# Systematic reviews of diagnostic test accuracy

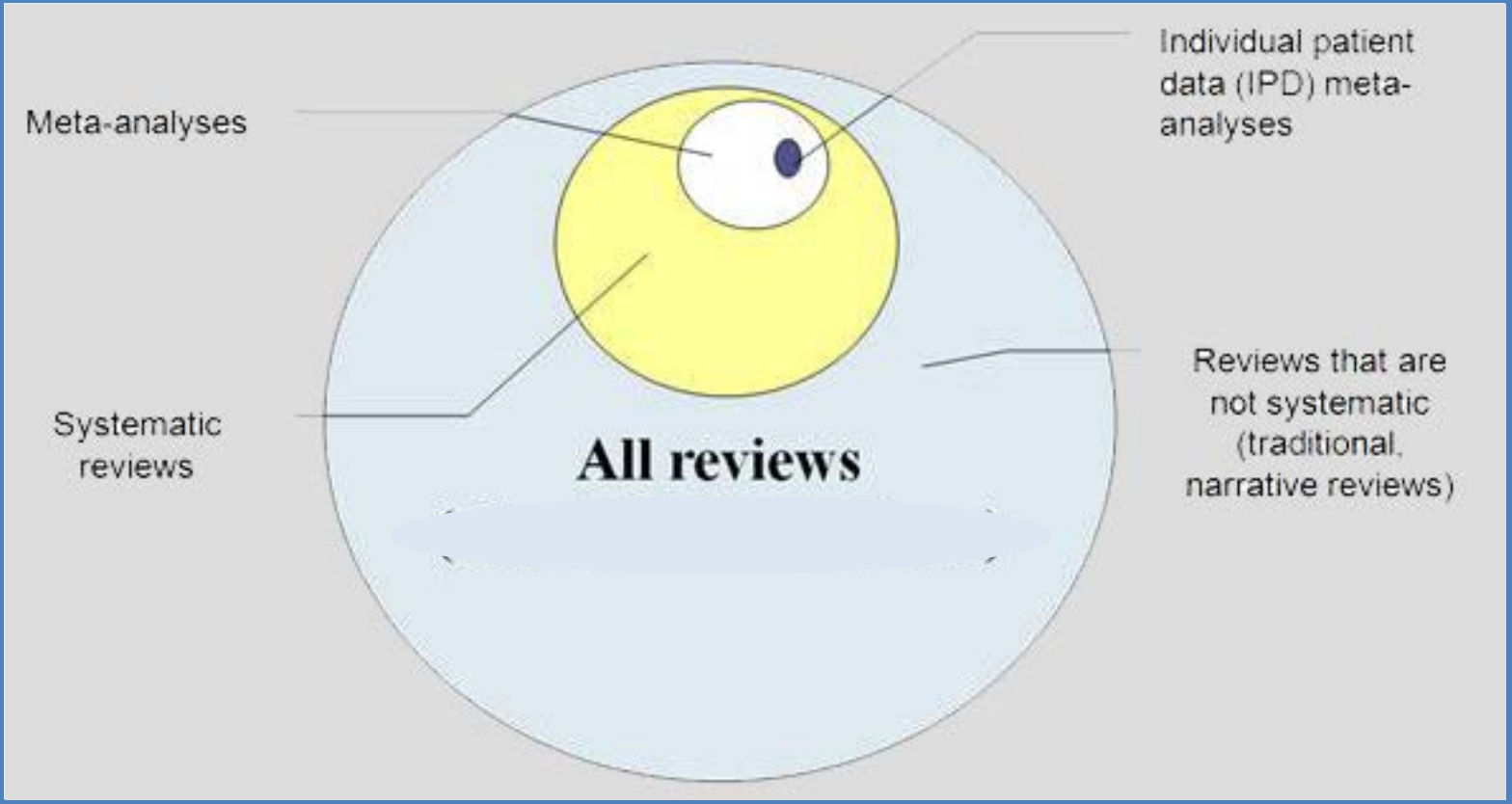
Karen R Steingart, MD, MPH
karenst@uw.edu
Montreal, July 2012

# Conflicts of interest

- I am an Editor with the Cochrane Infectious Diseases Group
- I am a member of the GRADE Working Group
- I have no financial interests to declare

# Overview

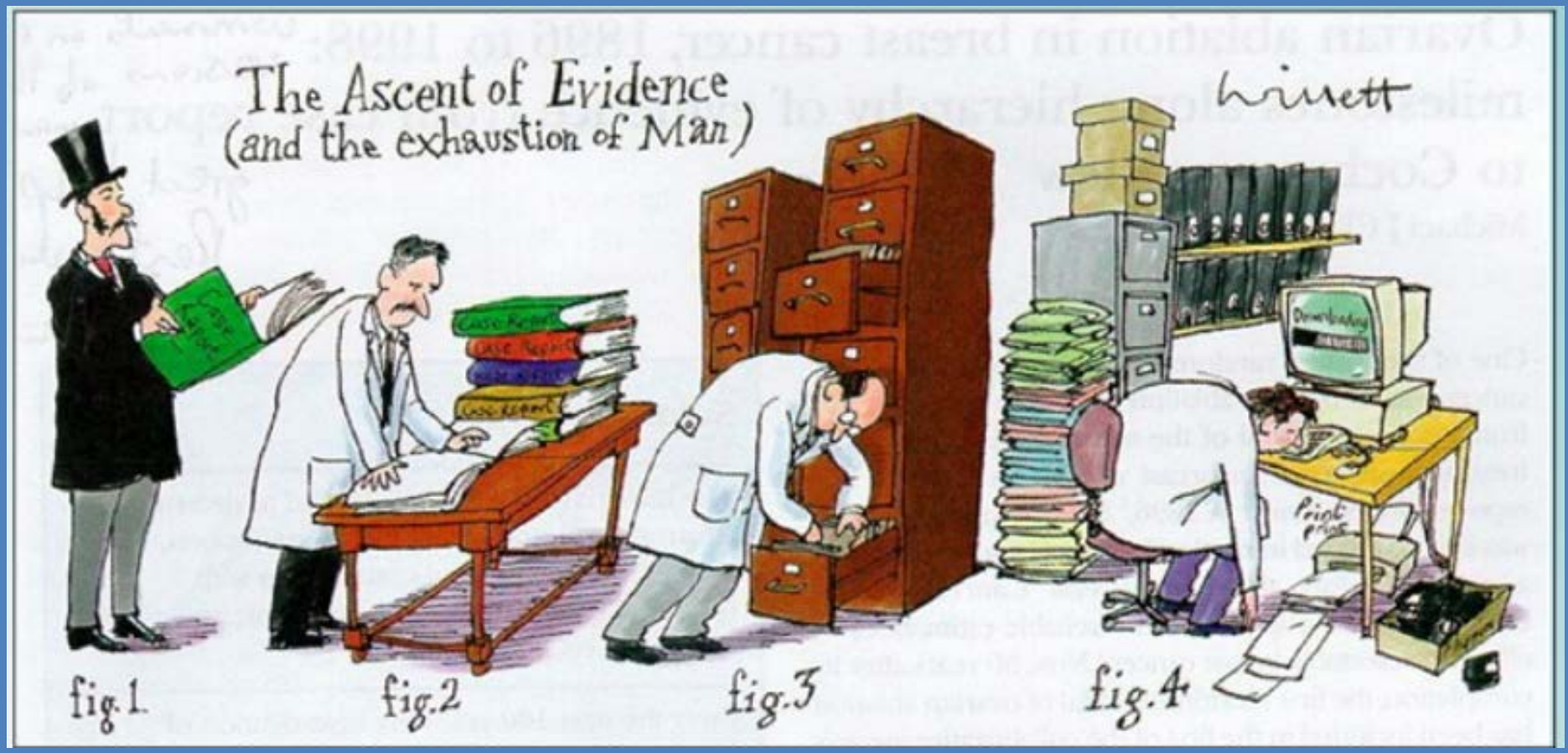
- Describe key steps in a systematic review/ meta-analysis of diagnostic test accuracy studies
- Demonstrate standard methods of meta-analysis of data from diagnostic studies
- Identify key references and tools for performing meta-analysis of diagnostic studies




A systematic review starts with a clearly formulated question and 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

Egg slide adapted from Madhu Pai

# Why systematic reviews?



# Why systematic reviews?

- Scientific summary of all available evidence
- Transparent and reproducible process
- Minimizes bias
- Studies can be formally compared to establish generalizability and consistency
- Heterogeneity can be identified & investigated
- Quantitative part (meta-analyses) may increase the precision of the overall result

# Why systematic reviews of diagnostic test accuracy?

- Increasing number of available tests
- Patients ask more questions
- Higher costs and burden on society
- Difficult to choose among tests
- New hypotheses can be generated about particular subgroups

# US Institute of Medicine, standards for systematic reviews

http://www.iom.edu/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx

REPORT BRIEF MARCH 2011

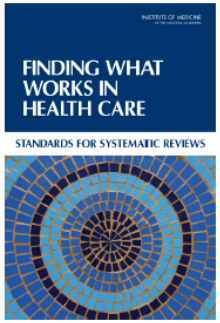
INSTITUTE OF MEDICINE
OF THE NATIONAL ACADEMIES

Advising the nation • Improving health

For more information visit www.iom.edu/srstandards

## Finding What Works in Health Care

### Standards for Systematic Reviews




**Healthcare decision makers** in search of reliable information comparing health interventions increasingly turn to systematic reviews for the best summary of the evidence. Systematic reviews identify, select, assess, and synthesize the findings of similar but separate studies and can help clarify what is known and not known about the potential benefits and harms of drugs, devices, and other healthcare services. Systematic reviews can be helpful for clinicians who want to integrate research findings into their daily practices, for patients to make well-informed choices about their own care, and for professional medical societies and other organizations that develop clinical practice guidelines.

In the *Medicare Improvement for Patients and Providers Act of 2008*, Congress directed the Institute of Medicine (IOM) to develop standards for conducting systematic reviews and to develop standards for clinical practice guidelines, which are evidence-based recommendations for clinicians to use when treating patients. The IOM formed two distinct committees to respond to this charge, and each committee assessed the relevant evidence and considered expert guidance to develop the standards. This report, *Finding What Works in Health Care: Standards for Systematic Reviews*, recommends standards for systematic reviews of the comparative effectiveness of medical or surgical interventions (see the insert for a list of the standards).

**Systematic reviews ... can help clarify what is known and not known about the potential benefits and harms of drugs, devices, and other healthcare services.**

# Expectations

- *1,000 to 2,000 person hours to perform a review*
  Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. JAMA 1999;282:634–635.

- *21 standards and 82 elements of performance*
  Institute of Medicine of the National Academies. Finding What Works in Healthcare. Standards for Systematic Reviews. National Academies Press: Washington, DC, 2011.

- *~ 100 elements of performance*
  Cochrane performance expectations

# Typical systematic reviewer

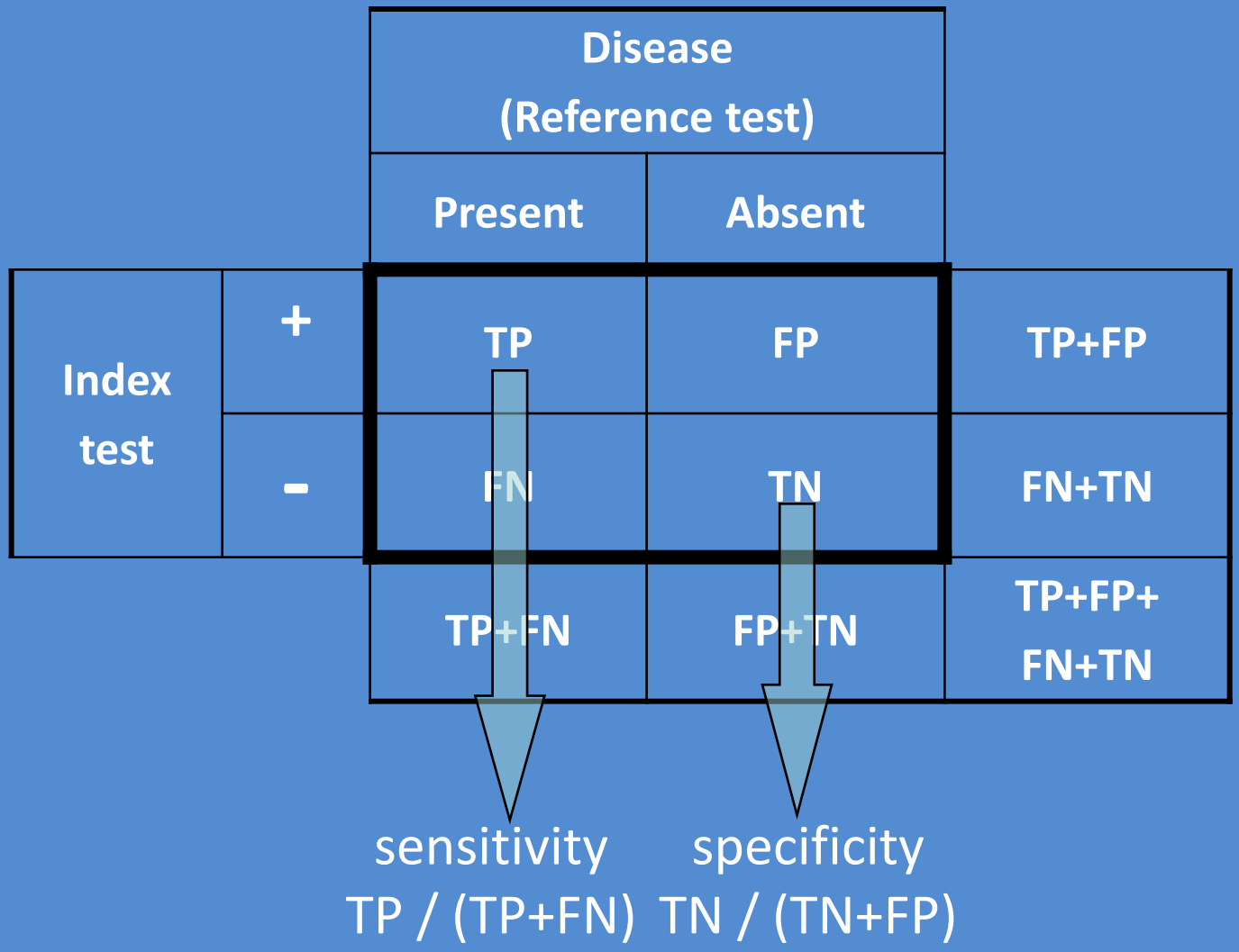
# What is diagnostic test accuracy?

