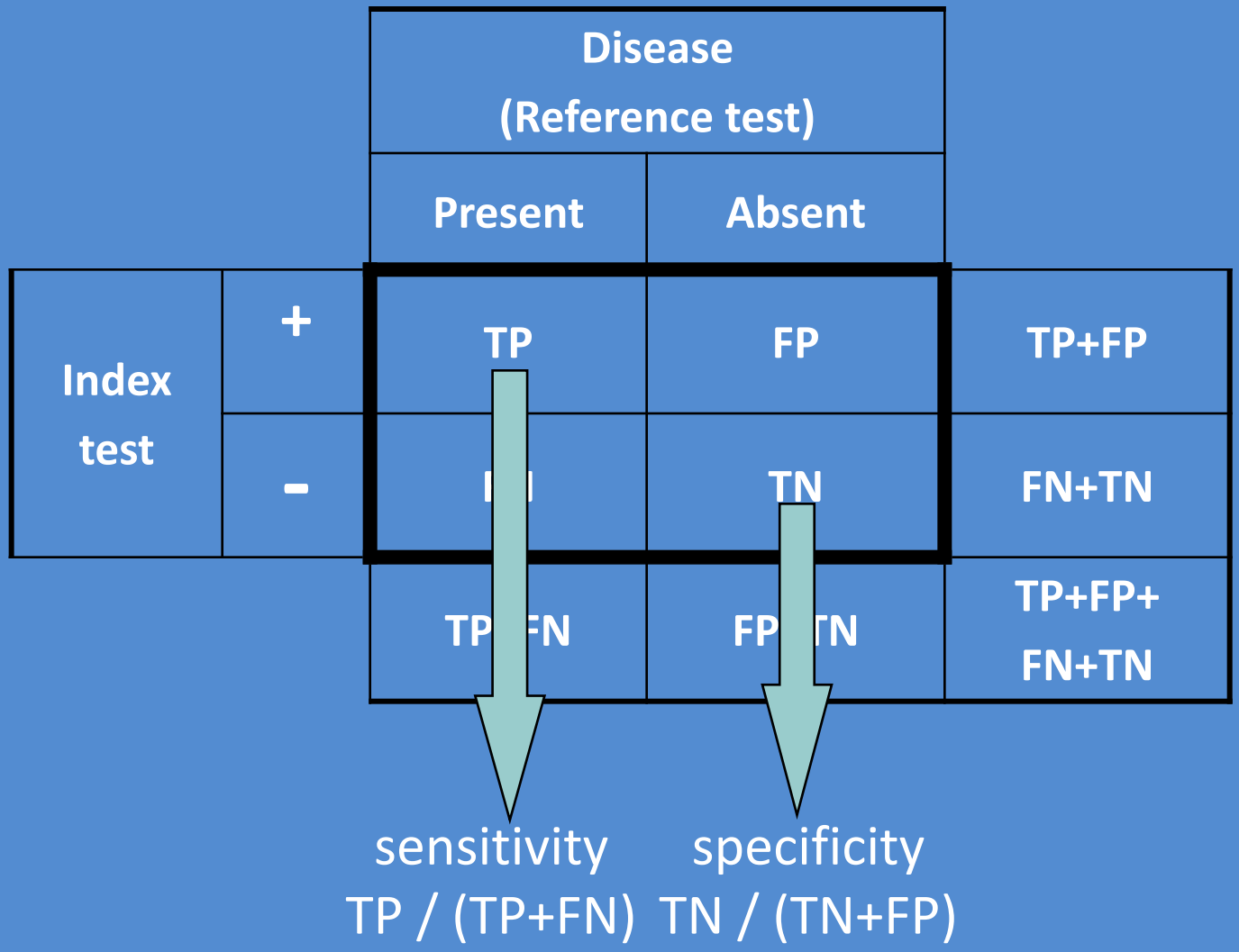
Systematic reviews of diagnostic test accuracy

Karen R Steingart, MD, MPH
Madhukar Pai, MD, PhD

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

2x2 Table – sensitivity and specificity



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

Test accuracy may not capture clinical impact

Table 1| **Attributes of the test-treatment pathway that affect patient health**

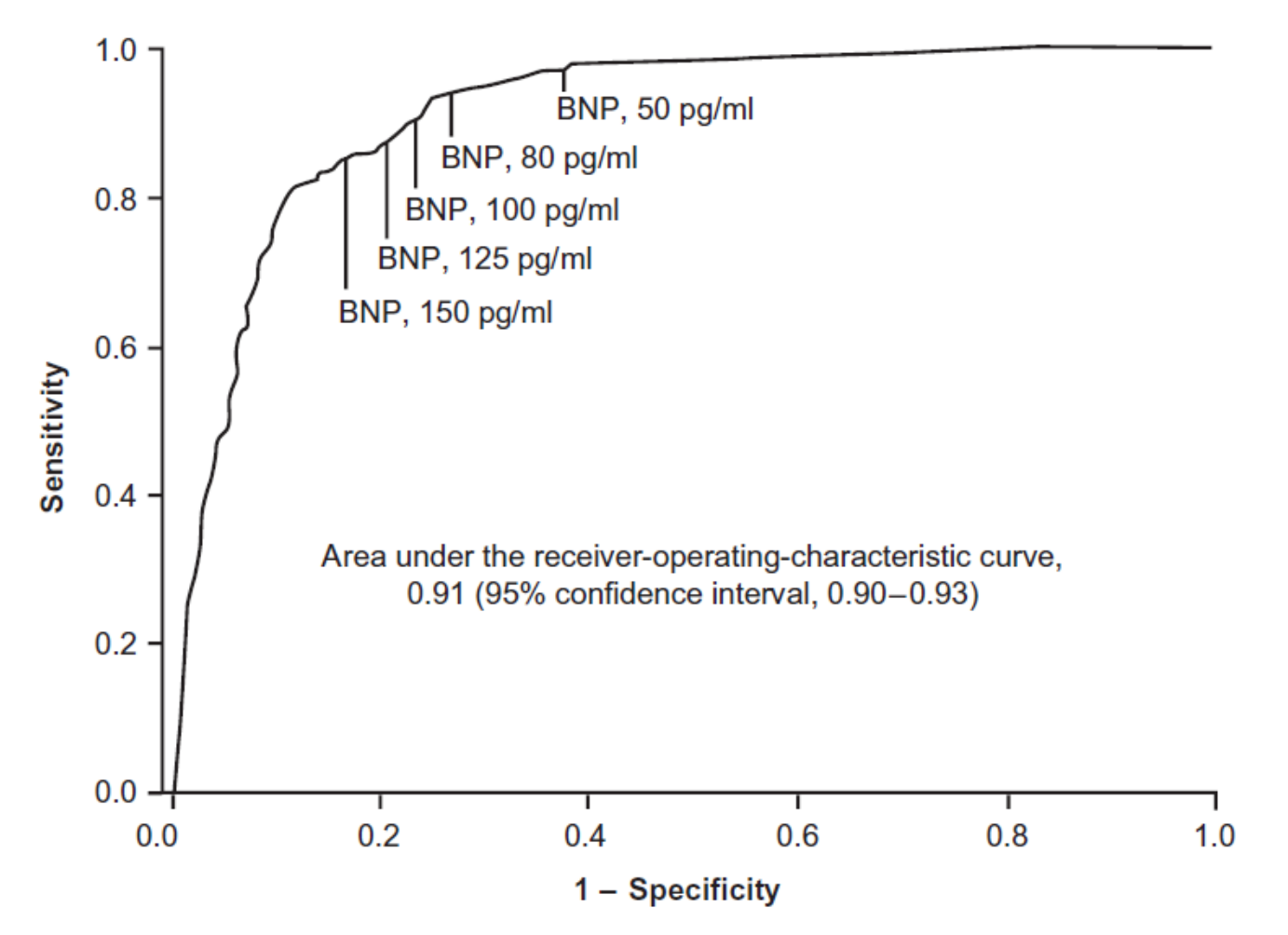
Pathway component and mechanism	Definition
(1) Diagnostic test delivered	
Timing of test	Speed with which a test is performed within the management strategy
Feasibility	Completion of test process. Reasons for non-completion are: patient acceptability (patient's refusal to have test), test was contraindicated (clinical reason not to administer test), and technical failure (ability of diagnostic equipment to produce data)
Test process	Patients' interaction with test procedure, potentially causing physical or psychological harms or benefits
(2) Test result produced	
Interpretability	Degree to which test data can be used to inform a diagnostic classification
Accuracy	Ability of a test to distinguish between patients who have disease and those who do not
Timing of results	Speed with which test results are available
(3) Diagnosis made	
Timing of diagnosis	Speed with which a diagnostic decision is made
Diagnostic yield	Degree to which the test contributes to a patient diagnosis in any form, including: provision of a definitive diagnosis, confirmation of a suspected diagnosis, ruling out a working diagnosis, and distinguishing between alternative diagnoses with different treatment implications. Diagnostic yield is different from accuracy because it also incorporates any other information used by a doctor to make a diagnosis (such as previous test results)
Diagnostic confidence	Degree of confidence that doctors and patients have in the validity or applicability of a test result
(4) Management decided	
Therapeutic yield	Degree to which diagnostic decisions affect treatment plans
Therapeutic confidence	Certainty with which doctors and patients pursue a course of treatment
(5) Treatment implemented	
Timing of treatment	Speed with which patients receive treatment
Treatment efficacy	Ability of the treatment intervention to improve patient outcomes
Adherence	Extent to which patients participate in the management plan, as advised by their doctor, to attain therapeutic goal

di Ruffano, BMJ 2012

Clinical impact of test results on diagnostic and treatment decisions, and eventually, patient outcomes

