

Systematic reviews of diagnostic test accuracy

Madhukar Pai, MD, PhD

Assistant Professor

Department of Epidemiology, Biostatistics & Occupational Health
McGill University, Montreal

Respiratory Epidemiology & Clinical Research Unit, Montreal
Chest Institute, Montreal



McGill



Evidence based diagnosis (“EBD”)

- Evidence based medicine (EBM) incorporates evidence based diagnosis
- EBD, in turn, requires synthesis of evidence on various diagnostic tests and algorithms
- Although diagnostic studies are very common, systematic reviews and meta-analyses of diagnostic studies are uncommon
 - The Cochrane Collaboration just published its first diagnostic review

First Cochrane review on diagnostic test accuracy [published in 2008]

Galactomannan detection for invasive aspergillosis in immunocompromized patients (Review)

Leeflang MM, Debets-Ossenkopp YJ, Visser CE, Scholten RJ, Hooft L, Bijlmer HA, Reitsma JB, Bossuyt PMM, Vandenbroucke-Grauls CM



**THE COCHRANE
COLLABORATION®**

Cochrane Website for Diagnostic Reviews



[Cochrane entities >](#)

Cochrane Diagnostic Test Accuracy Working Group



Welcome

This is the webpage for three related entities of the Cochrane Collaboration; the Diagnostic Test Accuracy Working Group, the Regional Support Units and the Diagnostic Test Accuracy Editorial Team. The combined roles of these entities is to implement the Cochrane Steering Group's decision to publish systematic reviews of diagnostic test accuracy on The Cochrane Library.

The aim of this website is to provide resources and information to all those involved in preparing Cochrane systematic reviews of the accuracy of diagnostic tests.

What do you need to know?

We will try to answer your questions in this website. Please read our FAQ and email us to ask more. Below are brief highlights of some of our activities and links

[Welcome](#)

[DTA Editorial Team](#)

[Regional Support Units](#)

[Contact us](#)

[Statistical analysis](#)

[Handbook for DTA Reviews](#)

[Workshops and events](#)

[Frequently Asked Questions](#)

[Editorial Process of Diagnostic Test Accuracy reviews](#)

[More about us](#)

Search

[Advanced](#)

[Tips](#)

First diagnostic review published

The first Cochrane Diagnostic Test Accuracy Review has been published in the Cochrane Database of Systematic Reviews, Issue 4, 2008.

<http://srdta.cochrane.org/en/index.html>

Meta-analyses of diagnostic studies

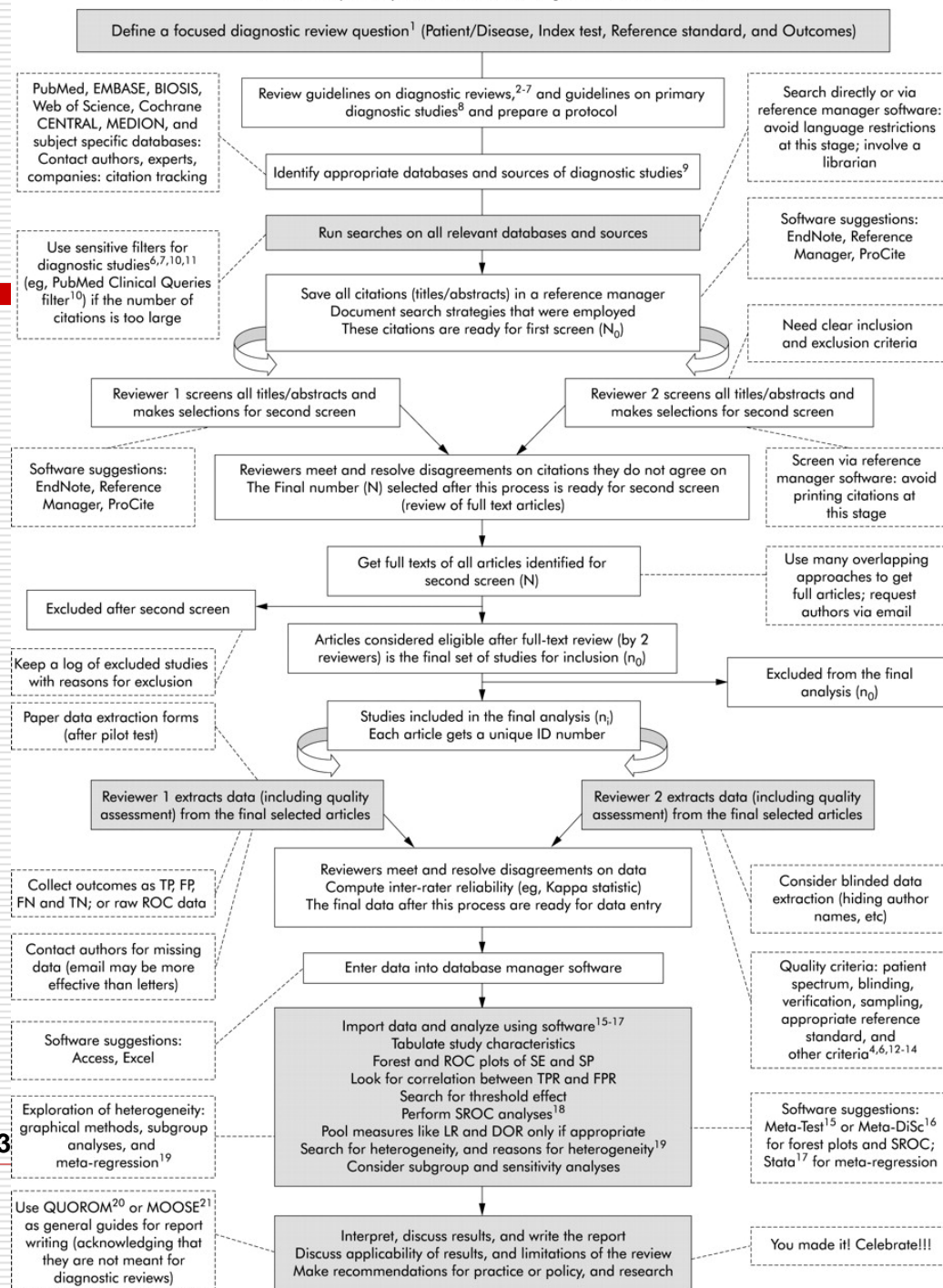
- Meta-analyses of diagnostic studies:
 - Not as common as meta-analyses of RCTs; becoming common
- The objectives of such meta-analyses:
 - Provide a summary of the overall accuracy of a diagnostic test
 - Appraise the quality of primary studies
 - Look for and explore reasons for heterogeneity
 - Evaluate the impact of quality and other study characteristics on diagnostic accuracy
 - Generate promising new research questions

How are meta-analyses of diagnostic studies different from RCT meta-analyses?

- ☐ They differ from meta-analyses of RCTs in some aspects:
 - Search terms for diagnostic studies not well defined
 - Quality of reporting of diagnostic studies is often poor
 - Sources of bias are different
 - ☐ Publication bias may be a bigger problem
 - Analysis
 - ☐ Methods not well developed (effect measure is a curve!)
 - ☐ Variability in thresholds across studies
 - ☐ Summary measures not always clinically meaningful
 - ☐ Heterogeneity is a bigger problem
 - ☐ Most meta-analysis software cannot handle diagnostic data
-

ROAD MAP FOR DIAGNOSTIC REVIEWS

A "road map" for systematic reviews of diagnostic test evaluations

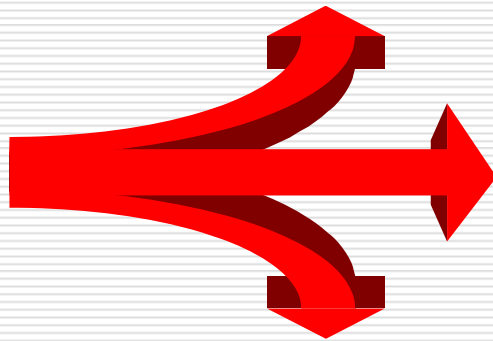


Pai, M. et al. Evid Based Med 2004;9:101-103

Step 1: Formulating a focused question: a 4-part review question

P - Who is the patient or what problem is being addressed?

I - What is the intervention or exposure?



C - What is the comparison group?

O - What is the outcome or endpoint?

+ study design

Richardson et al. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 1995;A-12

Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Ann Intern Med 1997;127:380-7.

Step 1: Formulating a focused review question

Test (intervention) Outcome

↓ ↓

Is Positron Emission Tomography (PET) helpful in diagnosing coronary disease?

Test (intervention) Outcome Patient

↓ ↓ ↓

Is PET a more sensitive and specific test in diagnosing coronary artery disease as compared to coronary angiography?

+ diagnostic studies

↑
Comparison

Step 2: Identify databases/sources of studies

- ☐ Electronic databases:
 - ☐ General: PubMed, Embase, Biosis, Web of Science, etc.
 - ☐ Subject-specific: AIDSLINE, CANCERLIT, PsycInfo, MEDION, etc.
 - ☐ Reference lists of included studies (citation tracking)
 - ☐ References lists of earlier reviews, commentaries
 - CDSR, DARE, MEDION, PubMed search with filters for systematic reviews
 - ☐ Personal communication with experts and authors
 - ☐ Contacting companies and test manufacturers
 - ☐ Hand-searching of key, high-yield journals
 - ☐ Grey literature
 - Dissertation abstracts, reports, conference proceedings, etc.
 - ☐ Sources of ongoing studies
 - companies, contacting experts
-

Need to search several databases

Systematic reviews of test accuracy should search a range of databases to identify primary studies

Penny Whiting^a, Marie Westwood^{b,*}, Margaret Burke^a, Jonathan Sterne^a, Julie Glanville^b

^a*MRC Health Services Research Collaboration, Department of Social Medicine, Canynge Hall, Whiteladies Road, Bristol, BS8 2PR, UK*

^b*Centre for Reviews and Dissemination, University of York, Heslington, York YO10 5DD, UK*

Accepted 1 May 2007

Abstract

Objective: To estimate the yield from searching a range of bibliographic databases and additional sources to identify test accuracy studies for systematic reviews.

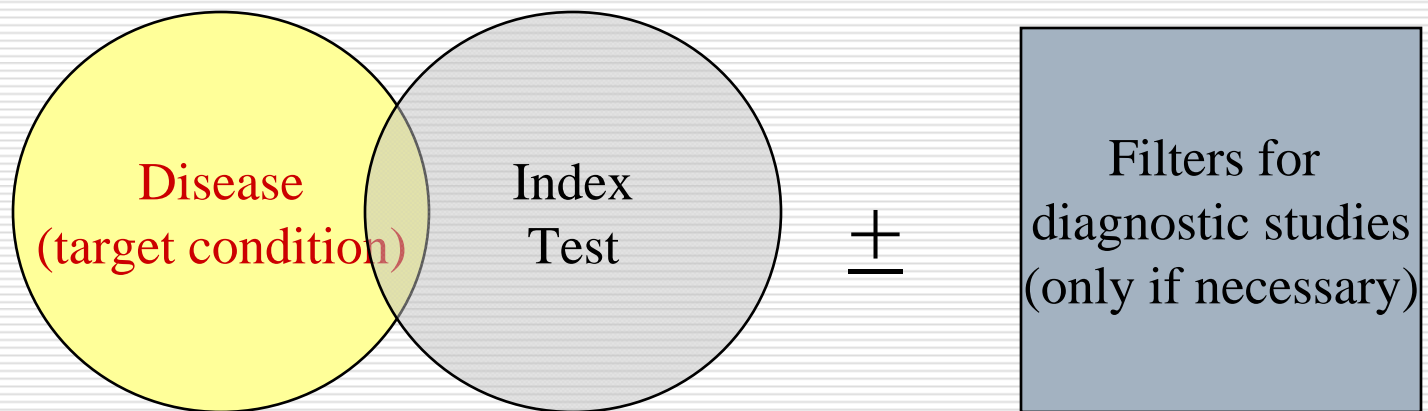
Study Design and Setting: We examined eight systematic reviews and their database searches: MEDLINE, EMBASE, BIOSIS, Science Citation Index, LILACS, Pascal, and CENTRAL. We used studies included in each systematic review as the “gold standard,” against which yield was estimated. For each database, we classified studies in each gold standard set as being (1) included in the database and identified by searches, (2) included and not identified, and (3) not included in the database.

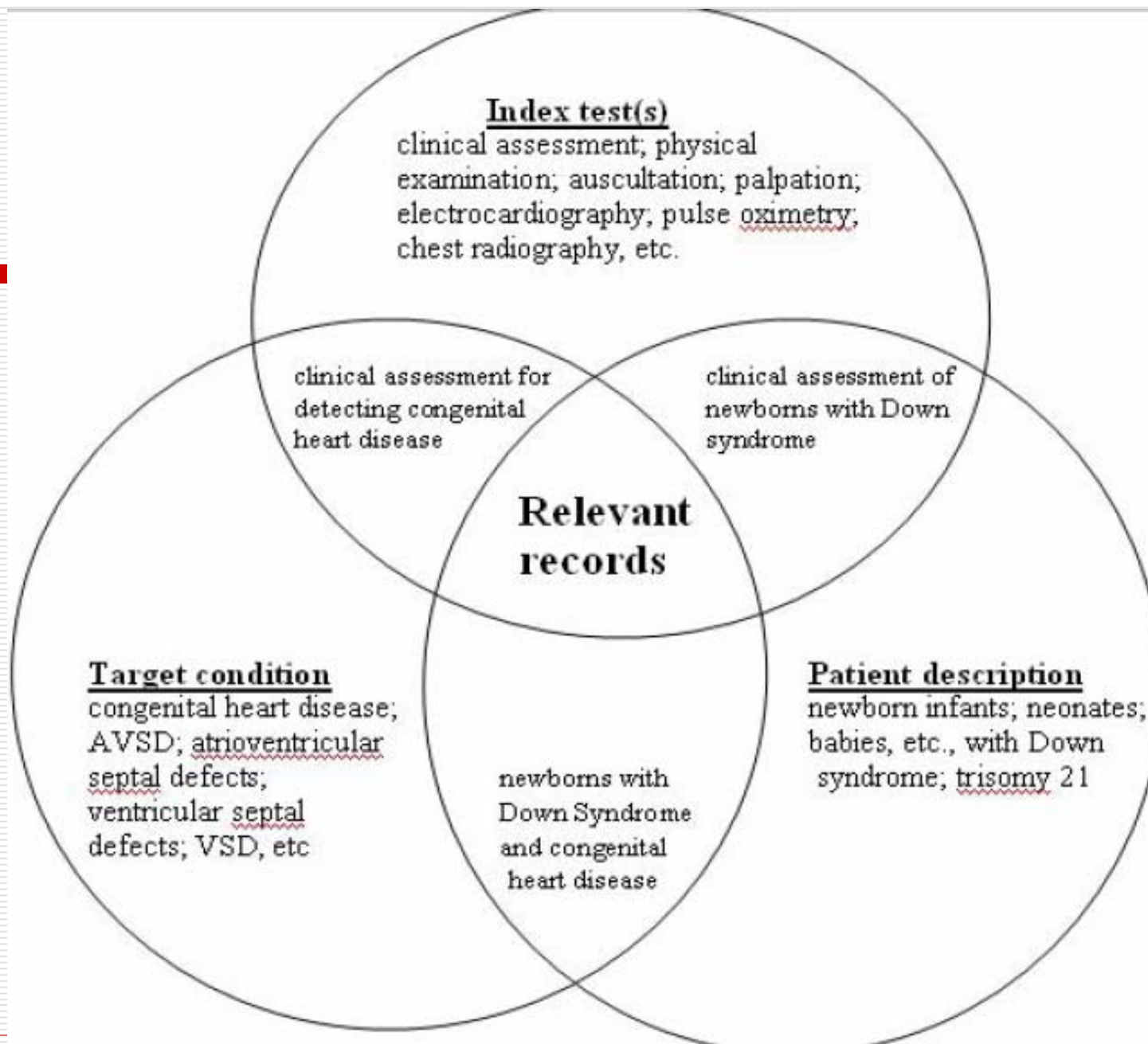
Results: No search identified all studies in any gold standard set. EMBASE, Science Citation Index, and BIOSIS contained studies that were not on MEDLINE. Over 20% of studies in the gold standard sets were not identified by searching MEDLINE. Six studies on LILACS were not on any other database. Eight gold standard studies were not included in any of the databases, and a further 22 were not identified by the electronic search strategies.