- Diagnosis
  - Does this patient have this disease at this point in time?
- Test accuracy
  - What proportion of those with the disease does the test detect? (sensitivity)
  - What proportion of those without the disease get negative test results? (specificity)
  - Requires 2×2 table of test vs reference standard

# Diagnostic accuracy

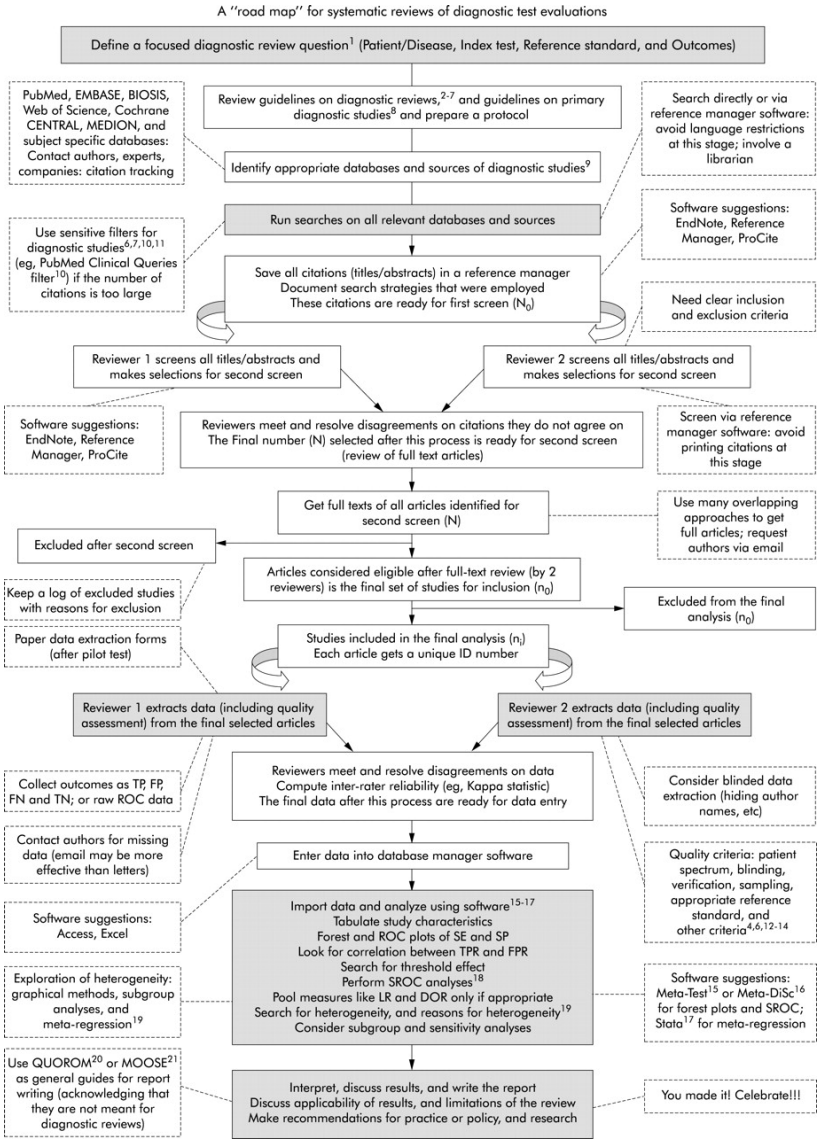
- Agreement between results of the index test and reference standard
- Many measures of agreement
- Focus on pairs of sensitivity & specificity

# 2x2 Table – sensitivity and specificity




P Bossuyt http://srdta.cochrane.org/presentations

# Road map for diagnostic accuracy reviews




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

# Key steps in a diagnostic test accuracy review

1. Framing focused questions
2. Searching for studies
3. Assessing study quality
4. Analyzing the data; undertaking meta-analyses
5. Drawing robust conclusions and informative presentation of results

# 1. Framing focused questions

# Begin with a well-framed question, PICO

The objectives of the review

**P**opulation

**I**ntervention

**C**omparison

**O**utcome

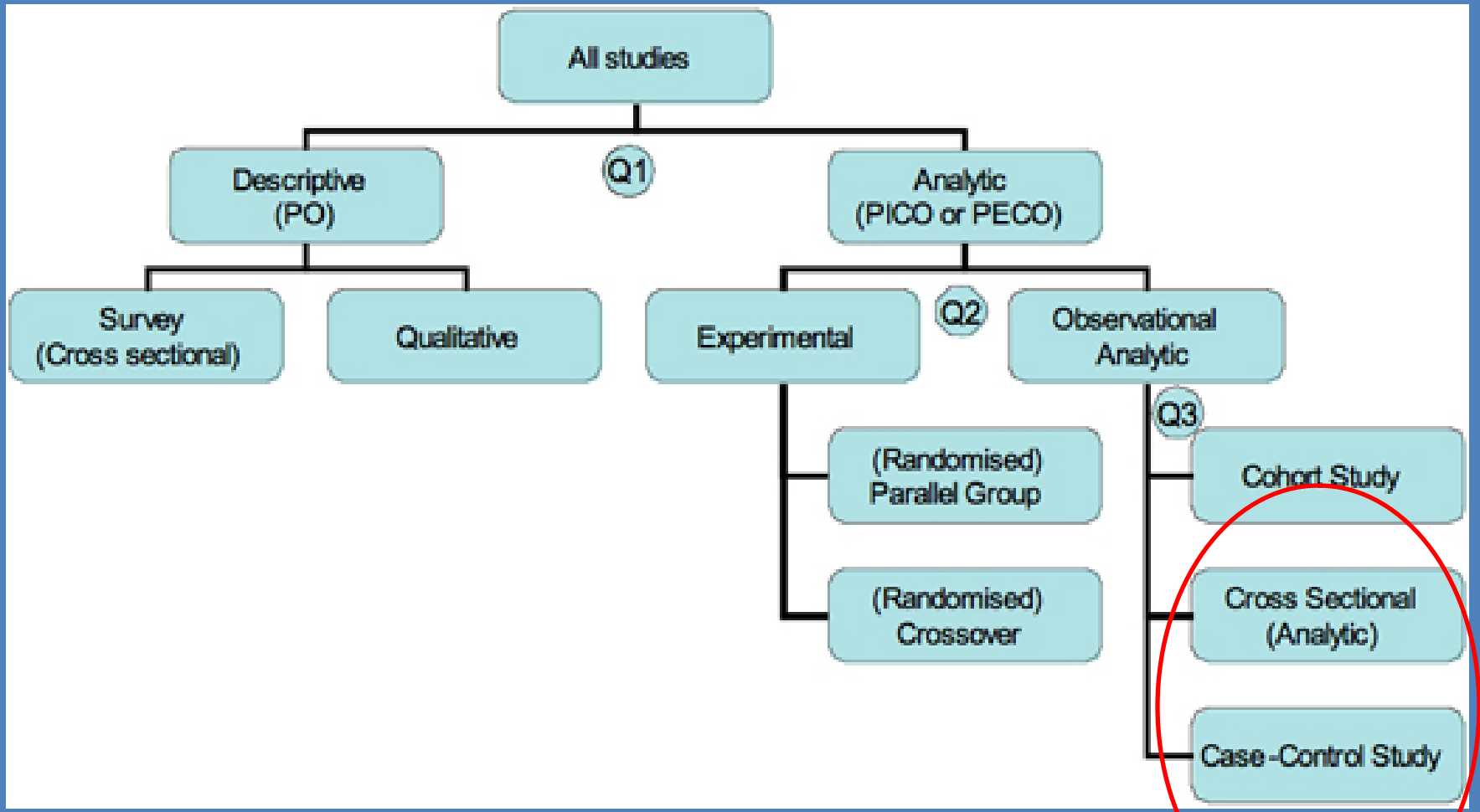
+ Study design

+ Purpose of the test/strategy

+ Reference standard

Richardson et al. ACP Journal Club 1995;A-12

# Overview of the study design tree




http://www.cebm.net/index.aspx?o=1043

# What is the purpose of the test?

- Triage
  - minimize use of invasive or expensive test
- Add-on
  - improve diagnosis beyond what is already done
- Replacement
  - replace test that is harmful or costly

Bossuyt et al. BMJ 2006

# What is the reference standard?

- …is usually the best test currently available
- …used to determine presence/absence of disease

- Indicators of diagnostic accuracy are calculated by comparing results of the index test with those of the reference standard
- Any discrepancy is assumed to arise from error in the index test
- The assumption of 100% accuracy for the reference standard rarely holds true in practice

# PICO or PPPICPTR for systematic review of diagnostic test accuracy?

- **P**atients, **P**resentation, **P**rior tests
- **I**ndex test, **C**omparator tests
- **P**urpose:  comparative question, role of test
- **T**arget condition, **R**eference standard

# Title, objective…

- [Index test] vs [comparator] for [target condition] in [participant description]

  - The GenoType® MTBDRsI test for detecting resistance to second-line antituberculosis drugs and XDR-TB

- Objective: To obtain summary estimates of the diagnostic accuracy of GenoType® MTBDRsI for the detection of resistance to each fluoroquinolone and injectable drug and ethambutol in patient specimens confirmed as TB

- Purpose: Replacement test or an add-on test
- Target condition: Resistance to any of the flouroquinolones, aminoglycosides and ethambutol
- Reference standard: Conventional drug susceptibility testing by solid or liquid culture+/- DNA sequencing

*G Theron and J Peter, Cochrane registered title 2012

# 2. Searching for studies

# Recommended standards for finding and assessing individual studies

**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

3.1.4 Search bibliographic databases

3.1.5 Search citation indexes

3.1.6 Search literature cited by eligible studies

3.1.7 Update the search at intervals appropriate to the pace of generation of new information for the research question being addressed

3.1.8 Search subject-specific databases if other databases are unlikely to provide all relevant evidence

3.1.9 Search regional bibliographic databases if other databases are unlik

provide all relevant evidence

# Sources of studies for diagnostic accuracy reviews

- MEDLINE, EMBASE, the Cochrane Register of Diagnostic Test Accuracy Studies (under development)
- Search related diagnostic test accuracy reviews (for example HTA database, DARE etc)
- Check references of relevant studies/reviews
- Use a highly sensitive (broad) search strategy
- Use a wide variety of search terms, both text words and database subject headings (MeSH terms)
- Routine use of search filters should generally be avoided

Bossuyt PM, Leeflang MM. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4 [updated September* 2008]. The Cochrane Collaboration, 2008

# Database searching, example

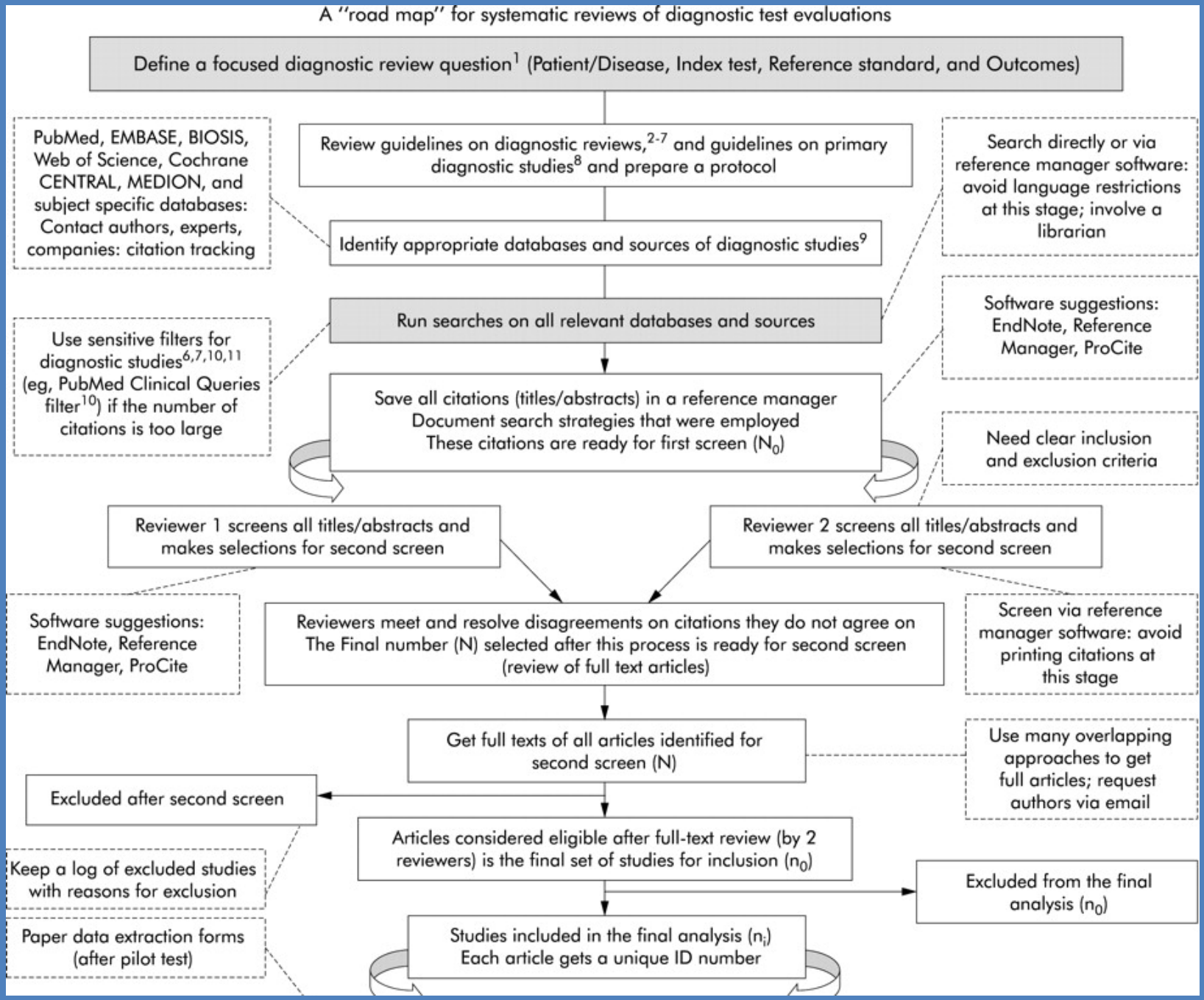
- We searched the following databases on 25 September 2011 using the search terms and strategy described in Appendix 1: Cochrane Infectious Diseases Group Specialized Register; MEDLINE; EMBASE; ISI Web of Knowledge; MEDION; LILACS; BIOSIS; and SCOPUS. We will also search the metaRegister of Controlled Trials (mRCT) and the search portal of the WHO International Clinical Trials Registry Platform, to identify ongoing trials. We updated the search in MEDLINE on 15 December 2011. Searches were limited to 2007 onward and performed without language restriction.

H Sohn. Cochrane Database of Systematic Reviews , Issue . Art. No.: CD009593. DOI: 10.1002/14651858.CD009593 .

# Search terms, example

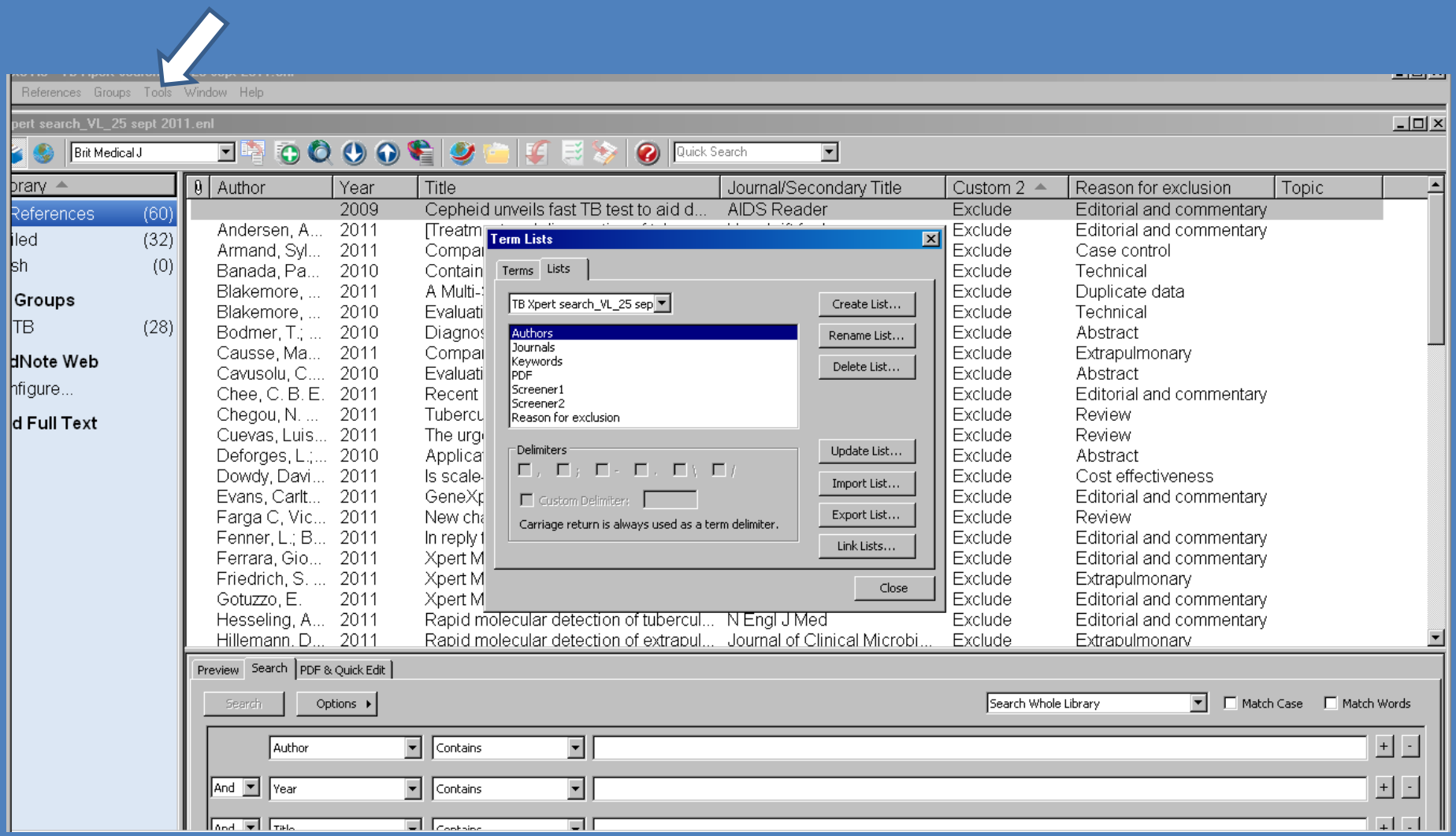
- The following search terms were used for MEDLINE searching (through PubMed): .Xpert*[Text Word] OR GeneXpert**[Text Word] OR Cepheid*[Text Word] OR near* patient in title or abstract AND Tuberculosis/ OR exp Tuberculosis, Pulmonary/ OR exp Tuberculosis, Multidrug-Resistant/ OR Mycobacterium tuberculosis/ OR TB[Text Word] OR tuberculosis[Text Word]

H Sohn. Cochrane Database of Systematic Reviews , Issue . Art. No.: CD009593. DOI: 10.1002/14651858.CD009593 .




