Systematic review and meta-analysis of diagnostic test accuracy studies

Karen R Steingart, MD, MPH
Montreal, July 2011
karenst@uw.edu
Disclosure

Coordinator, Evidence Synthesis and Policy Subgroup, NDWG, Stop TB Partnership

Editor, The Cochrane Infectious Diseases Group
Overview

• Describe key steps in a systematic review of diagnostic test accuracy studies

• Describe standard methods of meta-analysis of data from diagnostic studies

• Describe key tools used in carrying out a systematic review/meta-analysis of diagnostic test accuracy studies
  - QUADAS for quality assessment
  - STARD Statement for reporting
  - Computer software options for analysis and presentation
Definitions

• **Systematic review** is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyse data from the studies that are included in the review. A systematic review of diagnostic test accuracy studies concerns measures of test performance such as **sensitivity** and **specificity**.

• **Meta-analysis** is the use of statistical techniques in a systematic review to summarize (pool) the results of included studies. Not all **systematic reviews** include a meta-analysis.
Free Summary

Finding What Works in Health Care: Standards for Systematic Reviews

Jill Eden, Laura Levit, Alfred Berg, and Sally Morton, Editors; Committee on Standards for Systematic Reviews of Comparative Effectiveness Research; Institute of Medicine

Typical systematic reviewer

Institute of Medicine
• 21 Standards for Systematic Reviews
• 82 Elements of Performance
A systematic review/meta-analysis of data from diagnostic studies...

...appraises the quality of primary studies
...summarizes results from primary studies
...calculates an overall summary (if a meta-analysis is performed)
...looks for and investigates possible reasons for inconsistency in results (heterogeneity)
...evaluates the impact of quality and other study characteristics on diagnostic accuracy
...stimulates new research questions
“Road Map” for systematic reviews of diagnostic test evaluations

Pai M. Am Coll Phys JC 2004
Key steps in a systematic review of diagnostic test accuracy

1. Definition of the objectives of the review
2. Study selection
3. Assessment of study quality
4. Data extraction, analysis, and presentation
5. Interpretation of results

1. Definition of the objectives of the review

- Participants
- Index tests
- Comparator tests

+ Purpose of the test/strategy
  + Study design
  + Reference standard

Richardson et al. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 1995;A-12
What is the purpose of the test?

• Triage
  – To minimize use of invasive or expensive test

• Add-on
  – To improve diagnosis beyond what is already done

• Replacement
  – to eliminate tests with poorer performance compared with a current test, greater invasiveness, or increased cost
Objectives - 1

To determine the **diagnostic accuracy** of commercial serological tests for active pulmonary TB in adults and children

**Participants**: Adults and children suspected of having active pulmonary TB

**Index test**: Commercial serological test

**Comparator test**: No test or sputum microscopy
Objectives - 2

• **Purpose of the test:** We were interested in evaluating the use of a serological assay as a replacement test for, or an additional test after, smear microscopy.

• **Types of studies:** Randomized controlled trials and any observational design, including cross-sectional, case-control and cohort designs

• **Reference standard:** Culture
2. Study identification and selection

- Cochrane register of Diagnostic Test Accuracy Studies (under development)
- Search related diagnostic test accuracy reviews, for example DARE [http://www.york.ac.uk/inst/crd/](http://www.york.ac.uk/inst/crd/)
- Check references of relevant studies/reviews
- Use highly sensitive (broad) search strategy
- Reflect key concepts of the review (focus on index test and target condition) in search
- Use a wide variety of search terms, both text words and database subject headings (MeSH terms)

**TIPS:** Avoid language and human limits; routine use of search filters should generally be avoided!

Bossuyt PM, Leeflang MM. DTA Handbook Chapter 6: Developing Criteria for Including Studies.
Example: partial search strategy for Medline, ‘antibody or antigen based detection tests for the diagnosis of tuberculosis’

#4 Search antibodies, bacterial[mh] OR antibodies/blood OR antibodies/immunology OR antibody response OR "humoral immune" OR "humoral immunity" OR "humoral antibody" OR "immune based" OR "antibody detection" OR antigens, bacterial[mh] OR antigens/analysis OR antigens/blood OR antigens/cerebrospinal fluid OR antigens/immunology OR antigens/urine OR lipopolysaccharide* OR lipoarabinomannan OR "antigen detection" OR antigen[tiab] OR antigens[tiab]


#1 Search tuberculosis[mh] OR mycobacterium tuberculosis[mh]
• **Standard 3.1 Conduct a comprehensive systematic search for evidence**

• Required elements:
  
• **3.1.1** Work with a librarian or other information specialist trained in performing systematic reviews to plan the search strategy

• **3.1.2** Design the search strategy to address each key research question

• **3.1.3** Use an independent librarian or other information specialist to peer review the search strategy.....
Records identified through database searching (n = )

Records after duplicates removed (n = )

Records screened (n = )

Records excluded (n = )

Full-text articles assessed for eligibility (n = )

Studies included in qualitative synthesis (n = )

Studies included in quantitative synthesis (meta-analysis) (n = )

Full-text articles excluded, with reasons (n = )

www.prisma-statement.org
3. Assessment of study quality

Evidence of bias and variation in diagnostic accuracy studies

Anne W.S. Rutjes, Johannes B. Reitsma, Marcello Di Nisio, Nynke Smidt, Jeroen C. van Rijn, Patrick M.M. Bossuyt

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

Context: The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would invalidate the clinical application of most study results.