Conclusions: Systematic reviews of test accuracy studies should search a range of databases. Even searches designed to be very sensitive, that do not use study design filters, can fail to identify relevant studies. © 2008 Elsevier Inc. All rights reserved.

Overall search strategy

PICO \pm STUDY DESIGN FILTER





Example from Cochrane handbook

Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey

R Brian Haynes, Nancy L Wilczynski for the Hedges Team

Abstract

Objective To develop optimal search strategies in Medline for retrieving sound clinical studies on the diagnosis of health disorders.

Design Analytical survey.

Setting Medline, 2000.

Participants 170 journals for 2000 of which 161 were indexed in Medline.

selecting an optimal search strategy.³ Even clinicians who in principle support the use of evidence for patient care often do not have time to find and apply it in practice.⁶ When they do try, searches are not performed effectively.⁷

Search filters ("hedges") can improve the retrieval of clinically relevant and scientifically sound studies from Medline and similar databases.⁸⁻¹² For instance, when we searched Medline for studies on the diagnosis of arthritis from 1996 to the present using the term "arthritis", 7083 articles alone were

Search filters for Dx studies

BMC Medicine



Research article

Open Access

EMBASE search strategies for identifying methodologically sound diagnostic studies for use by clinicians and researchers


Nancy L Wilczynski¹, R Brian Haynes^{*1,2} and the Hedges Team

Address: ¹Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, L8N 3Z5 Canada and ²Health Information Research Unit, Department of Clinical Epidemiology and Biostatistics, Department of Medicine, McMaster University, Hamilton, Ontario, L8N 3Z5 Canada

Email: Nancy L Wilczynski - wilczyn@mcmaster.ca; R Brian Haynes^{*} - bhaynes@mcmaster.ca


^{*} Corresponding author

Search filters in PubMed (Clinical Queries)

**NCBI**

[All Databases](#) [PubMed](#) [Nucleotide](#) [Protein](#) [Genome](#) [Structure](#) [OMIM](#) [PMC](#) [Journals](#) [Books](#)

[About Entrez](#)
[Text Version](#)
Entrez PubMed
[Overview](#)
[Help](#)
[FAQ](#)
[Tutorials](#)

[New/Noteworthy](#) 

[E-Utilities](#)

PubMed Services
[Journals Database](#)
[MeSH Database](#)
[Single Citation Matcher](#)
[Batch Citation Matcher](#)
Clinical Queries
[Special Queries](#)
[LinkOut](#)
[My NCBI](#)

Related Resources
[Order Documents](#)
[NLM Mobile](#)
[NLM Gateway](#)
[TOXNET](#)
[Consumer Health](#)
[Clinical Alerts](#)
[ClinicalTrials.gov](#)
[PubMed Central](#)

[Privacy Policy](#)

PubMed Clinical Queries

This page provides the following specialized PubMed searches for clinicians:

- [Search by Clinical Study Category](#)
- [Find Systematic Reviews](#)
- [Medical Genetics Searches](#)

After running one of these searches, you may further refine your results using PubMed's [Limits](#) feature.

Results of searches on these pages are limited to specific clinical research areas. For comprehensive searches, use [PubMed](#) directly.

Search by Clinical Study Category

This search finds citations that correspond to a specific clinical study category. The search may be either broad and sensitive or narrow and specific. The search filters are based on the work of [Haynes RB et al](#). See the [filter table](#) for details.

Search

Category	Scope
<input type="radio"/> etiology	<input checked="" type="radio"/> narrow, specific search
<input type="radio"/> diagnosis	<input type="radio"/> broad, sensitive search
<input checked="" type="radio"/> therapy	
<input type="radio"/> prognosis	
<input type="radio"/> clinical prediction guides	

Find Systematic Reviews

For your topic(s) of interest, this search finds citations for systematic reviews, meta-analyses, reviews of clinical trials, evidence-based medicine, consensus development conferences, and guidelines.

For more information, see [Help](#). See also [related sources](#) for systematic review searching.

Search

Search filters for diagnostic studies

❑ PubMed "Clinical Queries" (Haynes BR et al):

Diagnosis	sensitive/ broad	98%/74%	(sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnos*[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp])
	Specific/ narrow	64%/98%	(specificity[Title/Abstract])

Haynes RB, Wilczynski NC for the Hedges Team. Optimal search strategies for retrieving scientifically strong studies of diagnosis from MEDLINE: analytical survey. BMJ. 2004 May 1;328(7447):1040

Use filters only if necessary


ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 234–240

epidemiology

Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies

M.M.G. Leeflang^{a,b,*}, R.J.P.M. Scholten^{a,b}, A.W.S. Rutjes^a, J.B. Reitsma^a, P.M.M. Bossuyt^a

^a*Academic Medical Center, Amsterdam, The Netherlands*

^b*Dutch Cochrane Centre, Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, J1B-210,
P.O. Box 22700, 1100 DE Amsterdam, The Netherlands*

Accepted 4 July 2005

Abstract

Objective: To determine the usefulness of methodological filters in search strategies for diagnostic studies in systematic reviews.

Study Design and Setting: We made an inventory of existing methodological search filters for diagnostic accuracy studies and applied them in PubMed to a reference set derived from 27 published systematic reviews in a broad range of clinical fields. Outcome measures were the fraction of not identified relevant studies and the reduction in the number of studies to read.

Results: We tested 12 search filters. Of the studies included in the systematic reviews, 2%–28% did not pass the sensitive search filters, 4%–24% did not pass the accurate filters, and 39%–42% did not pass the specific filters. Decrease in number-needed-to-read when a search filter was used in a search strategy for a diagnostic systematic review varied from 0% to 77%.

Conclusion: The use of methodological filters to identify diagnostic accuracy studies can lead to omission of a considerable number of relevant studies that would otherwise be included. When preparing a systematic review, it may be preferable to avoid using methodological filters. © 2006 Elsevier Inc. All rights reserved.

Step 3: Study selection by two reviewers

- ❑ Two reviewers independently screen the titles/abstracts for eligibility
 - Need clear, detailed inclusion/exclusion criteria!
 - ❑ Reviewers meet and resolve disagreements
 - When in doubt, err on the side of inclusion
 - ❑ Citations identified after first screen are eligible for second screen (full-text review)
 - ❑ Two reviewers screen full-text papers and select the final set of studies
-

Step 4: Data extraction and quality assessment

- ☐ Criteria for validity of diagnostic studies:
 - Study design
 - ☐ Cross-sectional study of a clinically indicated population or case-control
 - Verification
 - ☐ Complete, different reference tests, or partial
 - Blinding
 - ☐ Blinded or not
 - Patient selection
 - ☐ Consecutive or random or nonconsecutive
 - Data collection
 - ☐ Prospective or retrospective
 - Appropriateness of reference standard
 - Description of test
 - Description of study population
-

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

Jeroen G. Lijmer, MD

Ben Willem Mol, MD, PhD

Siem Heisterkamp, PhD

Gouke J. Bonsel, MD, PhD

Martin H. Prins, MD, PhD

Jan H. P. van der Meulen, MD, PhD

Patrick M. M. Bossuyt, PhD

DURING RECENT DECADES, THE number of available diagnostic tests has been rapidly increasing. As for all new medical technologies, new diagnostic tests should be thoroughly evaluated prior to their introduction into daily practice.

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.

Main Outcome Measures Relative diagnostic odds ratio (RDOR), which compared the diagnostic odds ratios of studies of a given test that lacked a particular methodological feature with those without the corresponding shortcomings in design.

Results Fifteen (6.8%) of 218 evaluations met all 8 criteria; 64 (30%) met 6 or more. Studies evaluating tests in a diseased population and a separate control group over-

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

An abridged version of this article appeared in the Feb. 14, 2006, issue of *CMAJ*.

ABSTRACT

Background: Studies with methodologic shortcomings can overestimate the accuracy of a medical test. We sought to determine and compare the direction and magnitude of the effects of a number of potential sources of bias and variation in studies on estimates of diagnostic accuracy.

Methods: We identified meta-analyses of the diagnostic accuracy of tests through an electronic search of the databases MEDLINE, EMBASE, DARE and MEDION (1999-2002). We in-

Although the number of test evaluations in the literature is increasing, much remains to be desired in terms of methodology. A series of surveys have shown that only a small number of studies of diagnostic accuracy fulfil essential methodologic standards.¹⁻³

Shortcomings in the design of clinical trials are known to affect results. The biasing effects of inadequate randomization procedures and differential dropout have been discussed and demonstrated in several publications.⁴⁻⁶ A growing understanding of the potential sources of bias and variation has led to the development of guidelines to help researchers and

Research article

Open Access

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

Research article

Open Access

Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies

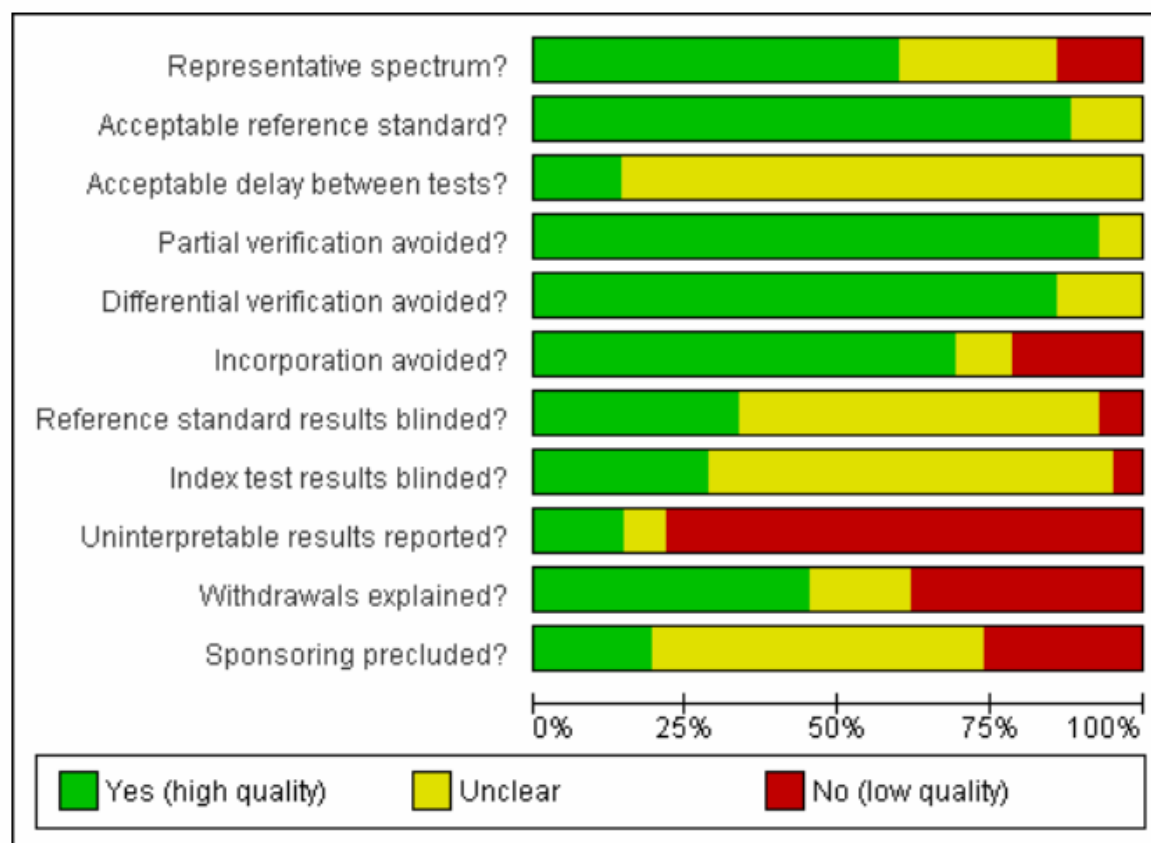
Penny F Whiting*¹, Marie E Weswood², Anne WS Rutjes³,
Johannes B Reitsma³, Patrick NM Bossuyt³ and Jos Kleijnen²

Table 2: The QUADAS tool

Item	Yes	No	Unclear
1. Was the spectrum of patients representative of the patients who will receive the test in practice?	()	()	()
2. Were selection criteria clearly described?	()	()	()
3. Is the reference standard likely to correctly classify the target condition?	()	()	()
4. 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?	()	()	()
5. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?	()	()	()
6. Did patients receive the same reference standard regardless of the index test result?	()	()	()
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?	()	()	()
8. Was the execution of the index test described in sufficient detail to permit replication of the test?	()	()	()
9. Was the execution of the reference standard described in sufficient detail to permit its replication?	()	()	()
10. Were the index test results interpreted without knowledge of the results of the reference standard?	()	()	()
11. Were the reference standard results interpreted without knowledge of the results of the index test?	()	()	()
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?	()	()	()
13. Were uninterpretable/ intermediate test results reported?	()	()	()
14. Were withdrawals from the study explained?	()	()	()

Example from a Cochrane review

Figure 2. Methodological quality graph: Review authors' judgments about each methodological quality item presented as percentages across all included studies.



