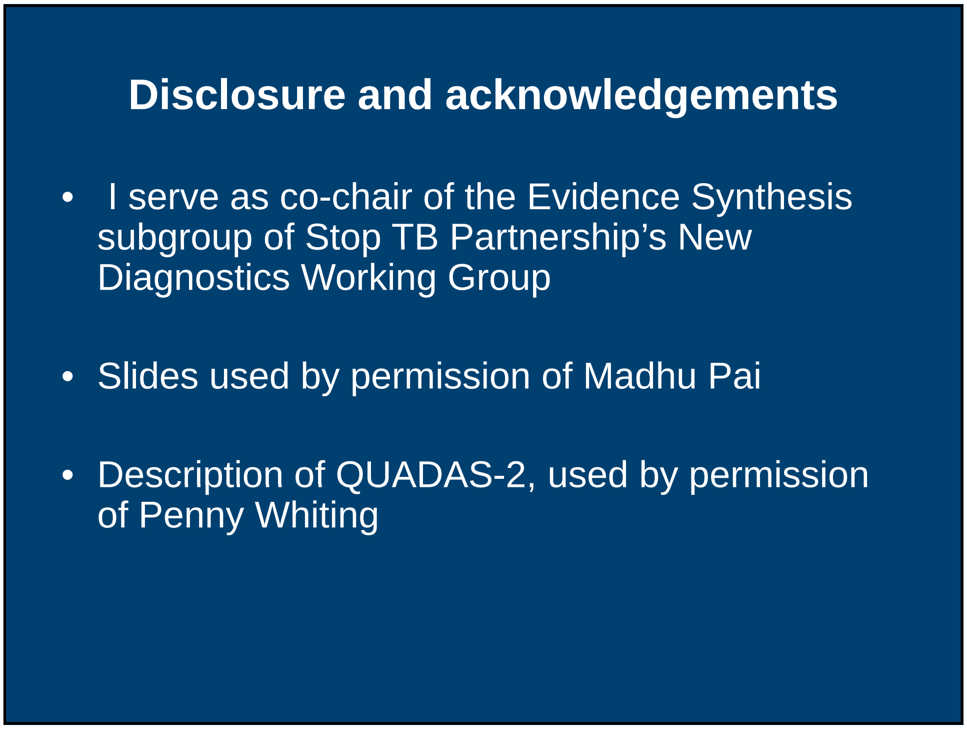
Bias in diagnostic research and sources of variation

Karen R Steingart, MD, MPH

karenst@uw.edu

Chennai, 13 December 2010



Disclosure and acknowledgements

- I serve as co-chair of the Evidence Synthesis subgroup of Stop TB Partnership's New Diagnostics Working Group
- Slides used by permission of Madhu Pai
- Description of QUADAS-2, used by permission of Penny Whiting

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

Overview

- Discuss major forms of bias and sources of variation in diagnostic studies
- Describe assessment of methodological quality of diagnostic accuracy studies

Diagnostic studies lack methodological rigor

Diagnostic studies in four prominent general medical journals

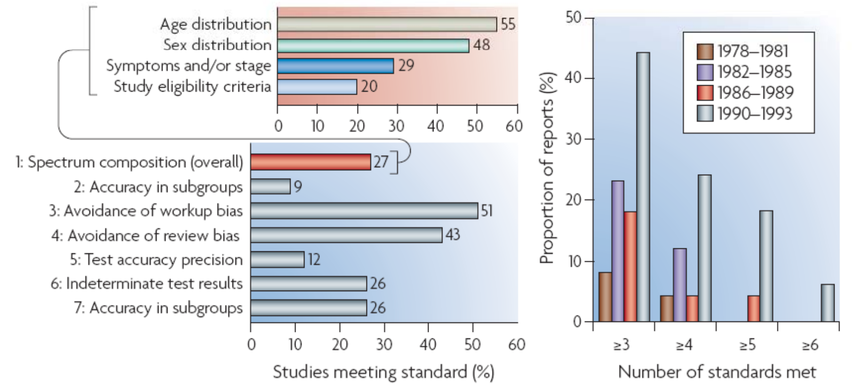


Figure 4 | **Proportion of diagnostic evaluations meeting accepted standards.** The seven standards are shown on the left. The data are taken from REF. 10.

Peeling et al. Nature Rev Micro 2006 (data are from Reid et al. JAMA. 1995 (274):645–651)

Lack of rigor: example from TB literature

12 meta-analyses; > 500 diagnostic studies

- 65% used prospective design
- 33% used consecutive or random sampling
- 72% used a cross-sectional design; 1/3 used case-control
- Blinding reported in 34%

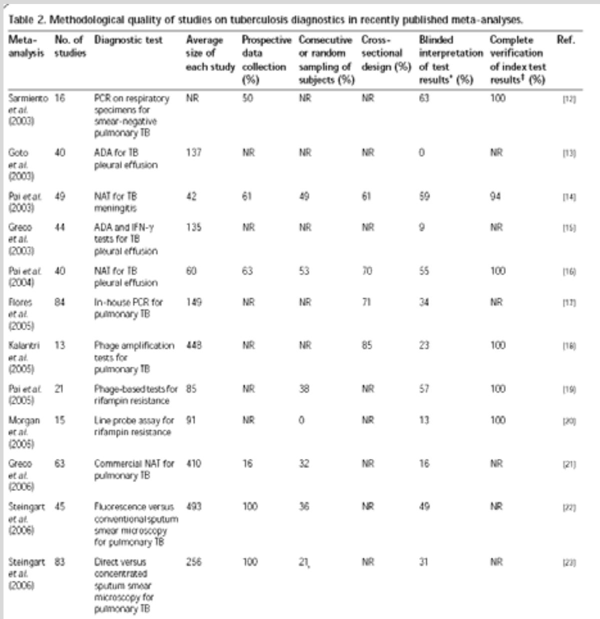
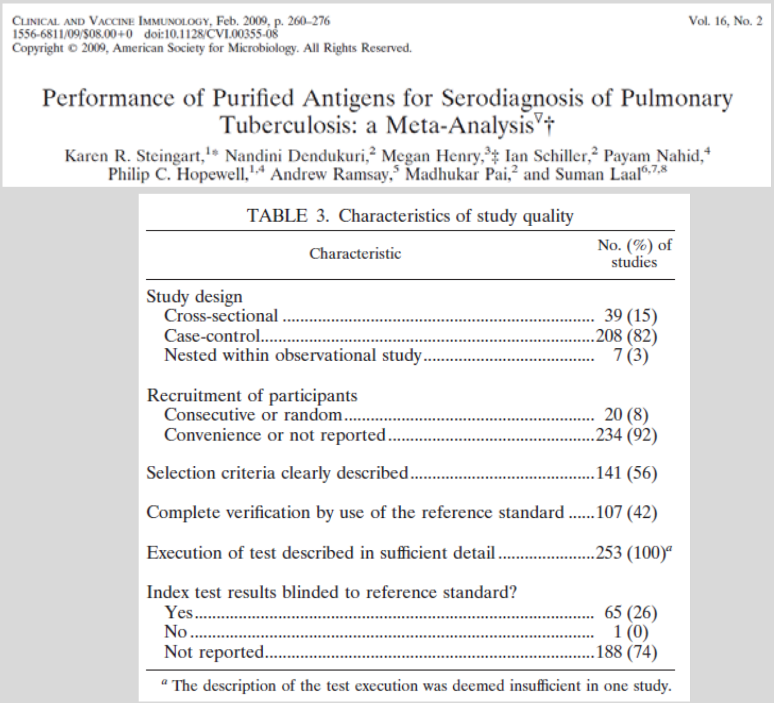