Objective: To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy.

Design and Setting: Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through a systematic search of the literature using MEDLINE, EMBASE, and DARE databases and the Cochrane Library (1996-1997). Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model.

Sources of Variation and Bias in Studies of Diagnostic Accuracy

A Systematic Review

Penny Whiting, MSc; Anne W.S. Rutjes, MSc; Johannes B. Reitsma, MD, PhD; Alfia S. Glas, MD, PhD; Patrick M.M. Bossuyt, PhD; and Jos Kleijnen, MD, PhD

Background: Studies of diagnostic accuracy are subject to different sources of bias and variation than studies that evaluate the effectiveness of an intervention. Little is known about the effects of these sources of bias and variation.

Purpose: To summarise the evidence on factors that can lead to data distortion.
Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests

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

<table>
<thead>
<tr>
<th>Study Characteristics</th>
<th>Relative Diagnostic Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-Control</td>
<td>3.0 (2.0-4.5)</td>
</tr>
<tr>
<td>Different Reference Tests</td>
<td>2.2 (1.5-3.3)</td>
</tr>
<tr>
<td>Partial Verification</td>
<td>1.0 (0.8-1.3)</td>
</tr>
<tr>
<td>Not Blinded</td>
<td>1.3 (1.0-1.9)</td>
</tr>
<tr>
<td>Nonconsecutive</td>
<td>0.9 (0.7-1.1)</td>
</tr>
<tr>
<td>Retrospective</td>
<td>1.0 (0.7-1.4)</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 (0.6-0.9)</td>
</tr>
</tbody>
</table>

Lijmer JG et al. JAMA.1999
Evidence of bias and variation in diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Study characteristics*</th>
<th>Lower estimate of diagnostic accuracy</th>
<th>Higher estimate of diagnostic accuracy</th>
<th>RDOR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe cases and healthy controls</td>
<td></td>
<td></td>
<td>4.9 (0.6-37.3)</td>
</tr>
<tr>
<td>Other case-control designs</td>
<td></td>
<td></td>
<td>1.1 (0.4-3.4)</td>
</tr>
<tr>
<td>Selection: referral for index test</td>
<td></td>
<td></td>
<td>0.5 (0.3-0.9)</td>
</tr>
<tr>
<td>Selection: other test results</td>
<td></td>
<td></td>
<td>0.9 (0.6-1.3)</td>
</tr>
<tr>
<td>Limited challenge</td>
<td></td>
<td></td>
<td>0.9 (0.6-1.3)</td>
</tr>
<tr>
<td>Increased challenge</td>
<td></td>
<td></td>
<td>1.0 (0.6-1.7)</td>
</tr>
<tr>
<td>Nonconsecutive sample</td>
<td></td>
<td></td>
<td>1.5 (1.0-2.1)</td>
</tr>
<tr>
<td>Random sample</td>
<td></td>
<td></td>
<td>1.7 (0.9-3.2)</td>
</tr>
<tr>
<td>Sampling not reported</td>
<td></td>
<td></td>
<td>0.9 (0.6-1.3)</td>
</tr>
<tr>
<td>Differential verification</td>
<td></td>
<td></td>
<td>1.6 (0.9-2.9)</td>
</tr>
<tr>
<td>Partial verification</td>
<td></td>
<td></td>
<td>1.1 (0.7-1.7)</td>
</tr>
<tr>
<td>Composite reference standard</td>
<td></td>
<td></td>
<td>0.9 (0.5-1.8)</td>
</tr>
<tr>
<td>Incorporation</td>
<td></td>
<td></td>
<td>1.4 (0.7-2.8)</td>
</tr>
</tbody>
</table>

Rutjes AWS et al. CMAJ. 2006
Case-control studies had a two-fold higher DOR than cross-sectional studies.
QUADAS (Quality Assessment of Diagnostic Accuracy Studies)

BMC Medical Research Methodology

Research article

The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews

Penny Whiting*1, Anne WS Rutjes2, Johannes B Reitsma2, Patrick MM Bossuyt2 and Jos Kleijnen1

Address: 1Centre for Reviews and Dissemination, University of York, England, UK and 2Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands

Email: Penny Whiting* - pfw2@york.ac.uk; Anne WS Rutjes - a.rutjes@amc.uva.nl; Johannes B Reitsma - j.reitsma@amc.uva.nl; Patrick MM Bossuyt - p.m.bossuyt@amc.uva.nl; Jos Kleijnen - jk13@york.ac.uk

* Corresponding author

Published: 10 November 2003
Received: 14 July 2003
Accepted: 10 November 2003

BMC Medical Research Methodology 2003, 3:25

This article is available from: http://www.biomedcentral.com/1471-2288/3/25

© 2003 Whiting et al; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.
QUADAS