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

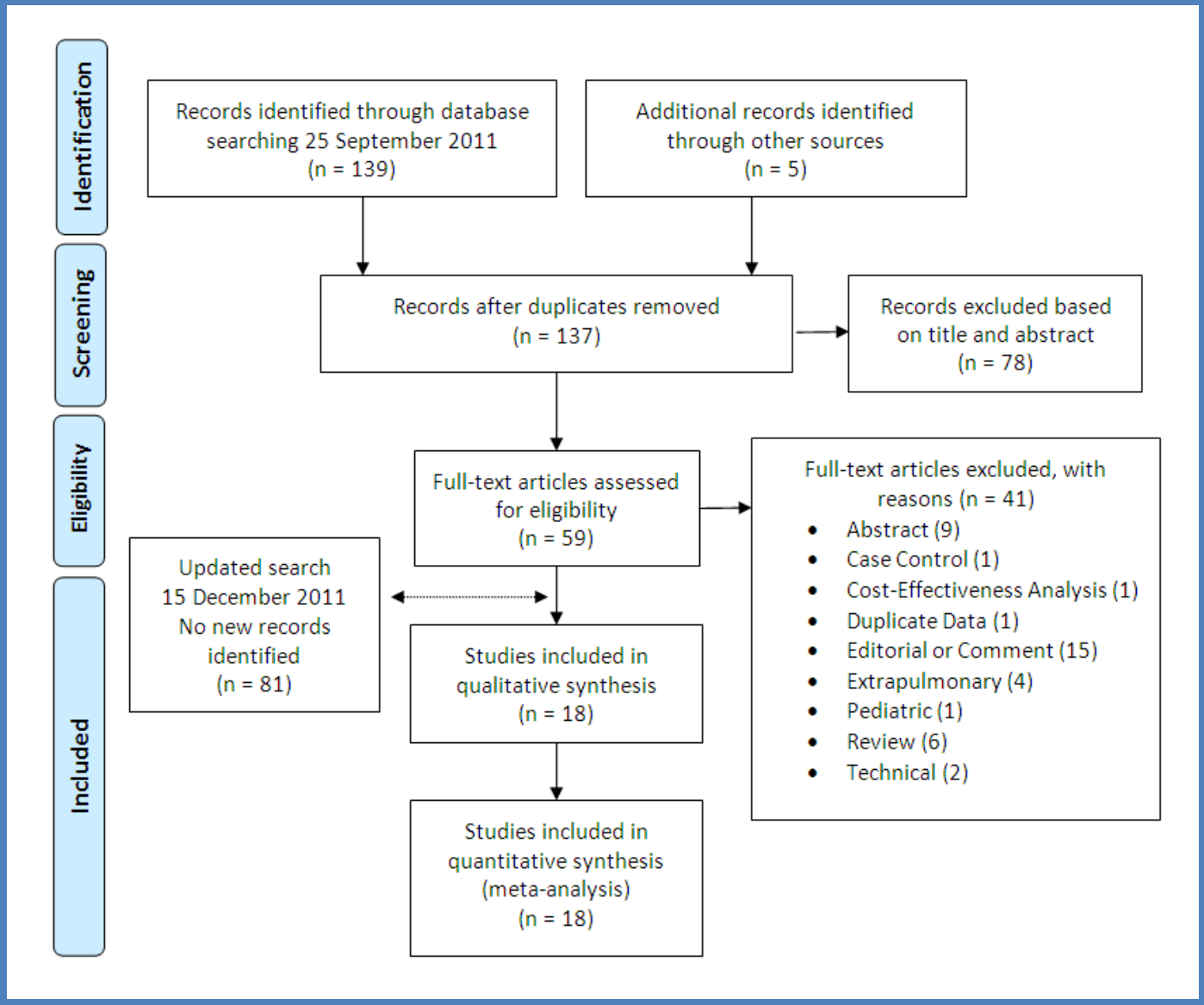
# Record judgments and ‘reason for exclusion’, example using EndNote



# How to create term lists linked to Screener and Reason for Exclusion

1) In EndNote go to Tools > ‘Define Term Lists’
2) Click ‘Create List’ and type ‘Screener1’
3) Make sure that the Screener1 list is highlighted and click on the ‘Terms’ tab
4) Click on ‘New Term’ to create the 3 terms in sequence: Include, Exclude, Further review
5) Click on the ‘Link Lists’ button
6) Find the custom field you are using as your Screener1 field and link it to the Screener1 list and click "OK“
7) Repeat these steps for ‘Screener2’ and the "Reason for Exclusion" list
8) As you begin to type in those fields, EndNote will auto-complete the rest of the text

# Xpert MTB/RIF Test



*The medical literature can be compared to a jungle. It is fast growing, full of deadwood, sprinkled with hidden treasure and infested with spiders and snakes. Morgan. Can Med Assoc J, 134,Jan 15, 1986*

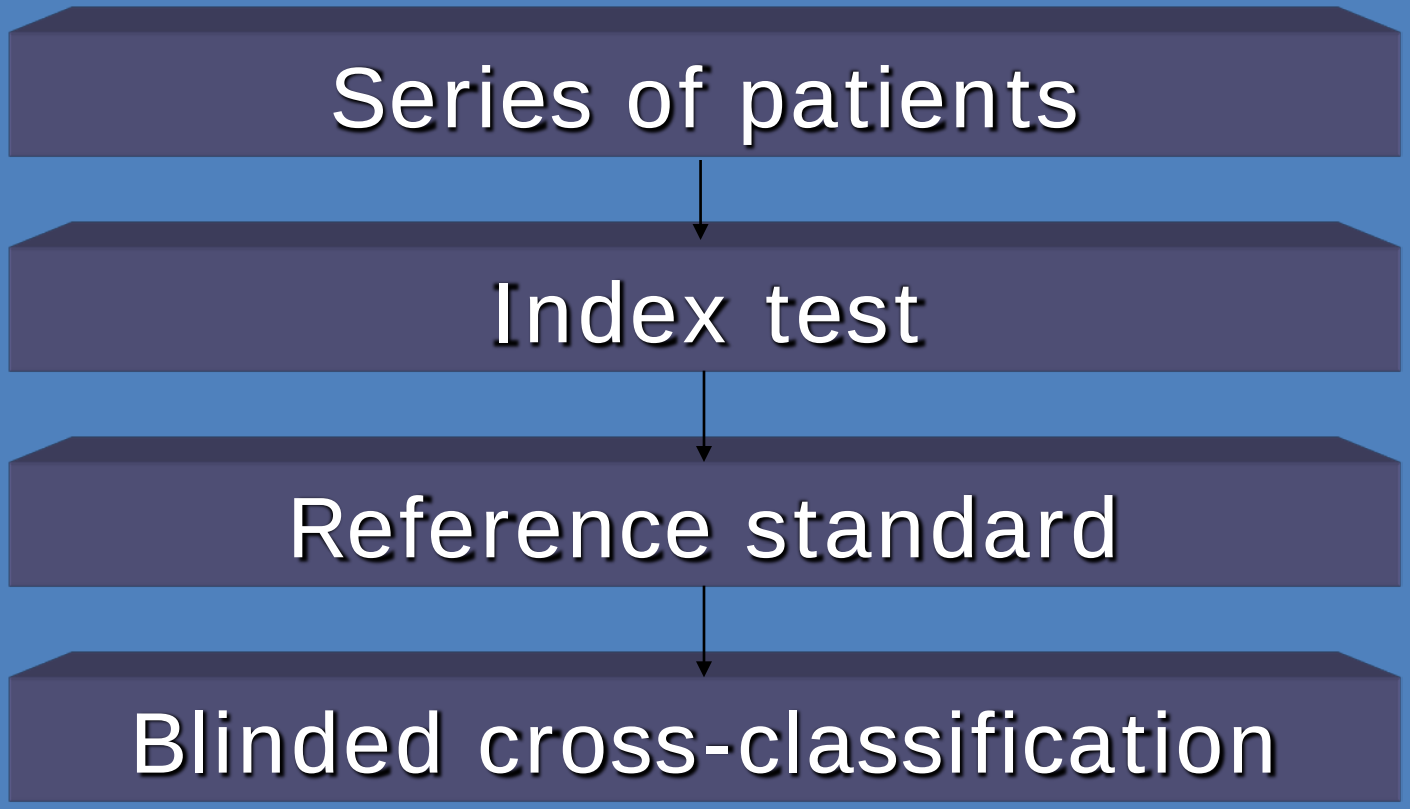
# 3. Assessing study quality

# Quality definition in diagnostic reviews

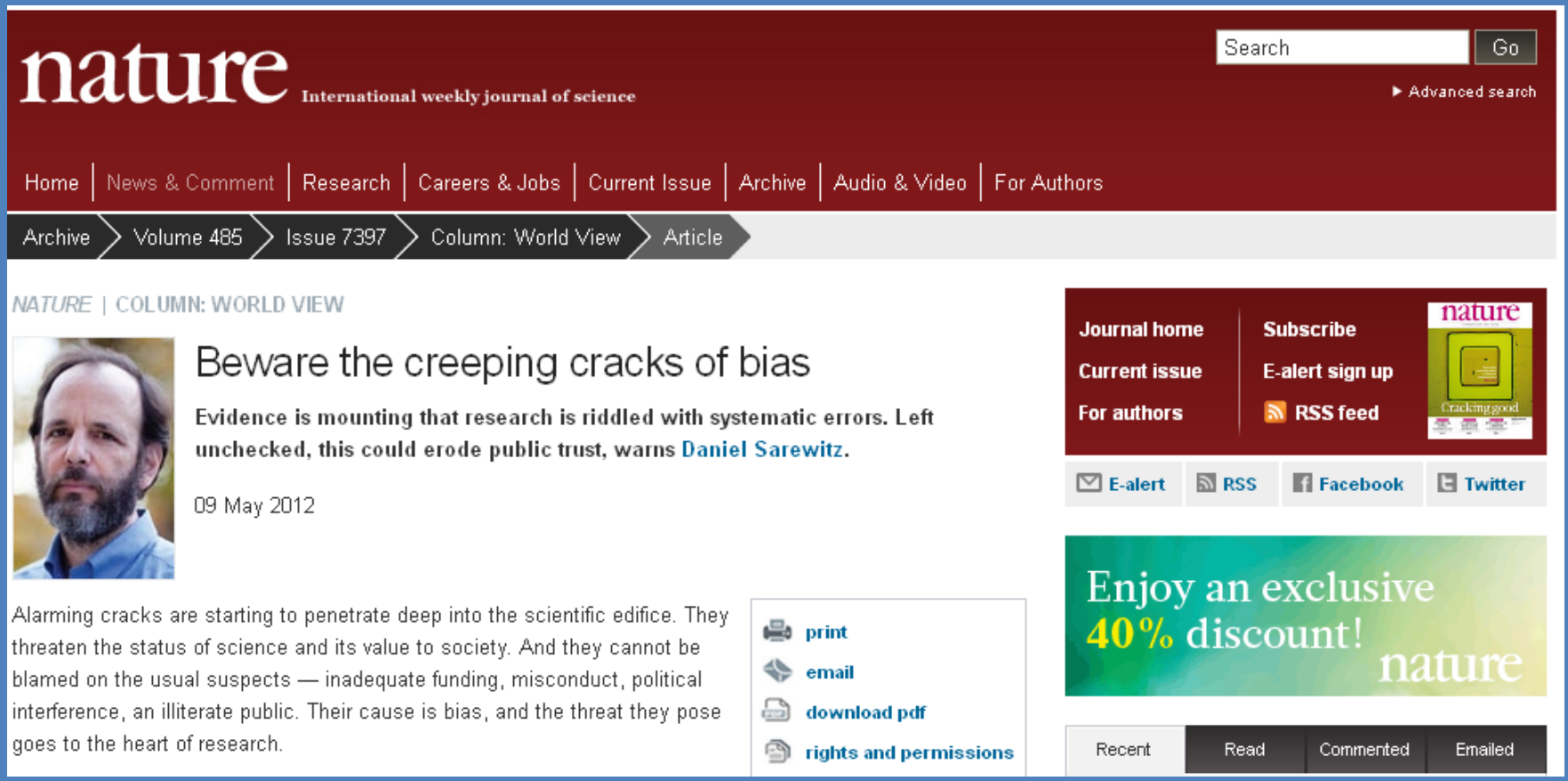
- Methodological quality of a study: the degree to which the design and conduct of a study match the study objectives

- Two components: risk of bias and applicability

# Diagnostic Accuracy Study: Ideal Design




P Bossuyt http://srdta.cochrane.org/presentations




***Nothing will corrode public trust more than a creeping awareness that scientists are unable to live up to the standards that they have set for themselves. Useful steps to deal with this threat may range from reducing the hype from universities and journals about specific projects, to strengthening collaborations between those involved in fundamental research and those who will put the results to use in the real world.***

# What is bias?

Bias is **any process** at **any stage** of inference tending to produce **results** that **differ systematically** from the **true** values. Murphy EA. *The logic of medicine,* 1976)

Bias is **any trend** in the **collection, analysis, interpretation, publication or review** of data that can lead to **conclusions** that are **systematically different** from the **truth** (Last J. *A dictionary of epidemiology*, 2001)

# A Catalogue of Bias, M. Tevfik Dorak (adapted from David Sackett)
# http://www.dorak.info/epi/bc.html

**Literature Review**

- Foreign language exclusion bias
- Literature search bias
- One-sided reference bias
- Rhetoric bias

**Study Design**

- Selection bias
- Sampling frame bias
  - Berkson (admission rate) bias
  - Centripetal bias
  - Diagnostic access bias
  - Diagnostic purity bias
  - Hospital access bias
  - Migrator bias
  - Prevalence-incidence (Neyman / selective survival; attrition) bias
  - Telephone sampling bias
- Nonrandom sampling bias
  - Autopsy series bias
  - Detection bias
  - Diagnostic work-up bias
  - Door-to-door solicitation bias
  - Previous opinion bias
  - Referral filter bias
  - Sampling bias
  - Self-selection bias
  - Unmasking bias
- Noncoverage bias
  - Early-comer bias
  - Illegal immigrant bias
  - Loss to follow-up (attrition) bias
  - Response bias
  - Withdrawal bias
- Noncomparability bias
  - Ecological (aggregation) bias
  - Healthy worker effect (HWE)
  - Lead-time bias
  - Length bias
  - Membership bias
  - Mimicry bias
  - Nonsimultaneous comparison bias
  - Sample size bias