Table 2. Methodological quality of studies on tuberculosis diagnostics in recently published meta-analyses.

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

Pai et al. Exp Rev Mol Diagn 2006



Clinical and Vaccine Immunology, Feb. 2009, p. 260–276 Vol. 16, No. 2
1556-6811/09/$08.00+0 doi:10.1128/CVI.00355-08
Copyright © 2009, American Society for Microbiology. All Rights Reserved.

Performance of Purified Antigens for Serodiagnosis of Pulmonary Tuberculosis: a Meta-Analysis[∇†]

Karen R. Steingart,[1*] Nandini Dendukuri,[2] Megan Henry,[3‡] Ian Schiller,[2] Payam Nahid,[4] Philip C. Hopewell,[1,4] Andrew Ramsay,[5] Madhukar Pai,[2] and Suman Laal[6,7,8]

TABLE 3. Characteristics of study quality

Characteristic	No. (%) of studies
Study design	
Cross-sectional	39 (15)
Case-control	208 (82)
Nested within observational study	7 (3)
Recruitment of participants	
Consecutive or random	20 (8)
Convenience or not reported	234 (92)
Selection criteria clearly described	141 (56)
Complete verification by use of the reference standard	107 (42)
Execution of test described in sufficient detail	253 (100)[a]
Index test results blinded to reference standard?	
Yes	65 (26)
No	1 (0)
Not reported	188 (74)

[a] The description of the test execution was deemed insufficient in one study.

"Bias is any process at any stage of inference which tends to produce results or conclusions that differ systematically from the truth." *

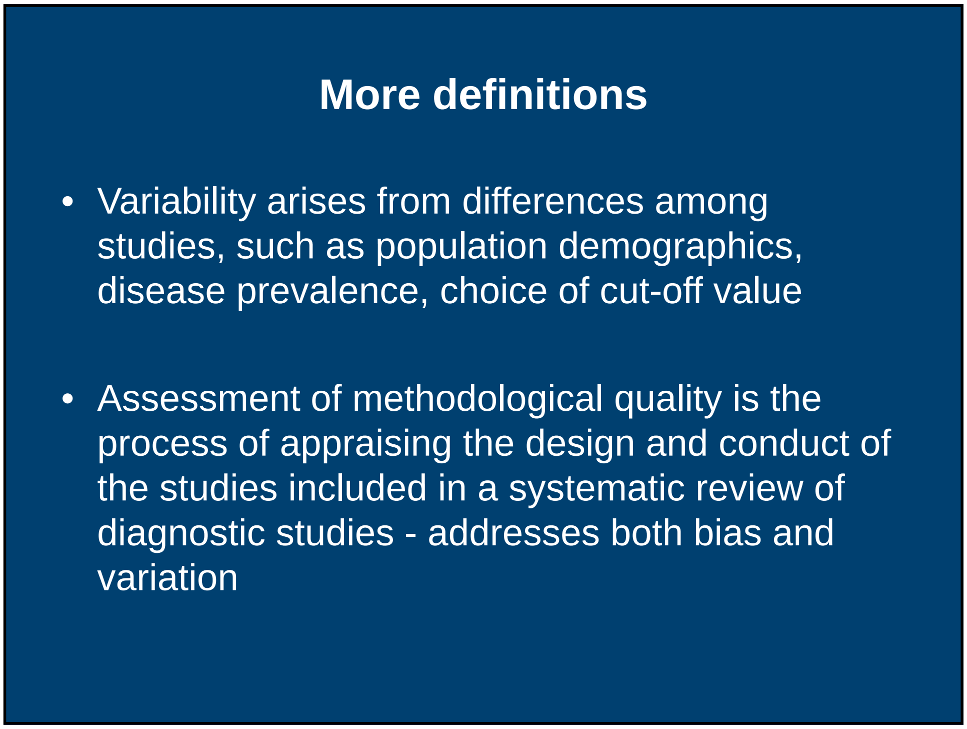
Biases

- can arise through problems in design, execution, analysis, and interpretation

- can lead to over or underestimates of test accuracy

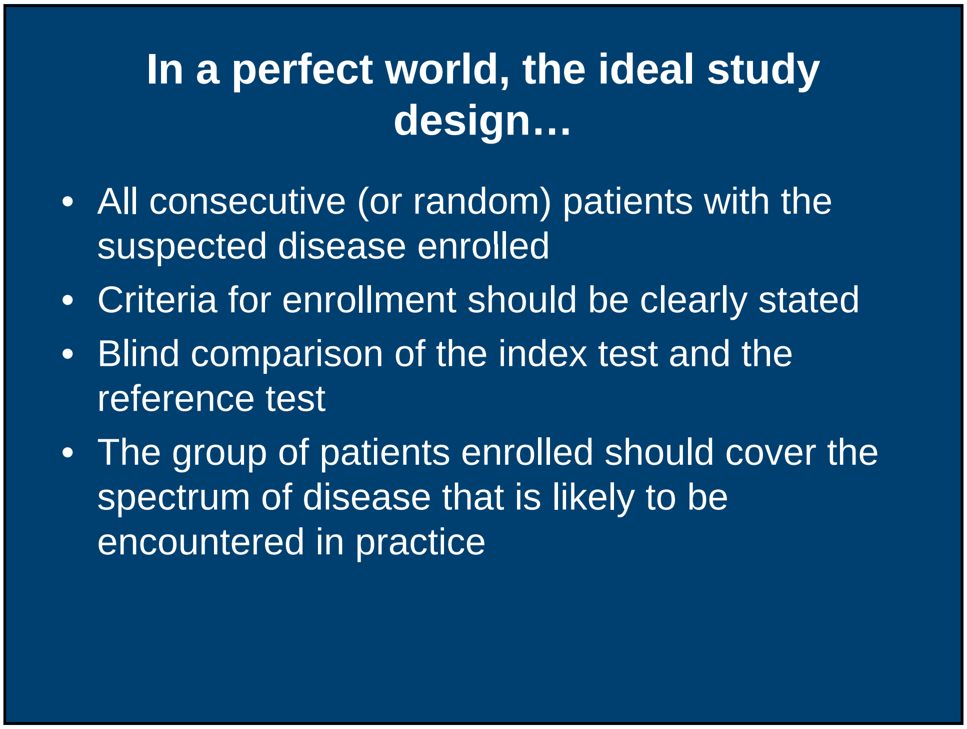
➡ Any factor that influences the assessment of disease status or test results can produce bias