"Improved accuracy is not always a necessary prerequisite for improving patient health, nor does it guarantee other downstream improvements"
[di Ruffano et al. *BMJ* 2012;344:e686]

Accuracy vs Impact:
Rapid measurement of B-type natriuretic peptide
in the emergency diagnosis of heart failure



Maisel et al, *N Engl J Med*. 2002 Jul 18;**347**(3):

Annals of Internal Medicine | ARTICLE

B-Type Natriuretic Peptide Testing, Clinical Outcomes, and Health Services Use in Emergency Department Patients With Dyspnea

A Randomized Trial

Hans-Gerhard Schneider, MBBS, MD; Louisa Lam, MPH; Amaali Lokuge, MBBS; Henry Krum, MBBS, PhD; Matthew T. Naughton, MBBS; Pieter De Villiers Smit, MBBS; Adam Bystrzycki, MBBS; David Eccleston, MBBS, PhD; Jacob Federman, MBBS; Genevieve Flannery, MBBS; and Peter Cameron, MBBS, MD

Background: B-type natriuretic peptide (BNP) is used to diagnose heart failure, but the effects of using the test on all dyspneic patients is uncertain.

Objective: To assess whether BNP testing alters clinical outcomes and health services use of acutely dyspneic patients.

Design: Randomized, single-blind study. Patients were assigned to a treatment group through randomized numbers in a sealed envelope. Patients were blinded to the intervention, but clinicians and those who assessed trial outcomes were not.

Setting: 2 Australian teaching hospital emergency departments.

Patients: 612 consecutive patients who presented with acute severe dyspnea from August 2005 to March 2007.

Intervention: BNP testing ($n = 306$) or no testing ($n = 306$).

Measurements: Admission rates, length of stay, and emergency department medications (primary outcomes); mortality and readmission rates (secondary outcomes).

Results: There were no between-group differences in hospital admission rates (85.6% [BNP group] vs. 86.6% [control group]; difference, −1.0 percentage point [95% CI, −6.5 to 4.5 percentage points]; $P = 0.73$), length of admission (median, 4.4 days [interquartile range, 2 to 9 days] vs. 5.0 days [interquartile range, 2 to 9 days]; $P = 0.94$), or management of patients in the emergency department. Test discrimination was good (area under the receiver-operating characteristic curve, 0.87 [CI, 0.83 to 0.91]). Adverse events were not measured.

Limitation: Most patients were very short of breath and required hospitalization; the findings might not apply for evaluating patients with milder degrees of breathlessness.

Conclusion: Measurement of BNP in all emergency department patients with severe shortness of breath had no apparent effects on clinical outcomes or use of health services. The findings do not support routine use of BNP testing in all severely dyspneic patients in the emergency department.

Primary Funding Source: Janssen-Cilag.

Ann Intern Med. 2009;150:365-371. www.annals.org
For author affiliations, see end of text.
ClinicalTrials.gov registration number: NCT00163709.

Example: POC CD4 counts

Accuracy

Evaluation of the PIMA Point-of-Care CD4 Analyzer in VCT Clinics in Zimbabwe

Sekesai Mtapuri-Zinyowera, PhD, MSc, Memory Chideme, BSc, MSc,* Douglas Mangwanya, BSc, MSc,† Owen Mugurungi, MD, MSc,† Stephano Gudukeya, BSc,‡ Karin Hatzold, MD, MPH,‡ Alexio Mangwiro, BSc,§ Gaurav Bhattacharya, MD, MPH,§ Jonathan Lehe, BA,§ and Trevor Peter, PhD, MPH§*

Abstract: Point-of-care (POC) CD4 testing was implemented at a stand-alone HIV voluntary testing and counseling centre in Harare, Zimbabwe. To validate the use of this new technology, paired blood samples were collected from 165 patients either by a nurse or a laboratory technician and tested using POC and conventional laboratory CD4 machines. Finger prick (capillary) blood was collected directly into the PIMA POC CD4 Analyzer cartridges and tested immediately, whereas venous blood collected into evacuated tubes was used for CD4 enumeration on a Becton Dickinson FACSCalibur. There was no significant difference in mean absolute CD4 counts between the POC PIMA and Becton Dickinson FACSCalibur platforms (+7.6 cells/μL; $P = 0.72$). Additionally, there was no significant difference in CD4 counts between the platforms when run by either a nurse (+18.0 cells/μL; $P = 0.49$), or a laboratory technicians (−3.1 cells/μL; $P = 0.93$). This study demonstrates that POC CD4 testing can be conducted in a voluntary testing and counseling setting for staging HIV-positive clients. Both nurses and laboratory technicians performed the test accurately, thereby increasing the human resources available for POC CD4 testing. By producing same-day results, POC CD4 facilitates immediate decision-making, patient management and referral and may help improve patient care and retention. POC CD4 may also alleviate testing burdens at traditional central CD4 laboratories, hence improving test access in both rural and urban environments.

Key Words: CD4, HIV, diagnosis, client-initiated testing, laboratory, PIMA, point-of-care, voluntary counseling and testing, VCT

(*J Acquir Immune Defic Syndr* 2010;55:1–7)

BACKGROUND

CD4 T-lymphocyte count is an important qualifying test for antiretroviral treatment (ART) in HIV-positive individuals and is also used to monitor treatment efficacy.[1–7] The scale up of public ART programs globally[8] has led to an increased demand for CD4 count tests, especially to assess treatment eligibility. Despite expansion of laboratory infrastructure and services, access to CD4 testing remains a bottleneck to ART scale-up. In Zimbabwe, an estimated 380,000 adults are in need of ART[9] and, by the end of 2009, an estimated 215,000 were on ART within the public sector.[10] There is clearly a need to increase access to ART services and improving CD4 access may help.

In Zimbabwe, the "New Start" voluntary testing and counseling (VCT) centers (also known as client-initiated testing and counselling centers) are established by the Ministry of Health and Child Welfare in partnership with Population Services International (PSI) and provide free rapid HIV testing services to more than 360,000 clients nationwide on an annual basis. Clients testing positive at VCT centers are then referred to Opportunistic Infection (OI) clinics for HIV care and ART if eligible. After enrollment at the OI clinics, patients are scheduled for a CD4 count test. Due to high demand, delays in CD4 testing can occur for 2–3 weeks on average. There is substantial loss-to-follow-up of patients between HIV diagnosis and registration at the OI clinics and delays in CD4 testing can result in further loss of patients who do not return or who die before initiating treatment. The situation is exacerbated in rural areas where more limited CD4 access creates a significant bottleneck to the scale up of ART.

Example: POC CD4 counts

Clinical impact

Effect of point-of-care CD4 cell count tests on retention of patients and rates of antiretroviral therapy initiation in primary health clinics: an observational cohort study

Ilesh V Jani, Nádia E Sitoe, Eunice R Alfai, Patrina L Chongo, Jorge I Quevedo, Beatriz M Rocha, Jonathan D Lehe, Trevor F Peter

Background Loss to follow-up of HIV-positive patients before initiation of antiretroviral therapy can exceed 50% in low-income settings and is a challenge to the scale-up of treatment. We implemented point-of-care counting of CD4 cells in Mozambique and assessed the effect on loss to follow-up before immunological staging and treatment initiation.