Leeflang et al. Galactomannan for aspergillosis...CDSR 2008

What about quality scores?

BMC Medical Research Methodology



Research article

Open Access

No role for quality scores in systematic reviews of diagnostic accuracy studies

Penny Whiting*¹, Roger Harbord¹ and Jos Kleijnen²

Address: ¹MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, Bristol, UK and ²Centre for Reviews and Dissemination, University of York, York, UK

Email: Penny Whiting* - penny.whiting@bristol.ac.uk; Roger Harbord - roger.harbord@bristol.ac.uk; Jos Kleijnen - jk13@york.ac.uk

* Corresponding author

The different methods of weighting individual items from the same quality assessment tool produced different quality scores. The different scoring schemes ranked different studies in different orders...

Step 5: Data analyses

- ☐ Enter data into Excel or ACCESS
- ☐ Import data and analyze using software
- ☐ Tabulate study characteristics
- ☐ Forest plots of sensitivity and specificity
- ☐ Avoid simple pooling of sens and spec
- ☐ Search for threshold effect; perform SROC analyses (HSROC is preferable)
- ☐ Search for heterogeneity, and reasons for heterogeneity
- ☐ Consider subgroup and sensitivity analyses
- ☐ If pooled estimates are needed, then use bivariate random effects regression

Tabulation of study characteristics

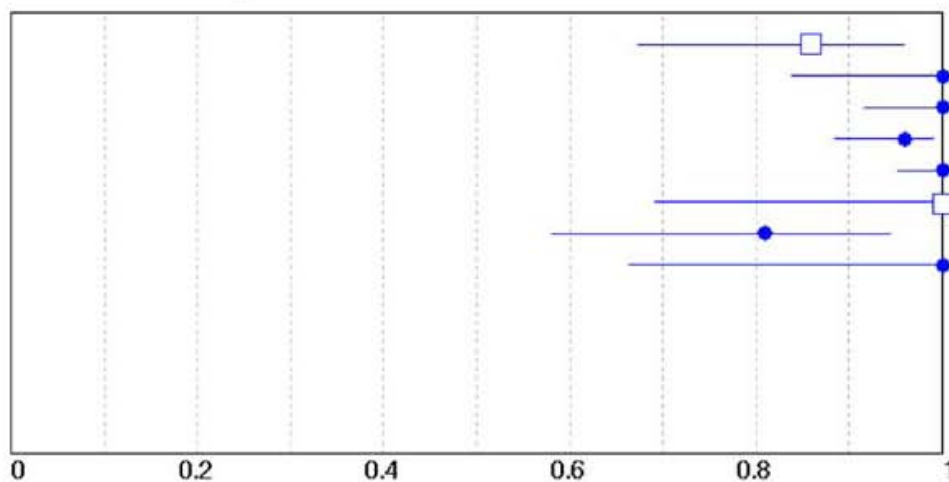
Table 1: Description of studies included in meta-analysis.

Author (year)	Country	Reference Test	Blinded to reference test?	Sample	Sample size (# resistant / # sensitive)	Sensitivity (95% CI)	Specificity (95% CI)
Ahmad (2002)	Kuwait	BACTEC 460	Not Specified	Isolate	29/12	0.97 (.82–1.0)	1.0 (.74–1.0)
De Oliveira (1998)	Brazil	Proportion	Not Specified	Isolate	113/15	0.97 (.92–.99)	1.0 (.78–1.0)
Gamboa (1998)	Spain	BACTEC 460	Not Specified	Isolate	46/13	1.0 (.92–1.0)	1.0 (.75–1.0)
Hirano (1999)	Japan	Proportion	Not Specified	Isolate	90/26	0.92 (.85–.97)	1.0 (.87–1.0)
Johansen (2003)	Denmark	BACTEC 460	Not Specified	Isolate	35/24	0.97 (.85–1.0)	1.0 (.86–1.0)
Jureen (2004)	Sweden	BACTEC 460	Not Specified	Isolate	27/26	1.0 (.87–1.0)	0.92 (.75–.99)
Lemus (2004)	Belgium	BACTEC 460, Proportion	Yes	Isolate	10/10	1.0 (.69–1.0)	1.0 (.69–1.0)
Rossau (1997)	Belgium	Proportion	Not Specified	Isolate	203/61	0.98 (.95–1.0)	1.0 (.94–1.0)
Sintchenko (1999)	Australia	BACTEC 460	Not Specified	Isolate	22/11	0.96 (.77–1.0)	1.0 (.72–1.0)
Somoskovi (2003)	USA	Proportion	Not Specified	Isolate	64/37	0.95 (.87–.99)	1.0 (.91–1.0)
Srivastava (2004)	India	MIC	Not Specified	Isolate	45/10	0.82 (.68–.92)	1.0 (.69–1.0)
Tracevska (2002)	Latvia	BACTEC 460	Not Specified	Isolate	34/19	1.0 (.90–1.0)	1.0 (.82–1.0)
Traore (2000)	Belgium	Proportion	Not Specified	Isolate	266/145	0.99 (.96–1.0)	1.0 (.98–1.0)
Watterson (1998)	England	BACTEC 460, Proportion	Not Specified	Isolate	16/16	1.0 (.80–1.0)	0.94 (.70–1.0)
De Beenhouwer (1995)	Belgium	Proportion	Not Specified	Clinical Specimen	21/46	0.91 (.70–1.0)	1.0 (.92–1.0)
Gamboa (1998)	Spain	BACTEC 460	Not Specified	Clinical Specimen	46/13	0.98 (.89–1.0)	1.0 (.75–1.0)
Johansen (2003)	Denmark	BACTEC 460	Not Specified	Clinical Specimen	26/21	1.0 (.87–1.0)	1.0 (.84–1.0)
Watterson (1998)	England	BACTEC 460, proportion	Yes	Clinical Specimen	10/24	0.80 (.44–.98)	1.0 (.86–1.0)

Morgan M, Kalantri SP, Flores L, Pai M. A commercial line probe assay for the rapid detection of rifampicin resistance in *Mycobacterium tuberculosis*: a systematic review and meta-analysis. *BMC Infect Dis* 2005;5:62.

Forest plots of sensitivity and specificity

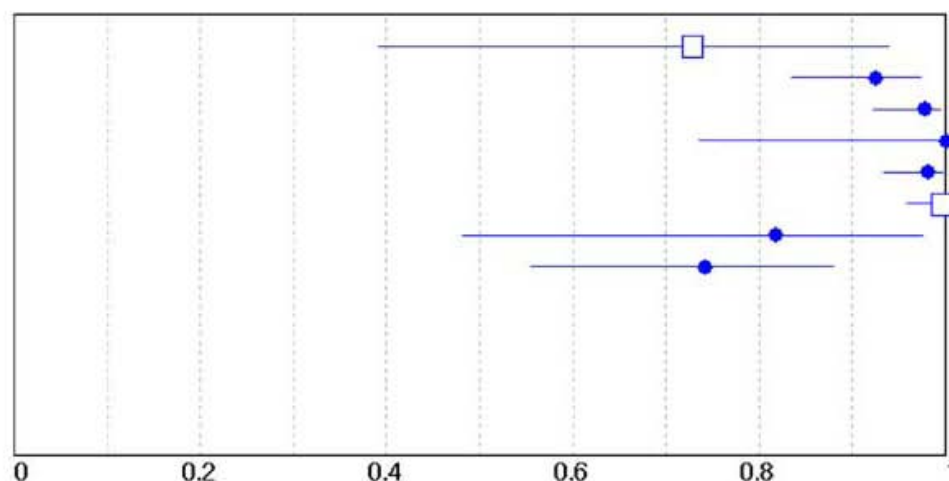
A Sensitivity



Sensitivity (95% CI)

Butt (2004)	0.86	(0.67 - 0.96)
Kisa (2003)	1.00	(0.84 - 1.00)
Albert (2002)	1.00	(0.92 - 1.00)
Krishnamurthy (2002)	0.96	(0.88 - 0.99)
Albert (2001)	1.00	(0.95 - 1.00)
Albert (2004)	1.00	(0.69 - 1.00)
Oguz (2002)	0.81	(0.58 - 0.95)
Aktepe (2001)	1.00	(0.66 - 1.00)

B Specificity

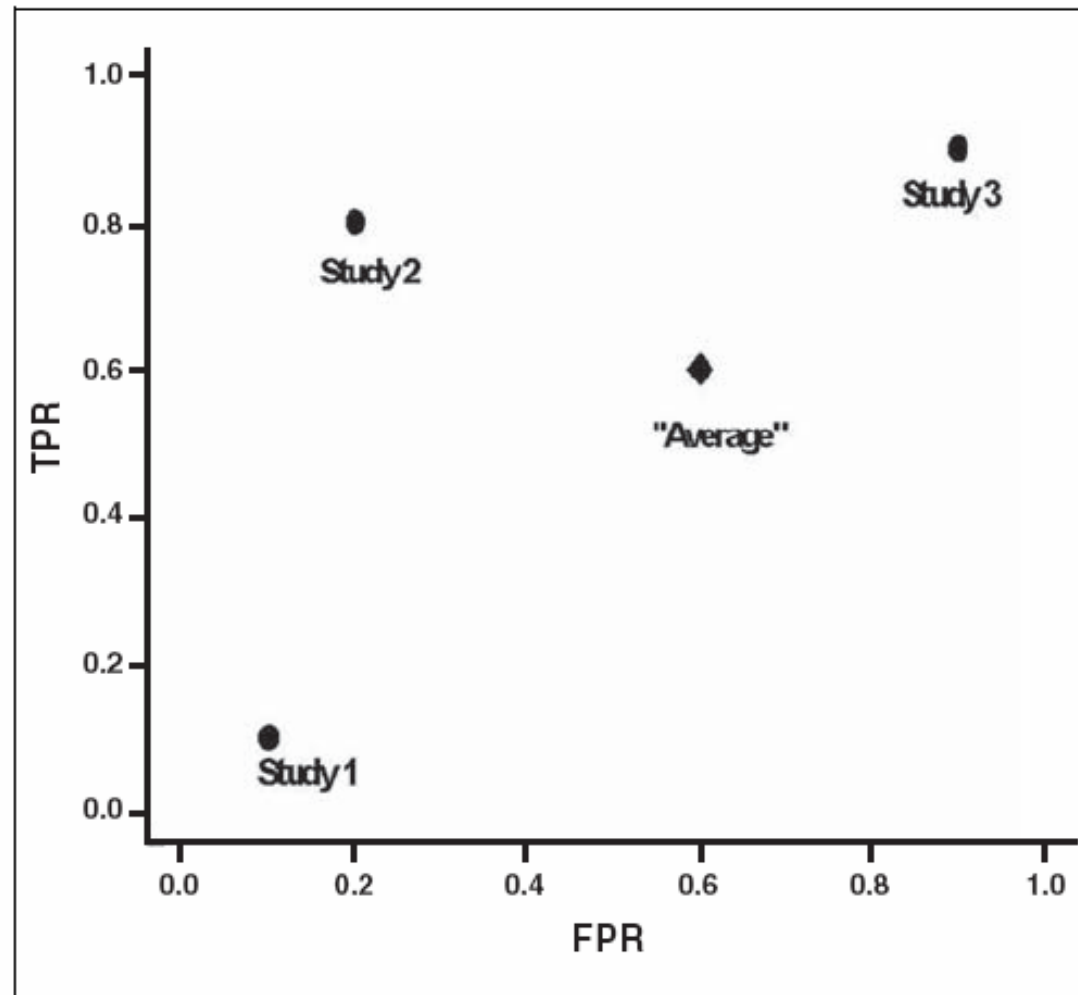


Specificity (95% CI)

Butt (2004)	0.73	(0.39 - 0.94)
Kisa (2003)	0.93	(0.83 - 0.98)
Albert (2002)	0.98	(0.92 - 1.00)
Krishnamurthy (2002)	1.00	(0.74 - 1.00)
Albert (2001)	0.98	(0.93 - 1.00)
Albert (2004)	0.99	(0.96 - 1.00)
Oguz (2002)	0.82	(0.48 - 0.98)
Aktepe (2001)	0.74	(0.55 - 0.88)

Why simple pooling of sens/spec can be misleading

Fig. 1—Graph shows that averaging sensitivities and specificities can be misleading. TPR = true-positive rate, FPR = false-positive rate.