*Murphy. The Logic of Medicine. Baltimore: John Hopkins University Press.1976.



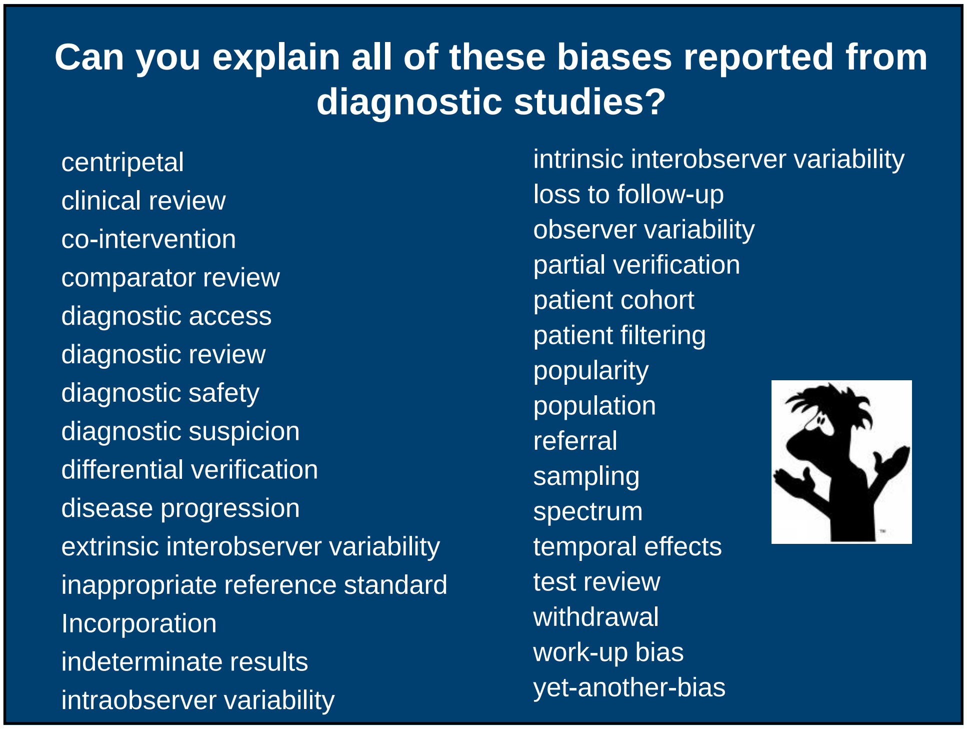
More definitions

- Variability arises from differences among studies, such as population demographics, disease prevalence, choice of cut-off value
- Assessment of methodological quality is the process of appraising the design and conduct of the studies included in a systematic review of diagnostic studies - addresses both bias and variation



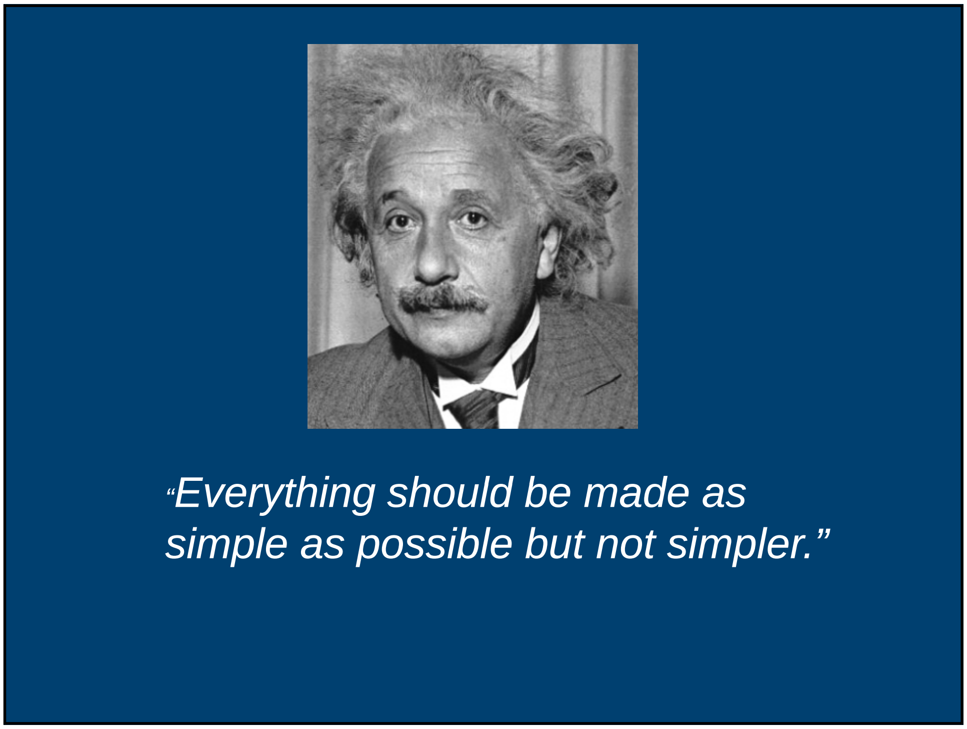
In a perfect world, the ideal study design…

- All consecutive (or random) patients with the suspected disease enrolled
- Criteria for enrollment should be clearly stated
- Blind comparison of the index test and the reference test
- The group of patients enrolled should cover the spectrum of disease that is likely to be encountered in practice



Can you explain all of these biases reported from diagnostic studies?