Methods In this observational cohort study, data for enrolment into HIV management and initiation of antiretroviral therapy were extracted retrospectively from patients' records at four primary health clinics providing HIV treatment and point-of-care CD4 services. Loss to follow-up and the duration of each preparatory step before treatment initiation were measured and compared with baseline data from before the introduction of point-of-care CD4 testing.

Findings After the introduction of point-of-care CD4 the proportion of patients lost to follow-up before completion of CD4 staging dropped from 57% (278 of 492) to 21% (92 of 437) (adjusted odds ratio [OR] 0·2, 95% CI 0·15–0·27). Total loss to follow-up before initiation of antiretroviral treatment fell from 64% (314 of 492) to 33% (142 of 437) (OR 0·27, 95% CI 0·21–0·36) and the proportion of enrolled patients initiating antiretroviral therapy increased from 12% (57 of 492) to 22% (94 of 437) (OR 2·05, 95% CI 1·42–2·96). The median time from enrolment to antiretroviral therapy initiation reduced from 48 days to 20 days ($p<0·0001$), primarily because of a reduction in the median time taken to complete CD4 staging, which decreased from 32 days to 3 days ($p<0·0001$). Loss to follow-up between staging and antiretroviral therapy initiation did not change significantly (OR 0·84, 95% CI 0·49–1·45).

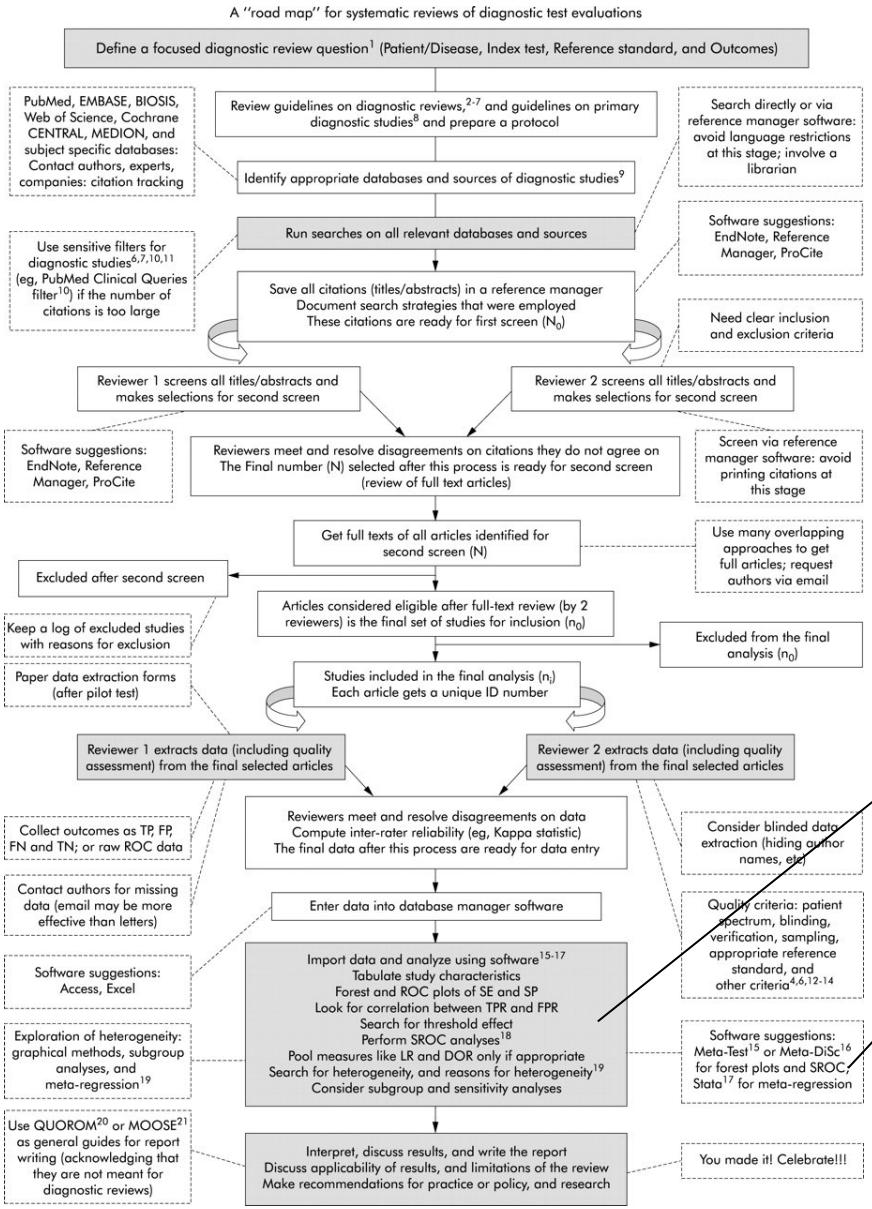
Interpretation Point-of-care CD4 testing enabled clinics to stage patients rapidly on-site after enrolment, which reduced opportunities for pretreatment loss to follow-up. As a result, more patients were identified as eligible for and initiated antiretroviral treatment. Point-of-care testing might therefore be an effective intervention to reduce pretreatment loss to follow-up.

Published Online September 26, 2011 DOI:10.1016/S0140-6736(11)61052-0

Instituto Nacional da Saúde, Maputo, Mozambique (I V Jani MD, N E Sitoe BSc, E R Alfai BSc, P L Chongo BSc); **and Clinton Health Access Initiative, Maputo, Mozambique** (J I Quevedo BSc, B M Rocha BA, J D Lehe BA, T F Peter PhD)

Correspondence to: Dr Ilesh V Jani, Instituto Nacional da Saúde, Av Eduardo Mondlane 1008, 2nd floor, Maputo, Mozambique **ilesh.jani@gmail.com**

Road map for diagnostic accuracy reviews



Use PRISMA for reporting of SR

Avoid simple pooling of sens and spec; need to use HSROC or bivariate random effects models; do not use funnel plots for publication bias

Stata (metandi command) for bivariate random effects pooling + HSROC)

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

EBM

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

Population

Intervention

Comparison

Outcome

+ Study design

+ Purpose of the test/strategy

+ Reference standard

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

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

Review | Annals of Internal Medicine

Accuracy of Rapid Influenza Diagnostic Tests

A Meta-analysis

Caroline Chartrand, MD, MSc; Mariska M.G. Leeflang, DVM, PhD; Jessica Minion, MD, MSc; Timothy Brewer, MD, MPH; and Madhukar Pai, MD, PhD

Purpose: To examine the accuracy of rapid influenza diagnostic tests (RIDTs) in adults and children with influenza-like illness and evaluate factors associated with higher accuracy.

OPEN ACCESS Freely available online | PLoS MEDICINE

Commercial Serological Tests for the Diagnosis of Active Pulmonary and Extrapulmonary Tuberculosis: An Updated Systematic Review and Meta-Analysis

Karen R. Steingart[1], Laura L. Flores[2,3], Nandini Dendukuri[4], Ian Schiller[4], Suman Laal[5,6,7], Andrew Ramsay[8], Philip C. Hopewell[2,3], Madhukar Pai[4]*

Types of studies. Diagnostic studies (with any study design) were included that evaluated serological tests for active TB (pulmonary and extrapulmonary TB) in patients who provided sera before or within 14 d of starting antituberculous treatment.

Participants. The participants constituted adults and children, with and without HIV infection, with suspected or confirmed active TB, from all clinical settings (clinic or hospital). The protocol for the current review included studies with at least ten TB cases. Studies could be performed in any country regardless of TB incidence or income status.

Index test. The index test was any commercial serological test for the diagnosis of active TB.

Comparator tests. There was either no test or smear microscopy used for comparison.

Target conditions. The target conditions were pulmonary and extrapulmonary TB.

Reference standards. Pulmonary TB required positivity on mycobacterial culture. (The previous review accepted positivity on either culture or smear microscopy as the reference standard [12].) Extrapulmonary TB required positivity on at least one of the following tests: culture, smear, or histopathological examination.

Outcomes. The outcomes were sensitivity and specificity.