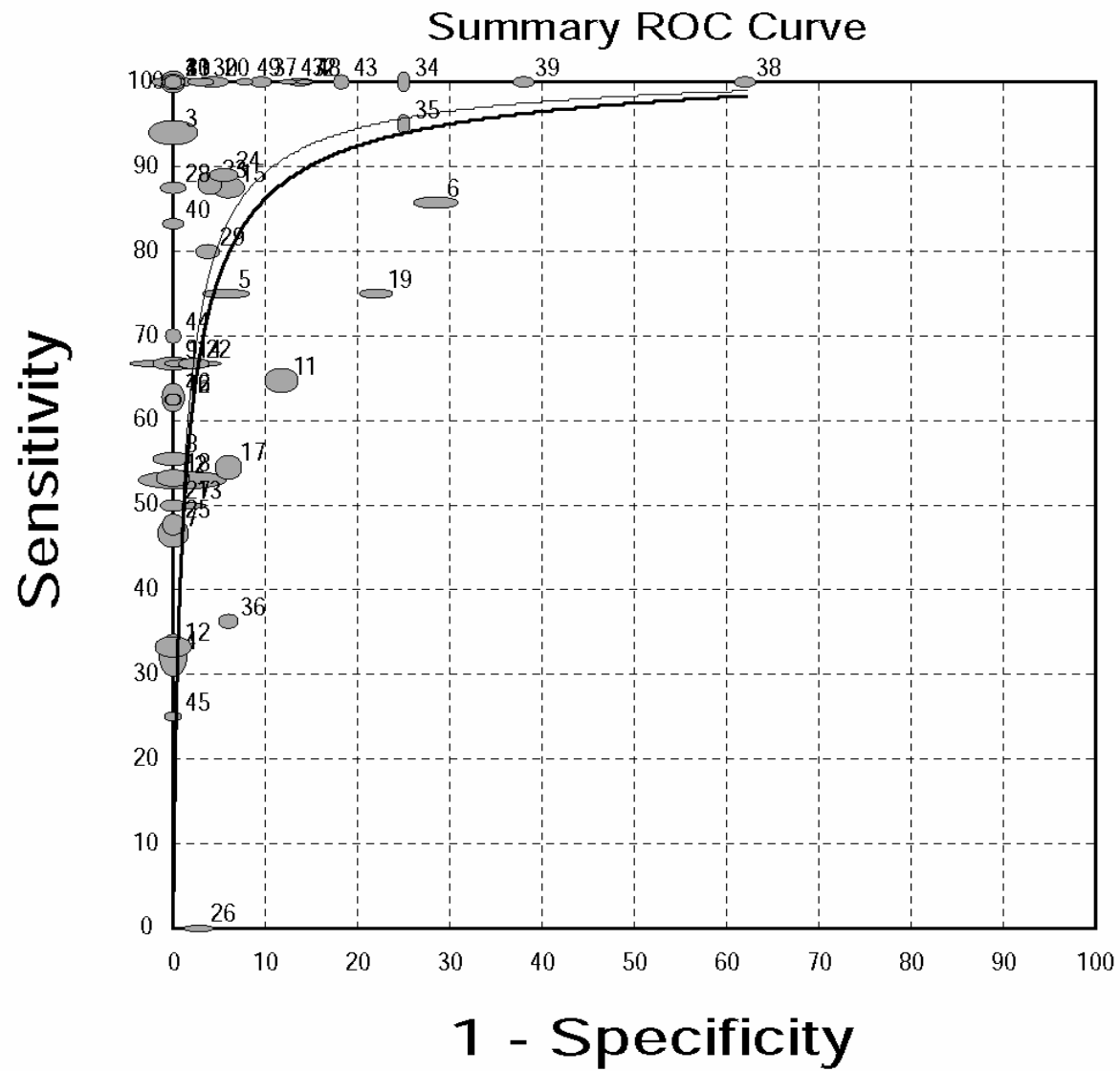


Gatsonis & Paliwal. Am J Roentgen 2006

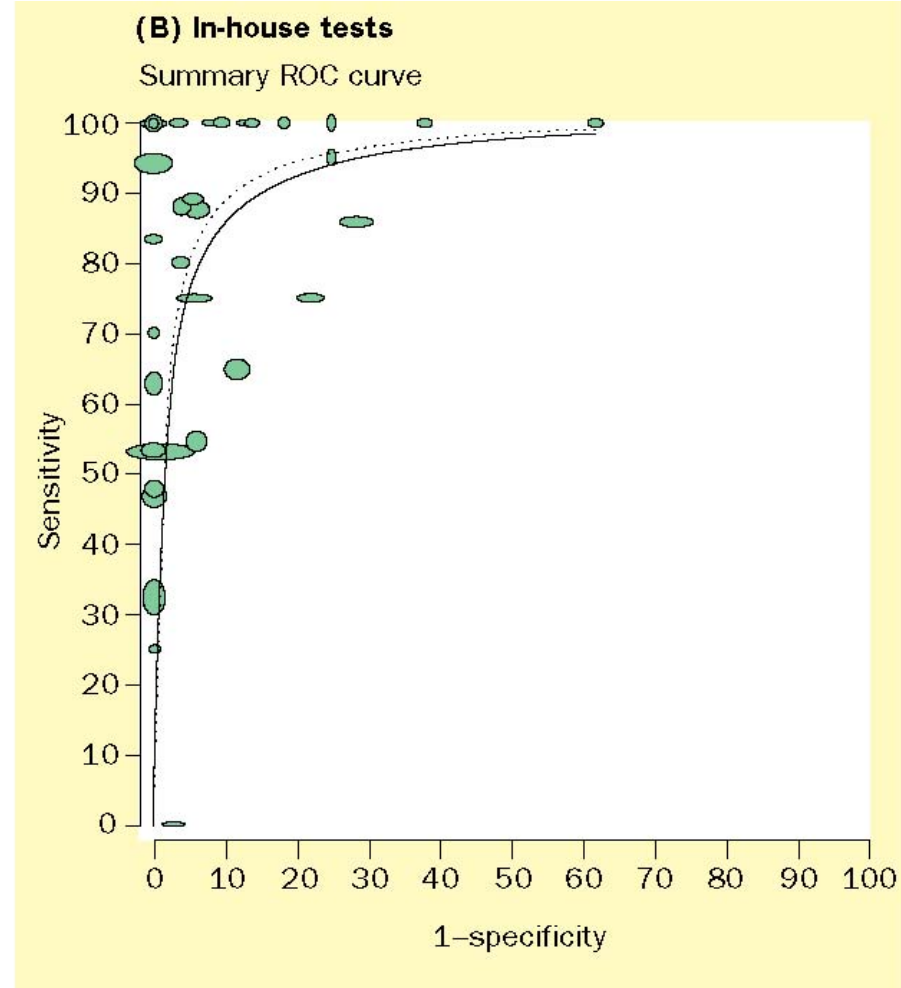
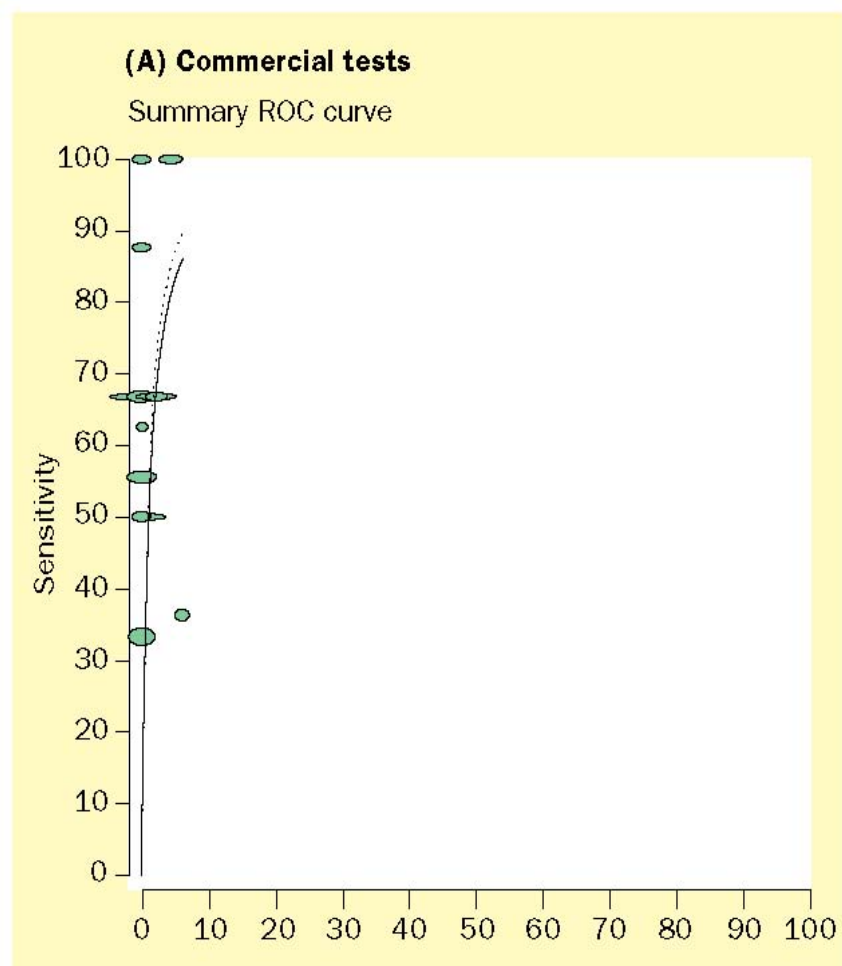
Summary ROC (SROC) Analysis

- ❑ Each study in the meta-analysis contributes a pair of numbers: TPR and FPR
- ❑ Since these measures are correlated and vary with the thresholds (cut points for determining test positives) employed, it is important to analyze them as pairs, and to also explore the effect of threshold on study results.
- ❑ A common approach to summarizing the joint distribution of Se and Sp is called Summary Receiver Operating Characteristic (SROC) curve (Littenberg & Moses 1993)
- ❑ Unlike a traditional ROC plot that explores the effect of varying thresholds on sensitivity and specificity in a single study, each data point in the SROC space represents a separate study
- ❑ The SROC curve and the area under it present a global summary of test performance, and display the trade off between sensitivity and specificity
- ❑ Q^* can also be used as a global summary of accuracy

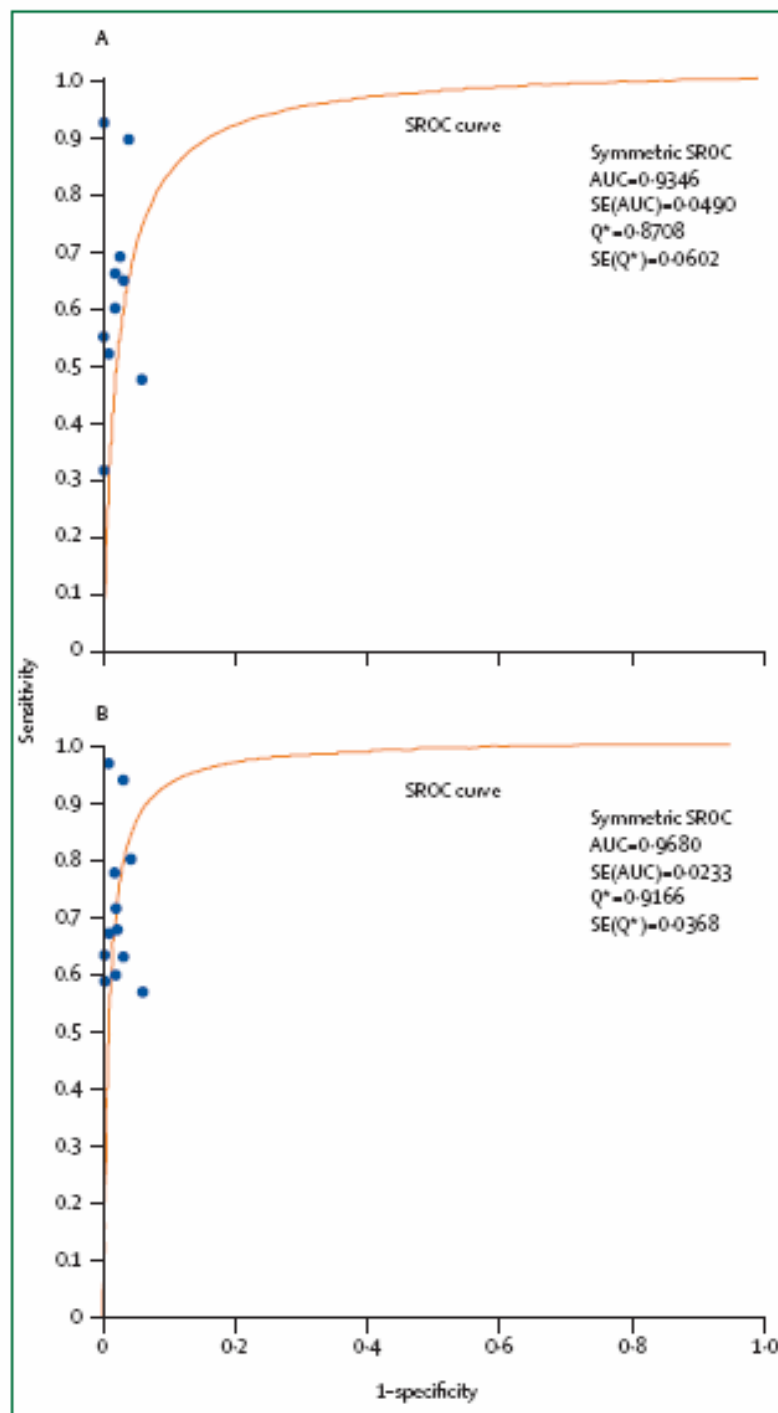
Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. *Med Decis Making* 1993; **13**: 313–321.



NAAT for TB meningitis

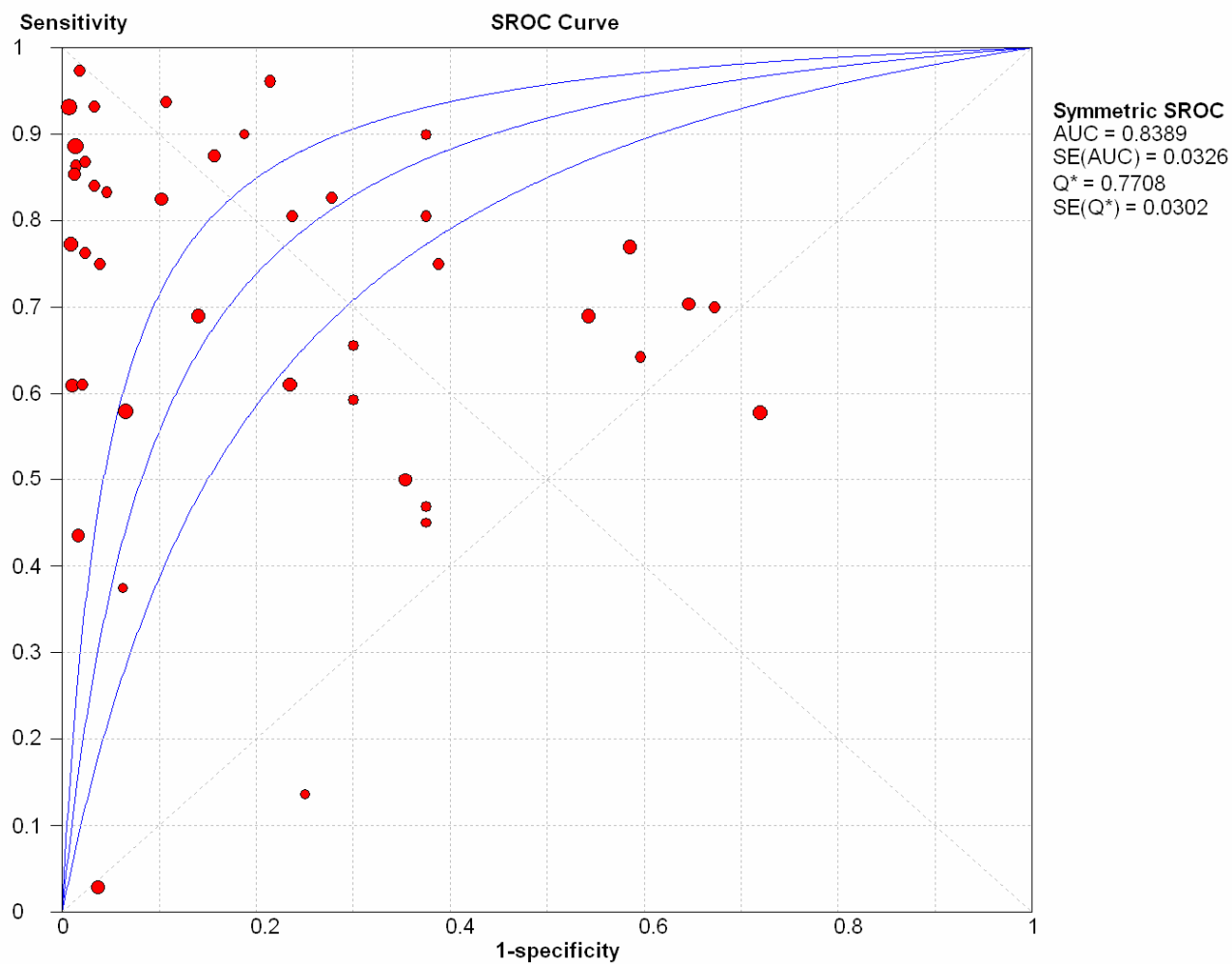


Pai M et al. Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis. *Lancet Infect Dis* 2003;3:633-43.