- centripetal
- clinical review
- co-intervention
- comparator review
- diagnostic access
- diagnostic review
- diagnostic safety
- diagnostic suspicion
- differential verification
- disease progression
- extrinsic interobserver variability
- inappropriate reference standard
- Incorporation
- indeterminate results
- intraobserver variability
- intrinsic interobserver variability
- loss to follow-up
- observer variability
- partial verification
- patient cohort
- patient filtering
- popularity
- population
- referral
- sampling
- spectrum
- temporal effects
- test review
- withdrawal
- work-up bias
- yet-another-bias



"Everything should be made as simple as possible but not simpler."

Sources of bias in diagnostic studies

- Bias due to an inappropriate/imperfect reference standard
- Spectrum bias
- Verification (work-up) bias
  - Partial verification bias
  - Differential verification bias
- Lack of blinding
- Incorporation bias
- Bias due to withdrawals, indeterminates, etc

An ideal reference standard…

- provides error-free classification of *all* participants
- verifies all test results
- both study test and reference standard can be performed within a short interval to avoid changes in target disease status

AP PHOTO

New gold standard: Phelps wins eighth medal

Michael Phelps won his record eighth gold medal at the Beijing Olympics as a member of the victorious U.S. 4x100-meter medley relay team, breaking a tie with Mark Spitz for most golds in a single games. full story

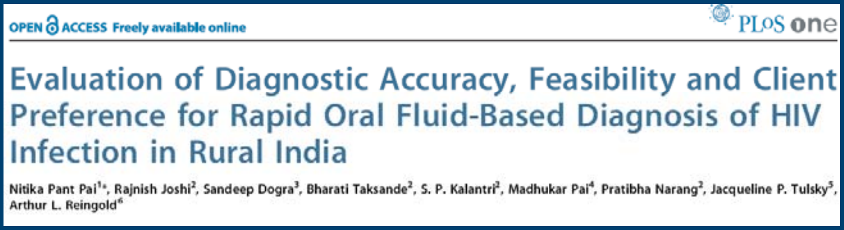
Bias due to inappropriate or imperfect reference standard

- The “gold standard” is the best performing test available, *but it is rarely perf*ect
- Imperfect reference standards are commonly used in diagnostic studies
- May lead to over or underestimation of test accuracy

Misclassification of disease status

- How accurately can you measure the following?
  - Depression
  - Tuberculosis in children
  - Latent TB infection
  - Dementia
  - Migraine
  - Attention deficit disorder
  - Cause of death
  - Irritable bowel syndrome
  - Chronic fatigue syndrome

Rarely, you get tests that are nearly perfect



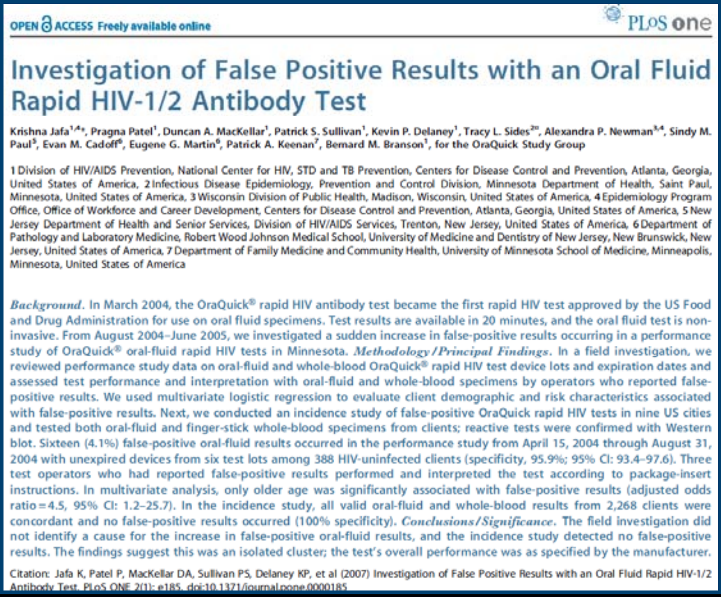
OPEN ACCESS Freely available online PLoS one

Evaluation of Diagnostic Accuracy, Feasibility and Client Preference for Rapid Oral Fluid-Based Diagnosis of HIV Infection in Rural India

Nitika Pant Pai[1*], Rajnish Joshi[2], Sandeep Dogra[3], Bharati Taksande[2], S. P. Kalantri[2], Madhukar Pai[4], Pratibha Narang[2], Jacqueline P. Tulsky[5], Arthur L. Reingold[6]

The OraQuick test on oral fluid specimens had better performance with a sensitivity of 100% (95% CI 98, 100) and a specificity of 100% (95% CI 99, 100), as compared to the OraQuick test on finger stick specimens with a sensitivity of 100% (95% CI 98, 100), and a specificity of 99.7% (95% CI 98.4, 99.9).

But even ‘nearly perfect’ tests run into problems!



OPEN ACCESS Freely available online PLoS one

Investigation of False Positive Results with an Oral Fluid Rapid HIV-1/2 Antibody Test

Krishna Jafa[1,4*], Pragna Patel[1], Duncan A. MacKellar[1], Patrick S. Sullivan[1], Kevin P. Delaney[1], Tracy L. Sides[2¤], Alexandra P. Newman[3,4], Sindy M. Paul[5], Evan M. Cadoff[6], Eugene G. Martin[6], Patrick A. Keenan[7], Bemard M. Branson[1], for the OraQuick Study Group

1 Division of HIV/AIDS Prevention, National Center for HIV, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, 2 Infectious Disease Epidemiology, Prevention and Control Division, Minnesota Department of Health, Saint Paul, Minnesota, United States of America, 3 Wisconsin Division of Public Health, Madison, Wisconsin, United States of America, 4 Epidemiology Program Office, Office of Workforce and Career Development, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America, 5 New Jersey Department of Health and Senior Services, Division of HIV/AIDS Services, Trenton, New Jersey, United States of America, 6 Department of Pathology and Laboratory Medicine, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey, New Brunswick, New Jersey, United States of America, 7 Department of Family Medicine and Community Health, University of Minnesota School of Medicine, Minneapolis, Minnesota, United States of America