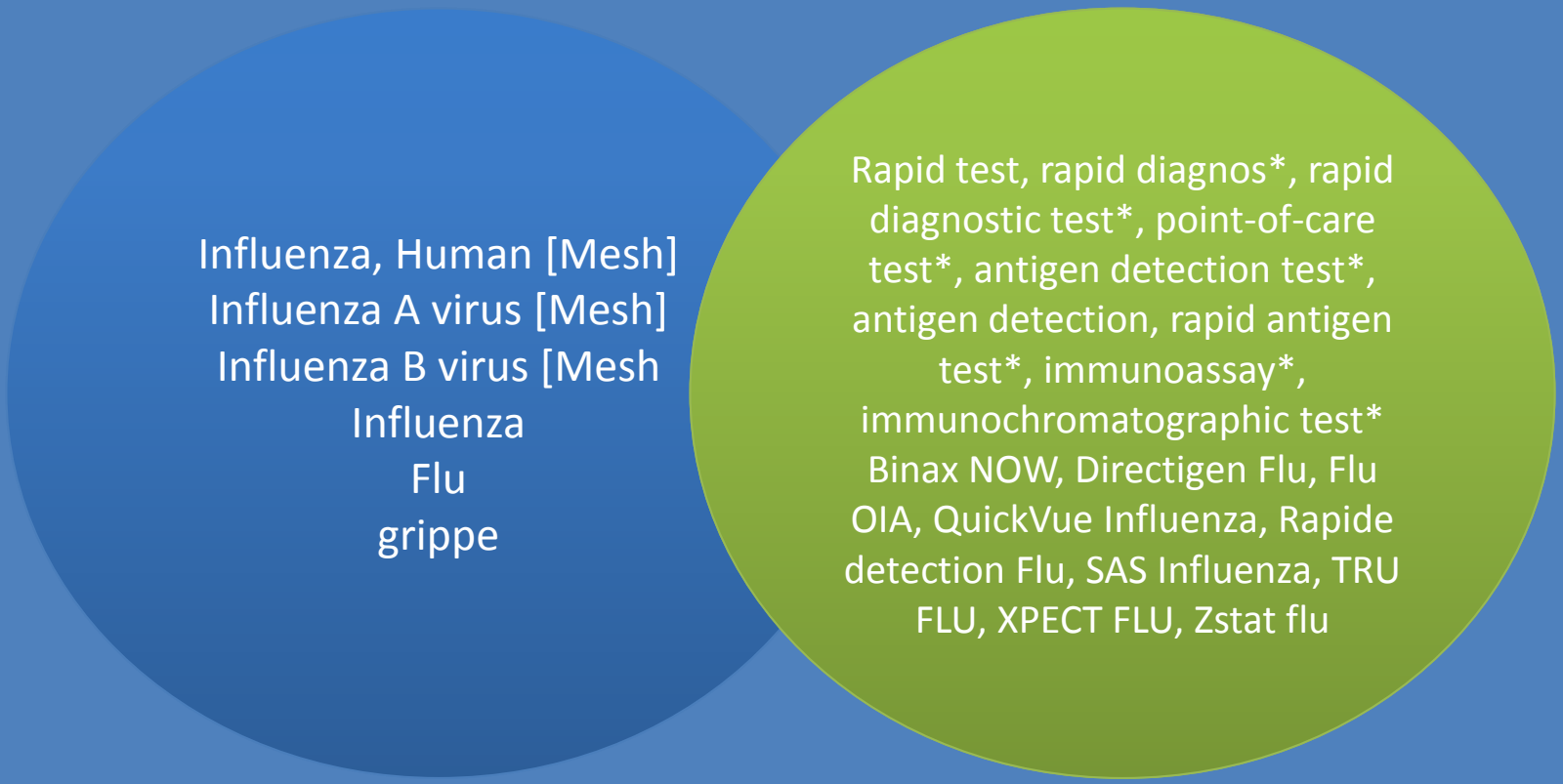
2. Searching for studies

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

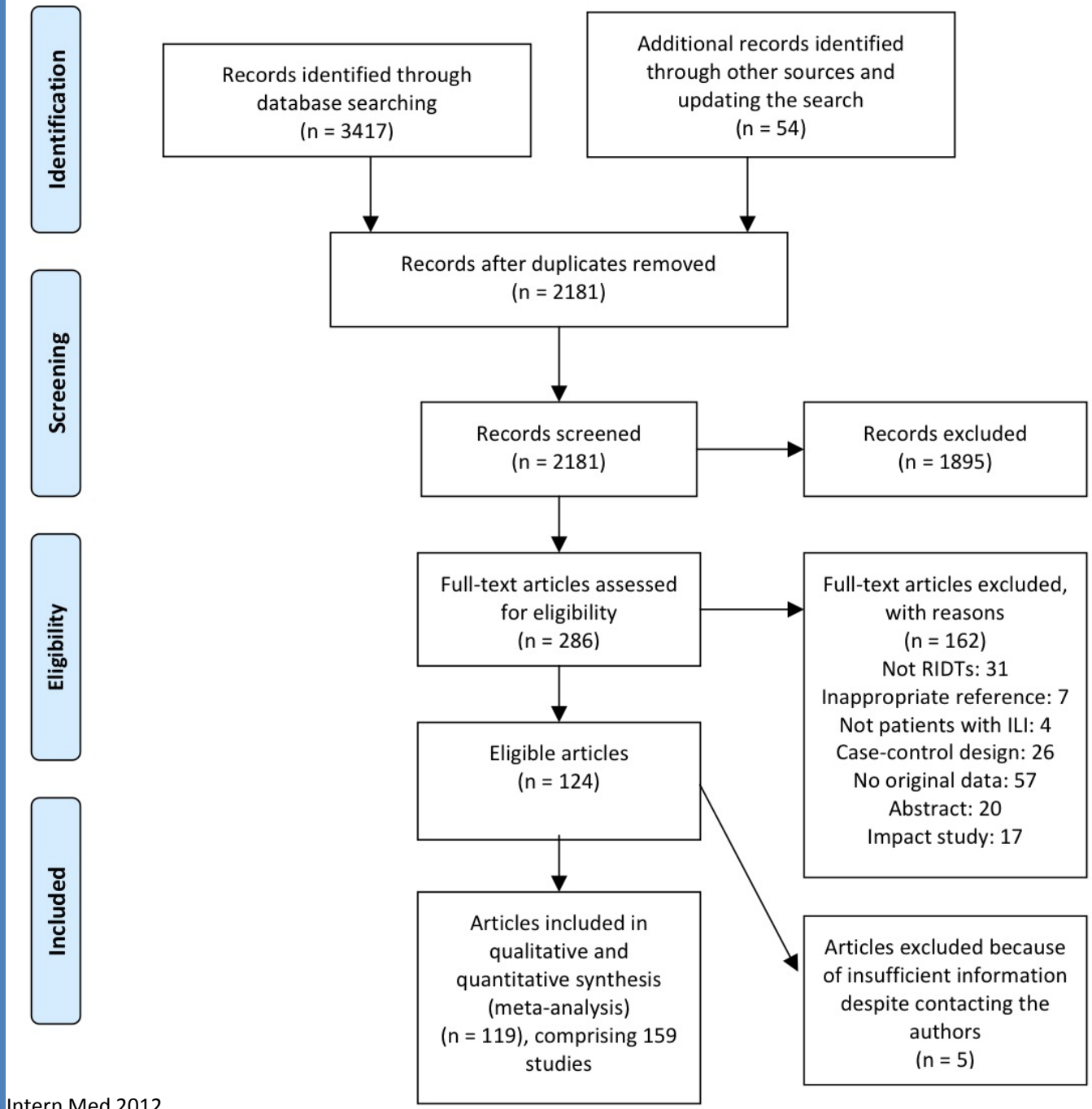
Bossuyt PM, Leeflang MM. *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 0.4 [updated September* 2008]. The Cochrane Collaboration, 2008

Influenza rapid tests: Search strategy



Databases: MEDLINE via Pubmed, EMBASE, Biosis et Web of Science
March 2010, updated december 2011