**Study Execution**

- Bogus control bias
- Contamination bias
- Compliance bias

**Data Collection**

- Instrument bias
  - Case definition bias
  - Diagnostic vogue bias
  - Forced choice bias
  - Framing bias
  - Insensitive measure bias
  - Juxtaposed scale bias
  - Laboratory data bias
  - Questionnaire bias
  - Scale format bias
  - Sensitive question bias
  - Stage bias
  - Unacceptability bias
  - Underlying/contributing cause of death bias
  - Voluntary reporting bias
- Data source bias
  - Competing death bias
  - Family history bias
  - Hospital discharge bias
  - Spatial bias
- Observer bias
  - Diagnostic suspicion bias
  - Exposure suspicion bias
  - Expectation bias
  - Interviewer bias
  - Therapeutic personality bias
- Subject bias
  - Apprehension bias
  - Attention bias (Hawthorne effect)
  - Culture bias
  - End-aversion bias (end-of-scale or central tendency bias)
  - Faking bad bias
  - Faking good bias
  - Family information bias
  - Interview setting bias
  - Obsequiousness bias
  - Positive satisfaction bias
  - Proxy respondent bias
- Recall bias
  - Reporting bias
  - Response fatigue bias
  - Unacceptable disease bias
  - Unacceptable exposure bias
  - Underlying cause (rumination bias)
  - Yes-saying bias
- Data handling bias
  - Data capture error
  - Data entry bias
  - Data merging error
  - Digit preference bias
  - Record linkage bias

**Analysis**

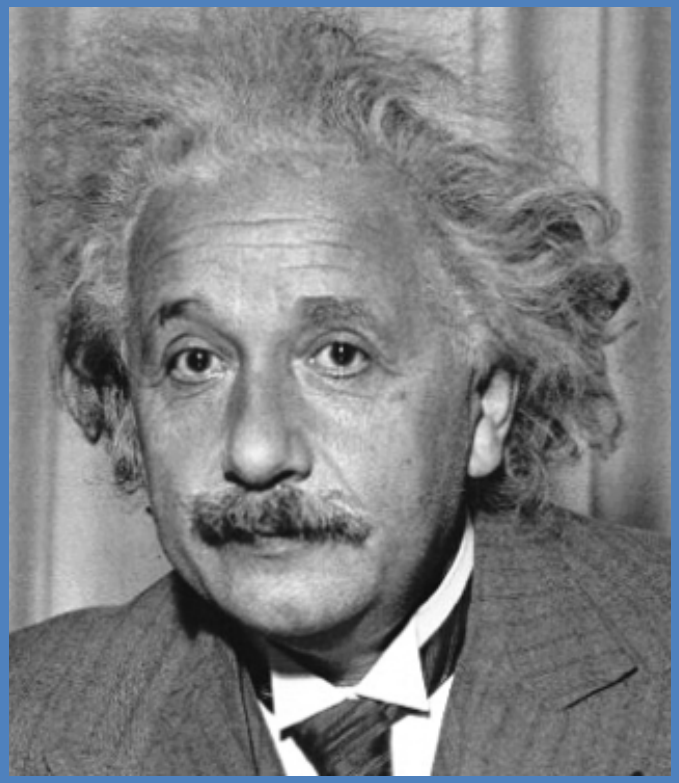
- Confounding bias
  - Latency bias
  - Multiple exposure bias
  - Nonrandom sampling bias
  - Standard population bias
  - Spectrum bias
- Analysis strategy bias
  - Distribution assumption bias
  - Enquiry unit bias
  - Estimator bias
  - Missing data handling bias
  - Outlier handling bias
  - Overmatching bias
  - Scale degradation bias
- Post hoc analysis bias
  - Data dredging bias
  - Post hoc significance bias
  - Repeated peeks bias

**Interpretation of Results**

- Assumption bias
- Cognitive dissonance bias
- Correlation bias
- Generalization bias
- Magnitude bias
- Significance bias
- Underexhaustion bias

**Publication**

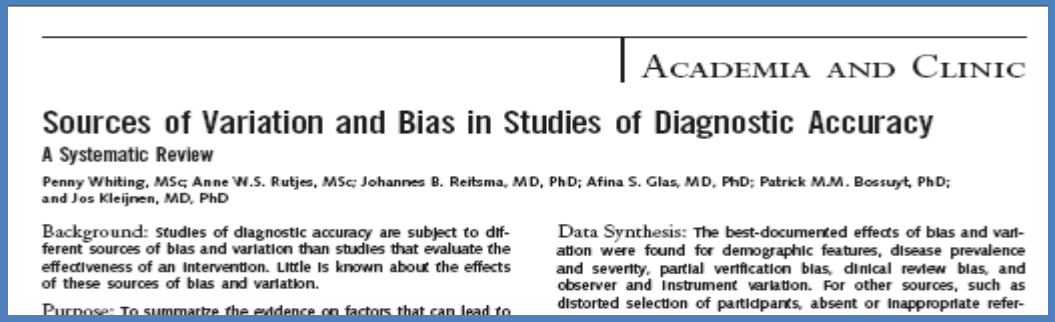
- All's well literature bias
- Positive result bias
- Hot topic bias



*“Everything should be made as simple as possible but not simpler.”*

# Sources of bias in diagnostic studies: 3 key issues

- Inclusion of right spectrum of patients
- Verification of patients
  - choice of reference standard
  - complete verification
- Independent assessment of index test and reference standard (blinding)



ACADEMIA AND CLINIC

Sources of Variation and Bias in Studies of Diagnostic Accuracy

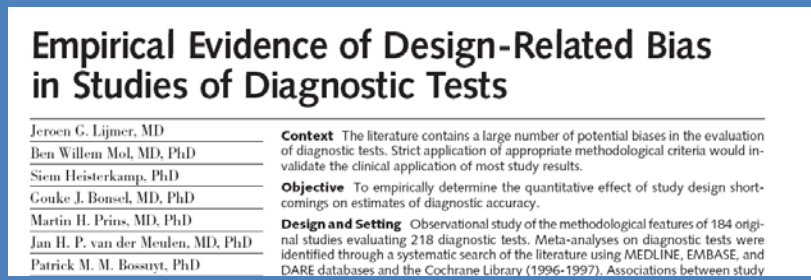
A Systematic Review

Penny Whiting, MSc; Anne W.S. Rutjes, MSc; Johannes B. Reitsma, MD, PhD; Afina 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 summarize the evidence on factors that can lead to

Data Synthesis: The best-documented effects of bias and variation were found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation. For other sources, such as distorted selection of participants, absent or inappropriate refer-



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

**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

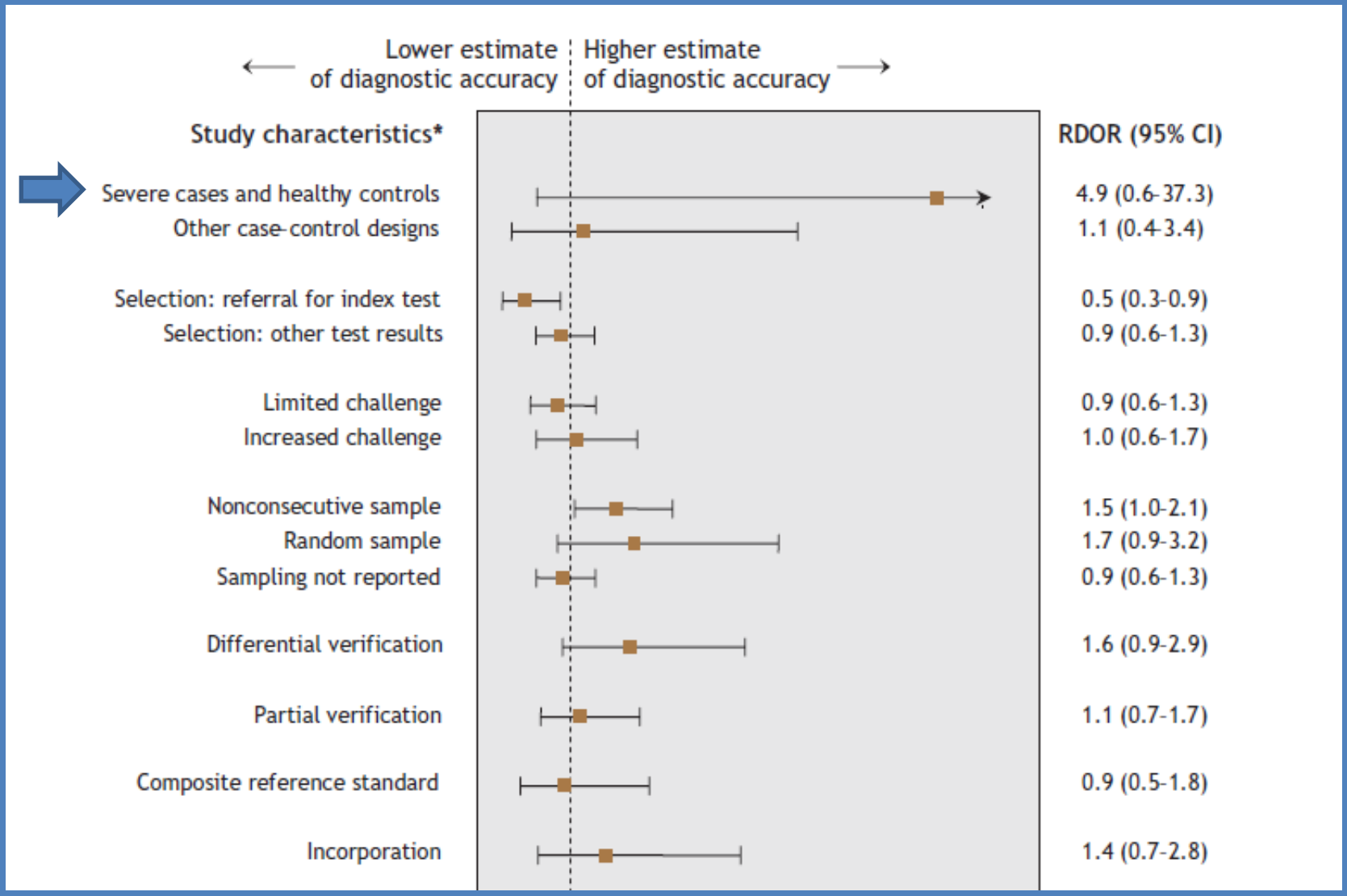
RESEARCH

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*.

# Effects of study design, A Rutges CMAJ 2006



# Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis

Madhukar Pai, Laura L Flores, Nitika Pai, Alan Hubbard, Lee W Riley, and John M Colford Jr

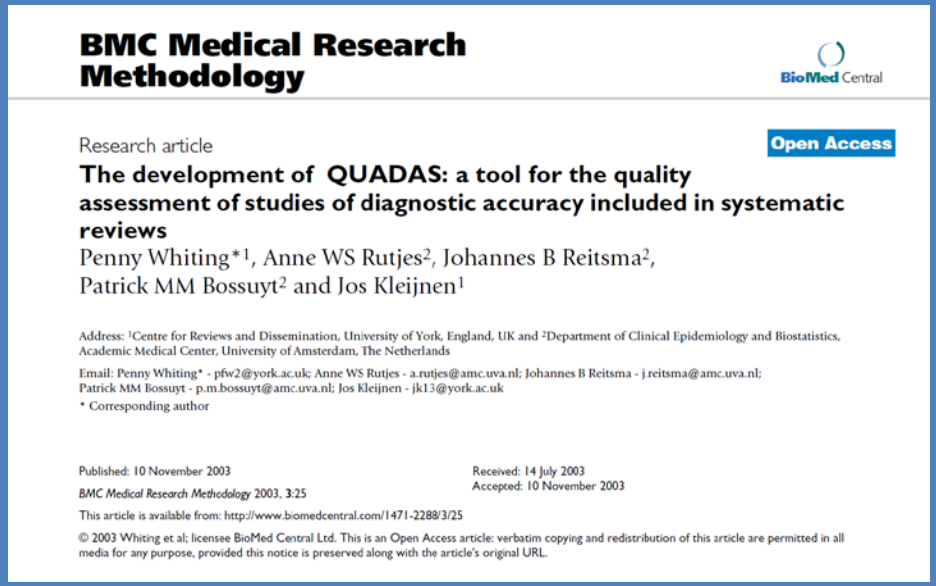
The Lancet Infect Dis 2003

**Table 4. Stratified analyses for the evaluation of heterogeneity among studies with in-house tests**

| Subgroup | Number of studies | Summary diagnostic odds ratio* (95% CI) | Test for heterogeneity† p value |
|---|---|---|---|
| **Study design** | | | |
| Case-control | 19 | 86·5 (39·3, 190·2) | 0·03 |
| Cross-sectional | 16 | 43·3 (22·5, 83·3) | 0·94 |
| **Blinded interpretation of test and/or reference standard results** | | | |
| Yes | 21 | 46·9 (24·9, 88·6) | 0·16 |
| No | 14 | 82·3 (39·8, 170·2) | 0·70 |
| **Consecutive or random sampling of participants** | | | |
| Yes | 18 | 63·3 (32·8, 122·4) | 0·20 |
| No | 17 | 46·8 (23·6, 92·8) | 0·42 |
| **Prospective data collection** | | | |
| Yes | 18 | 59·9 (28·1, 127·6) | 0·12 |
| No | 17 | 55·2 (29·9, 101·6) | 0·59 |

*Random effects model. †$\chi^2$ test for heterogeneity. CI=confidence interval.

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



**BMC Medical Research Methodology**

BioMed Central

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 Rutjes[2], Johannes B Reitsma[2], Patrick MM Bossuyt[2] and Jos Kleijnen[1]

Address: [1]Centre for Reviews and Dissemination, University of York, England, UK and [2]Department 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

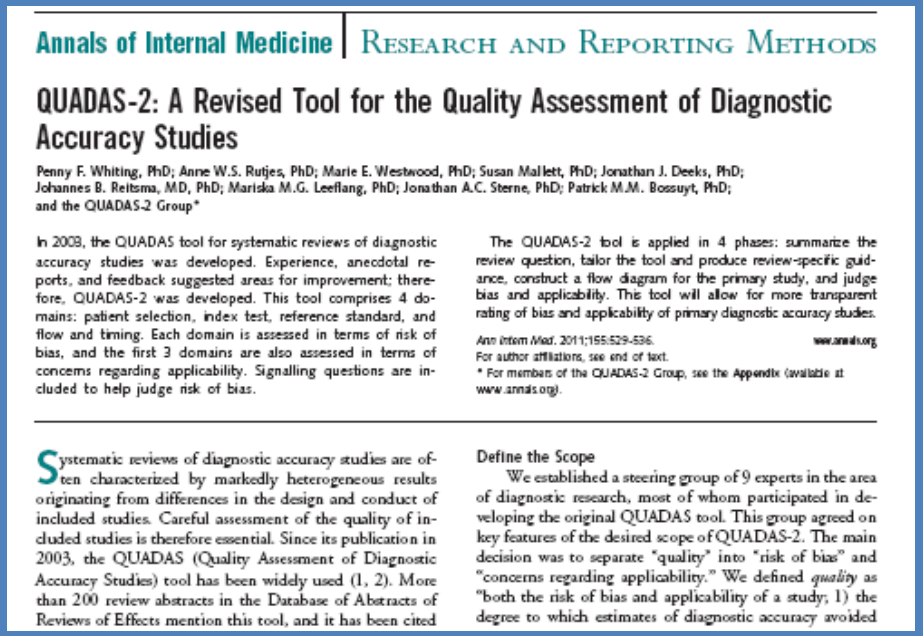
*BMC Medical Research Methodology* 2003, 3:25

Received: 14 July 2003
Accepted: 10 November 2003

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



# QUADAS, 2003



**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.S. Rutjes, PhD; Marie E. Westwood, PhD; Susan Mallett, PhD; Jonathan J. Deeks, PhD; Johannes B. Reitsma, M.D, PhD; Mariska M.G. Leeflang, PhD; Jonathan A.C. Sterne, PhD; Patrick M.M. Bossuyt, PhD; and the QUADAS-2 Group*

In 2003, 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 question, 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.

*Ann Intern Med.* 2011;155:529-536. www.annals.org

For author affiliations, see end of text.