Background. In March 2004, the OraQuick® rapid HIV antibody test became the first rapid HIV test approved by the US Food and Drug Administration for use on oral fluid specimens. Test results are available in 20 minutes, and the oral fluid test is non-invasive. From August 2004–June 2005, we investigated a sudden increase in false-positive results occurring in a performance study of OraQuick® oral-fluid rapid HIV tests in Minnesota. *Methodology/Principal Findings*. In a field investigation, we reviewed performance study data on oral-fluid and whole-blood OraQuick® rapid HIV test device lots and expiration dates and assessed test performance and interpretation with oral-fluid and whole-blood specimens by operators who reported false-positive results. We used multivariate logistic regression to evaluate client demographic and risk characteristics associated with false-positive results. Next, we conducted an incidence study of false-positive OraQuick rapid HIV tests in nine US cities and tested both oral-fluid and finger-stick whole-blood specimens from clients; reactive tests were confirmed with Western blot. Sixteen (4.1%) false-positive oral-fluid results occurred in the performance study from April 15, 2004 through August 31, 2004 with unexpired devices from six test lots among 388 HIV-uninfected clients (specificity, 95.9%; 95% CI: 93.4–97.6). Three test operators who had reported false-positive results performed and interpreted the test according to package-insert instructions. In multivariate analysis, only older age was significantly associated with false-positive results (adjusted odds ratio = 4.5, 95% CI: 1.2–25.7). In the incidence study, all valid oral-fluid and whole-blood results from 2,268 clients were concordant and no false-positive results occurred (100% specificity). *Conclusions/Significance*. The field investigation did not identify a cause for the increase in false-positive oral-fluid results, and the incidence study detected no false-positive results. The findings suggest this was an isolated cluster; the test's overall performance was as specified by the manufacturer.

Citation: Jafa K, Patel P, MacKellar DA, Sullivan PS, Delaney KP, et al (2007) Investigation of False Positive Results with an Oral Fluid Rapid HIV-1/2 Antibody Test. PLoS ONE 2(1): e185. doi:10.1371/journal.pone.0000185

Health Technology Assessment 2007; Vol. 11: No. 50

Evaluation of diagnostic tests when there is no gold standard. A review of methods

AWS Rutjes, JB Reitsma, A Coomarasamy, KS Khan and PMM Bossuyt

December 2007

Health Technology Assessment
NHS R&D HTA Programme
www.hta.ac.uk

ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 797–806

Journal of Clinical Epidemiology

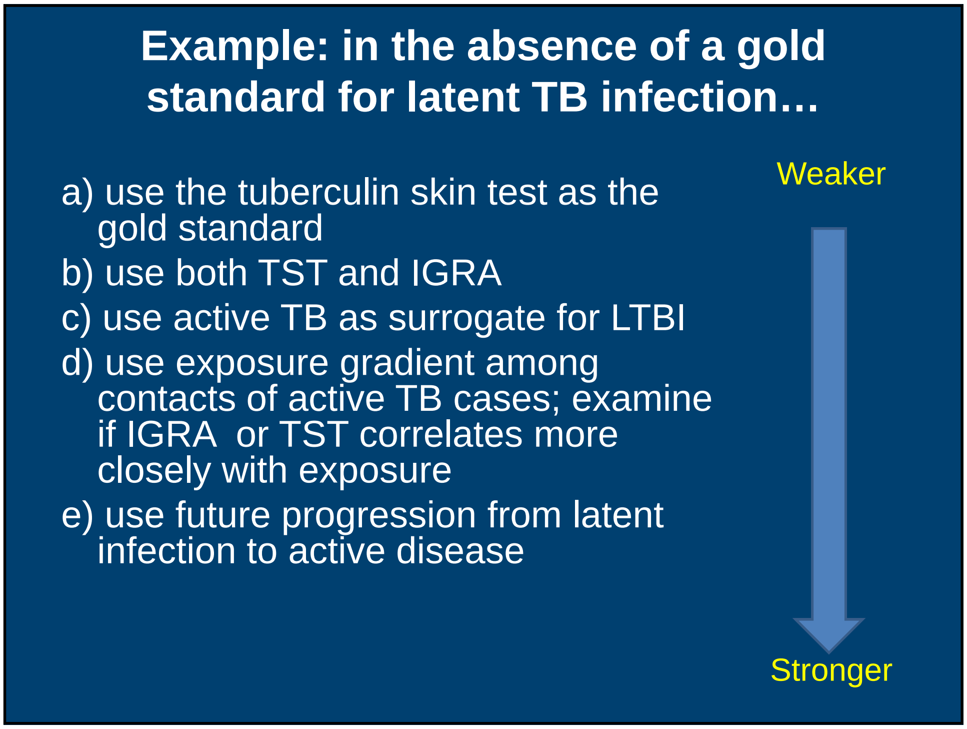
A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard

Johannes B. Reitsma[a,*], Anne W.S. Rutjes[a], Khalid S. Khan[b], Arri Coomarasamy[b], Patrick M. Bossuyt[a]

What if the reference standard is imperfect or missing?

Methods for diagnostic research where reference standard is imperfect or missing

1. Adjust for missing data on reference standard
2. Correct for imperfections in reference standard (based on previous research about the degree of imperfection)
3. Combine multiple pieces of information to construct a reference standard
4. Validate the index test results with other relevant clinical characteristics