- Created in 2003
- Recommended by The Cochrane Collaboration
- Covers risk of bias, sources of variation, reporting quality
- Checklist with 14 items (11 core items)
- Items rated as ‘yes’, ‘no’, or ‘unclear’
- ‘Yes’ indicates absence of bias

Dr. Atul Gawande, the author of "The Checklist Manifesto"
<table>
<thead>
<tr>
<th>Item</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Was the spectrum of patients representative of the patients who will receive the test in practice?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>2. Is the reference standard likely to correctly classify the target condition?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>3. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>4. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>5. Did patients receive the same reference standard regardless of the index test result?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>6. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>7. Were the index test results interpreted without knowledge of the results of the reference standard?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>8. Were the reference standard results interpreted without knowledge of the results of the index test?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>9. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>10. Were uninterpretable/intermediate test results reported?</td>
<td>Yes No Unclear</td>
</tr>
<tr>
<td>11. Were withdrawals from the study explained?</td>
<td>Yes No Unclear</td>
</tr>
</tbody>
</table>
QUADAS-2 (under development)

- Domain list (not a checklist)
- Domains are
  - patient selection
  - index test
  - reference standard
  - flow and timing
- First 3 domains assessed for applicability
- *Signalling questions* are used for judgments of risk of bias
- **Key difference from QUADAS-1:** includes both risk of bias and applicability
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?

Applicability: Do the included patients and setting match the review question?
• Created in 2003
• Objective is to improve the accuracy and completeness of reporting of diagnostic accuracy studies
• 25 item checklist
• > 200 journals encourage the use of STARD Statement in their instructions for authors
<table>
<thead>
<tr>
<th>Item #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identify the article as a study of diagnostic accuracy (recommended MeSH heading ‘sensitivity and specificity’).</td>
</tr>
<tr>
<td>2</td>
<td>State the research questions or study aims, such as estimating the diagnostic accuracy or comparing accuracy between tests or across participant groups. Describe:</td>
</tr>
<tr>
<td>3</td>
<td>The study population: the inclusion and exclusion criteria, the setting and the locations where the data were collected.</td>
</tr>
<tr>
<td>4</td>
<td>Participant recruitment: was the recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?</td>
</tr>
<tr>
<td>5</td>
<td>Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.</td>
</tr>
</tbody>
</table>
Clinical Diagnostic Utility of IP-10 and LAM Antigen Levels for the Diagnosis of Tuberculous Pleural Effusions in a High Burden Setting

Keertan Dheda¹,²,³*, Richard N. Van-Zyl Smit¹, Leonardo A. Sechi⁴, Motasim Badri¹, Richard Meldau¹, Gregory Symons¹, Hoosein Khalfy¹, Igshaan Carr¹, Alice Maredza¹, Rodney Dawson¹, Helen Wainright¹, Andrew Whitelaw⁵,⁶, Eric D. Bateman¹, Alimuddin Zumla³

Statistical analysis

Categorical variables were compared using the $\chi^2$ test or Fisher exact test and continuous variables were compared using t-student test, whenever appropriate. Non-parametric tests (Mann-Whitney) were used for non-normally distributed variables. Concordance between tests was measured using the kappa co-efficient. Diagnostic accuracy, including 95% confidence intervals, was assessed using sensitivity, specificity, predictive values and area under the ROC in the TB and non-TB sub-groups. The study report was prepared using the Standards for Reporting of Diagnostic Accuracy (STARD initiative) format (19).
## QUADAS Item Scored ‘Yes’

<table>
<thead>
<tr>
<th>QUADAS Item Scored ‘Yes’</th>
<th>Tuberculosis (N=45) 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 blinding of index test result</td>
<td>6 (13)</td>
</tr>
<tr>
<td>Absence of blinding of reference test result</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>
4. Data extraction, analysis, and presentation

- Extract paired estimates of sensitivity and specificity
- Visually examine results of individual studies
- Calculate overall summary estimates using HSROC/bivariate meta-analysis
- Look for and investigate possible reasons for heterogeneity
### Measures of test performance

<table>
<thead>
<tr>
<th></th>
<th>Disease present</th>
<th>Disease absent</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Index test +</strong></td>
<td>True positives</td>
<td>False positives</td>
<td>TP + FP</td>
</tr>
<tr>
<td><strong>Index test -</strong></td>
<td>False negative</td>
<td>True negatives</td>
<td>FN + TN</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>TP + FN</td>
<td>TN + FP</td>
<td>TP + FP + FN + TN</td>
</tr>
</tbody>
</table>

**Sensitivity** = $\frac{TP}{TP+FN}$

**Specificity** = $\frac{TN}{FP + TN}$

**Positive predictive value** = $\frac{TP}{TP + FP}$

**Negative predictive value** = $\frac{TN}{FN + TN}$

**Likelihood ratio positive** = $\frac{Sensitivity}{1 – Specificity}$

**Likelihood ratio negative** = $(1 – Sensitivity)/Specificity$

**Prevalence** (proportion of people with disease in population to whom the test has been applied) = $\frac{TP + FN}{TP + FP + FN + TN}$
Forest plots of sensitivity and specificity, anda-TB IgG for the diagnosis of pulmonary TB, smear-positive patients