Conventional vs. fluorescence
microscopy for diagnosis
of tuberculosis

Steingart et al. Lancet Infect Dis 2006



NAAT for TB
lymphadenitis

Daley et al. IJTLD
2007

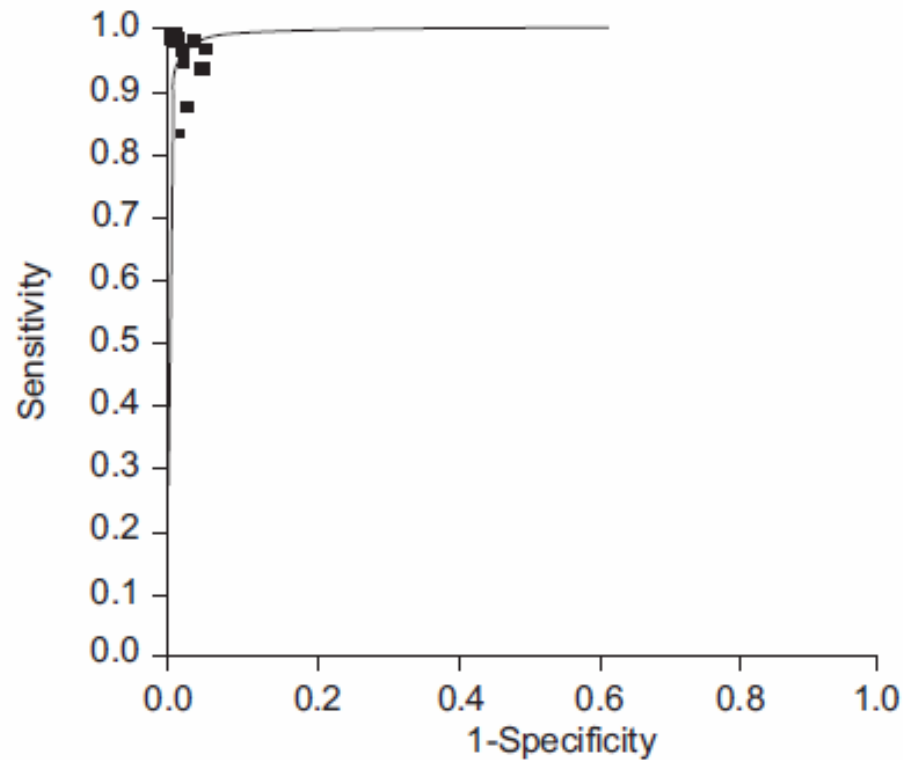
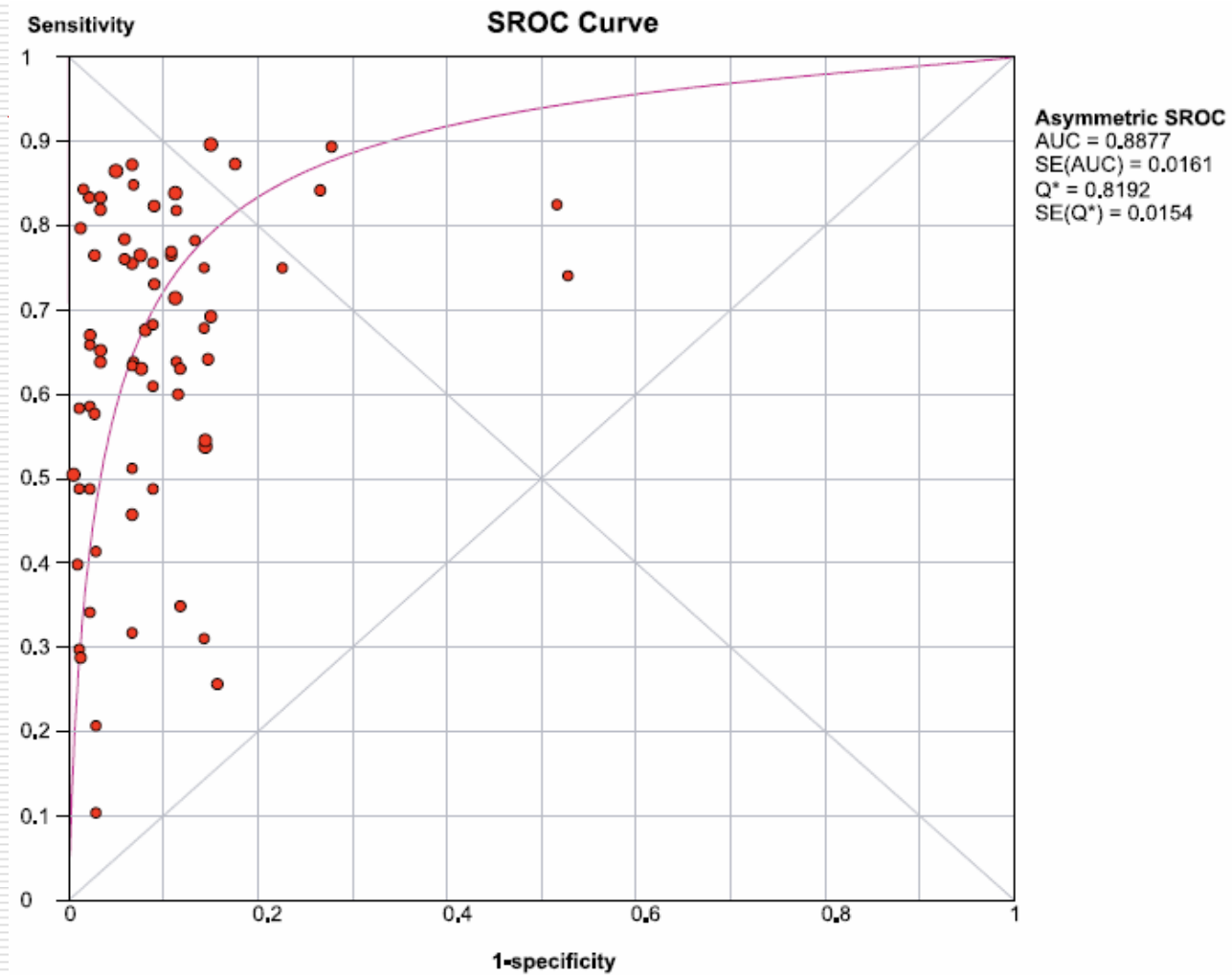


FIGURE 3. Summary receiver operating characteristic (SROC) plot for rifampicin resistance (all 14 studies, regardless of specimen type or assay version). ■: each study in the meta-analysis size proportional to size of study; —: regression line that summarises the overall diagnostic accuracy. Area under the curve (AUC) 0.9949; SE of AUC 0.0023; point of the SROC curve where the sensitivity and specificity are equal (Q^*) 0.9722; SE of Q^* 0.0073.

Line Probe Assay
for Rifampin
resistance

Ling et al. ERJ
2008

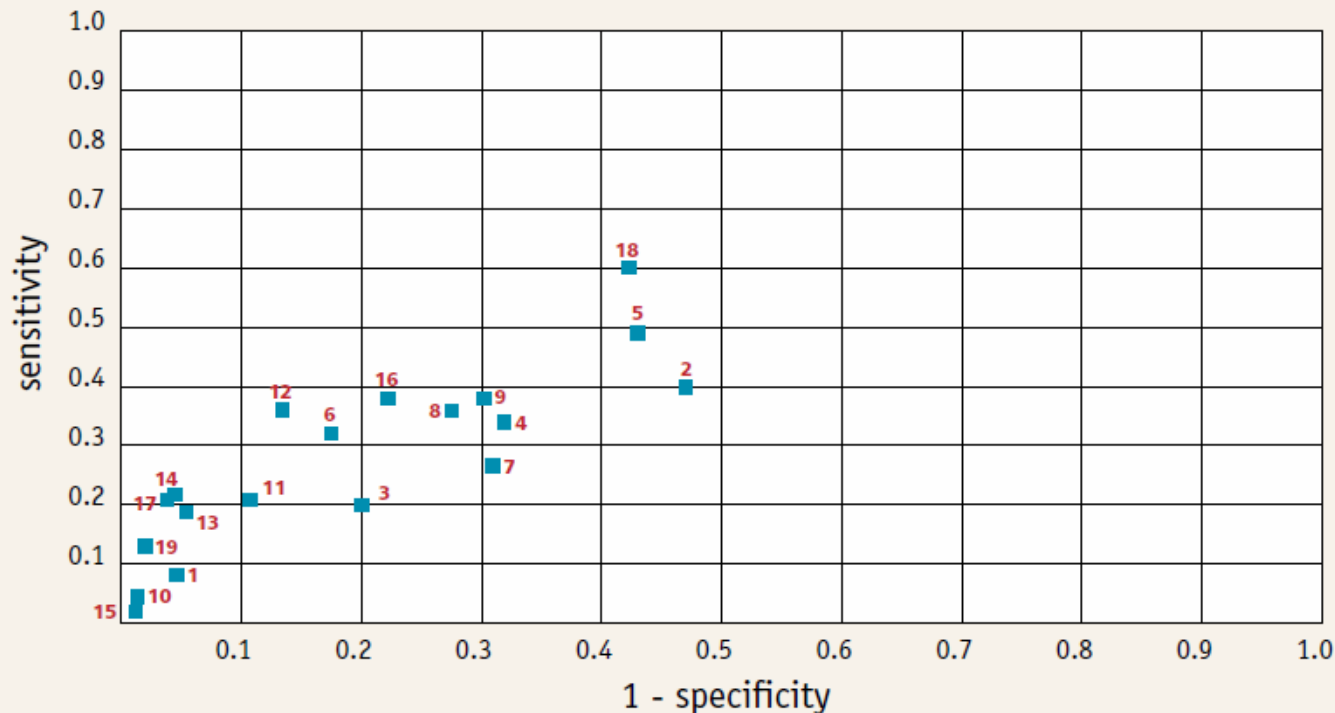


Serological tests
for pulmonary TB

Steingart et al.
PLoS Med 2007

WHO/TDR evaluation of 19 commercial serological tests for TB

Figure 4. ROC curve of commercial rapid tests for the diagnosis of pulmonary tuberculosis (all patients, n=355)



1. ABP Diagnostics 2. Advanced Diagnostics 3. Products 6. Chembio Diagnostic Systems 7. CTK Biotech American Bionostica 4. Ameritek USA 5. Bio-Medical 8. Hema Diagnostic Systems 9. Laboratorios Silanes

Heterogeneity

- Heterogeneity in measures of test accuracy is an important concern with all meta-analyses
 - A bigger concern with diagnostic meta-analyses?
 - Exploration of heterogeneity may be the most important contribution of a diagnostic meta-analysis
 - Sources of heterogeneity include variability in:
 - Disease
 - Index tests
 - Reference standards
 - Thresholds used
 - Populations and disease spectrum
 - Study quality
 - Random error
-

Line Probe Assay for Rifampin Resistance

Sens & Spec
are homogenous

D.I. LING ET AL.

GENOTYPE MTBDR ASSAY FOR DRUG-RESISTANT TUBERCULOSIS

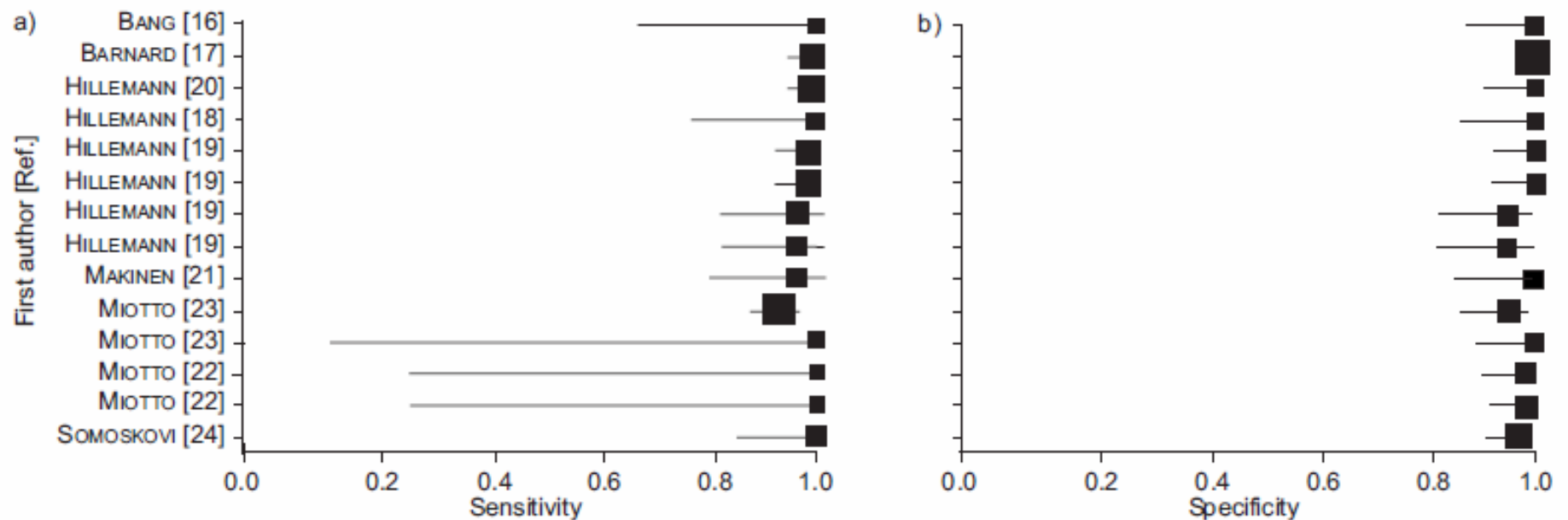
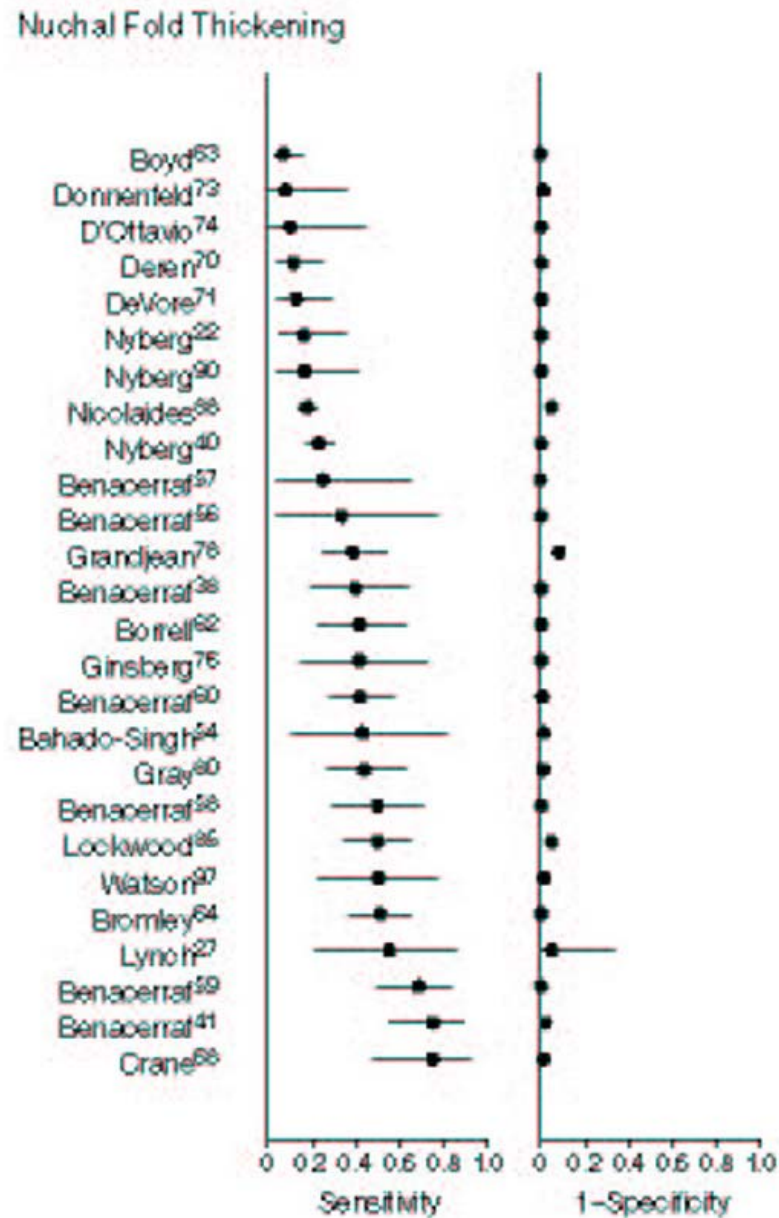