Example: in the absence of a gold standard for latent TB infection…

a) use the tuberculin skin test as the gold standard

b) use both TST and IGRA

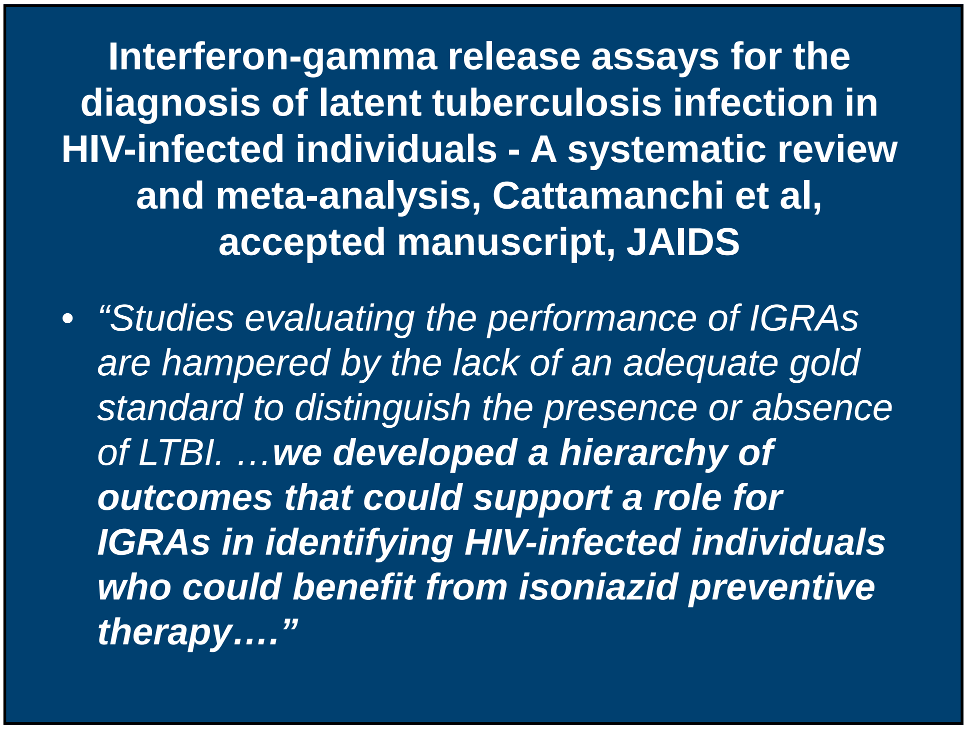
c) use active TB as surrogate for LTBI

d) use exposure gradient among contacts of active TB cases; examine if IGRA or TST correlates more closely with exposure

e) use future progression from latent infection to active disease

Weaker

Stronger



Interferon-gamma release assays for the diagnosis of latent tuberculosis infection in HIV-infected individuals - A systematic review and meta-analysis, Cattamanchi et al, accepted manuscript, JAIDS

- *"Studies evaluating the performance of IGRAs are hampered by the lack of an adequate gold standard to distinguish the presence or absence of LTBI. …**we developed a hierarchy of outcomes that could support a role for IGRAs in identifying HIV-infected individuals who could benefit from isoniazid preventive therapy….**"*

Spectrum bias (a form of selection bias)

Could the selection of patients have introduced bias?

- Extreme case, case-control design where study enrolls patients with definite disease and healthy controls, estimates of accuracy may be inflated
- However, the use of a case-control design does not always produce biased estimates of accuracy, for example enrolling *diseased* controls will reduce the potential for bias

Example: spectrum bias

CLINICAL AND VACCINE IMMUNOLOGY, Feb. 2009, p. 260–276 Vol. 16, No. 2
1556-6811/09/$08.00+0 doi:10.1128/CVI.00355-08
Copyright © 2009, American Society for Microbiology. All Rights Reserved.

Performance of Purified Antigens for Serodiagnosis of Pulmonary Tuberculosis: a Meta-Analysis[▽]†

Karen R. Steingart,[1*] Nandini Dendukuri,[2] Megan Henry,[3]‡ Ian Schiller,[2] Payam Nahid,[4] Philip C. Hopewell,[1,4] Andrew Ramsay,[5] Madhukar Pai,[2] and Suman Laal[6,7,8]

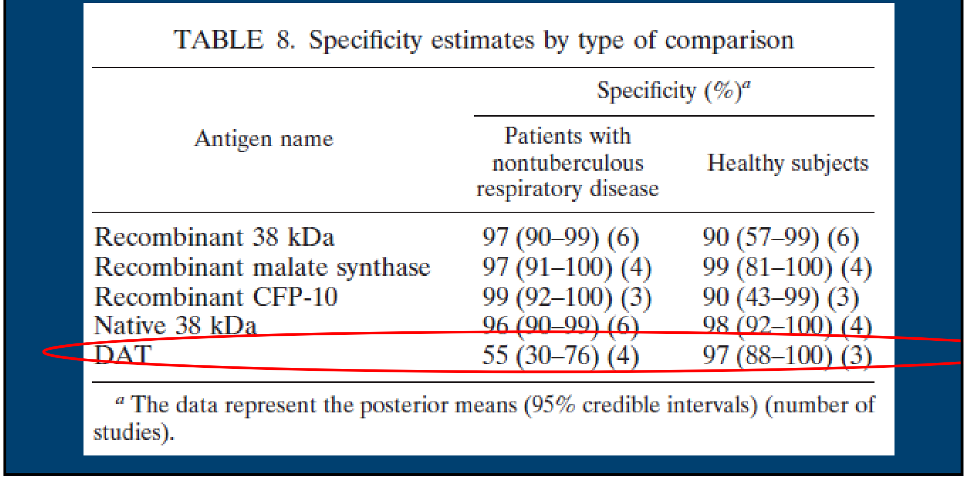
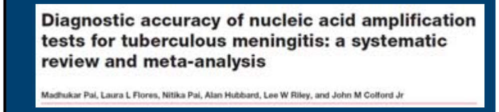


TABLE 8. Specificity estimates by type of comparison

Antigen name	Specificity (%)[a]	
	Patients with nontuberculous respiratory disease	Healthy subjects
Recombinant 38 kDa	97 (90–99) (6)	90 (57–99) (6)
Recombinant malate synthase	97 (91–100) (4)	99 (81–100) (4)
Recombinant CFP-10	99 (92–100) (3)	90 (43–99) (3)
Native 38 kDa	96 (90–99) (6)	98 (92–100) (4)
DAT	55 (30–76) (4)	97 (88–100) (3)

[a] The data represent the posterior means (95% credible intervals) (number of studies).

Example: spectrum bias - NAAT for tuberculous meningitis



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

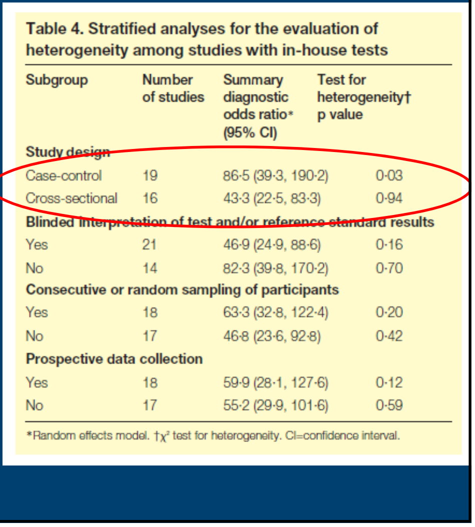


Table 4. Stratified analyses for the evaluation of heterogeneity among studies with in-house tests