<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alifano 1994</td>
<td>35</td>
<td>2</td>
<td>7</td>
<td>92</td>
<td>0.83 [0.69, 0.93]</td>
<td>0.98 [0.93, 1.00]</td>
</tr>
<tr>
<td>Alifano 1996 (a)</td>
<td>28</td>
<td>3</td>
<td>5</td>
<td>41</td>
<td>0.85 [0.68, 0.95]</td>
<td>0.93 [0.81, 0.99]</td>
</tr>
<tr>
<td>Kalantri 2005 (a)</td>
<td>84</td>
<td>0</td>
<td>21</td>
<td>40</td>
<td>0.80 [0.71, 0.87]</td>
<td>1.00 [0.91, 1.00]</td>
</tr>
<tr>
<td>Okuda 2004 (a)</td>
<td>28</td>
<td>10</td>
<td>6</td>
<td>101</td>
<td>0.82 [0.65, 0.93]</td>
<td>0.91 [0.84, 0.96]</td>
</tr>
<tr>
<td>Traunmuller 2005</td>
<td>32</td>
<td>21</td>
<td>6</td>
<td>58</td>
<td>0.84 [0.69, 0.94]</td>
<td>0.73 [0.62, 0.83]</td>
</tr>
<tr>
<td>Wu 2004 (a)</td>
<td>58</td>
<td>4</td>
<td>34</td>
<td>30</td>
<td>0.63 [0.52, 0.73]</td>
<td>0.88 [0.73, 0.97]</td>
</tr>
<tr>
<td>Wu 2005</td>
<td>35</td>
<td>19</td>
<td>30</td>
<td>40</td>
<td>0.54 [0.41, 0.66]</td>
<td>0.68 [0.54, 0.79]</td>
</tr>
</tbody>
</table>

Extracted data are presented as TP, FP, FN, TN
Data shown in the graph are also displayed numerically
Each study result is given a box for a point estimate
Horizontal line is the confidence interval (CI); measures how the result of the study varies with chance
  - The wider the CI, the less confident we are in the result
Judge whether results are consistent depending if CIs overlap

<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alifano 1994</td>
<td>35</td>
<td>2</td>
<td>7</td>
<td>92</td>
<td>0.83 [0.69, 0.93]</td>
<td>0.98 [0.93, 1.00]</td>
</tr>
<tr>
<td>Alifano 1996 (a)</td>
<td>28</td>
<td>3</td>
<td>5</td>
<td>41</td>
<td>0.85 [0.68, 0.95]</td>
<td>0.93 [0.81, 0.99]</td>
</tr>
<tr>
<td>Kalantri 2005 (a)</td>
<td>84</td>
<td>0</td>
<td>21</td>
<td>40</td>
<td>0.80 [0.71, 0.87]</td>
<td>1.00 [0.91, 1.00]</td>
</tr>
</tbody>
</table>
http://ims.cochrane.org/revman
Select diagnostic accuracy review
Enter review title
Ready to enter studies
Enter studies
Assessment of methodological quality
Enter data
Forest plots of sensitivity and specificity, anda-TB IgG for the diagnosis of pulmonary TB, smear-positive patients

<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alifano 1994</td>
<td>35</td>
<td>2</td>
<td>7</td>
<td>92</td>
<td>0.83 [0.69, 0.93]</td>
<td>0.98 [0.93, 1.00]</td>
</tr>
<tr>
<td>Alifano 1996 (a)</td>
<td>28</td>
<td>3</td>
<td>5</td>
<td>41</td>
<td>0.85 [0.68, 0.95]</td>
<td>0.93 [0.81, 0.99]</td>
</tr>
<tr>
<td>Kalantri 2005 (a)</td>
<td>84</td>
<td>0</td>
<td>21</td>
<td>40</td>
<td>0.80 [0.71, 0.87]</td>
<td>1.00 [0.91, 1.00]</td>
</tr>
<tr>
<td>Okuda 2004 (a)</td>
<td>28</td>
<td>10</td>
<td>6</td>
<td>101</td>
<td>0.82 [0.65, 0.93]</td>
<td>0.91 [0.84, 0.96]</td>
</tr>
<tr>
<td>Traumuller 2005</td>
<td>32</td>
<td>21</td>
<td>6</td>
<td>58</td>
<td>0.84 [0.69, 0.94]</td>
<td>0.73 [0.62, 0.83]</td>
</tr>
<tr>
<td>Wu 2004 (a)</td>
<td>58</td>
<td>4</td>
<td>34</td>
<td>30</td>
<td>0.63 [0.52, 0.73]</td>
<td>0.88 [0.73, 0.97]</td>
</tr>
<tr>
<td>Wu 2005</td>
<td>35</td>
<td>19</td>
<td>30</td>
<td>40</td>
<td>0.54 [0.41, 0.66]</td>
<td>0.68 [0.54, 0.79]</td>
</tr>
</tbody>
</table>

Software

**Meta-DiSc: a software for meta-analysis of test accuracy data**

Javier Zamora*¹, Victor Abaira¹, Alfonso Muriel¹, Khalid Khan² and Arri Coomarasamy²