FIGURE 2. Forest plot of sensitivity (a) and specificity (b) estimates for rifampin resistance (all 14 studies, regardless of specimen type or assay version). ■: point estimates of sensitivity and specificity from each study (proportionate to size of the study); —: 95% confidence intervals. See also table 4.

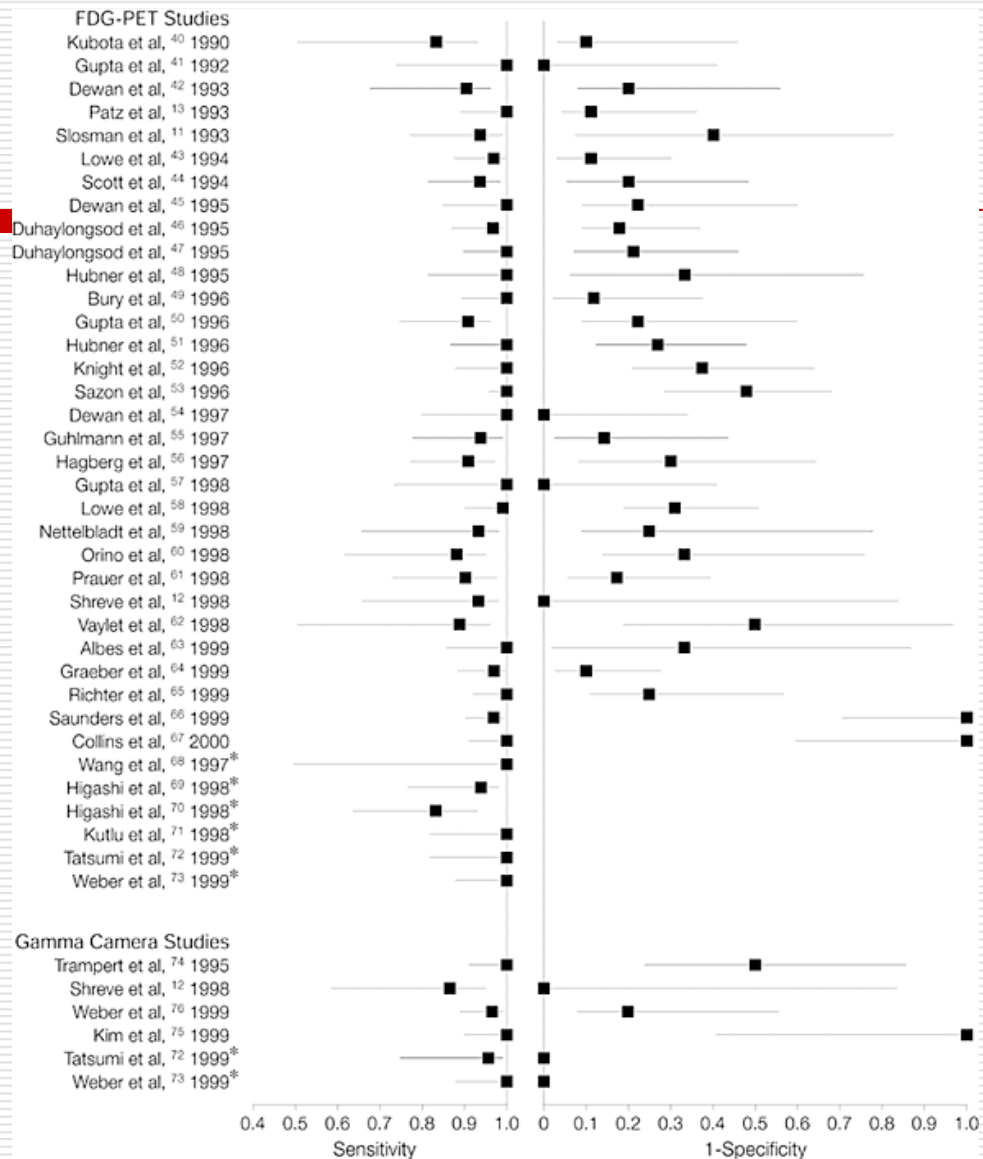
Ultrasound for Down's syndrome

Sensitivity heterogeneous



PET for lung nodules

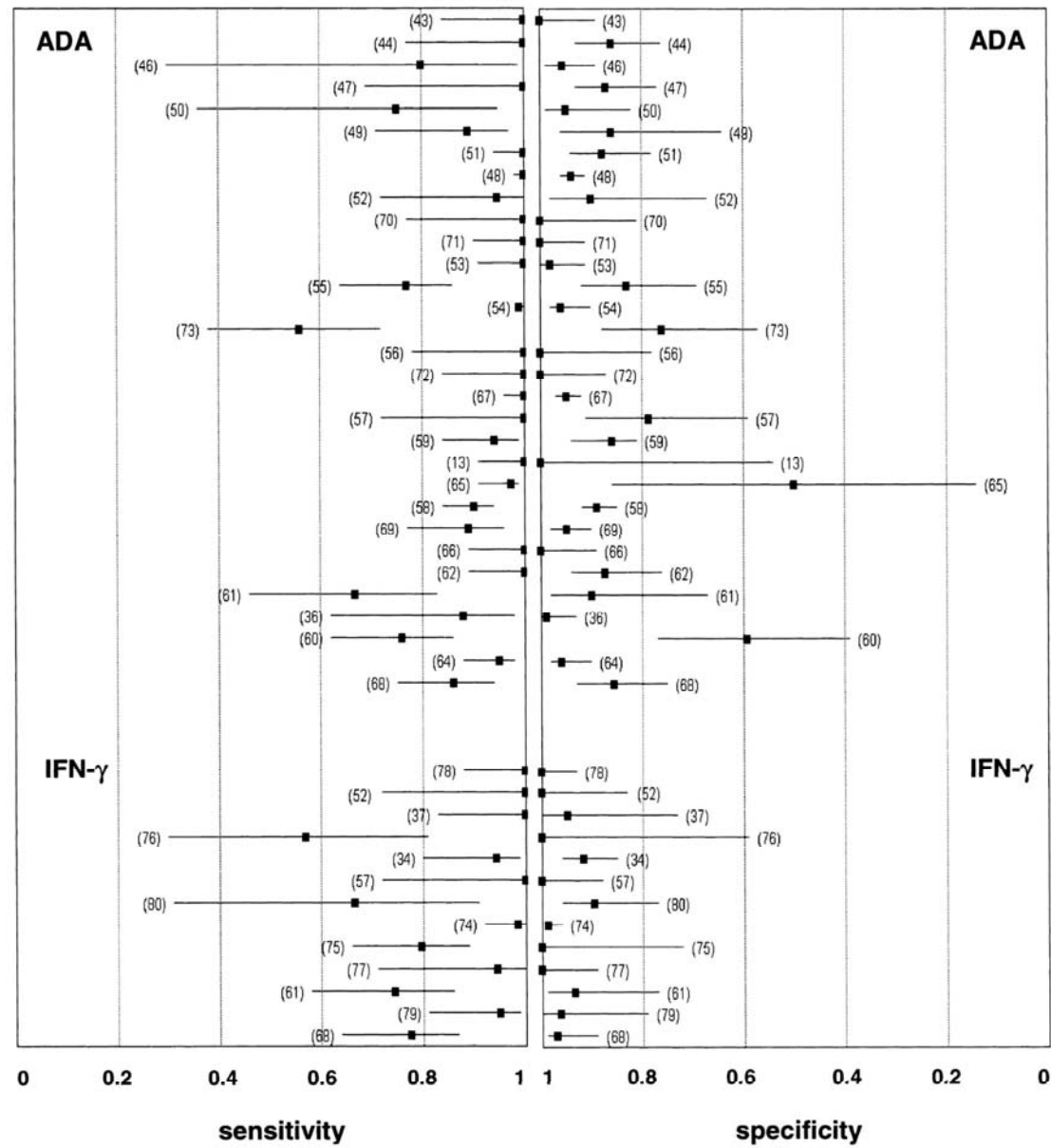
Specificity heterogeneous



Gould et al. JAMA. 2001;285:914-924.

ADA and IFN- γ for pleural TB

Both sensitivity & specificity heterogeneous



Heterogeneity

- Approaches to heterogeneity:
 - Avoid simple pooling of accuracy measures
 - Do subgroup (stratified) analyses
 - Meta-regression analysis
 - Extension of the SROC analysis
 - Outcome variable: Diagnostic Odds Ratio (DOR)
 - Covariates in the model: study-level factors that may be responsible for the heterogeneity (e.g. study quality)
-

Diagnostic Meta-analysis Software

- Specialized packages:

- Meta-DiSc (J Zamora):
 - Windows-based, public domain
- Dr-ROC (M Mitchell)
 - Windows-based, commercial

- General packages:

- STATA: metandi [by Roger Harbord]
 - SAS: programs exist for bivariate regression and HSROC [Rietsma et al]
 - R programs
-

Software

Open Access

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

Javier Zamora*¹, Victor Abraira¹, 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 Abraira - Victor.abraira@hrc.es; Alfonso Muriel - Alfonso.muriel@hrc.es; Khalid Khan - k.s.khan@bham.ac.uk; Arri Coomarasamy - arricoomar@blueyonder.co.uk

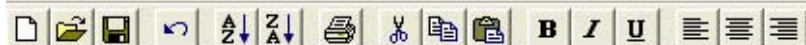
* Corresponding author

Published: 12 July 2006

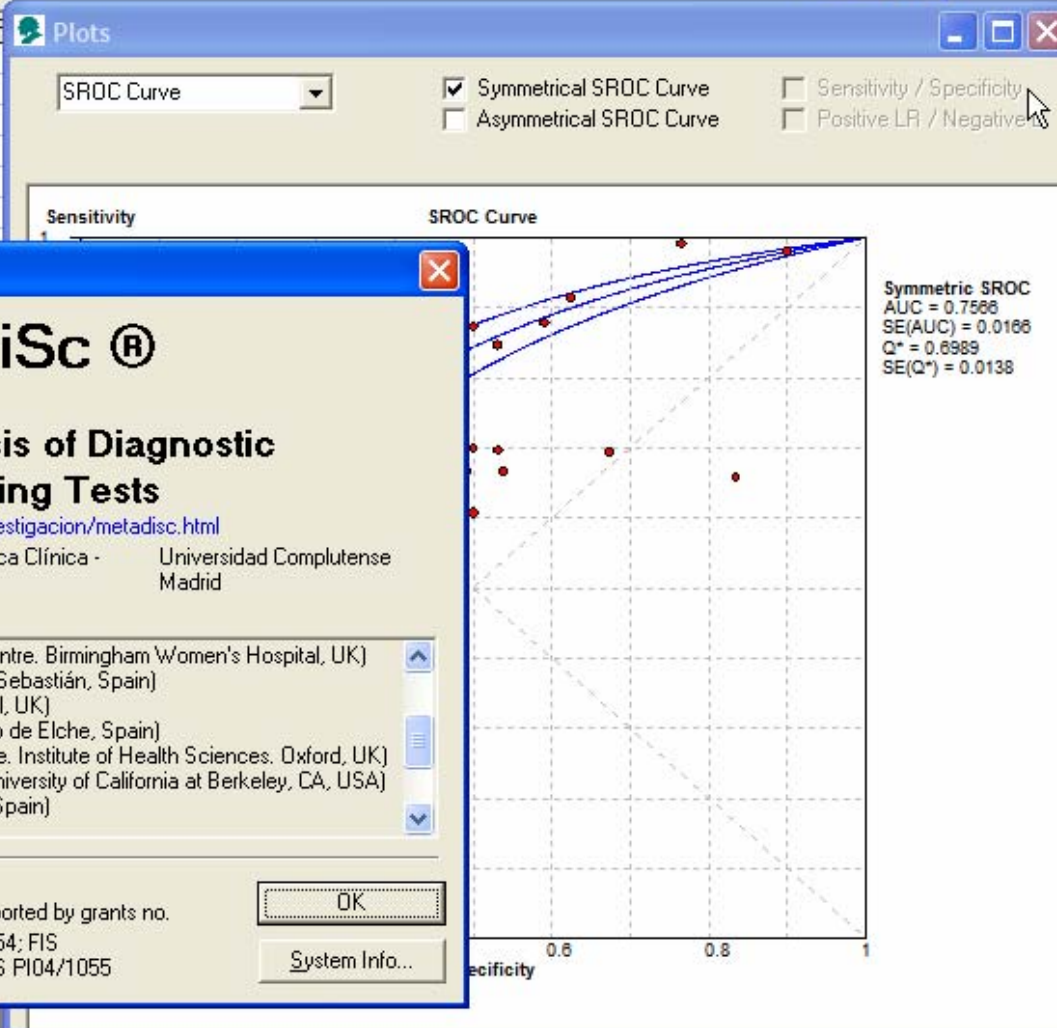
BMC Medical Research Methodology 2006, 6:31 doi:10.1186/1471-2288-6-31

Received: 31 March 2006

Accepted: 12 July 2006



No.	StudyId	Author	TP	FP
1	Ref 1.	Ajons-van K		31
2	Ref 2.	Alloub		8
3	Ref 3.	Anderson 1		70
4	Ref 4.	Anderson 2		65
5	Ref 5.	Anderson 3		20
6	Ref 6.	Andrews		35
7	Ref 7.			20
8	Ref 8.			
9	Ref 9.			
10	Ref 10.			
11	Ref 11.			
12	Ref 12.			
13	Ref 13.			
14	Ref 14.			
15	Ref 15.			
16	Ref 16.			
17	Ref 17.			
18	Ref 18.			
19	Ref 19.			
20	Ref 20.			
21	Ref 21.			
22	Ref 22.			