Chartrand C et al. Annals of Int Med 2012



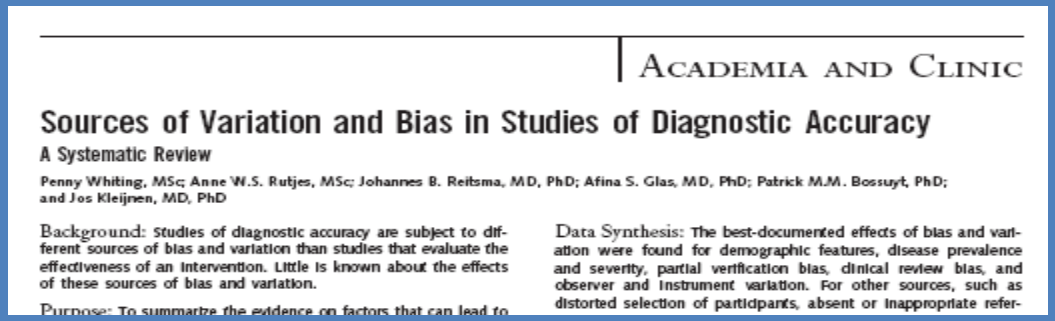
Chartrand C et al. Ann Intern Med 2012

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

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

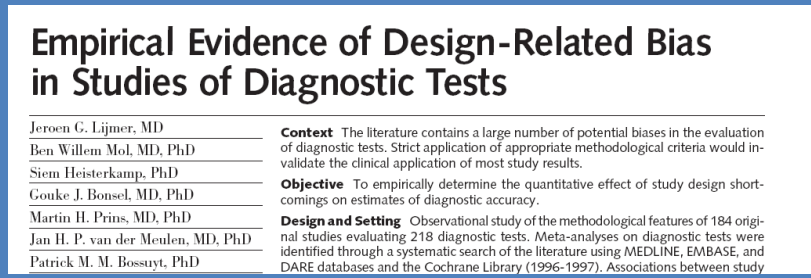
A Systematic Review

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

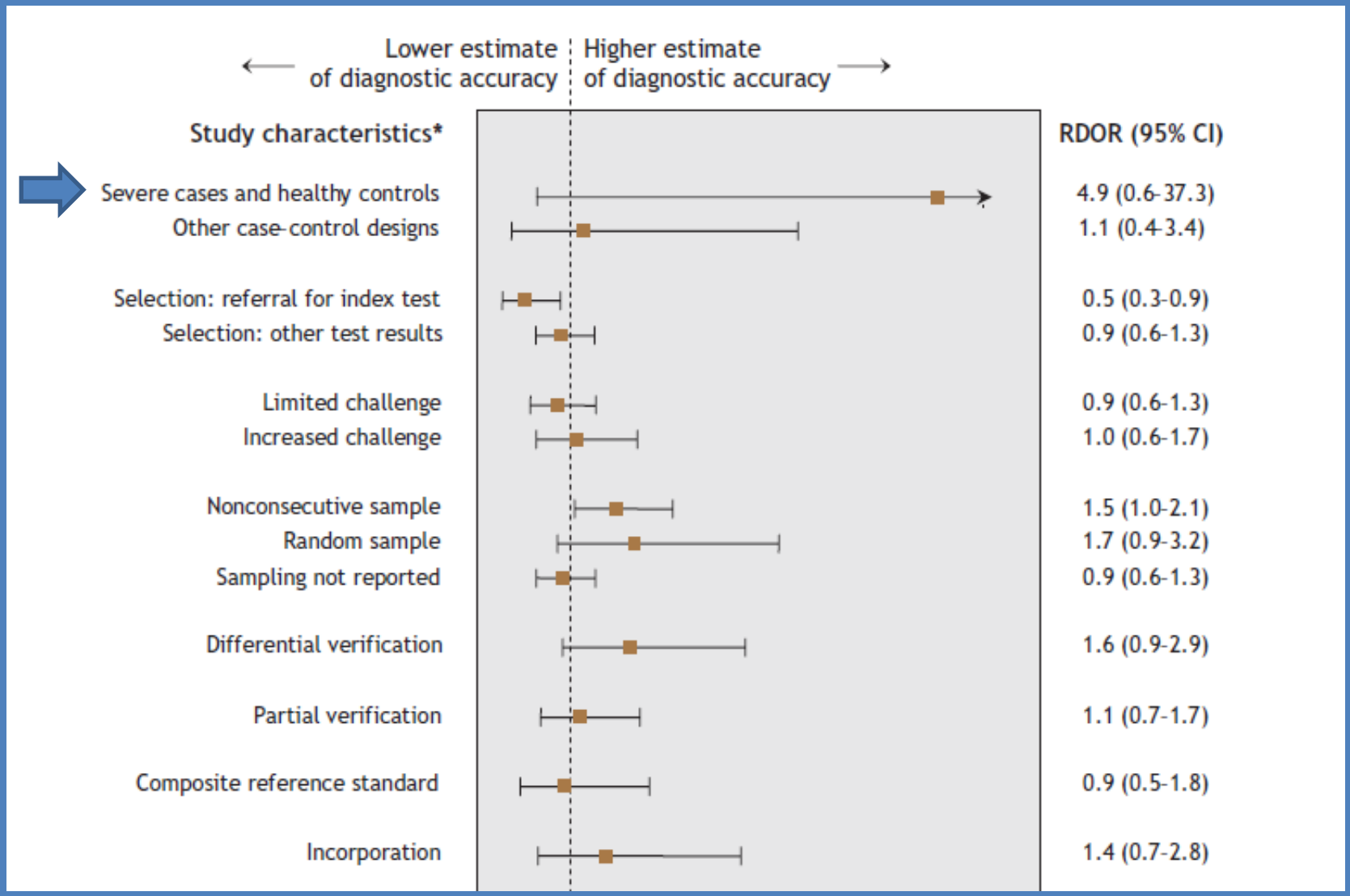
RESEARCH

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

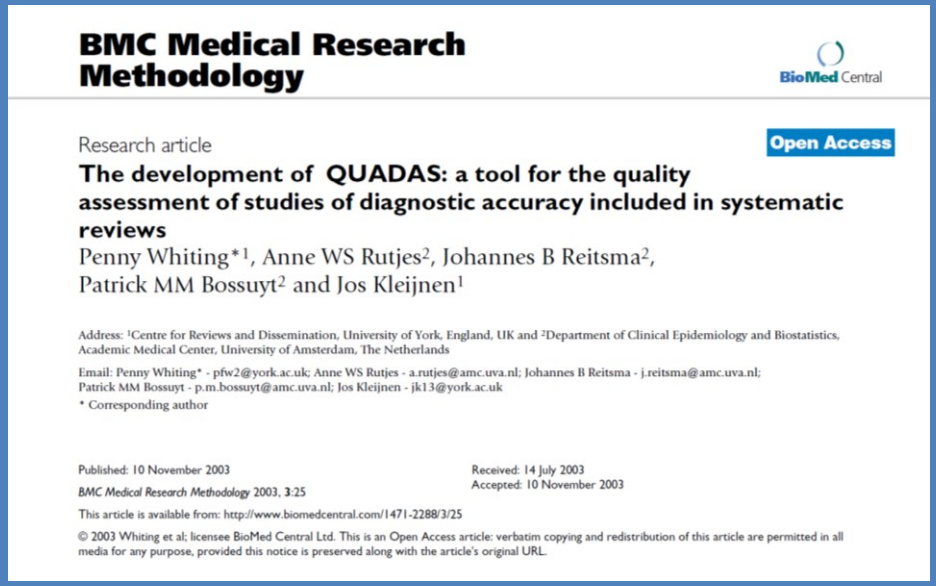
The Lancet Infect Dis 2003

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. †χ^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

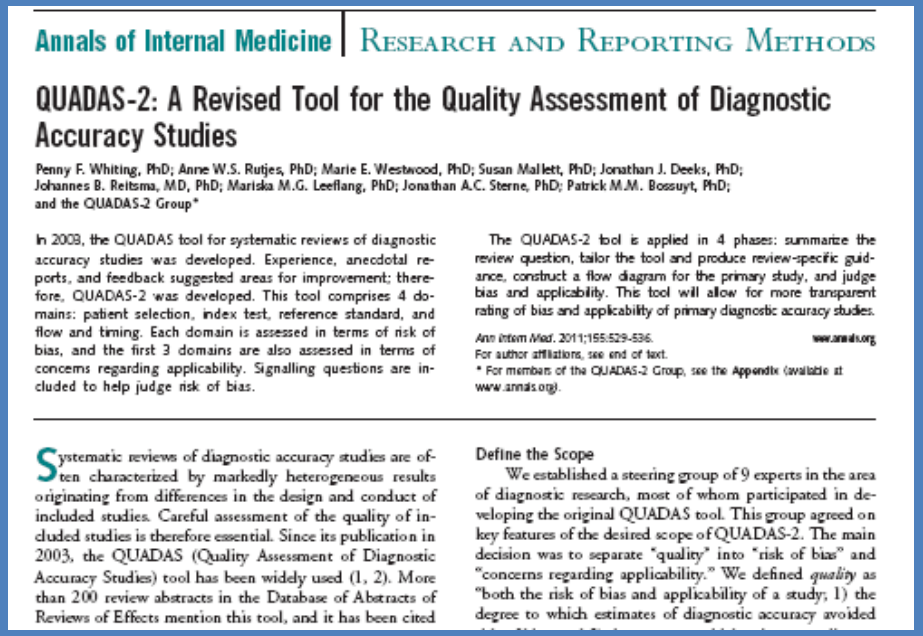
Published: 10 November 2003

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, MD, 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