Address: ¹Clinical Biostatistics Unit, Ramón y Cajal Hospital, Madrid, Ctra. Colmenar km 9.100 Madrid 28034, Spain and ²University of Birmingham and Birmingham Women's Hospital, Edgbaston, Birmingham, UK

Email: Javier Zamora* - javier.zamora@hrc.es; Victor Abaira - Victor.abaira@hrc.es; Alfonso Muriel - Alfonso.muriel@hrc.es; Khalid Khan - k.s.khan@bham.ac.uk; Arri Coomarasamy - arrcoomar@blueyonder.co.uk

* Corresponding author

Published: 12 July 2006


Received: 31 March 2006

Accepted: 12 July 2006

This article is available from: http://www.biomedcentral.com/1471-2288/6/31

© 2006 Zamora et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Enter data from excel

<table>
<thead>
<tr>
<th>No.</th>
<th>Author</th>
<th>StudyId</th>
<th>Smear</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alifano 1994</td>
<td></td>
<td></td>
<td>35</td>
<td>2</td>
<td>7</td>
<td>92</td>
</tr>
<tr>
<td>2</td>
<td>Alifano 1996</td>
<td>a</td>
<td></td>
<td>28</td>
<td>3</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>3</td>
<td>Kalantri 2005</td>
<td>a</td>
<td></td>
<td>84</td>
<td>0</td>
<td>21</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>Okuda 2004</td>
<td>a</td>
<td></td>
<td>28</td>
<td>10</td>
<td>6</td>
<td>101</td>
</tr>
<tr>
<td>5</td>
<td>Traunmuller 2005</td>
<td></td>
<td></td>
<td>32</td>
<td>21</td>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>6</td>
<td>Wu 2004</td>
<td>a</td>
<td></td>
<td>58</td>
<td>4</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>Wu 2005</td>
<td></td>
<td></td>
<td>35</td>
<td>19</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Select plot and characteristics
Export plot

Sensitivity (95% CI)

- Alifano 1994: 0.83 (0.69 - 0.93)
- Alifano 1996: 0.85 (0.68 - 0.95)
- Kalantri 2005: 0.80 (0.71 - 0.87)
- Okuda 2004: 0.82 (0.65 - 0.93)
- Traunmuller 2005: 0.84 (0.69 - 0.94)
- Wu 2004: 0.63 (0.52 - 0.73)
- Wu 2005: 0.54 (0.41 - 0.66)

Pooled Sensitivity = 0.73 (0.69 to 0.78)
Chi-square = 27.47; df = 6 (p = 0.0001)
Inconsistency (I-square) = 78.2%
Forest plots of sensitivity and specificity, anda-TB IgG for the diagnosis of pulmonary TB, smear-positive patients

<table>
<thead>
<tr>
<th>Study</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alifano 1994</td>
<td>35</td>
<td>2</td>
<td>7</td>
<td>92</td>
<td>0.83 [0.69, 0.93]</td>
<td>0.98 [0.93, 1.00]</td>
</tr>
<tr>
<td>Alifano 1996 (a)</td>
<td>28</td>
<td>3</td>
<td>5</td>
<td>41</td>
<td>0.85 [0.68, 0.95]</td>
<td>0.93 [0.81, 0.99]</td>
</tr>
<tr>
<td>Kalantri 2005 (a)</td>
<td>84</td>
<td>0</td>
<td>21</td>
<td>40</td>
<td>0.80 [0.71, 0.87]</td>
<td>1.00 [0.91, 1.00]</td>
</tr>
<tr>
<td>Okuda 2004 (a)</td>
<td>28</td>
<td>10</td>
<td>6</td>
<td>101</td>
<td>0.82 [0.65, 0.93]</td>
<td>0.91 [0.84, 0.96]</td>
</tr>
<tr>
<td>Traunmuller 2005</td>
<td>32</td>
<td>21</td>
<td>6</td>
<td>58</td>
<td>0.84 [0.69, 0.94]</td>
<td>0.73 [0.62, 0.83]</td>
</tr>
<tr>
<td>Wu 2004 (a)</td>
<td>58</td>
<td>4</td>
<td>34</td>
<td>30</td>
<td>0.63 [0.52, 0.73]</td>
<td>0.88 [0.73, 0.97]</td>
</tr>
<tr>
<td>Wu 2005</td>
<td>35</td>
<td>19</td>
<td>30</td>
<td>40</td>
<td>0.54 [0.41, 0.66]</td>
<td>0.68 [0.54, 0.79]</td>
</tr>
</tbody>
</table>

Meta-analysis of diagnostic test accuracy studies

- Provides summaries of the results of included studies
  - estimate of the average diagnostic accuracy of a test
  - the uncertainty of this average
  - the variability of study findings around the estimates
"Se ha dicho que un tipo con una pierna sumergida en hielo y la otra en agua hirviendo está cómodo - en promedio.” JM Yancey

It has been said that a fellow with one leg frozen in ice and the other leg in boiling water is comfortable - on average.
Challenges with meta-analysis of diagnostic studies