About Meta-DiSc

Meta-DiSc ®
Version 1.4

Meta-analysis of Diagnostic and Screening Tests

<http://www.hrc.es/investigacion/metadisc.html>

Unidad de Bioestadística Clínica - Universidad Complutense
Hospital Ramón y Cajal Madrid

Acknowledgments

- Arri Coomarasamy (Education Resource Centre, Birmingham Women's Hospital, UK)
- José I. Emparanza (Hospital Donostia, San Sebastián, Spain)
- Khalid Khan (Birmingham Women's Hospital, UK)
- Jaime Latour (Hospital General Universitario de Elche, Spain)
- Sue Mallett (Centre for Statistics in Medicine, Institute of Health Sciences, Oxford, UK)
- Madhukar Pai (Division of Epidemiology, University of California at Berkeley, CA, USA)
- José I. Pijoan (Hospital de Cruces, Bilbao, Spain)

Authors ©:
Javier Zamora,
Alfonso Muriel,
Víctor Abraira

Partially supported by grants no.
FIS PI02/0954; FIS
G03/090; FIS PI04/1055

OK
System Info...

R functions for diagnostic meta analysis

Gillian Raab and Francesca Chappell, Napier Univesity

[back to Gillian Raab's personal page](#)

This page explains how to use the functions to carry out a diagnostic meta-anal as discussed in our draft paper "When are summary ROC curves appropriate for diagnostic meta analyses?" which can be accessed [here](#) as a pdf file. It includes a [flow chart](#) that suggests how the functions might be used.

These notes are intended to help people who have never used R before to use the functions to analyse their data. Some notes for R users are [here](#). They are written for a Windows implementatiation of R, windows or XP. Users of other systems should contact me for the source code and advice.

First few steps - do once only to install programs

- Download the current version of the R program for Windows (2.7.1 as I write this) from the Web site <http://www.R-project.org>. This should be easy by following prompts but click here [Rinstall help](#) for details of the steps to take. When you have downloaded and run the installation file you will find the current version of R in your start menu and as a desktop icon.
- Download the files for the libraries *DiagMeta* and *lme4* and *Matrix* to somewhere on your computer. These are

[DiagMeta_1.01.zip](#) our library of functions for Diagnostic meta analyses.

[lme4_0.99875-9.zip](#) Douglas Bates's library for non-linear fitting and

[Matrix_0.999375-9.zip](#) and the matrix package it requires.

- Start R and you will have a command window open with a prompt like this `>`. There are also menus along the top. Go to the packages menu and select the last option "Install package(s) from local zip files". Click on this and install the packages from the zip files.
- You now have everything you need to use the DiagMeta package any time you start R.

<http://www2.napier.ac.uk/depts/fhls/diagmeta/>

Limitations of current analytic methods

- ❑ Simple pooling of sens and spec is not recommended
 - ❑ SROC analysis has become default
 - Easy to do, but not easy to interpret
 - Based on fixed effects model
 - ❑ Approaches that overcome these problems:
 - Hierarchical SROC [Rutter & Gatsonis 2001]
 - Bivariate random effects regression [van Houwelingen et al, 1993; Reitsma et al, 2005]
 - Both models incorporate random effects
 - ❑ Allows for variability in thresholds, and between and within study variation in accuracy
 - ❑ Allow for inclusion of covariates
-

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

Carolyn M. Rutter^{1,*†} and Constantine A. Gatsonis²

¹*Group Health Cooperative, Center for Health Studies, 1730 Minor Avenue, Suite 1600, Seattle, WA 98101, U.S.A.*

²*Center for Statistical Sciences, Brown University, Box G-H, Providence, RI 02912, U.S.A.*



ELSEVIER

Journal of Clinical Epidemiology 58 (2005) 982–990

**Journal of
Clinical
Epidemiology**

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

Johannes B. Reitsma^{a,*}, Afina S. Glas^a, Anne W.S. Rutjes^a, Rob J.P.M. Scholten^b,
Patrick M. Bossuyt^a, Aeilko H. Zwinderman^a

^a*Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam,
PO Box 22700, 1100 DE Amsterdam, The Netherlands*

^b*Dutch Cochrane Centre, Academic Medical Center, University of Amsterdam, The Netherlands*

Accepted 21 February 2005

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

BY ROGER M. HARBORD

*MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol,
Canynges Hall, Whiteladies Road, Bristol BS8 2PR, UK*
e-mail: roger.harbord@bristol.ac.uk Tel: +44 (0) 117 928 7289 Fax: +44 (0) 117 928 7325

JONATHAN J. DEEKS

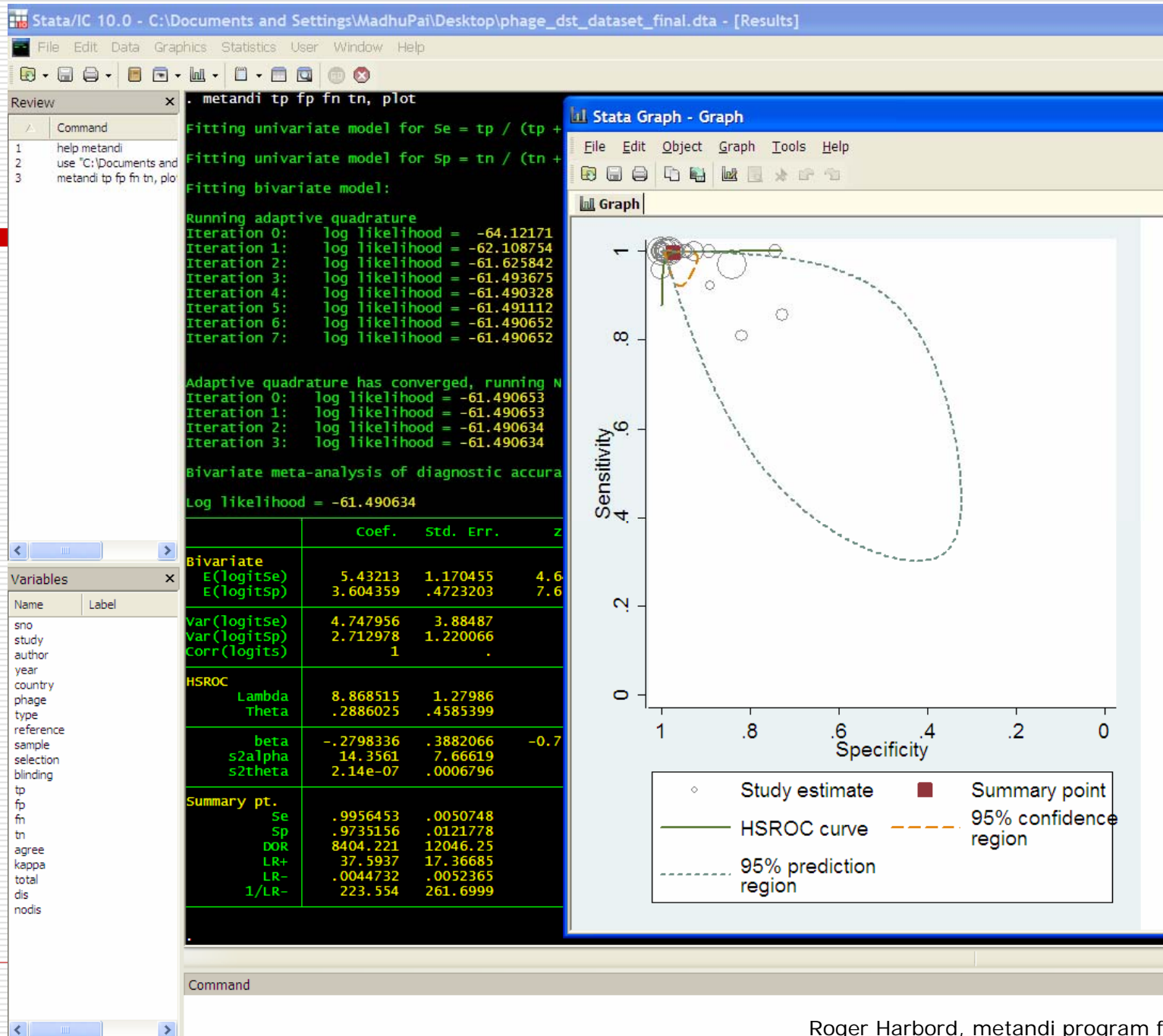
Centre for Statistics in Medicine, Oxford, UK

MATTHIAS EGGER

Department of Social and Preventive Medicine, University of Berne, Switzerland

PENNY WHITING, JONATHAN A.C. STERNE

MRC Health Services Research Collaboration, Department of Social Medicine, University of Bristol, UK



Roger Harbord, metandi program for STATA

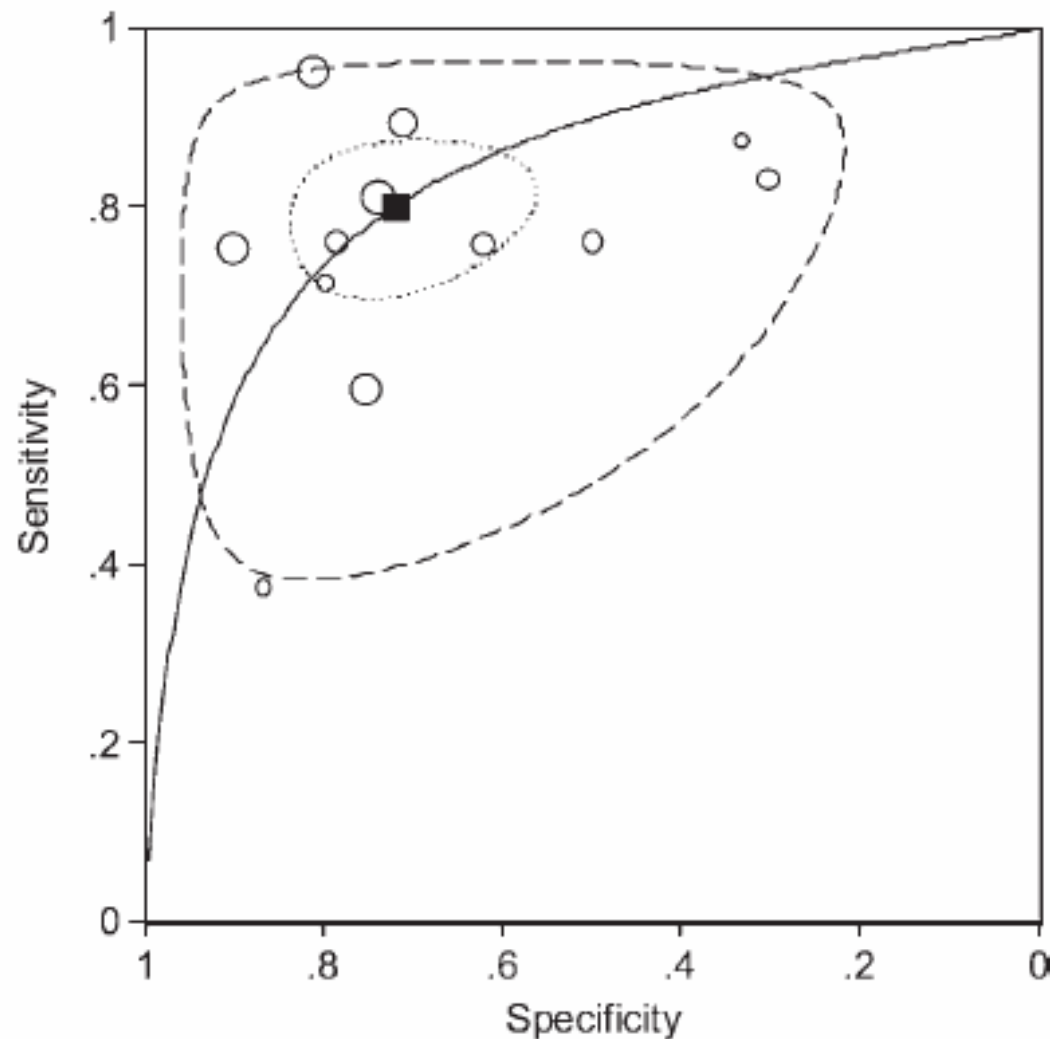


Figure 1 Example of a plot produced by the metandi user-written command for the Stata statistical software package. The open circles show the results of individual studies. The summary ROC curve is shown by the solid line. The solid square shows estimated summary sensitivity and specificity, with the 95% confidence region shown by the dotted line. The dashed line represents the region within which 95% of future studies are predicted to lie

Challenges in diagnostic reviews

- ❑ Identifying published and unpublished studies
 - Publication bias is a major concern
 - Search terms for Dx studies are not well defined
- ❑ Available studies are mostly focused on accuracy
- ❑ Quality assessment is hampered by poor reporting; most diagnostic studies are of poor quality
- ❑ Heterogeneity is almost always found
- ❑ Synthesis of accuracy measures:
 - Lack of meaningful effect measures
 - Need to go beyond SROC

Effect measures in diagnostic studies

Metric	Definition	Advantages	Disadvantages
Accuracy	$(TP + TN)/N$	Intuitive	Depends on prevalence
Sensitivity	TP/N_D	Does not depend on prevalence	Applies only to diseased persons
Specificity	TN/N_W	Does not depend on prevalence	Applies only to nondiseased persons
Positive predictive value	TP/N_P	Clinical relevance	Depends on prevalence
Negative predictive value	TN/N_N	Clinical relevance	Depends on prevalence
Positive likelihood ratio	$(TP/N_D)/(FP/N_W)$	Does not depend on prevalence	Applies only to positive tests
Negative likelihood ratio	$(FN/N_D)/(TN/N_W)$	Does not depend on prevalence	Applies only to negative tests
Odds ratio	$TP \times TN / FN \times FP$	Does not depend on prevalence; combines sensitivity and specificity	Values FP and FN errors equally; not intuitive
Area under curve	Area under ROC curve	Does not depend on prevalence; combines sensitivity and specificity	Lack of clinical interpretation

* FN = false-negative; FP = false-positive; N = sample size; $N_D = TP + FN$; $N_N = TN + FN$; $N_P = TP + FP$; $N_W = TN + FP$; ROC = receiver-operating characteristic; TN = true-negative; TP = true-positive.

