUADAS - 2

- Define the question:

<table>
<thead>
<tr>
<th>Patients:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index test:</td>
</tr>
<tr>
<td>Comparator test (if applicable):</td>
</tr>
<tr>
<td>Target condition:</td>
</tr>
<tr>
<td>Reference Standard:</td>
</tr>
</tbody>
</table>

- Two reviewers working independently
- Transparent process
- Goal is to achieve consensus
Domain 1: Patient Selection

**Risk of bias: Could the selection of patients have introduced bias?**

- **Signalling Question 1:** Were eligibility criteria defined?
- **Signalling Question 2:** Was an unselected sample of patients enrolled?

Whiting P, QUADAS2, DRAFT

---

Domain 2: Index Test - DRAFT

**Risk of bias: Could methods used to interpret or conduct the index test have introduced bias?**

- **Signalling Question 1:** Were the index test results interpreted without knowledge of the results of the reference standard?
- **Signalling Question 2:** Did the study pre-specify the threshold?
  - Selecting the threshold to maximise the sensitivity and/or specificity of the test may lead to overoptimistic measures of test performance

Whiting P, QUADAS2, DRAFT
Domain 3: Reference Standard

Risk of bias: Could methods used to conduct or interpret reference standard have introduced bias?

• Signalling Question 1: Is the reference standard likely to correctly classify the target condition?

• Signalling Question 2: Were the reference standard results interpreted without knowledge of the results of the index test?

Whiting P, QUADAS2, DRAFT

Domain 4: Flow and timing

Risk of bias: Could the patient flow have introduced bias?

• Signalling Question 1: Was there a short interval between the index test and reference standard?

• Signalling Question 2: Did all patients receive a reference standard?

• Signalling Question 3: Were all patients included in the analysis?

Whiting P, QUADAS2, DRAFT
Applicability

• **Patient selection:** Do the included patients and setting match the review question?

• **Index test:** Does the test technology, execution and interpretation match the question?

• **Reference Standard:** Does the target condition as defined by the reference standard match the question?

Whiting P, QUADAS2, DRAFT

Methodological quality summary: review authors’ judgments about each methodological quality item for each included study, created in RevMan http://ims.cochrane.org/revman
Quality and Reporting of Diagnostic Accuracy Studies in TB, HIV and Malaria: Evaluation Using QUADAS and STARD Standards

Patricila Scurlfontela1, Nithika Pant Pai1, Ian Schiller2, Nandini Dendukuri2, Andrew Ramsay3, Madhukar Pai1

1Department of Microbiology, Ecology and Immunology, McGill University, Montreal, Canada. 2Department of Medical Malariology, McGill University, Montreal, Canada. 3Biostatistical Program for Research and Training in Tropical Diseases, World Health Organization, Geneva, Switzerland. 4Infectious Disease Epidemiology and Clinical Research Unit, Montreal Children’s Hospital, Montreal, Canada.

Abstract

Background: Poor methodological quality and reporting are known concerns with diagnostic accuracy studies. In 2009, the QUADAS tool and the STARD standards were published for evaluating the quality and improving the reporting of diagnostic studies, respectively. However, it is unclear whether these tools have been applied to diagnostic studies of infectious diseases. We performed a systematic review on the methodological and reporting quality of diagnostic studies in TB, malaria and HIV.

Methods: We identified diagnostic accuracy studies of commercial tests for TB, malaria and HIV through a systematic search of the literature using PubMed and EMBASE (2004–2008). Original studies that reported sensitivity and specificity data were included. Two reviewers independently extracted data on study characteristics and diagnostic accuracy, and used QUADAS and STARD to evaluate the quality of methods and reporting, respectively.

Results: Ninety (35%) of 258 articles met inclusion criteria. All studies had design deficiencies. Study quality indicators that were met in less than 25% of the studies included adequate description of withdrawals (8%), reference test execution (10%), absence of index test review bias (10%), and description of withdrawals (10%).

Conclusions: Recently published diagnostic accuracy studies on commercial tests for TB, malaria and HIV have moderate to low quality and are poorly reported. The more frequent use of tools such as QUADAS and STARD may be necessary to improve the methodological and reporting quality of future diagnostic accuracy studies in infectious diseases.

Quality of TB accuracy studies using QUADAS

<table>
<thead>
<tr>
<th>Quality item</th>
<th>45 studies n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequate spectrum composition</td>
<td>26 (58)</td>
</tr>
<tr>
<td>Adequate reference standard</td>
<td>44 (98)</td>
</tr>
<tr>
<td>Absence of disease progression bias</td>
<td>42 (93)</td>
</tr>
<tr>
<td>Absence of partial verification bias</td>
<td>44 (98)</td>
</tr>
<tr>
<td>Absence of differential verification bias</td>
<td>42 (93)</td>
</tr>
<tr>
<td>Absence of incorporation bias</td>
<td>45 (100)</td>
</tr>
<tr>
<td>Absence of index test review bias</td>
<td>6 (13)</td>
</tr>
<tr>
<td>Absence of reference test review bias</td>
<td>7 (16)</td>
</tr>
<tr>
<td>Absence of clinical review bias</td>
<td>14 (31)</td>
</tr>
<tr>
<td>Report of uninterpretable results</td>
<td>9 (20)</td>
</tr>
<tr>
<td>Description of withdrawals</td>
<td>3 (7)</td>
</tr>
</tbody>
</table>
Initiative to improve reporting of diagnostic accuracy studies

STARD Statement
STandards for the Reporting of Diagnostic accuracy studies

Objective of the STARD Initiative

The objectives of the STARD initiative is to improve the accuracy and comprehensiveness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential biases in the study (differential validity) and to evaluate its generalizability (external validity).

The STARD statement consists of a checklist of items and recommends the use of a flow diagram which describes the design of the study and the flow of patients.

News

April 2009
- More than 280 biomedical journals encourage the use of the STARD statement in their instructions for authors.

Last update 20 April 2009

http://www.stard-statement.org/

Thank you!

Questions?