- Meta-analysis methods for diagnostic test accuracy have to deal with two summary statistics (for example, sensitivity and specificity) simultaneously rather than one.
- Meta-analysis methods allow studies to be combined that have used tests at different thresholds.
- Considerable heterogeneity in results of test accuracy studies is to be expected.
- Random effects models are required to describe the variability in test accuracy across studies.
Calculating an overall summary

A unification of models for meta-analysis of diagnostic accuracy studies

ROGER M. HARBORD
Medical Research Council Health Services Research Collaboration,
Department of Social Medicine, University of Bristol, Canynge Hall,
Whiteladies Road, Bristol BS8 2PR, UK
roger.harbord@bristol.ac.uk

JONATHAN J. DEEKS
Centre for Statistics in Medicine, Oxford, UK

MATTHIAS EGGER
Department of Social and Preventive Medicine, University of Bern, Switzerland

PENNY WHITING, JONATHAN A. C. STERNE
Medical Research Council Health Services Research Collaboration,
Department of Social Medicine, University of Bristol, UK

A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations

Carolyn M. Rutter1,*,1 and Constantine A. Gatsonis2
1Group Health Cooperative, Center for Health Studies, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, U.S.A.
2Center for Statistical Sciences, Brown University, Box G-H, Providence, R1 02912, U.S.A.

Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews

Johannes B. Reitsma*, A., Afina S. Ghas*, Anne W.S. Rutjes*, Rob J.P.M. Scholten*, Patrick M. Bossuyt*, Aeilko H. Zwinderman†
*Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1180 DD Amsterdam, The Netherlands
†Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, The Netherlands
Accepted 21 February 2003
Hierarchical models are recommended

- Account for the patterns of correlation between sensitivity and specificity across studies caused by the relationship between sensitivity and specificity within each study.
- Random effects methods are recommended when data are heterogeneous (this is the rule with diagnostic studies).
- Separate pooling of sensitivity and specificity not recommended.
Metandi in Stata

The Stata Journal (2009)
9, Number 2, pp. 211–229

metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression

Roger M. Harbord
Department of Social Medicine
University of Bristol
Bristol, UK
roger.harbord@bristol.ac.uk

Penny Whiting
Department of Social Medicine
University of Bristol
Bristol, UK

Abstract. Meta-analysis of diagnostic test accuracy presents many challenges. Even in the simplest case, when the data are summarized by a $2 \times 2$ table from each study, a statistically rigorous analysis requires hierarchical (multilevel) models that respect the binomial data structure, such as hierarchical logistic regression. We present a Stata package, metandi, to facilitate the fitting of such models in Stata. The commands display the results in two alternative parameterizations and produce a customizable plot. metandi requires either Stata 10 or above (which has the new command xtmelogit), or Stata 8.2 or above with gllamm installed.

Keywords: st0163, metandi, metandiplot, diagnosis, meta-analysis, sensitivity and specificity, hierarchical models, generalized mixed models, gllamm, xtmelogit, receiver operating characteristic (ROC), summary ROC, hierarchical summary ROC
Copy and paste data from Excel into Stata

<table>
<thead>
<tr>
<th>author</th>
<th>newid</th>
<th>year</th>
<th>smear_status</th>
<th>tp</th>
<th>fp</th>
<th>fn</th>
<th>tn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ali-Tave</td>
<td>a</td>
<td>1994</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>92</td>
</tr>
<tr>
<td>Ali-Tave</td>
<td>a</td>
<td>1996</td>
<td>1</td>
<td>28</td>
<td>3</td>
<td>5</td>
<td>41</td>
</tr>
<tr>
<td>Kalantari</td>
<td>a</td>
<td>2005</td>
<td>1</td>
<td>64</td>
<td>6</td>
<td>21</td>
<td>49</td>
</tr>
<tr>
<td>Okuda</td>
<td>a</td>
<td>2004</td>
<td>1</td>
<td>28</td>
<td>10</td>
<td>6</td>
<td>161</td>
</tr>
<tr>
<td>Troumoller</td>
<td>2005</td>
<td></td>
<td>1</td>
<td>32</td>
<td>21</td>
<td>6</td>
<td>58</td>
</tr>
<tr>
<td>Wu</td>
<td>a</td>
<td>2004</td>
<td>1</td>
<td>58</td>
<td>4</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>Wu</td>
<td>a</td>
<td>2005</td>
<td>1</td>
<td>15</td>
<td>10</td>
<td>30</td>
<td>40</td>
</tr>
</tbody>
</table>
Enter commands
Metandi output
The resulting graph (figure 2) shows the following summaries, together with circles showing the individual study estimates:

- A summary curve from the HSROC model
- A summary operating point, i.e., summary values for sensitivity and specificity
- A 95% confidence region for the summary operating point
- A 95% prediction region (confidence region for a forecast of the true sensitivity and specificity in a future study)
Figure 3a-b. Hierarchical Summary Receiver Operating Characteristics (HSROC) Plot of Studies that Reported both Sensitivity and Specificity in Active TB Suspects.