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

Systematic reviews of diagnostic accuracy studies are often characterized by markedly heterogeneous results originating from differences in the design and conduct of included studies. Careful assessment of the quality of included studies is therefore essential. Since its publication in 2003, the QUADAS (Quality Assessment of Diagnostic Accuracy Studies) tool has been widely used (1, 2). More than 200 review abstracts in the Database of Abstracts of Reviews of Effects mention this tool, and it has been cited

### Define the Scope

We established a steering group of 9 experts in the area of diagnostic research, most of whom participated in developing the original QUADAS tool. This group agreed on key features of the desired scope of QUADAS-2. The main decision was to separate "quality" into "risk of bias" and "concerns regarding applicability." We defined *quality* as "both the risk of bias and applicability of a study; 1) the degree to which estimates of diagnostic accuracy avoided

# QUADAS-2, 2011

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

# QUADAS-2

- Domain list
  - patient selection
  - index test
  - reference standard
  - flow and timing
- First 3 domains assessed for applicability
- *Signalling questions* used for judgments of risk of bias

# Patient selection - 1

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

- Was a consecutive or random sample of patients enrolled? Yes, No, Unclear
- Was a case control design avoided? Yes, No, Unclear

- *Could the selection of patients have introduced bias?*

  Record Low, High, or Unclear

# Patient selection - 2

*Are there concerns regarding applicability?*

- Do the included patients and setting match the question? (demographic and clinical features, prior testing, intended use of index test, and setting)
- Record Low, High, or Unclear

# Suggested displays – QUADAS-2

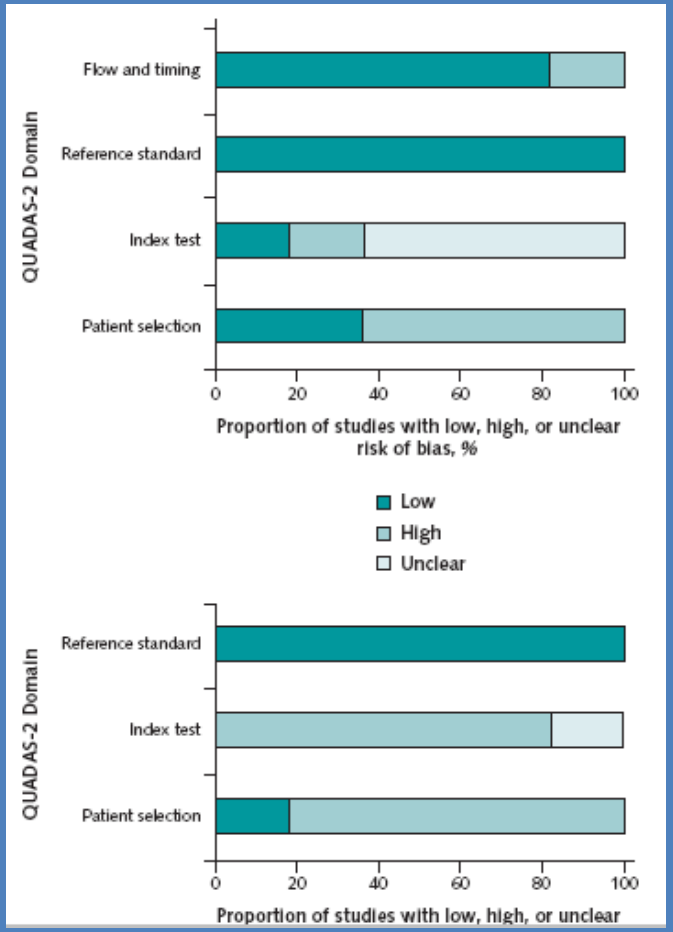


*Table 2.* **Suggested Tabular Presentation for QUADAS-2 Results**

| Study | Risk of Bias | | | | Applicability Concerns | | |
|---|---|---|---|---|---|---|---|
| | Patient Selection | Index Test | Reference Standard | Flow and Timing | Patient Selection | Index Test | Reference Standard |
| 1 | ☺ | ☺ | ☺ | ☺ | ☹ | ☺ | ☺ |
| 2 | ☺ | ☺ | ☺ | ☺ | ☹ | ☺ | ☺ |
| 3 | ☹ | ☹ | ☹ | ☹ | ☹ | ☹ | ☹ |
| 4 | ☹ | ☹ | ☹ | ☹ | ☹ | ☹ | ☹ |
| 5 | ☹ | ? | ☹ | ☹ | ☹ | ☹ | ☹ |
| 6 | ☹ | ? | ☺ | ☺ | ☹ | ? | ☺ |
| 7 | ☹ | ? | ☹ | ☹ | ☹ | ☹ | ☹ |
| 8 | ☹ | ? | ☹ | ☹ | ☹ | ? | ☹ |
| 9 | ☹ | ? | ☹ | ☹ | ☹ | ☹ | ☹ |
| 10 | ☹ | ? | ☺ | ☹ | ☹ | ☺ | ☺ |
| 11 | ☹ | ? | ☹ | ☹ | ☹ | ☹ | ☹ |

☺ = low risk; ☹ = high risk; ? = unclear risk.

http://www.bris.ac.uk/quadas/



# Problems with quality assessment

- Not as straightforward as it might seem
- No quality scores or cut-offs for ‘good’ quality
- Quality assessment is subjective
- Statistical incorporation of quality problematic with limited number of studies
- Hampered by poor reporting

OPEN ACCESS Freely available online

PLoS one

# Quality and Reporting of Diagnostic Accuracy Studies in TB, HIV and Malaria: Evaluation Using QUADAS and STARD Standards

**Patricia Scolari Fontela[1], Nitika Pant Pai[2], Ian Schiller[2], Nandini Dendukuri[2], Andrew Ramsay[3], Madhukar Pai[1,4]***

**1** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, **2** Department of Medicine, Division of Clinical Epidemiology, McGill University, Montreal, Canada, **3** Special Programme for Research and Training in Tropical Diseases, World Health Organization, Geneva, Switzerland, **4** Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, Montreal, Canada

## Abstract

*Background:* Poor methodological quality and reporting are known concerns with diagnostic accuracy studies. In 2003, the QUADAS tool and the STARD standards were published for evaluating the quality and improving the reporting of diagnostic studies, respectively. However, it is unclear whether these tools have been applied to diagnostic studies of infectious diseases. We performed a systematic review on the methodological and reporting quality of diagnostic studies in TB, malaria and HIV.

*Methods:* We identified diagnostic accuracy studies of commercial tests for TB, malaria and HIV through a systematic search of the literature using PubMed and EMBASE (2004–2006). Original studies that reported sensitivity and specificity data were included. Two reviewers independently extracted data on study characteristics and diagnostic accuracy, and used QUADAS and STARD to evaluate the quality of methods and reporting, respectively.

*Findings:* Ninety (38%) of 238 articles met inclusion criteria. All studies had design deficiencies. Study quality indicators that were met in less than 25% of the studies included adequate description of withdrawals (6%) and reference test execution (10%), absence of index test review bias (19%) and reference test review bias (24%), and report of uninterpretable results (22%). In terms of quality of reporting, 9 STARD indicators were reported in less than 25% of the studies: methods for calculation and estimates of reproducibility (0%), adverse effects of the diagnostic tests (1%), estimates of diagnostic accuracy between subgroups (10%), distribution of severity of disease/other diagnoses (11%), number of eligible patients who did not participate in the study (14%), blinding of the test readers (16%), and description of the team executing the test and management of indeterminate/outlier results (both 17%). The use of STARD was not explicitly mentioned in any study. Only 22% of 46 journals that published the studies included in this review required authors to use STARD.

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

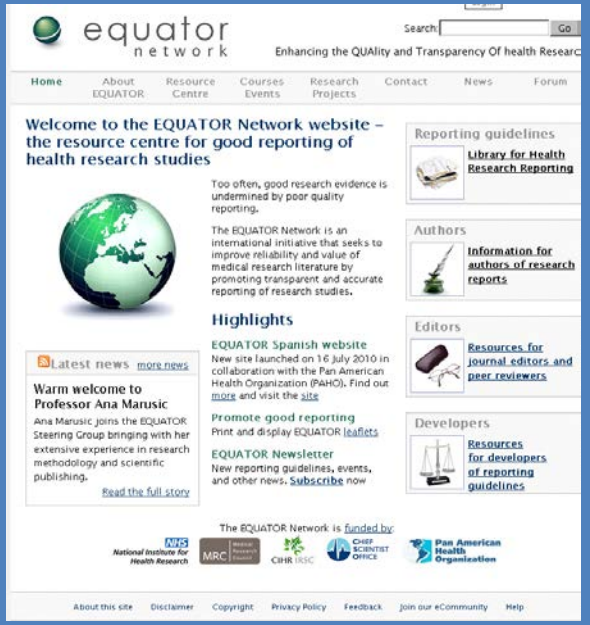
# Quality of TB accuracy studies using QUADAS

| Quality item | 45 studies n (%) |
|---|---|
| Adequate spectrum composition | 26 (58) |
| Adequate reference standard | 44 (98) |
| Absence of disease progression bias | 42 (93) |
| Absence of partial verification bias | 44 (98) |
| Absence of differential verification bias | 42 (93) |
| Absence of incorporation bias | 45 (100) |
| Absence of index test review bias | 6 (13) |
| Absence of reference test review bias | 7 (16) |
| Absence of clinical review bias | 14 (31) |
| Report of uninterpretable results | 9 (20) |
| Description of withdrawals | 3 (7) |

Fontela et al. PLoS One 2009

# Initiatives to improve quality and reporting

- STARD: reporting of diagnostic studies
- PRISMA: reporting of systematic reviews/meta-analyses of RCTs
- STROBE: reporting of observational studies
- MOOSE: reporting of meta-analyses of observational studies
- **AMSTAR: assessing quality of systematic reviews**

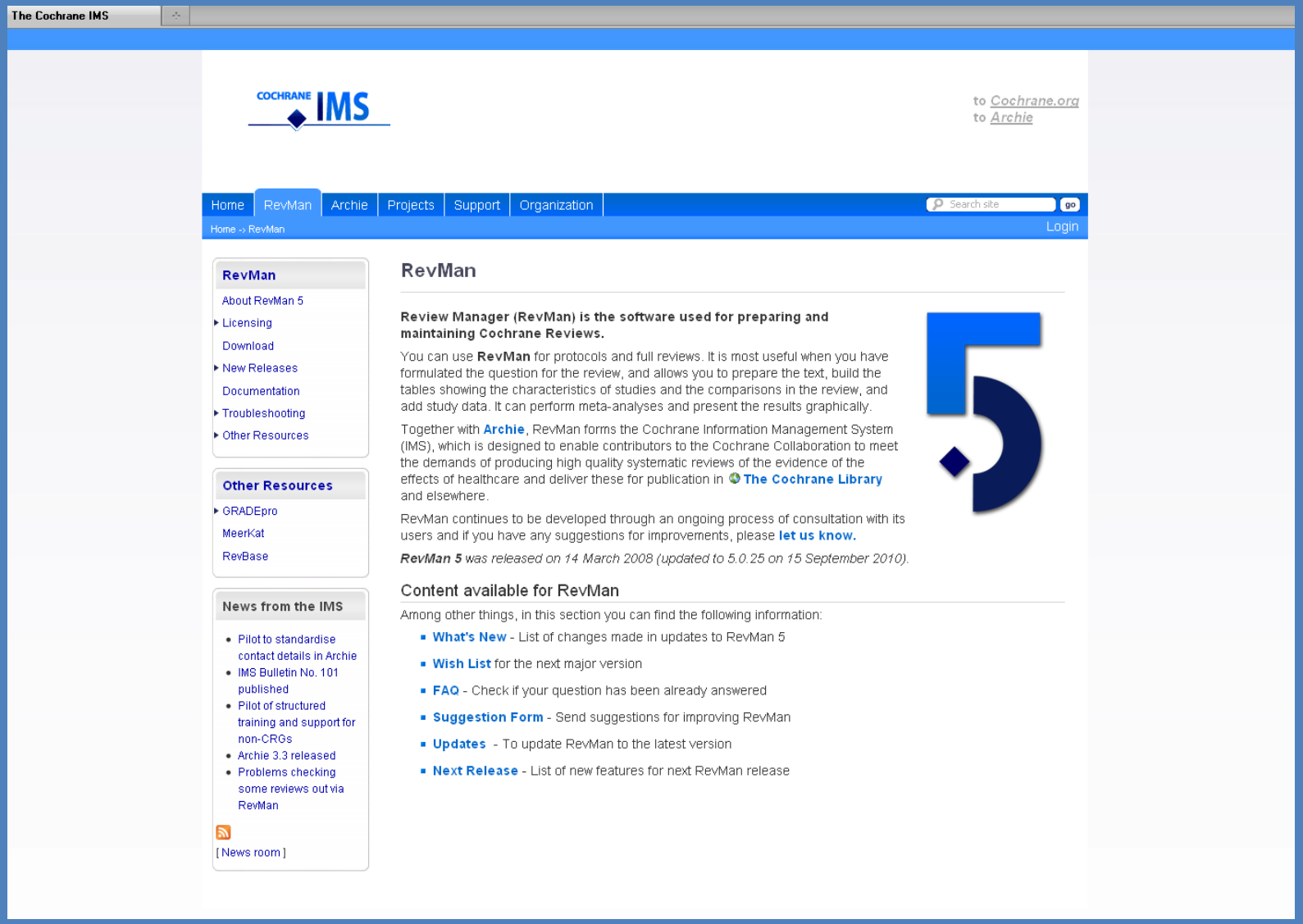



www.equator-network.org/

# 4. Analyzing the data; undertaking meta-analyses

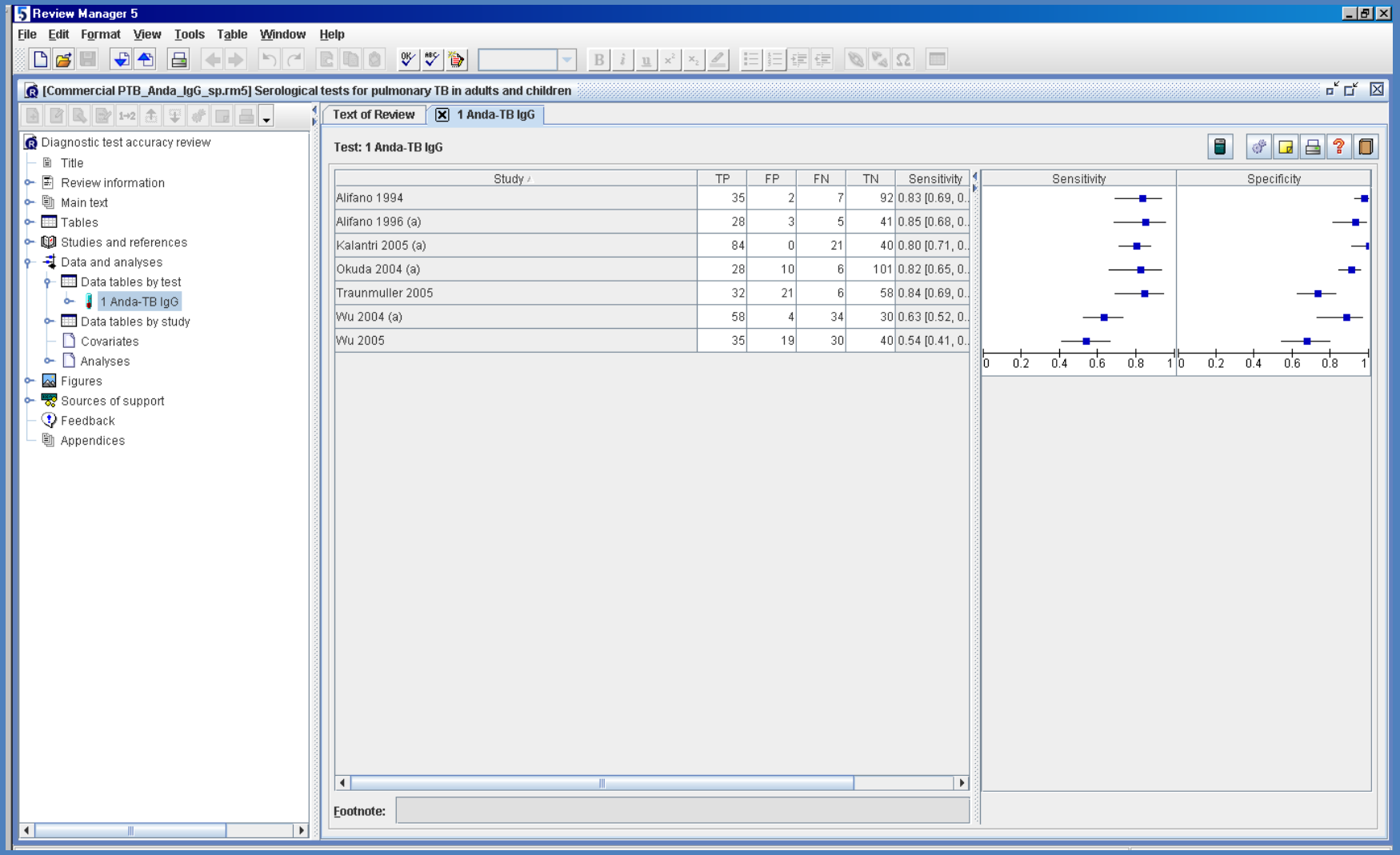
# Key steps

- Extract TP, FP, FN, and TN to determine 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