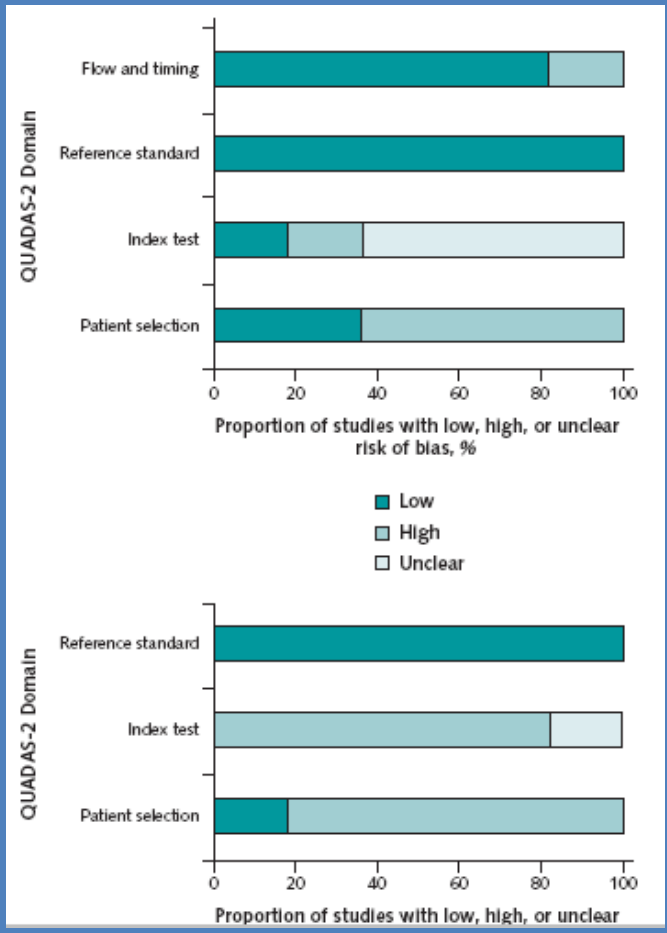
QUADAS-2, 2011

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/

In general, diagnostic studies are poorly done and reported (contacting authors is helpful)

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.

4. Analyzing the data; undertaking meta-analyses

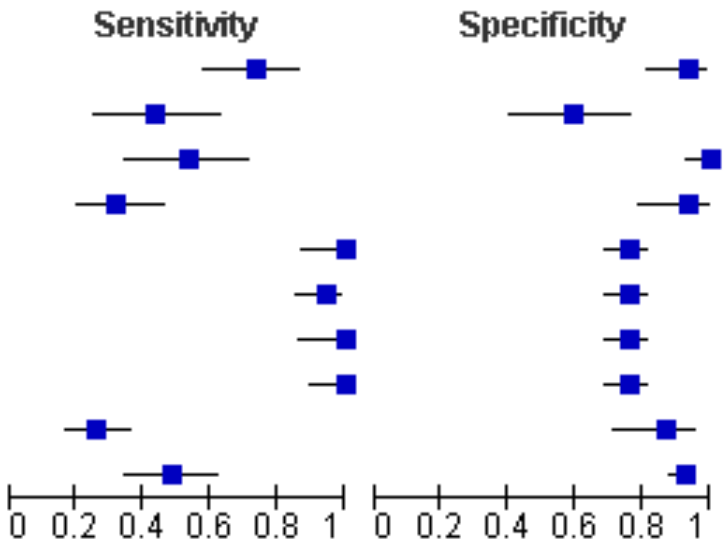
Key steps

- 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

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

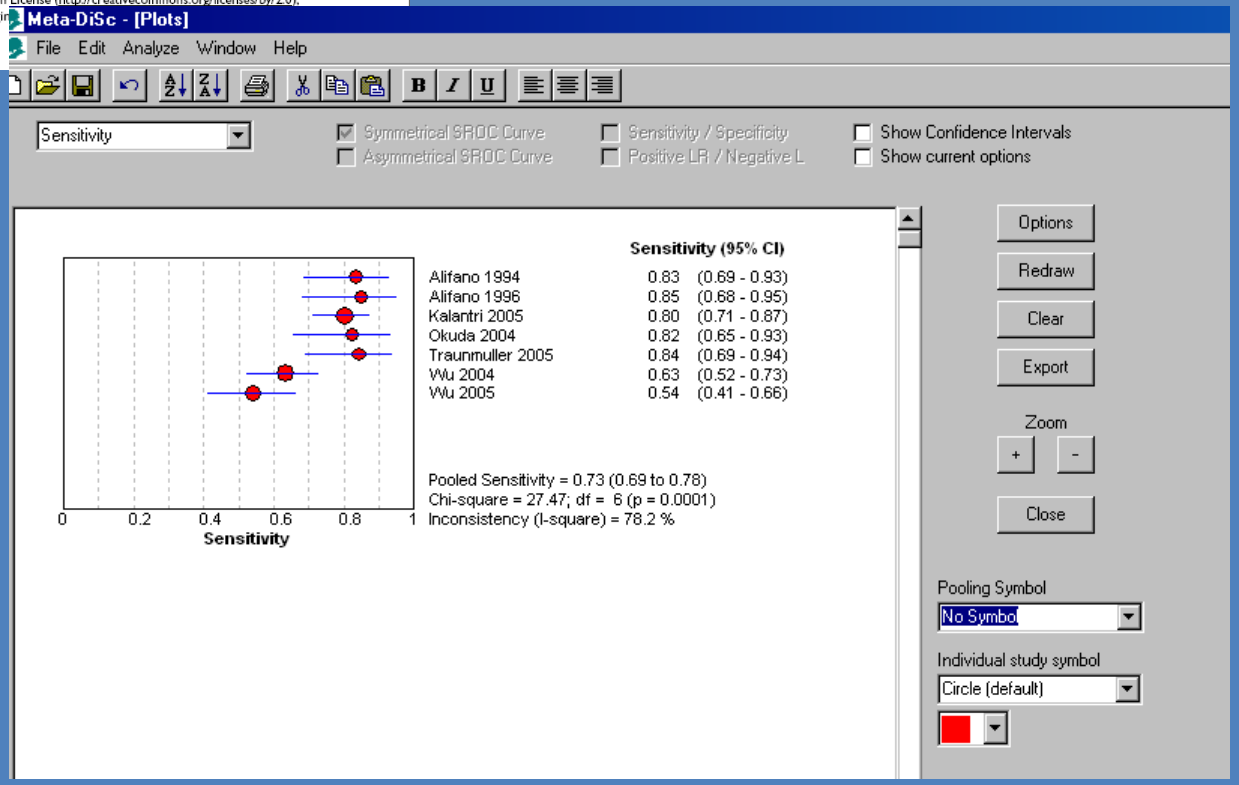
* Corresponding author

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



Statistical models for meta-analysis of diagnostic studies

- Simple, separate pooling of sens and spec should not be done
- Two recommended 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