The summary curves from the HSROC model contain a summary operating point (red square) representing summarized sensitivity and specificity point estimates for individual study estimates (open circles). The 95% confidence region is delineated by the area within the orange dashed line.

Metcalfe J et al. JID, in press
Summary HSROC plots of sensitivity and specificity for anda-TB IgG in smear-positive and smear-negative pulmonary TB patients

- X axis displays specificity
- Y axis displays sensitivity
- Circle for each study
- Width of the circles is proportional to the number of patients in each study
- 236 studies published before 2009
- Bivariate random effects and HSROC models used in only 22% and 5% of studies, respectively.

Uptake of newer methodological developments and the deployment of meta-analysis in diagnostic test research: a systematic review

Brian H Willis¹, Muireann Quigley²

Abstract

Background: The last decade has seen a number of methodological developments in meta-analysis of diagnostic test studies. However, it is unclear whether such developments have permeated the wider research community and on which applications they are being deployed. The objective was to assess the uptake and deployment of the main methodological developments in the meta-analysis of diagnostic tests, and identify the tests and target disorders most commonly evaluated by meta-analysis.

Methods: Design - systematic review. Data Sources - Medline, EMBASE, CINAHL, Cochrane, PsychInfo, Global health, HMIC, and AMED were searched for studies published before 31st December 2008. Selection criteria - studies were included if they satisfied all of the following: evaluated a diagnostic test; measured test performance; searched two or more databases; stated search terms and inclusion criteria; used a statistical method to summarise performance. Data extraction - included the following data items: year; test; reference standard; target disorder; setting; statistical and quality methods.

Results: 236 studies were included. Over the last 5 years the number of meta-analyses published has increased, but the uptake of new statistical methods lags behind. Pooling the sensitivity and specificity and using the SROC remain the preferred methods for analysis in 70% of studies, with the bivariate random effects and HSROC model being used in only 22% and 5% of studies respectively. In contrast, between 2006 and 2008 the QUADAS tool was used in 40% of studies. Broadly, radiological imaging was the most frequent category of tests analysed (36%), with cancer (22%) and infection (21%) being the most common categories of target disorder. Nearly 80% of tests analysed were those normally used in specialist settings.

Conclusion: Although quality assessment in meta-analyses has improved with the introduction of QUADAS, uptake of the newer statistical methods is still lagging behind. Furthermore, the focus of secondary research seems to be in evaluating specialist tests in specialist settings, in contrast to the more routine tests and settings encountered in the majority of clinical practice.
Heterogeneity

• Refers to variation in results among studies
• The variability is often greater than would be expected from within study sampling error alone
• May be explained by variation in
  – Patient characteristics
  – Test methods
  – Study design
  – Other factors
  – Chance
What can we do with heterogeneity?

- Double check the data
- Ignore it (Don’t do this)
- Describe it, but not meta-analyze it
- Encompass it (random effects methods)
- Explore it
Exploring heterogeneity

- Subgroup analyses - are homogeneous analyses with respect to important potential confounders such as patient spectrum, test methods, study design

- Meta-regression analysis – is a form of linear regression
  - aims to relate the size of effect to one or more characteristics of the studies involved
The results show a high degree of variability in accuracy across studies.

Table 4. Diagnostic Odds Ratio (DOR) Estimates from Subgroup Analysis

<table>
<thead>
<tr>
<th>Study Characteristic (n)</th>
<th>DOR</th>
<th>Chi² test of heterogeneity</th>
<th>P value for heterogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DIRECTION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospective (108)</td>
<td>255.63 (199.23, 328.01)</td>
<td>678.67</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Retrospective (9)</td>
<td>315.65 (99.68, 999.57)</td>
<td>150.21</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Both (8)</td>
<td>371.42 (161.83, 852.49)</td>
<td>31.40</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>STUDY DESIGN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross Sectional (124)</td>
<td>269.56 (212.30, 342.26)</td>
<td>869.08</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>RECRUITMENT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consecutive (43)</td>
<td>220.90 (154.41, 316.00)</td>
<td>180.24</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Convenient (24)</td>
<td>347.98 (225.63, 536.67)</td>
<td>91.71</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Both (5)</td>
<td>298.50 (90.72, 982.18)</td>
<td>40.54</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Random (2)</td>
<td>278.72 (3.12, 24901.4)</td>
<td>9.73</td>
<td>0.002</td>
</tr>
<tr>
<td>Not Reported (51)</td>
<td>284.91 (184.02, 441.13)</td>
<td>529.38</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>VERIFICATION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete (123)</td>
<td>264.79 (208.66, 336)</td>
<td>863.88</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>BLINDING</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both (8)</td>
<td>163.93 (69.91, 384.42)</td>
<td>25.49</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Commercial Nucleic-Acid Amplification Tests for Diagnosis of Pulmonary Tuberculosis in Respiratory Specimens: Meta-Analysis and Meta-Regression: Example Meta-regression

Table 6. Results from Meta-Regression Analysis Using the Restricted Maximum Likelihood Method