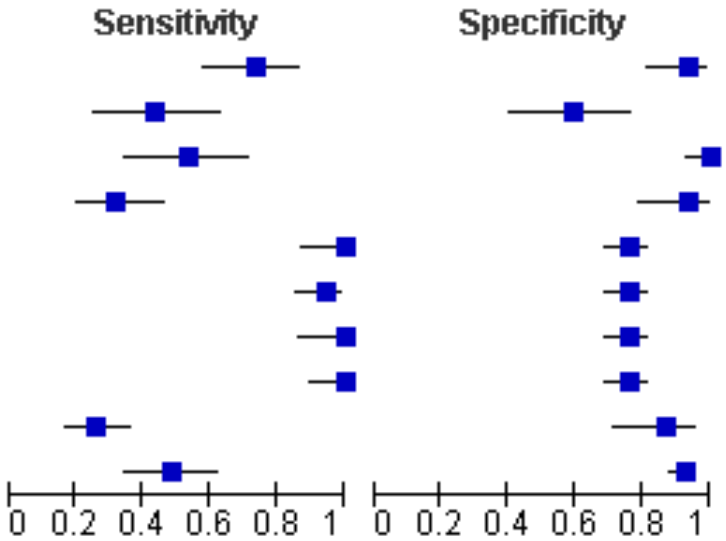

http://ims.cochrane.org/revman

# Cut and paste data from excel



# Forest plot – diagnostic test accuracy review

| Study | TP | FP | FN | TN | Sensitivity | Specificity |
|---|---|---|---|---|---|---|
| Alifano 1998b | 31 | 3 | 11 | 41 | 0.74 [0.58, 0.86] | 0.93 [0.81, 0.99] |
| Banerjee 2003a | 13 | 13 | 17 | 19 | 0.43 [0.25, 0.63] | 0.59 [0.41, 0.76] |
| Caminero 1993 | 16 | 0 | 14 | 48 | 0.53 [0.34, 0.72] | 1.00 [0.93, 1.00] |
| Caminero 1994 | 18 | 2 | 38 | 29 | 0.32 [0.20, 0.46] | 0.94 [0.79, 0.99] |
| Gevaudan 1992b | 26 | 47 | 0 | 147 | 1.00 [0.87, 1.00] | 0.76 [0.69, 0.82] |
| Gevaudan 1992f | 53 | 47 | 3 | 147 | 0.95 [0.85, 0.99] | 0.76 [0.69, 0.82] |
| Gevaudan 1992j | 25 | 47 | 0 | 147 | 1.00 [0.86, 1.00] | 0.76 [0.69, 0.82] |
| Gevaudan 1992n | 34 | 47 | 0 | 147 | 1.00 [0.90, 1.00] | 0.76 [0.69, 0.82] |
| Kunter 2003a | 23 | 5 | 65 | 32 | 0.26 [0.17, 0.37] | 0.86 [0.71, 0.95] |
| Luh 1996 | 28 | 17 | 30 | 207 | 0.48 [0.35, 0.62] | 0.92 [0.88, 0.96] |




One row is displayed for each study
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 = confidence interval

Steingart, PLoS Med 2011

**BMC Medical Research Methodology**

BioMed Central

Software

Open Access

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

Javier Zamora*[1], Victor Abraira[1], Alfonso Muriel[1], Khalid Khan[2] and Arri Coomarasamy[2]

Address: [1]Clinical Biostatistics Unit, Ramón y Cajal Hospital, Madrid, Ctra. Colmenar km 9.100 Madrid 28034, Spain and [2]University 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

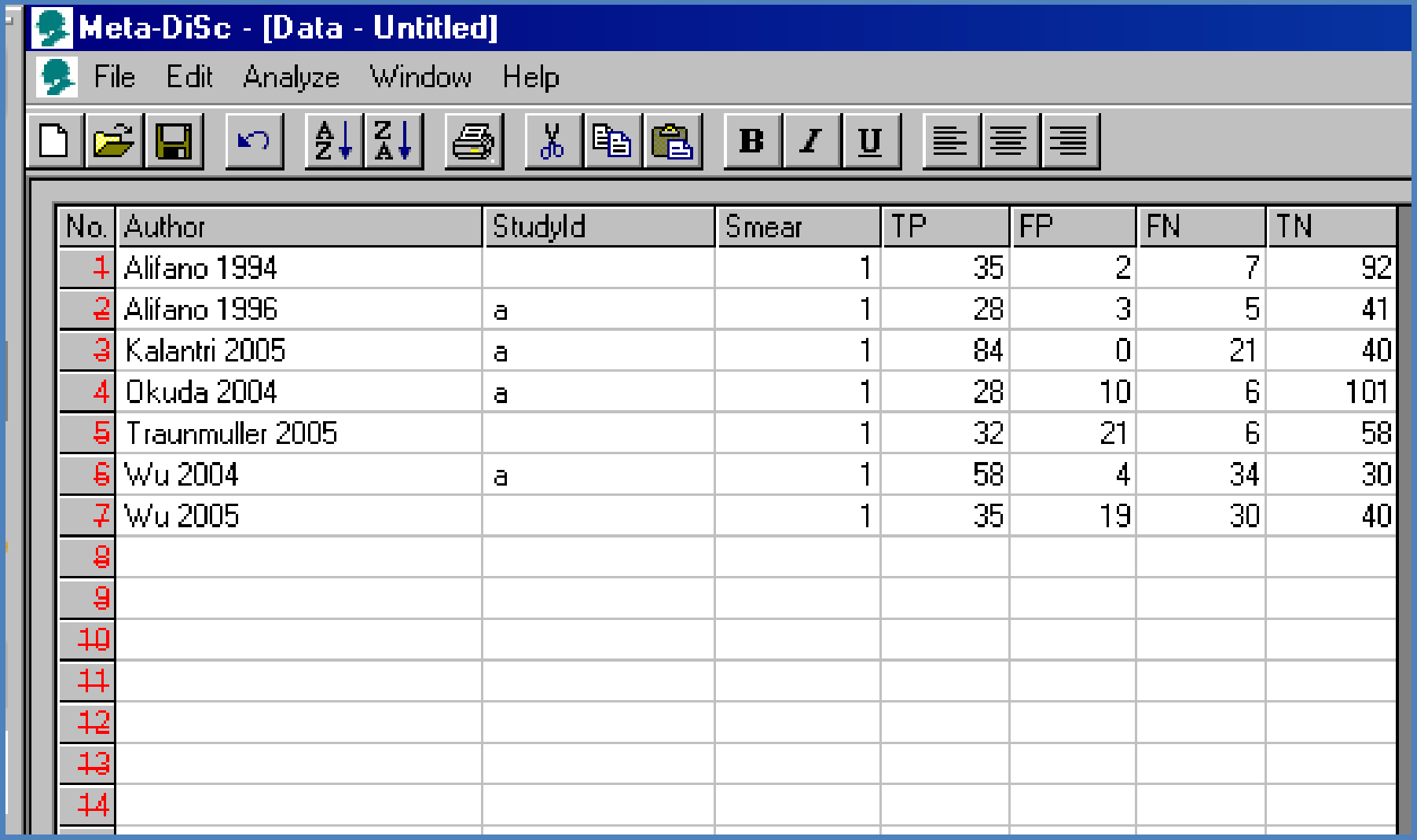
Published: 12 July 2006

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

Received: 31 March 2006
Accepted: 12 July 2006

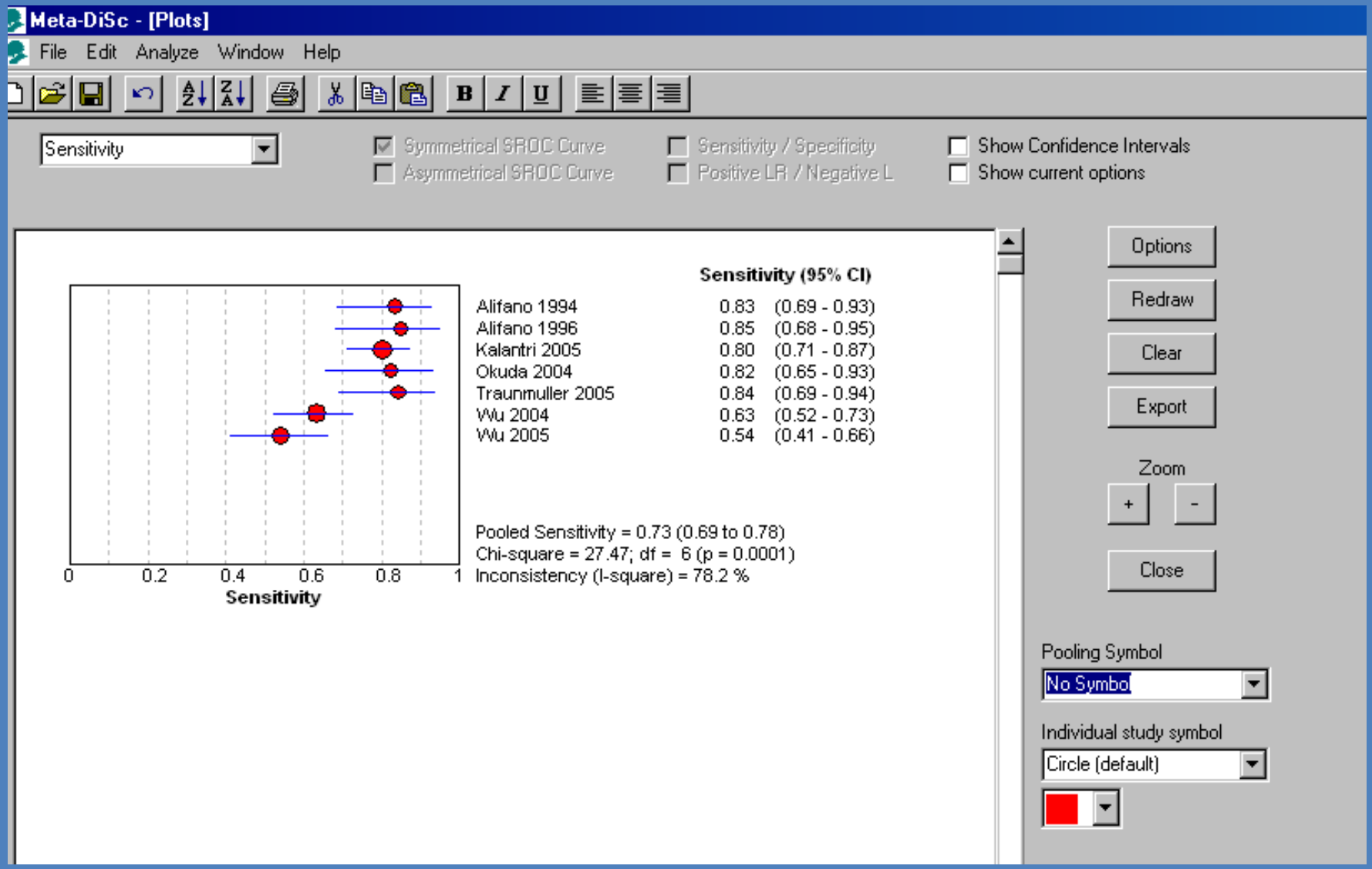
This article is available from: http://www.biomedcentral.com/1471-2288/6/31

# Cut and paste data from excel worksheet into MetaDisc

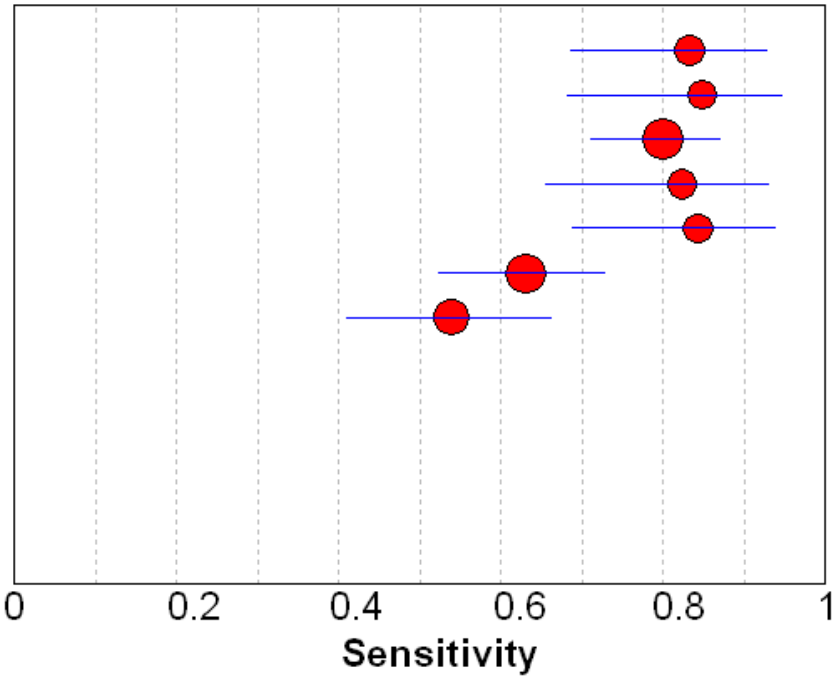



| No. | Author | StudyId | Smear | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|
| 1 | Alifano 1994 | | 1 | 35 | 2 | 7 | 92 |
| 2 | Alifano 1996 | a | 1 | 28 | 3 | 5 | 41 |
| 3 | Kalantri 2005 | a | 1 | 84 | 0 | 21 | 40 |
| 4 | Okuda 2004 | a | 1 | 28 | 10 | 6 | 101 |
| 5 | Traunmuller 2005 | | 1 | 32 | 21 | 6 | 58 |
| 6 | Wu 2004 | a | 1 | 58 | 4 | 34 | 30 |
| 7 | Wu 2005 | | 1 | 35 | 19 | 30 | 40 |
| 8 | | | | | | | |
| 9 | | | | | | | |
| 10 | | | | | | | |
| 11 | | | | | | | |
| 12 | | | | | | | |
| 13 | | | | | | | |
| 14 | | | | | | | |