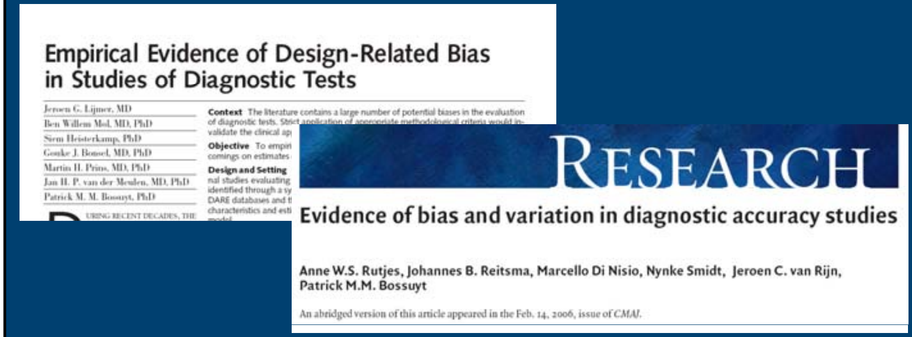
Subgroup	Number of studies	Summary diagnostic odds ratio* (95% CI)	Test for heterogeneity† p value
Study design			
Case-control	19	86·5 (39·3, 190·2)	0·03
Cross-sectional	16	43·3 (22·5, 83·3)	0·94
Blinded interpretation of test and/or reference standard results			
Yes	21	46·9 (24·9, 88·6)	0·16
No	14	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

Pai et al. Lancet Infect Dis 2003

Empirical evidence of sources of bias in diagnostic studies



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

D URING RECENT DECADES, THE

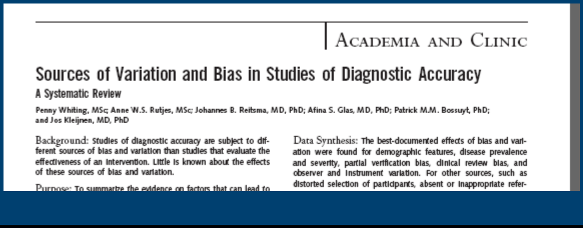
Context The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would in-
validate the clinical ap
Objective To empir
comings on estimates
Design and Setting
nal studies evaluating
identified through a sy
DARE databases and t
characteristics and esti
model

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*.



ACADEMIA AND CLINIC

Sources of Variation and Bias in Studies of Diagnostic Accuracy

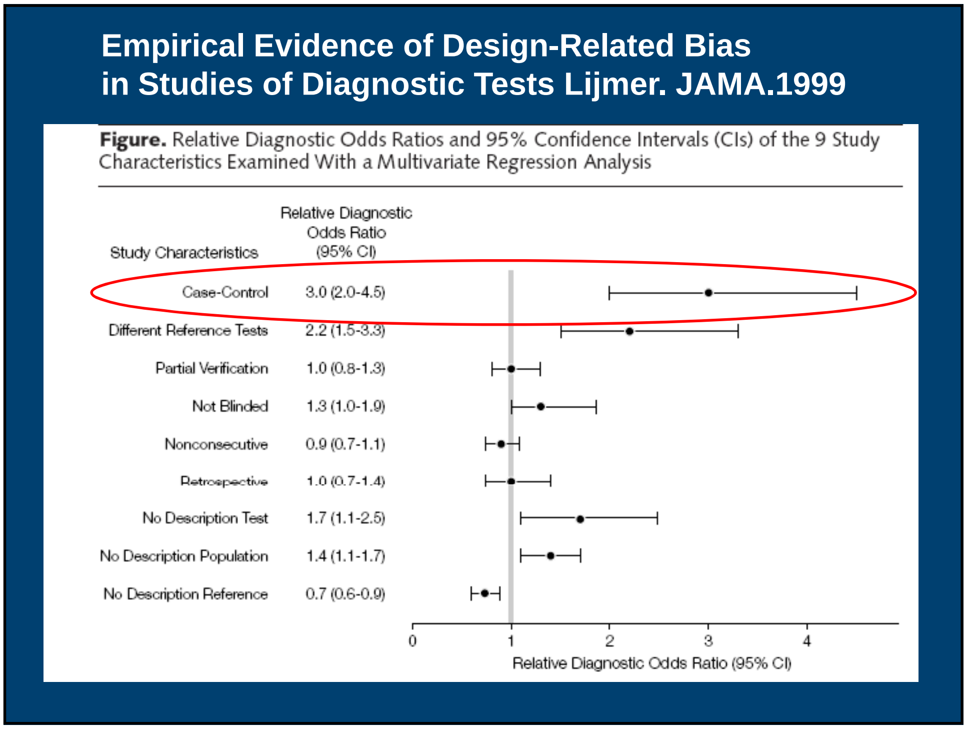
A Systematic Review

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.

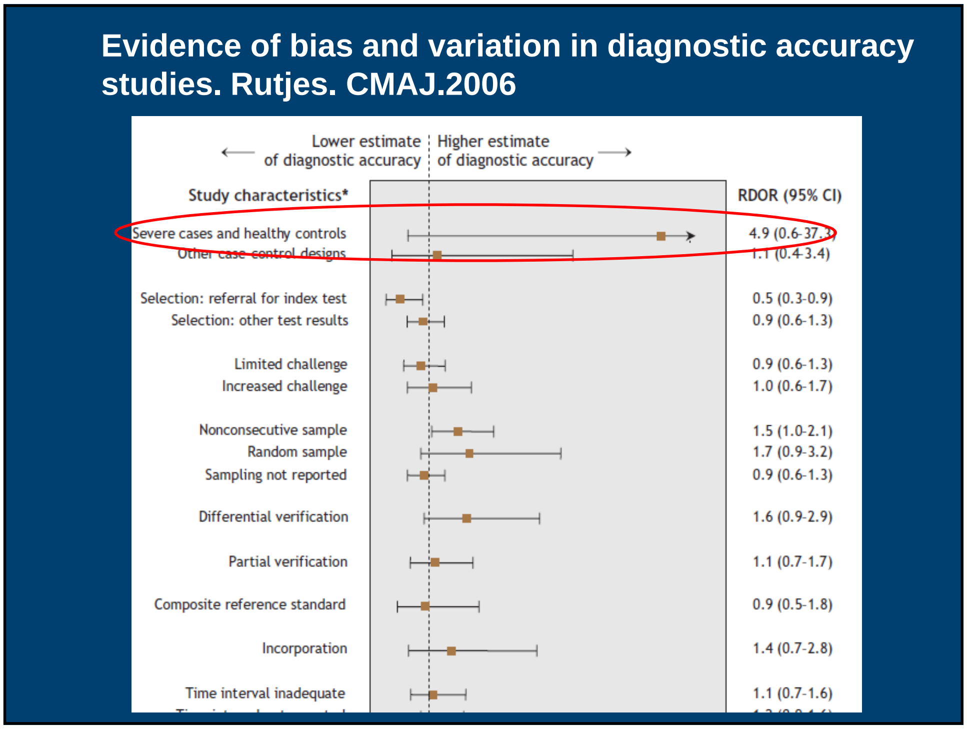
Purpose: To summarize the evidence on factors that can lead to