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Model Coefficient</th>
<th>Relative Diagnostic Odds Ratio (95% CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold Effect (S)</td>
<td>−0.21</td>
<td>—</td>
<td>0.01</td>
</tr>
<tr>
<td>Retrospective/Both (17) vs Prospective Design (108)</td>
<td>0.13</td>
<td>1.14 (0.56, 2.33)</td>
<td>0.71</td>
</tr>
<tr>
<td>Some Convenient Sampling/NR (80) vs Consecutive/Random Sampling (45)</td>
<td>0.38</td>
<td>1.46 (0.87, 2.43)</td>
<td>0.15</td>
</tr>
<tr>
<td>No Blinding/NR (105) vs Any Blinding (20)</td>
<td>0.25</td>
<td>1.29 (0.65, 2.58)</td>
<td>0.47</td>
</tr>
<tr>
<td>FDA-Approved NAATs (92) vs Not FDA-Approved NAATs (33)</td>
<td>−0.06</td>
<td>0.95 (0.53, 1.68)</td>
<td>0.85</td>
</tr>
<tr>
<td>Respiratory Specimens (95) vs Sputum Specimens (30)</td>
<td>0.64</td>
<td>1.89 (1.01, 3.52)</td>
<td>0.05</td>
</tr>
<tr>
<td>Culture Reference Standard (105) vs Clinical Reference/Both (20)</td>
<td>0.34</td>
<td>1.40 (0.70, 2.81)</td>
<td>0.34</td>
</tr>
<tr>
<td>Resolved Data (37) vs Unresolved Data (88)</td>
<td>−0.05</td>
<td>0.95 (0.54, 1.66)</td>
<td>0.86</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0001536.t006

Determined using ‘Metareg’ command in Stata
5. Interpretation of results

• What are the consequences of using the test in terms of the numbers of TP, FP, FN, and TN?
• How applicable are the results?
• To what extent were the primary studies biased? If serious study limitations were identified, could these impact the results?
• What are the implications for research?
**Table 2. GRADE Summary of Findings – Role of IGRAs for evaluation of patients with pulmonary TB in low- and middle-income countries**

**Review question:** What is the diagnostic accuracy of commercial IGRAs for pulmonary tuberculosis?

**Patients/population:** Adult pulmonary TB suspects and confirmed cases in low- and middle-income countries

**Setting:** Outpatients and inpatients

**Index test:** Commercial Interferon-gamma Release Assays (QuantiFERON-TB Gold In-Tube [QFT-GIT], Cellestis, Australia and T-SPOT.TB [T-SPOT], Oxford Immunotec, United Kingdom)

**Importance:** Rapid, accurate, simple test could supplement microscopy and expand testing to peripheral health centers

**Reference standard:** Microbiologic (culture or smear-microscopy) or clinical diagnosis of pulmonary TB

**Studies:** Cross-sectional or cohort

<table>
<thead>
<tr>
<th>Outcomes: TP, TN, FP, FN</th>
<th>Effect % (95% CI)</th>
<th>No. of participants (studies)</th>
<th>What do these results mean given 10% prevalence among suspects being screened for TB?</th>
<th>What do these results mean given 30% prevalence among suspects being screened for TB?</th>
<th>Quality of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subgroups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-SPOT.TB, HIV-Infected</td>
<td>Sensitivity 78% (56, 91) Specificity 55% (45, 64)</td>
<td>549 (5)</td>
<td>With a prevalence of 10%, 100/1000 will have TB. Of these, 78 (TP) will be identified; 22 (FN) will be missed by T-SPOT.TB. Of the 900 patients without TB, 495 (TN) will not be treated; 405 (FP) will be unnecessarily treated.</td>
<td>With a prevalence of 30%, 300/1000 will have TB. Of these, 234 (TP) will be identified; 66 (FN) will be missed by T-SPOT.TB. Of the 700 patients without TB, 385 (TN) will not be treated; 315 (FP) will be unnecessarily treated.</td>
<td>Very Low</td>
</tr>
</tbody>
</table>

Metcalfe J et al. JID, in press
Summary

• Described essential steps in a systematic review of diagnostic test accuracy studies

• Described HSROC/bivariate meta-analysis of data from diagnostic studies

• Described key tools used in carrying out systematic reviews/meta-analysis of diagnostic test accuracy studies
ARE WE THERE YET !?!
Going beyond diagnostic accuracy...

• Link evidence on diagnostic test accuracy to clinical practice
• Use results of systematic reviews of diagnostic test accuracy as inputs into decision analyses
• Go beyond summary ROC curves to describe test performance in terms of the expected downstream benefits and harms of using a test
References and tools for systematic reviews of diagnostic test accuracy

• Leeflang. Ann Intern Med. 2008;149:889-897
• Zamora. BMC Medical Research Methodology 2006, 6:31
• Cochrane Diagnostic Test Accuracy Working Group http://srdta.cochrane.org/
• http://www.teachepi.org/ Dr Pai’s website for learning and teaching epidemiology
• http://www.tbevidence.org/ Evidence-based TB diagnosis
• RevMan http://ims.cochrane.org/revman
• Meta-analysis in Stata… Ed. Jonathan Sterne 2009
special thanks to

- Mariska Leeflang
- Madhu Pai
- Penny Whiting
- Many others