Systematic review and meta-analysis of diagnostic test accuracy studies

Karen R Steingart, MD, MPH
Madhukar Pai, MD, PhD
[madhukar.pai@mcgill.ca]

There are now 60+ systematic reviews in TB diagnostics!
These systematic reviews have contributed to several new policies.

Are these the same or different?

- Traditional, narrative review
- Systematic review
- Overview
- Meta-analysis
- Pooled analysis
Types of review articles

All reviews (also called overviews)

Meta-analyses

Individual patient data (IPD) meta-analyses

Reviews that are not systematic (traditional, narrative reviews)

Systematic reviews


In practice, not all meta-analyses are conducted as part of systematic reviews
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.

All systematic reviews are not meta-analyses!

• “…it is always appropriate and desirable to systematically review a body of data, but it may sometimes be inappropriate, or even misleading, to statistically pool results from separate studies. Indeed, it is our impression that reviewers often find it hard to resist the temptation of combining studies even when such meta-analysis is questionable or clearly inappropriate.”

Elements of a Systematic Review

- Formulate the review question & write a protocol
- Search for and include primary studies
- Assess study quality
- Extract data
- Analyze data
- Interpret results & write a report

Systematic reviews are a lot of work!

Road map for diagnostic reviews

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
- Outcome

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

Overall search strategy

PICO ± STUDY DESIGN FILTER

Filters for diagnostic studies (only if necessary)
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

- Major databases: PubMed, Embase, Biosis, Web of Science
- Check references of relevant studies/reviews
- Contact industry, if relevant
- 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)
- Two reviewers should, ideally, screen citations and identify eligible studies

- 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’

- Search antibodies, bacteria OR antibody/blood OR antibodies/immunology OR antibody react* OR
  "hemolysin immune" OR "hemolysin immunity" OR "hemonal antibody" OR "immune-based" OR "antibody
detection" OR antigens, bacteria OR antigens/analysis OR antigens/blood OR antigens/extravascular fluid OR
antigens/immunology OR antigens/human OR lipopoly saccharide* OR lipopolysaccharide OR "antigen
detection" OR antigen[tiab] OR antigen[ti]

- Search immunologic test(s) OR immunochemistry[mh] OR serology[ti] OR serological[ti] OR serodiagnosis[tiab]

reproducibility of results[ti]

- Search tuberculosis[ti] OR mycobacterium tuberculosis[ti]

3. Assessment of study quality

Evidence of bias and variation in diagnostic accuracy studies

Anja W.E. Rutjes, Johanna B. Reitsma, Marcello Di Nisio, Mykola Smit, Jesper C. van Rijn, Patrick H.M. Bossuyt

Sources of Variation and Bias in Studies of Diagnostic Accuracy

Academia and Clinic

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

Methods: To determine the problems on factors that may lead to...
### Evidence of bias and variation in diagnostic accuracy studies

<table>
<thead>
<tr>
<th>Study characteristics*</th>
<th>Lower estimate</th>
<th>Higher estimate</th>
<th>RDOR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Severe cases and healthy controls</td>
<td>1.1 (0.9-1.3)</td>
<td>4.9 (4.6-5.3)</td>
<td></td>
</tr>
<tr>
<td>Other (case-control designs)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection: referral for index test</td>
<td></td>
<td></td>
<td>0.5 (0.2-0.9)</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></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

---

### 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 Reitsma3, Patrick MM Bossuyt2 and Jos Kleijnen1

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

Email: Penny Whiting* - pwhi@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 - j.kleijnen@amc.uva.nl

* Corresponding author

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

This article is available from: http://www.biomedcentral.com/1471-2288/1/25
© 2003 Whiting 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.
QUADAS-2

Annals of Internal Medicine  Research and Reporting Methods

QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies

Penny F. Whiting, PhD; Anne W.J. Rutjes, PhD; Maartje E. Westwood, PhD; Susan Mallett, PhD; Jonathan J. Deeks, PhD; Johannes B. Reitsma, MD, PhD; Mariska M.G. Leeflang, PhD; Jonathan A.C. Sterne, PhD; Patrick M.M. Bossuyt, PhD; and the QUADAS-2 Group*

In 2009, the QUADAS tool for systematic reviews of diagnostic accuracy studies was developed. Experience, anecdotal reports, and feedback suggested areas for improvement; therefore, QUADAS-2 was developed. This tool comprises 4 domains: patient selection, index test, reference standard, and flow and timing. Each domain is assessed in terms of risk of bias, and the first 3 domains are also assessed in terms of concerns regarding applicability. Signalling questions are included to help judge risk of bias.

The QUADAS-2 tool is applied in 4 phases: summarize the review questions, tailor the tool and produce review-specific guidance, construct a flow diagram for the primary study, and judge bias and applicability. This tool will allow for more transparent rating of bias and applicability of primary diagnostic accuracy studies.

For author affiliations, see end of text.
* For members of the QUADAS-2 Group, see the Appendix (available at www.annals.org).