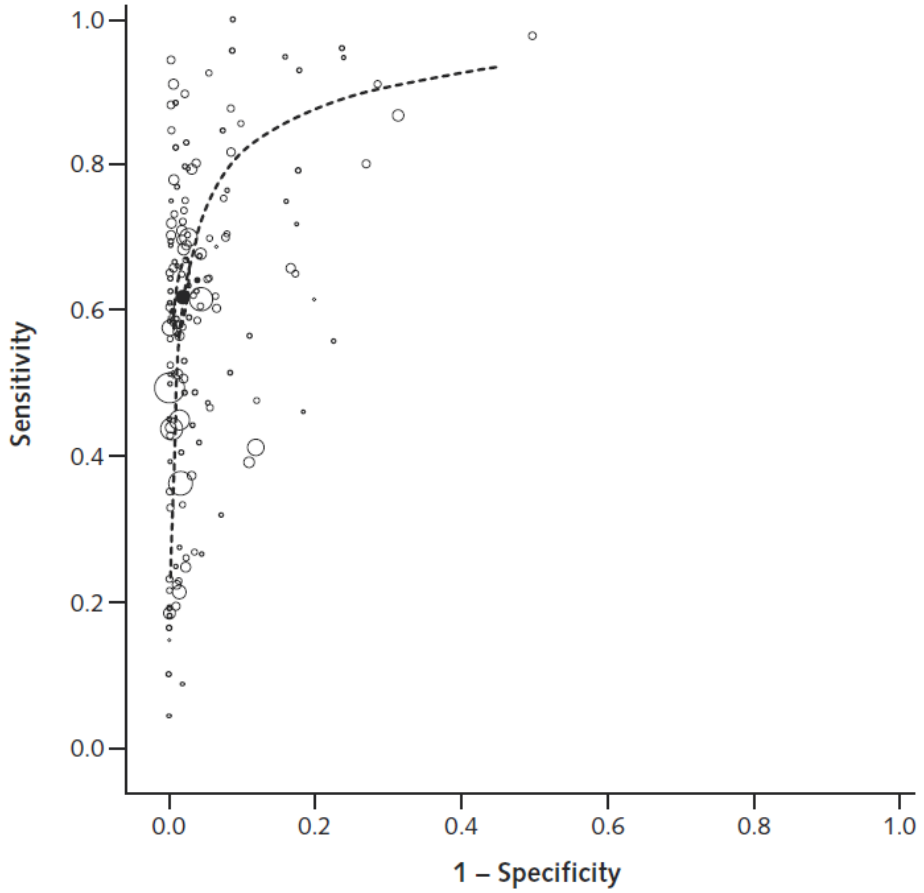
Bivariate model vs HSROC model

- Where studies report a common threshold (or cut-off) for a positive result, use the bivariate model

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

Influenza rapid tests

Figure 2. **Hierarchical summary receiver-operating characteristic curve plot of rapid influenza diagnostic test studies.**



Sensitivity: 62.3% (57.9 – 66.6)
Specificity: 98.2% (97.5 – 98.7)
LR+: 34.5 (23.8 – 45.2)
LR-: 0.38 (0.34 – 0.43)

Chartrand C et al. Ann Intern Med 2012

Stata command, metandi

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

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

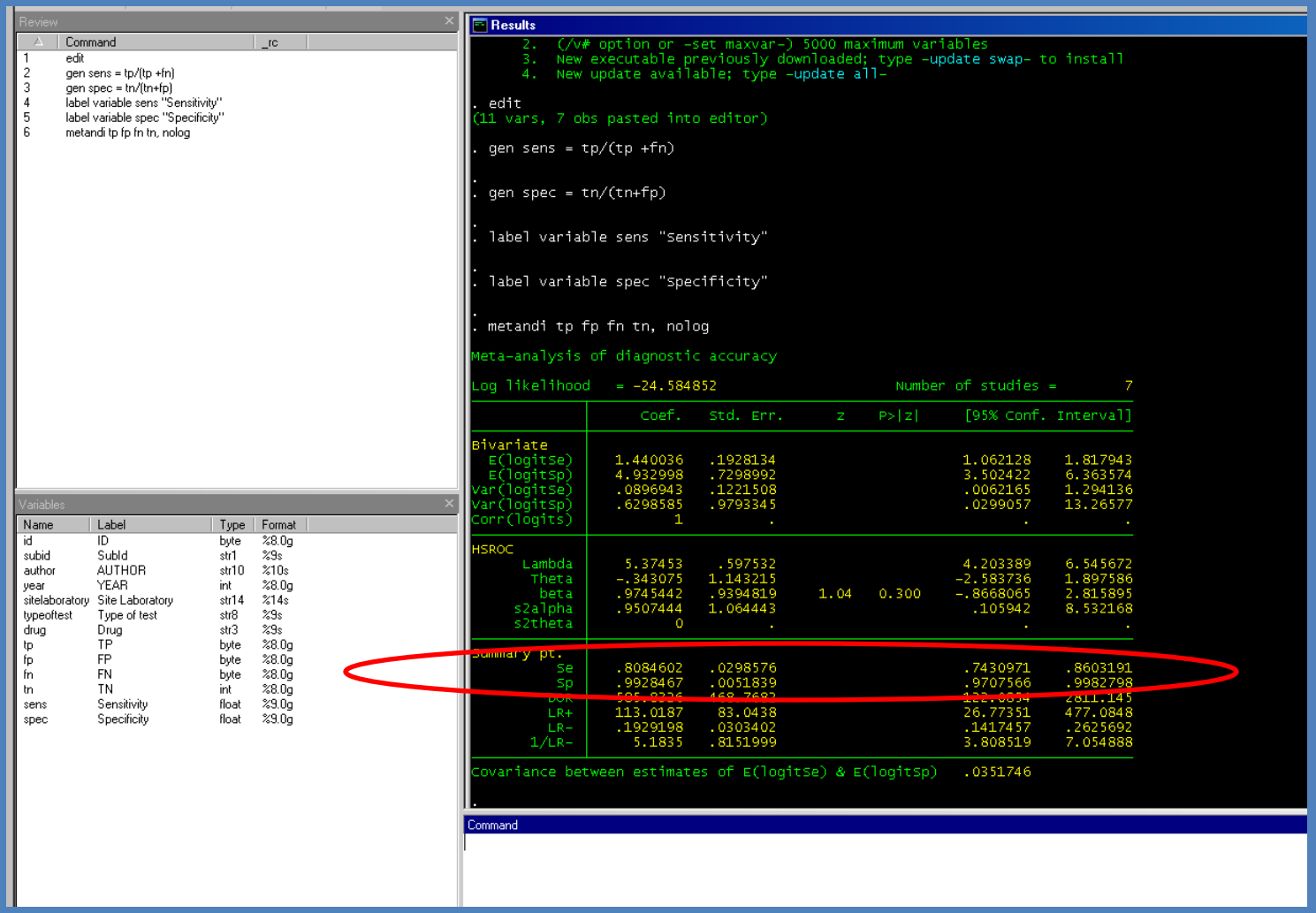
Roger M. Harbord
Department of Social Medicine
University of Bristol
Bristol, UK
roger.harbord@bristol.ac.uk

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 2×2 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

Stata output



Review

	Command	_rc
1	edit	
2	gen sens = tp/(tp +fn)	
3	gen spec = tn/(tn+fp)	
4	label variable sens "Sensitivity"	
5	label variable spec "Specificity"	
6	metandi tp fp fn tn, nolog	

Variables

Name	Label	Type	Format
id	ID	byte	%8.0g
subid	SubId	str1	%9s
author	AUTHOR	str10	%10s
year	YEAR	int	%8.0g
sitelaboratory	Site Laboratory	str14	%14s
typeoftest	Type of test	str8	%9s
drug	Drug	str3	%9s
tp	TP	byte	%8.0g
fp	FP	byte	%8.0g
fn	FN	byte	%8.0g
tn	TN	int	%8.0g
sens	Sensitivity	float	%9.0g
spec	Specificity	float	%9.0g

Results

```
      2.  (/v# option or -set maxvar-) 5000 maximum variables
      3.  New executable previously downloaded; type -update swap- to install
      4.  New update available; type -update all-

. edit
(11 vars, 7 obs pasted into editor)

. gen sens = tp/(tp +fn)

.
. gen spec = tn/(tn+fp)

.
. label variable sens "Sensitivity"

.
. label variable spec "Specificity"

.
. metandi tp fp fn tn, nolog

Meta-analysis of diagnostic accuracy

Log likelihood   = -24.584852                    Number of studies =         7
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Bivariate						
E(logitSe)	1.440036	.1928134			1.062128	1.817943
E(logitSp)	4.932998	.7298992			3.502422	6.363574
Var(logitSe)	.0896943	.1221508			.0062165	1.294136
Var(logitSp)	.6298585	.9793345			.0299057	13.26577
Corr(logits)	1	.			.	.
HSROC						
Lambda	5.37453	.597532			4.203389	6.545672
Theta	-.343075	1.143215			-2.583736	1.897586
beta	.9745442	.9394819	1.04	0.300	-.8668065	2.815895
s2alpha	.9507444	1.064443			.105942	8.532168
s2theta	0	.			.	.
Summary pt.						
Se	.8084602	.0298576			.7430971	.8603191
Sp	.9928467	.0051839			.9707566	.9982798
DOR	585.8336	468.7683			122.8854	2811.145
LR+	113.0187	83.0438			26.77351	477.0848
LR-	.1929198	.0303402			.1417457	.2625692
1/LR-	5.1835	.8151999			3.808519	7.054888