# Select plot and characteristics



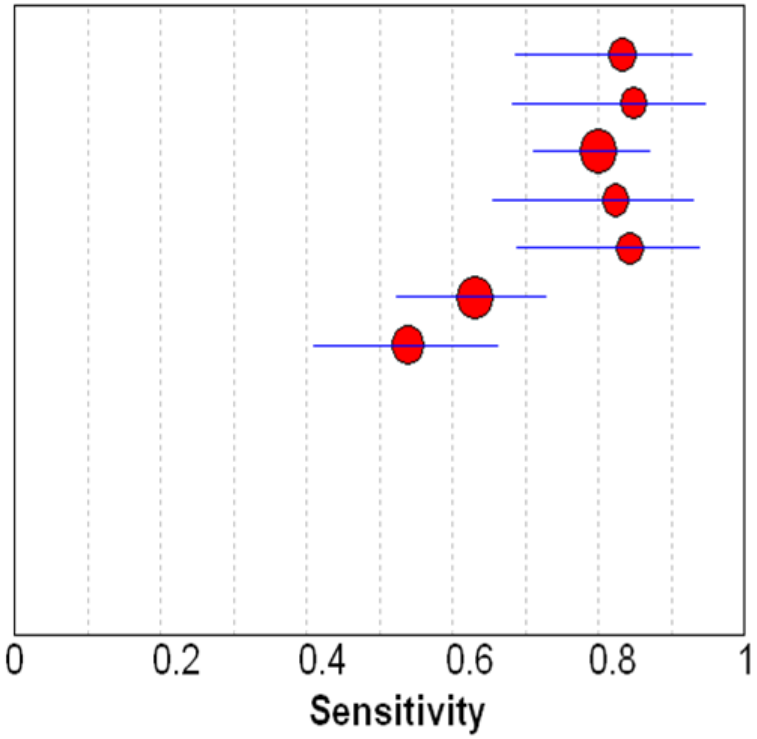
# Export plot




| | 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 %

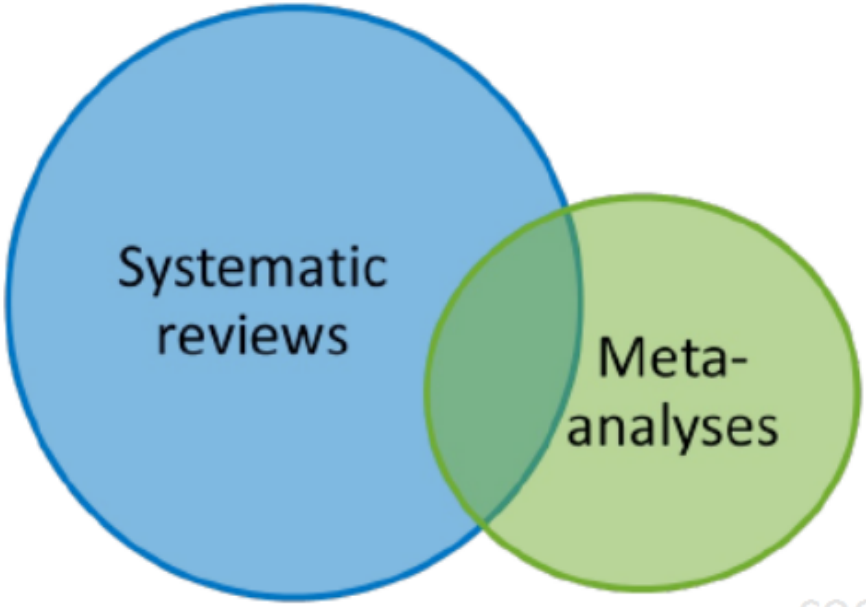
# Edit the plot




| | 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) |

# What is a meta-analysis?

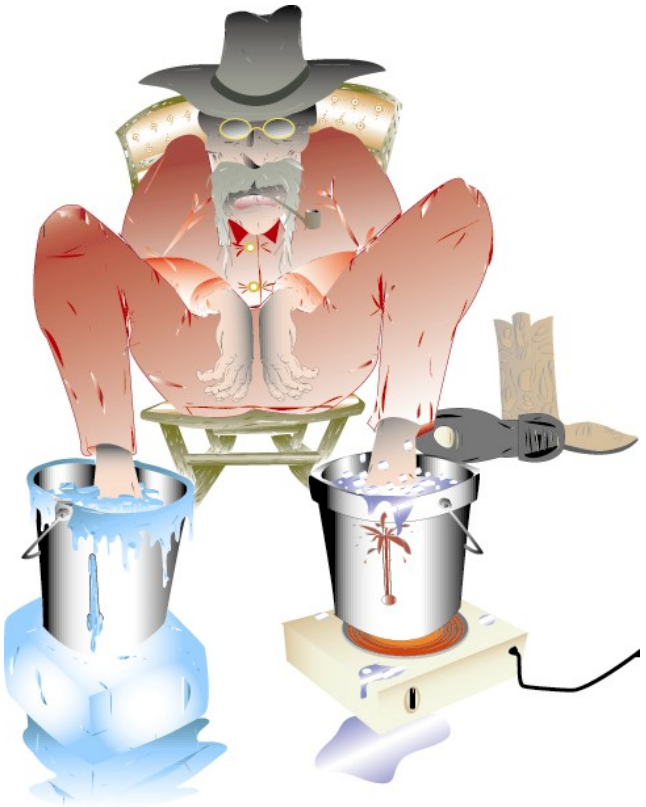
- combines the results from two or more studies
- estimates an 'average' or 'common' effect
- optional part of a systematic review




Source: Julian Higgins

"Il a été dit que un homme avec une jambe dans la glace et l'autre jambe dans l'eau bouillante est à l'aise - en moyenne. 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.



# Improved statistical models for meta-analysis of diagnostic studies

- To repair statistical shortcomings
  - hierarchical structure
  - random effects model
- Two different approaches:

  – hierarchical summary ROC model (HSROC, Gatsonis and Rutter 2001)

  – bivariate regression of sensitivity and specificity (Bivariate, Reitsma 2005)

## The models are ‘hierarchical’ because they involve statistical distributions at two levels

- At the lower level, they model the cell counts in the 2×2 tables extracted from each study using binomial distributions and logistic (log-odds) transformations of proportions

- At the second (higher) level, the models assume random study effects to account for heterogeneity in diagnostic test accuracy between studies beyond that accounted for by sampling variability at the lower level

Macaskill P et al. Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration 2010.

STATISTICS IN MEDICINE
*Statist. Med.* 2001; **20**:2865–2884 (DOI: 10.1002/sim.942)

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

Carolyn M. Rutter[1,*,†] and Constantine A. Gatsonis[2]

[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

*Biostatistics* (2007), **8**, 2, *pp.* 239–251
doi:10.1093/biostatistics/kxl004
Advance Access publication on May 11, 2006

# 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 Berne, Switzerland*

PENNY WHITING, JONATHAN A. C. STERNE

*Medical Research Council Health Services Research Collaboration, Department of Social Medicine, University of Bristol, UK*

# Bivariate model vs HSROC model

- Where studies report a common threshold for a positive result, use the bivariate  model

- Where studies report several different thresholds, use the HSROC model

# Stata command, metandi

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

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

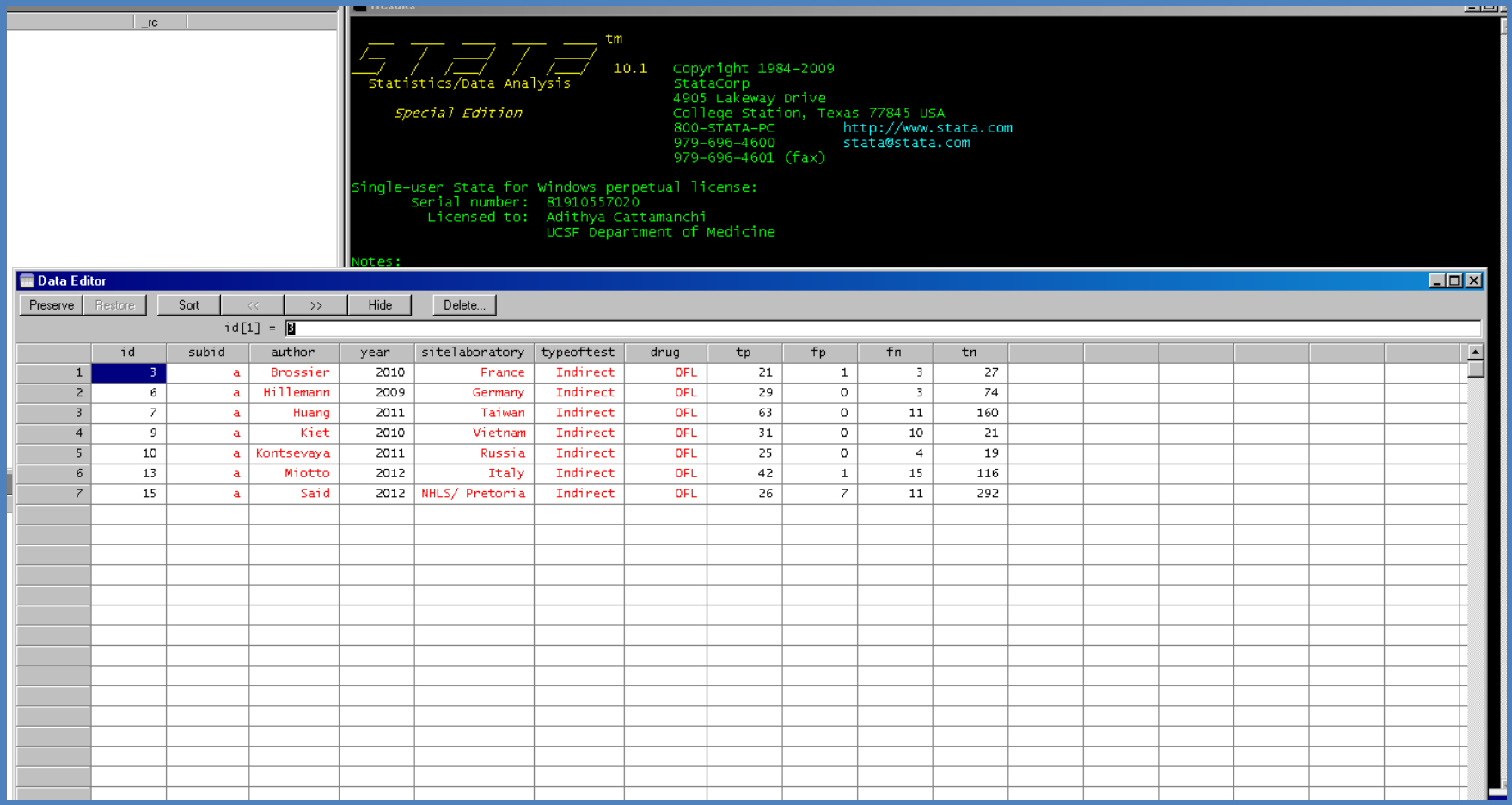
Roger M. Harbord
Department of Social Medicine
University of Bristol
Bristol, UK
roger.harbord@bristol.ac.uk

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.

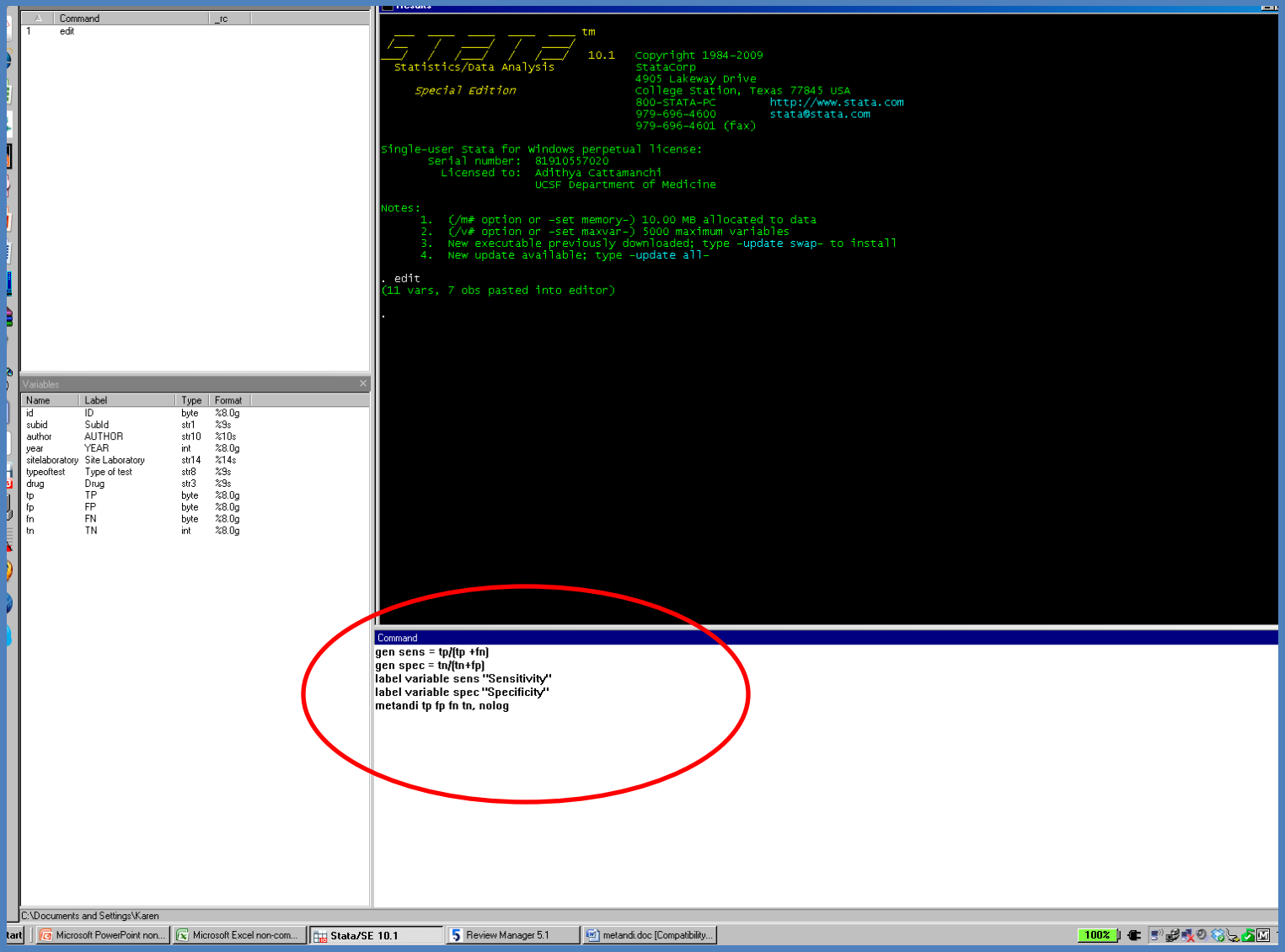
**Keywords:** st0163, metandi, metandiplot, diagnosis, meta-analysis, sensitivity and specificity, hierarchical models, generalized mixed models, gllamm, xtmelogit, receiver operating characteristic (ROC), summary ROC, hierarchical summary ROC