Are sensitivity and specificity the most meaningful measures?

Table 1. Hierarchy of Diagnostic Evaluation and the Number of Studies Available for Different Levels of Diagnostic Test in a Technology Assessment of Magnetic Resonance Spectroscopy for Brain Tumors*

Level	Description	Examples of Study Purpose or Measures	Studies Available, <i>n</i>	Patients, <i>n</i>
1	Technical feasibility and optimization	Ability to produce consistent spectra	85	2434
2	Diagnostic accuracy	Sensitivity and specificity	8	461
3	Diagnostic thinking impact	Percentage of times clinicians' subjective assessment of diagnostic probabilities changed after the test	2	32
4	Therapeutic choice impact	Percentage of times therapy planned before MRS changed after the test	2	105
5	Patient outcome impact	Percentage of patients who improved with MRS diagnosis compared with those without MRS (e.g., survival, quality of life)	0	0
6	Societal impact	Cost-effectiveness analysis (e.g., use to detect tumor in asymptomatic population)	0	0

* MRS = magnetic resonance spectroscopy.

Redundancy of Single Diagnostic Test Evaluation

Karel G.M. Moons,^{1,2,3} Gerri-Anne van Es,⁴ Bowine C. Michel,⁵ Harry R. Büller,⁶
J. Dik F. Habbema,³ and Diederick E. Grobbee¹

Moons et al. Epidemiology 1999

Diagnostic research

Diagnostic studies as multivariable,
prediction research

K G M Moons, D E Grobbee

Patient outcomes in diagnostic research

Moons et al. JECH 2002

Opinion

Test Research versus Diagnostic Research

Moons et al. Clin Chem 2004

Diagnostic trials lack methodologic rigor

Diagnostic studies in 4 general medical journals

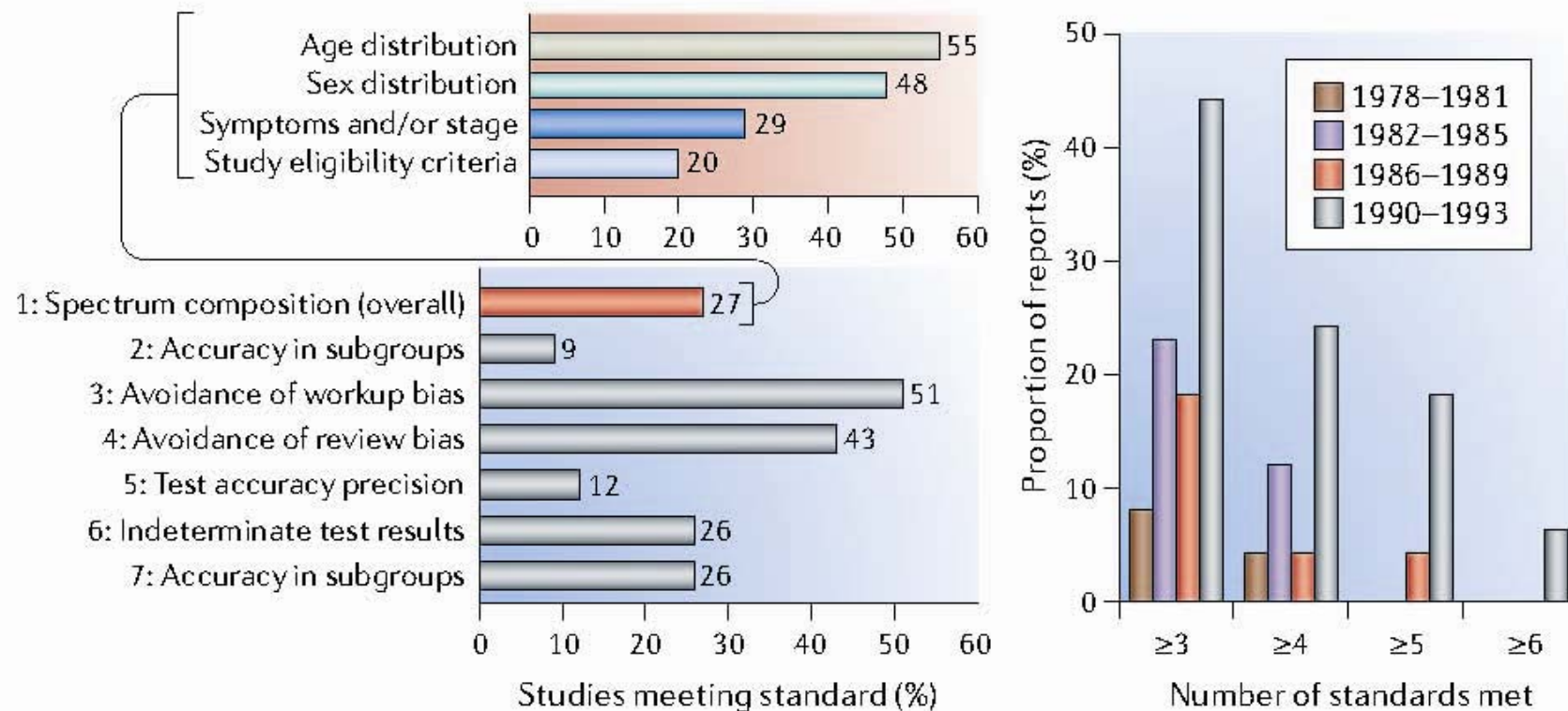


Figure 4 | **Proportion of diagnostic evaluations meeting accepted standards.** The seven standards are shown on the left. The data are taken from REF. 10.

Lack of rigor: example from TB literature

12 meta-analysis with
over 500 diagnostic
studies

- 65% used prospective design
- 33% used consecutive or random sampling
- 72% used a cross-sectional design, a third used case-control
- Blinding was reported in 34% of the trials.

Pai M, et al. Exp Rev Mol Diagn 2006.

Table 2. Methodological quality of studies on tuberculosis diagnostics in recently published meta-analyses.

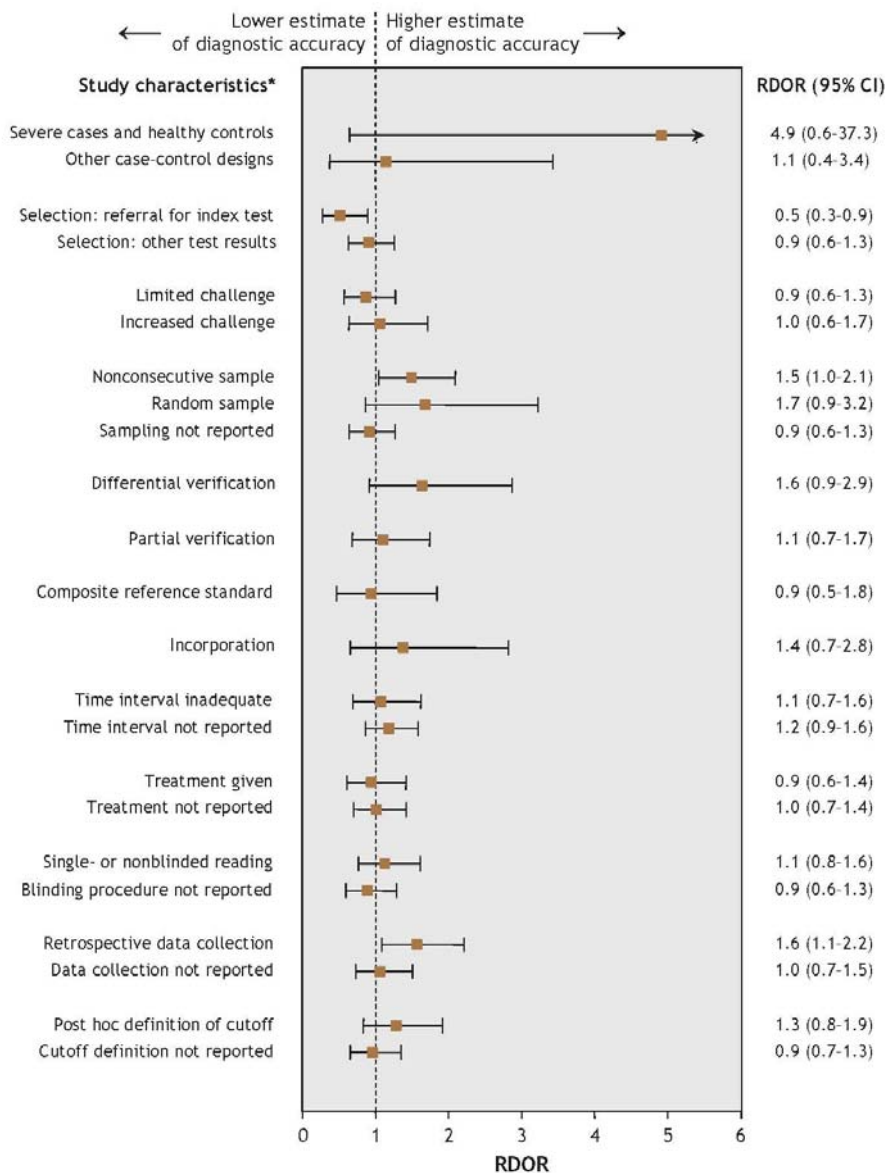
Meta-analysis	No. of studies	Diagnostic test	Average size of each study	Prospective data collection (%)	Consecutive or random sampling of subjects (%)	Cross-sectional design (%)	Blinded interpretation of test results* (%)	Complete verification of index test results† (%)	Ref.
Sarmiento et al. (2003)	16	PCR on respiratory specimens for smear-negative pulmonary TB	NR	50	NR	NR	63	100	[12]
Goto et al. (2003)	40	ADA for TB pleural effusion	137	NR	NR	NR	0	NR	[13]
Pai et al. (2003)	49	NAT for TB meningitis	42	61	49	61	59	94	[14]
Graco et al. (2003)	44	ADA and IFN- γ tests for TB pleural effusion	135	NR	NR	NR	9	NR	[15]
Pai et al. (2004)	40	NAT for TB pleural effusion	60	63	53	70	55	100	[16]
Flores et al. (2005)	84	In-house PCR for pulmonary TB	149	NR	NR	71	34	NR	[17]
Kalantri et al. (2005)	13	Phage amplification tests for pulmonary TB	448	NR	NR	85	23	100	[18]
Pai et al. (2005)	21	Phage-based tests for rifampin resistance	85	NR	38	NR	57	100	[19]
Morgan et al. (2005)	15	Line probe assay for rifampin resistance	91	NR	0	NR	13	100	[20]
Graco et al. (2006)	63	Commercial NAT for pulmonary TB	410	16	32	NR	16	NR	[21]
Steingart et al. (2006)	45	Fluorescence versus conventional sputum smear microscopy for pulmonary TB	493	100	36	NR	49	NR	[22]
Steingart et al. (2006)	83	Direct versus concentrated sputum smear microscopy for pulmonary TB	256	100	21	NR	31	NR	[23]

*At least single blind. †By reference standard.

ADA: Adenosine deaminase; IFN: Interferon; NAT: Nucleic acid amplification test; NR: Not reported; TB: Tuberculosis.

Study quality affect study results

487 diagnostic studies



*See Appendix 2 for descriptions of the study characteristics.

Fig. 2: Effects of study design characteristics on estimates of diagnostic accuracy. RDOR = relative diagnostic odds ratio (adjusted RDORs were estimated in a multivariable random-effects meta-epidemiologic regression model).

Study quality vs. study reporting

Data from a meta-analysis of NAAT for TB meningitis (Pai et al. *Lancet Infect Dis* 2003)

Characteristic	Before contact	After contact % [N = 49]
Blinding	% [N = 49]	
Double blind	12	35
Single blind	14	24
Unblinded	0	10
Not reported	74	31
Sampling		
Consecutive/random	18	49
Not consecutive/random	6	20
Not reported	76	31
Data collection		
Prospective	51	61
Retrospective	0	4
Both	2	10
Not reported	47	25

A systematic review finds that diagnostic reviews fail to incorporate quality despite available tools

Penny Whiting^{a,d,*}, Anne W.S. Rutjes^b, Jacqueline Dinnes^c,
Johannes B. Reitsma^b, Patrick M.M. Bossuyt^b, Jos Kleijnen^a

^a*Centre for Reviews and Dissemination, University of York, United Kingdom*

^b*Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, The Netherlands*

^c*Wessex Institute for Health Research and Development, University of Southampton, United Kingdom*

^d*MRC Health Services Research Collaboration, University of Bristol, United Kingdom*

Abstract

Background and Objective: To review existing quality assessment tools for diagnostic accuracy studies and to examine to what extent quality was assessed and incorporated in diagnostic systematic reviews.

Methods: Electronic databases were searched for tools to assess the quality of studies of diagnostic accuracy or guides for conducting, reporting or interpreting such studies. The Database of Abstracts of Reviews of Effects (DARE; 1995–2001) was used to identify systematic reviews of diagnostic studies to examine the practice of quality assessment of primary studies.

Results: Ninety-one quality assessment tools were identified. Only two provided details of tool development, and only a small proportion provided any indication of the aspects of quality they aimed to assess. None of the tools had been systematically evaluated. We identified 114 systematic reviews, of which 58 (51%) had performed an explicit quality assessment and were further examined. The majority of reviews used more than one method of incorporating quality.

Conclusion: Most tools to assess the quality of diagnostic accuracy studies do not start from a well-defined definition of quality. None has been systematically evaluated. The majority of existing systematic reviews fail to take differences in quality into account. Reviewers should consider quality as a possible source of heterogeneity. © 2005 Elsevier Inc. All rights reserved.