```
Covariance between estimates of E(logitSe) & E(logitSp)   .0351746

.
```

Command

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

Heterogeneity: very common in diagnostic SRs

- 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

Example: subgroup analysis

Table 2. **Accuracy Estimates From Subgroup Analyses**

Characteristic	Pooled Sensitivity (95% CI), %	*P* Value	Pooled Specificity (95% CI), %	*P* Value
Population				
Children (60 studies)	66.6 (61.6–71.7)	<0.001	98.2 (97.5–99.0)	0.135
Adults (33 studies)	53.9 (47.9–59.8)	Reference	98.6 (98.0–98.9)	Reference
Virus type				
Influenza A (72 studies)	64.6 (59.0–70.1)	0.62	99.1 (98.7–99.4)	<0.001
Influenza B (27 studies)	52.2 (45.0–59.3)	0.050	99.8 (99.7–99.9)	<0.001
Influenza A and B (47 studies)	62.3 (55.2–69.4)	Reference	96.1 (94.4–97.8)	Reference
Study conducted during the H1N1 pandemic				
Yes (41 studies)	56.3 (48.7–63.9)	0.065	98.9 (98.3–99.5)	0.022
No (74 studies)	65.0 (59.7–70.4)	Reference	97.5 (96.6–98.5)	Reference
Index test*				
BinaxNOW (17 studies)†	57.0 (45.9–67.5)	0.028‡	98.6 (96.9–99.3)	0.057‡
Directigen Flu A (10 studies)	76.7 (63.8–86.0)	0.49‡	97.2 (92.6–99.0)	0.62‡
Directigen Flu A+B (30 studies)	57.2 (48.8–65.2)	0.011‡	99.3 (98.8–99.6)	<0.001‡
QuickVue Influenza (16 studies)	69.0 (58.1–78.2)	0.66‡	95.8 (91.3–98.0)	0.82‡
QuickVue Influenza A+B (21 studies)	48.8 (39.0–58.8)	<0.001‡	98.4 (96.8–99.2)	0.064‡
Reference standard				
RT-PCR (67 studies)	53.9 (48.2–59.6)	<0.001	98.8 (98.3–99.3)	0.002
Culture (48 studies)	72.3 (66.8–77.9)	Reference	96.7 (95.2–98.3)	Reference

Chartrand C et al. Ann Intern Med 2012

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 D et al. PLoS ONE 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

Exploration of heterogeneity – urine LAM ELISA for TB

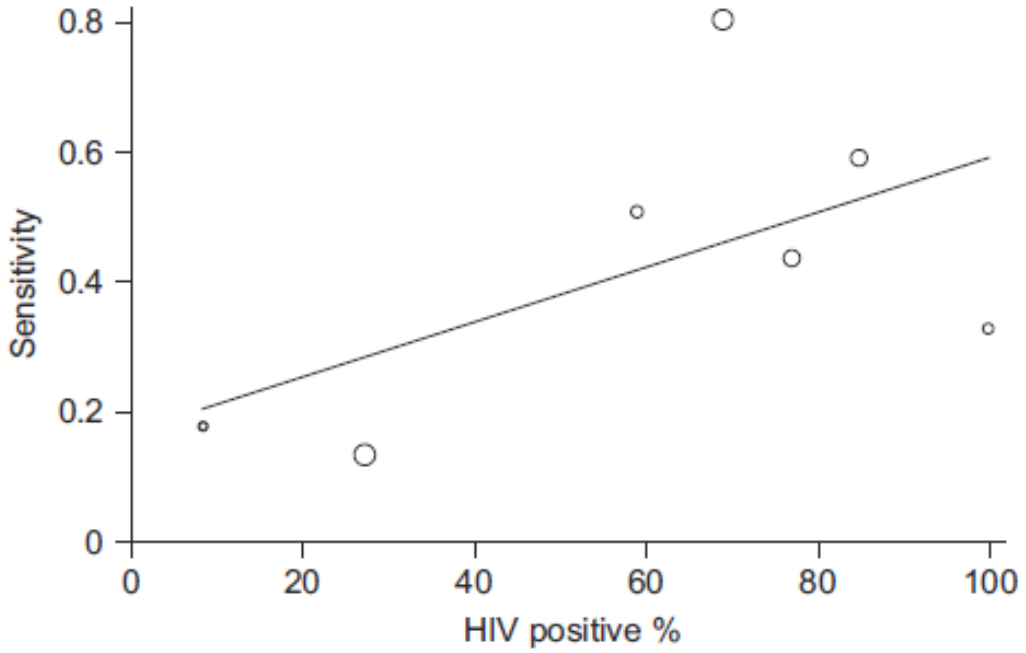


FIGURE 4. Linear regression of sensitivity on the proportion of HIV-positive subjects included in studies. Open circles represent studies reporting HIV prevalence; sizes of the circles depend on the precision of each study estimate (*i.e.* the inverse of its within-study variance). The line represents fitted values for the linear regression equation: sensitivity=0.17 (SE 0.18)+0.0042 (SE 0.0027) × %HIV. 95% CIs[#]: α= -0.30-0.64; β= -0.0027-0.011. Logistic model (not displayed): logit(sensitivity)= -1.63 (SE 0.87)+0.021 (SE 0.013) × %HIV. 95% CIs[#]: α= -3.86–0.60; β= -0.012–0.053. [#]: not statistically significant.

Minion J et al. ERJ 2011

Publication bias

- Formal assessment of publication bias using methods such as funnel plots or regression tests is not recommended for diagnostic test accuracy studies

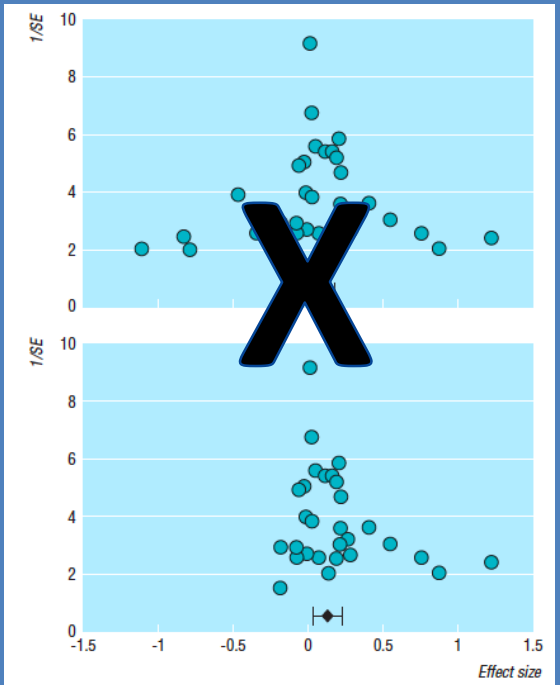


Fig 1 Typical funnel plot generated from 35 simulated studies (top) and same data with five missing studies showing a typical manifestation of publication bias (bottom)

5. Drawing robust conclusions and informative presentation of results

- summary of findings tables

Issues to discuss

- 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 were the limitations of the SR itself?
- 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 **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

Some general limitations of diagnostic SRs

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

Keeping systematic reviews updated!

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

Madhukar Pai, Lee W Riley, and John M Colford Jr

2004

Annals of Internal Medicine | ARTICLE

Meta-analysis: New Tests for the Diagnosis of Latent Tuberculosis Infection: Areas of Uncertainty and Recommendations for Research

Dick Menzies, MD, MSc; Madhukar Pai, MD, PhD; and George Comstock, MD, DrPH

2007

Annals of Internal Medicine | REVIEW

Systematic Review: T-Cell–based Assays for the Diagnosis of Latent Tuberculosis Infection: An Update

Madhukar Pai, MD, PhD; Alice Zwerling, MSc; and Dick Menzies, MD, MSc

2008

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

Molebogeng X Rangaka, Katalin A Wilkinson, Judith R Glynn, Daphne Ling, Dick Menzies, Judith Mwansa-Kambafwile, Katherine Fielding, Robert J Wilkinson, Madhukar Pai

2012

References and Tools

- Cochrane Diagnostic Test Accuracy Working Group http://srdta.cochrane.org/
  - Cochrane Handbook for Diagnostic Test Accuracy Reviews
  - Macaskill P. Analyzing and Presenting Results. Chapter 10. Cochrane Handbook for Diagnostic Test Accuracy Reviews
  - http://training.cochrane.org/authors/dta-reviews
- Leeflang. Ann Intern Med. 2008;149:889-897
- Whiting PF et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36.
- www.tbevidence.org
- RevMan http://ims.cochrane.org/revman