# Import/paste data from excel worksheet into Stata data editor

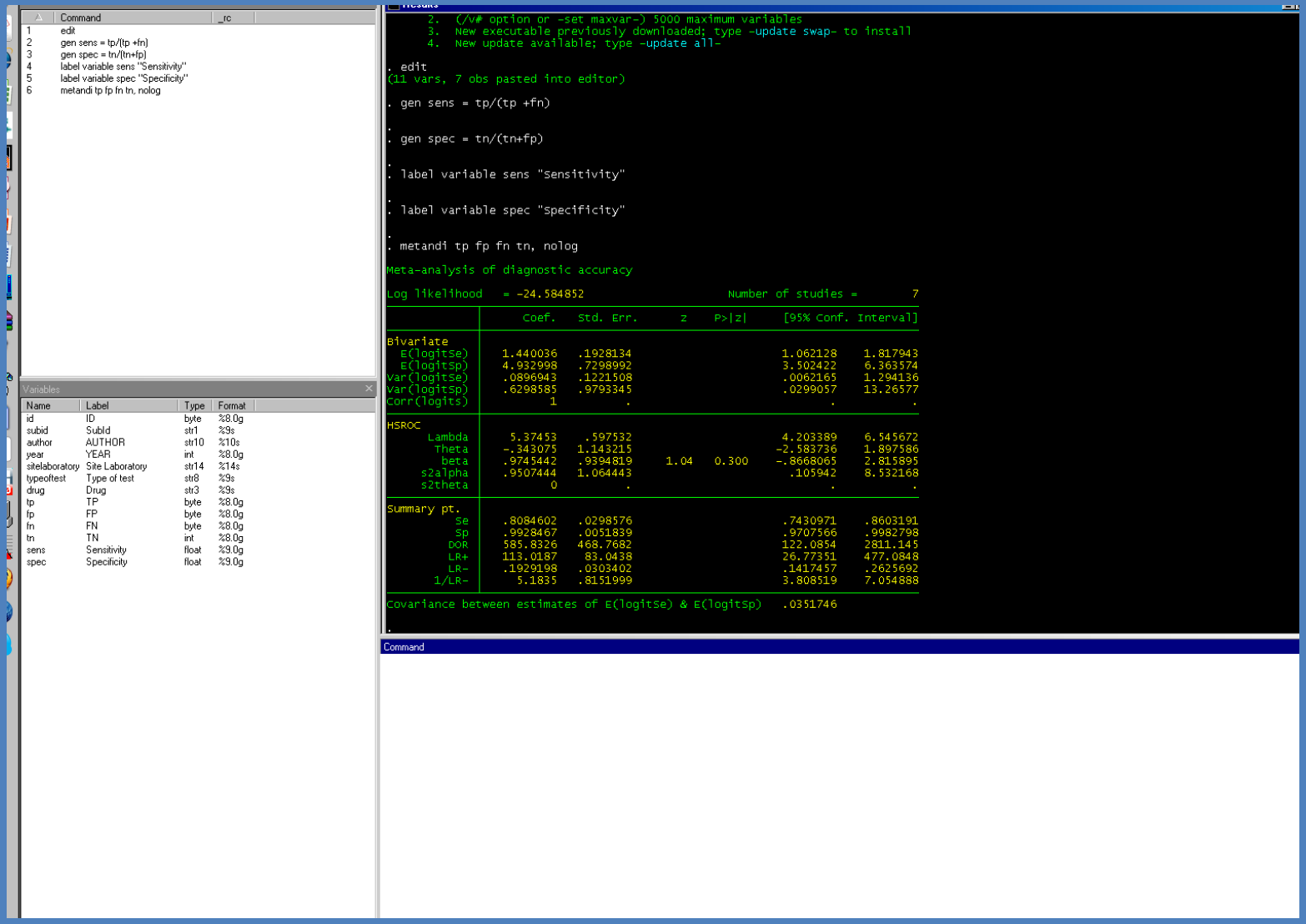



| | id | subid | author | year | sitelaboratory | typeoftest | drug | tp | fp | fn | tn |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | a | Brossier | 2010 | France | Indirect | OFL | 21 | 1 | 3 | 27 |
| 2 | 6 | a | Hillemann | 2009 | Germany | Indirect | OFL | 29 | 0 | 3 | 74 |
| 3 | 7 | a | Huang | 2011 | Taiwan | Indirect | OFL | 63 | 0 | 11 | 160 |
| 4 | 9 | a | Kiet | 2010 | Vietnam | Indirect | OFL | 31 | 0 | 10 | 21 |
| 5 | 10 | a | Kontsevaya | 2011 | Russia | Indirect | OFL | 25 | 0 | 4 | 19 |
| 6 | 13 | a | Miotto | 2012 | Italy | Indirect | OFL | 42 | 1 | 15 | 116 |
| 7 | 15 | a | Said | 2012 | NHLS/ Pretoria | Indirect | OFL | 26 | 7 | 11 | 292 |

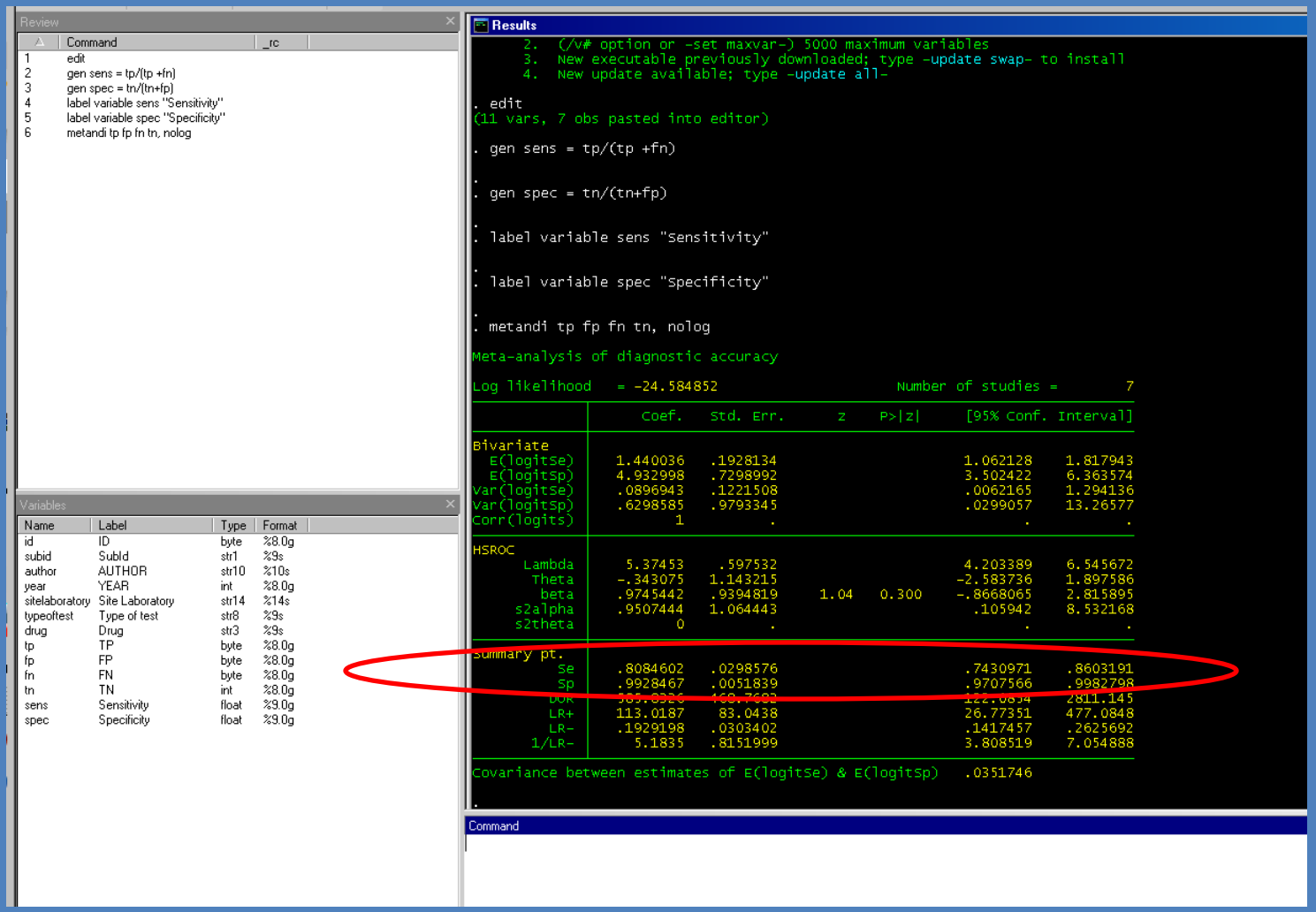
# Close the data editor, type/paste metandi syntax in lower right ‘Command’ screen



# Hit enter and receive Stata output



# Read Stata output




Pooled sensitivity = 80.8% (95% CI 74.3, 86,0)
Pooled specificity = 99.3% (95% CI 97.1, 99.8)

# Heterogeneity

- Refers to variation in results among studies
- May be caused by variation in
  - test thresholds (unique to meta-analyses of diagnostic tests)
  - prevalence of disease
  - patient spectrum
  - study quality
  - chance variation

# Variation due to threshold differences

- Explicit threshold differences
  – studies have used different cut-off values

to define positive test results

- Implicit threshold differences
  – differences in observers
  – differences in equipment
- Consequence: negative correlation arises between sensitivity and specificity

# Exploring heterogeneity

- Subgroup analysis
- Meta-regression analysis

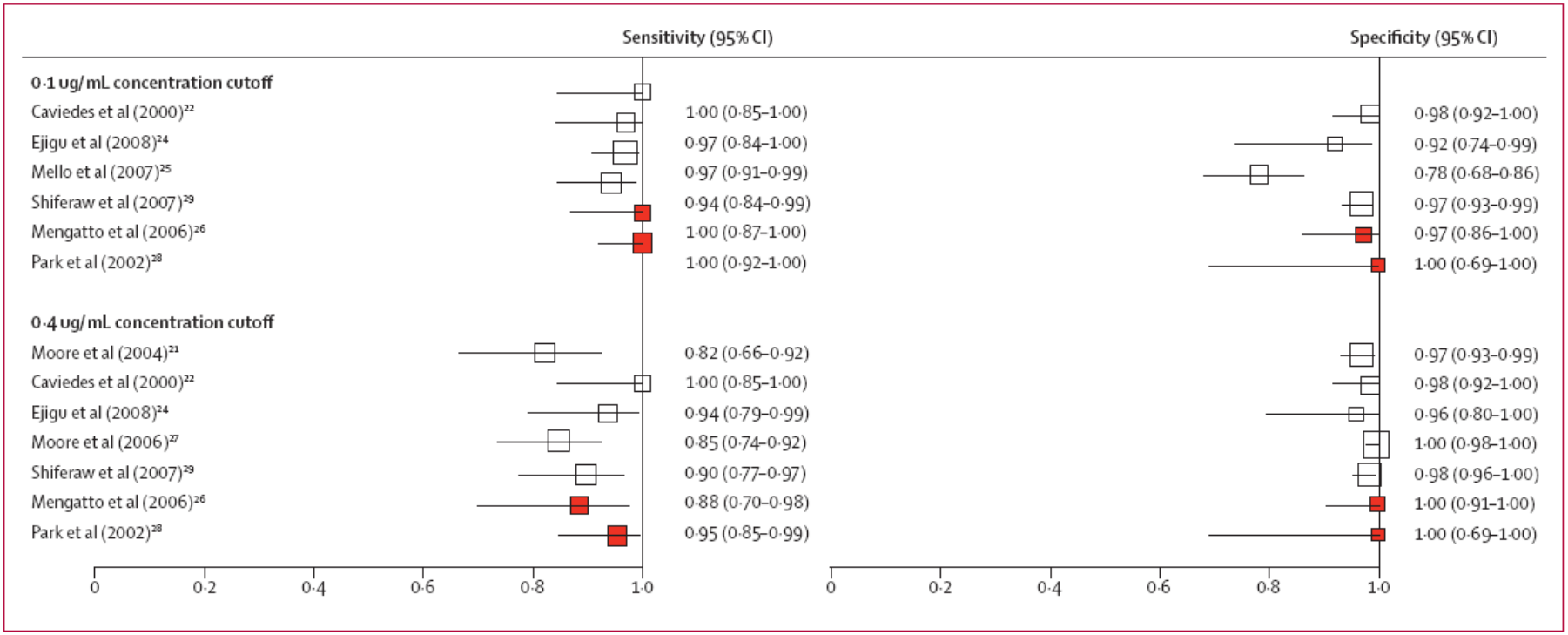



***Figure 4:*** **Forest plot of accuracy of the microscopic-observation drug susceptibility assay for detection of isoniazid resistance**
Open squares represent studies using direct inoculation with specimens from patients; red squares represent studies using indirect inoculation with isolates. Size of the square is proportionate to the size of the study.

Use of the 0·1 µg/mL isoniazid cutoff was associated with higher sensitivity, but lower specificity than the 0·4 µg/mL cutoff. J Minion, TLID 2010

5. Drawing robust conclusions and informative presentation of results

- summary of findings tables

# Some questions

- 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 future research?

## GRADE Summary of Findings Table for Xpert MTB/RIF Assay

Review question: What is the diagnostic accuracy of Xpert MTB/RIF assay for diagnosis of pulmonary TB and detection of rifampicin resistance?
Patients/population: Adult pulmonary TB suspects (for diagnosis of pulmonary TB); Confirmed TB cases (for detection of rifampicin resistance)
Setting: Clinical centers and laboratories
Index test: Xpert MTB/RIF assay
Importance: Compared with sputum smear microscopy and conventional drug susceptibility testing, near point-of-care tests, such as Xpert MTB/RIF assay, have considerable advantages for scaling up programmatic management by offering speed of diagnosis, standardized testing, potential for high throughput, and fewer requirements for laboratory bio-safety
Reference standard: Conventional drug susceptibility testing by solid or liquid culture
Studies: Cross-sectional or cohort

| **Outcomes: TP, TN, FP, FN**<br><br>**Diagnostic accuracy for diagnosis of pulmonary TB** | **Effect % (95% CI)** | **No. of Participants (Studies)** | **What do these results mean given 5% prevalence among suspects being screened for TB?** | **What do these results mean given 15% prevalence among suspects being screened for TB?** | **What do these results mean given 30% prevalence among suspects being screened for TB?** | **Quality of Evidence** |
|---|---|---|---|---|---|---|
| **All patients** | Pooled sensitivity ##.# (95% CI ##.#, ##.#) and pooled specificity ##.#% (95% CI ##.#, ##.#) | #### (18) | With a prevalence of 5%, 50/1000 will have pulmonary TB. Of these, ## (TP) will be identified; ## (FN) will be missed. Of the 950 patients without TB, ## (TN) will not be treated; ## (FP) will be unnecessarily treated | With a prevalence of 15%, 150/1000 will have pulmonary TB. Of these, ## (TP) will be identified; ## (FN) will be missed. Of the 850 patients without TB, ## (TN) will not be treated; ## (FP) will be unnecessarily treated | With a prevalence of 30%, 300/1000 will have pulmonary TB. Of these, ## (TP) will be identified; ## (FN) will be missed. Of the 700 patients without TB, ## (TN) will not be treated; ## (FP) will be unnecessarily treated | Moderate ⊕⊕⊕◯ |
| **Smear positive patients** | Pooled sensitivity ##.#% (95% CI ##.#, ##.#) and pooled specificity ##.#% (95% CI ##.#, ##.#) | #### (##) | With a prevalence of 5%, 50/1000 will have pulmonary TB. Of these, ## (TP) will be identified; ## (FN) will be missed. Of the 950 patients without TB, ## (TN) will not be treated; ## (FP) will be unnecessarily | With a prevalence of 15%, 150/1000 will have pulmonary TB. Of these, ## (TP) will be identified; ## (FN) will be missed. Of the 850 patients without TB, ## (TN) will not be treated; ## (FP) will be unnecessarily | With a prevalence of 30%, 300/1000 will have pulmonary TB. Of these, ## (TP) will be identified; ## (FN) will be missed. Of the 700 patients without TB, ## (TN) will not be treated; ## (FP) will be unnecessarily | Moderate ⊕⊕⊕◯ |

Steingart draft template

# Summary

- Diagnostic tests should be evaluated in consecutively or randomly selected patients with diagnostic uncertainty who are representative of those in whom the test will be used in practice; an appropriate reference standard should be used to establish diagnosis

- Currently, searching electronic databases is challenging. Searches based upon index test and target condition, which are designed to maximize sensitivity, are recommended

- Test accuracy studies are often poorly reported, hampering data extraction and quality assessment

- Though sometimes unable to provide a definitive estimate of test accuracy, systematic reviews can highlight gaps in the evidence base and aid in the design of future studies

# References and tools for meta-analysis


- Leeflang. Ann Intern Med. 2008;149:889-897
- Rutter and Gatsonis. Stat Med. 2001; 20:2865–2884
- Reitsma. J Clin Epidemiol. 2005; 982–990
- 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

# *With special thanks to*

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

Diagnostic Test Accuracy Review Training, Montreal, May 2009

# Meta-regression

- Is a form of linear regression in which studies are the unit of analysis
- Aims to relate the size of effect to one or more characteristics of the studies involved
- DOR is the dependent variable
- Covariates that might be associated with the variability in DOR are the independent variables

- Tip: Specify covariates that you want to explore in advance

# The threshold effect (-0.21) was significant (p = 0.01). This was also seen in the SROC plot, Ling 2008.

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

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

doi:10.1371/journal.pone.0001536.t006

Determined using ‘Metareg’ command in Stata

# Diagnostic Odds Ratio (DOR) and Relative DOR

- DOR = odds of a positive result in diseased individuals versus odds of a positive result in non-diseased individuals
- Combines both likelihood ratios  DOR = LR+/LR-
- DOR = 1 means the test cannot discriminate between people with and without disease
- RDOR (relative DOR) = ratio of 2 DORs
- RDOR = 1 means a particular covariate (e.g. blinded study design) does not affect the overall DOR