<table>
<thead>
<tr>
<th>QUADAS Item Scored &quot;Yes&quot;</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>
• 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

DARDS FOR REPORTING OF DIAGNOSTIC ACCURACY (STARD) CHECKLIST

<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>
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

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).
# Measures of test accuracy

<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 (TP)</td>
<td>False positives (FP)</td>
<td>TP + FP</td>
</tr>
<tr>
<td><strong>Index test -</strong></td>
<td>False negative (FN)</td>
<td>True negatives (TN)</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 = $\frac{\text{Sensitivity}}{1 - \text{Specificity}}$
- Likelihood ratio negative = $\frac{1 - \text{Sensitivity}}{\text{Specificity}}$
- Prevalence (proportion of people with disease in population to whom the test has been applied) = $\frac{TP + FN}{TP + FP + FN + TN}$

---

### Other measures of test performance that can be summarized

- Reproducibility of diagnostics
- Predictive value for disease progression (e.g. IGRAs)
- Cost-effectiveness of diagnostics
- Diagnostic delays
- Yield of specimen collection approaches
Within-Subject Variability of Interferon-g Assay Results for Tuberculosis and Boosting Effect of Tuberculin Skin Testing: A Systematic Review

Reproducibility

Pred. value

Specimen collection

Sputum induction for the diagnosis of pulmonary tuberculosis: a systematic review and meta-analysis

Steingart 2010 unpublished

Forest plots of sensitivity and specificity, anda-TB IgG for the diagnosis of pulmonary TB, smear-positive patients

Forest plots of sensitivity and specificity, urine LAM ELISA

SROC curve – GenoType MTBDRplus for RIF resistance

Minion J et al. ERJ 2011

Ling D et al. ERJ 2008
Software options

RevMan 5
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
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
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

---

**Figures**

**Figure 3:** A and B, Hierarchical summary receiver operating characteristic (HSROC) plot of studies that reported both sensitivity and specificity among persons with suspected active tuberculosis. 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 in the orange dashed line.

*Metcalfe JZ et al, J Infect Dis 2011*
Summary ROC plots for Anda-TB IgG for diagnosis of TB:

- (A) smear+ and (B) smear- pulmonary TB patients.
- Red squares are pooled sensitivity and specificity values.

Steingart unpublished

- 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

Summary HSROC plots of sensitivity and specificity for anda-TB IgG in smear-positive and smear-negative pulmonary TB patients


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?

- 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</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>235.63</td>
<td>678.67</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Retrospective (6)</td>
<td>371.43</td>
<td>31.48</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Both (8)</td>
<td>371.43</td>
<td>31.48</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>STUDY DESIGN</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross sectional</td>
<td>289.58</td>
<td>140.88</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>RECRUITMENT</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consecutive (43)</td>
<td>220.90</td>
<td>180.24</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Convenient (24)</td>
<td>247.58</td>
<td>91.71</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Both (5)</td>
<td>280.30</td>
<td>40.04</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Random (2)</td>
<td>278.73</td>
<td>9.73</td>
<td>0.32</td>
</tr>
<tr>
<td>Not reported (11)</td>
<td>289.51</td>
<td>529.28</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>VERIFICATION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complete (122)</td>
<td>244.79</td>
<td>163.68</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>RUNNING</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both (8)</td>
<td>163.93</td>
<td>25.49</td>
<td>0.00</td>
</tr>
</tbody>
</table>


### 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>Overall Effect (6)</td>
<td>0.21</td>
<td>1.23 (1.01, 1.48)</td>
<td>0.03</td>
</tr>
<tr>
<td>Retrospective/Both (17) vs Prospective/Diagnostic (16)</td>
<td>0.13</td>
<td>1.14 (0.58, 2.23)</td>
<td>0.71</td>
</tr>
<tr>
<td>Some Convenient Sampling/Not (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/No (100) vs Any Blinding (50)</td>
<td>0.23</td>
<td>1.29 (0.70, 2.39)</td>
<td>0.47</td>
</tr>
<tr>
<td>FDA-Approved NAATs (12) vs Not FDA-Approved NAATs (33)</td>
<td>0.05</td>
<td>1.05 (0.53, 1.66)</td>
<td>0.85</td>
</tr>
<tr>
<td>Respiratory Specimen (8) vs Sputum Specimen (20)</td>
<td>0.64</td>
<td>1.89 (1.91, 3.52)</td>
<td>0.00</td>
</tr>
<tr>
<td>Culture Reference Standard (10) 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>Nocardia Data (37) vs Unrecognized Data (68)</td>
<td>0.05</td>
<td>0.95 (0.53, 1.66)</td>
<td>0.80</td>
</tr>
</tbody>
</table>

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?

- Systematic reviews should avoid making policy recommendations
Some general limitations of TB diagnostic SRs

- Literature search strategies are imperfect and studies can be missed
- Publication bias is always a concern
- Poor quality studies or poorly reported studies
- Unexplained heterogeneity
- Not enough studies on clinical impact of tests
- Industry supported studies or COI of study authors
- COI of systematic reviewers
- Keeping up to date in rapidly evolving fields
Keeping systematic reviews updated!

Interferon-γ assays in the immunodiagnosis of tuberculosis: a systematic review

Marikani Pai, Lai N, Niu and Linder M Cutbush Jr

Annals of Internal Medicine 2004


Erik Leeflang, et al.

Annals of Internal Medicine 2007


Annals of Internal Medicine 2008

Predictive value of interferon-γ release assays for incident active tuberculosis: a systematic review and meta-analysis


Annals of Internal Medicine 2012

References and tools for systematic reviews of diagnostic test accuracy

- 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
- RevMan http://ims.cochrane.org/revman