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-

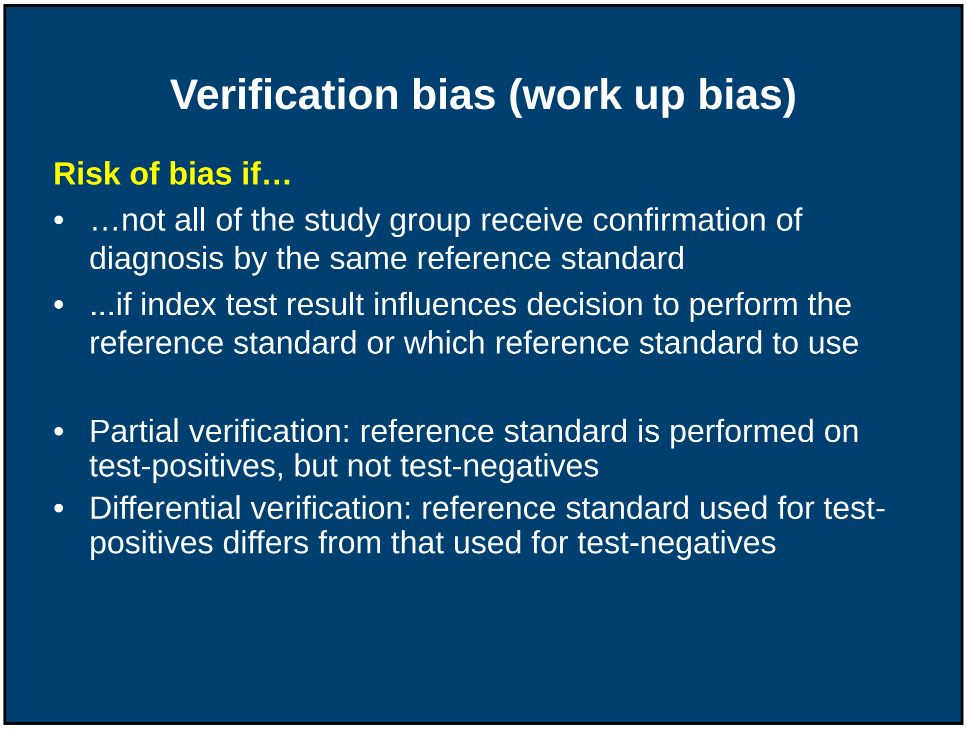


Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests Lijmer. JAMA.1999

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



Evidence of bias and variation in diagnostic accuracy studies. Rutjes. CMAJ.2006

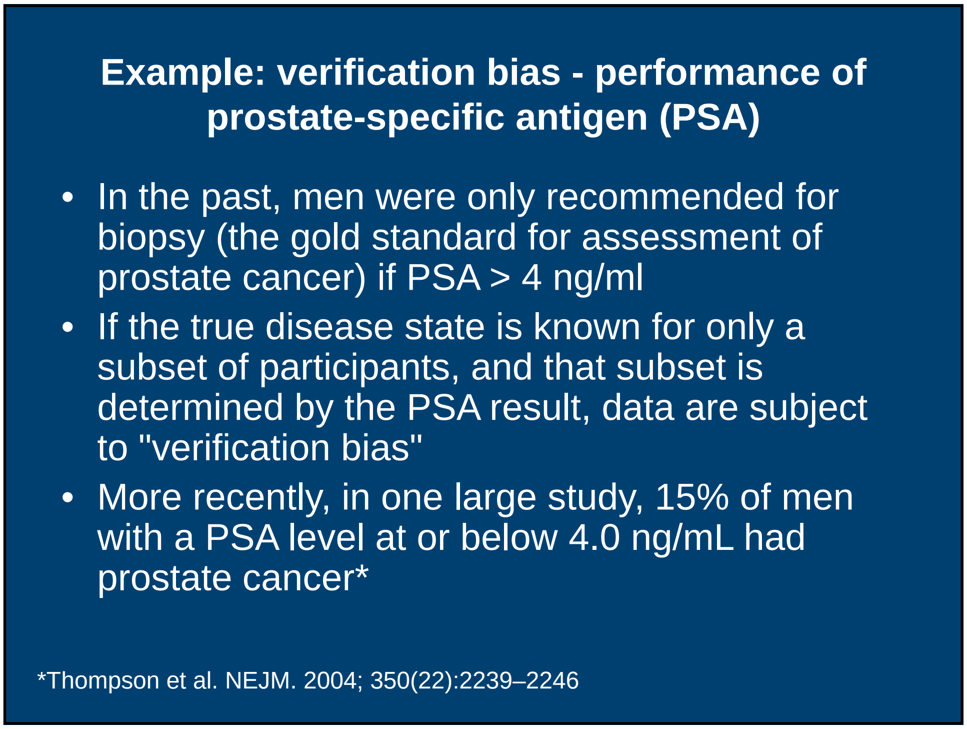


Verification bias (work up bias)

Risk of bias if…

- …not all of the study group receive confirmation of diagnosis by the same reference standard
- ...if index test result influences decision to perform the reference standard or which reference standard to use

- Partial verification: reference standard is performed on test-positives, but not test-negatives
- Differential verification: reference standard used for test-positives differs from that used for test-negatives



Example: verification bias - performance of prostate-specific antigen (PSA)

- In the past, men were only recommended for biopsy (the gold standard for assessment of prostate cancer) if PSA > 4 ng/ml
- If the true disease state is known for only a subset of participants, and that subset is determined by the PSA result, data are subject to "verification bias"
- More recently, in one large study, 15% of men with a PSA level at or below 4.0 ng/mL had prostate cancer*

*Thompson et al. NEJM. 2004; 350(22):2239–2246

Empirical evidence of verification bias reported in 3 systematic reviews of diagnostic accuracy studies

Question?	Lijmer	Whiting	Rutjes
Did investigators perform the same gold standard on all patients regardless of the study test results?	Different gold standard used for some patients RDOR 2.2 (95% CI 1.5,3.3)	Inappropriate gold standard (some empirical support)	Different gold standard used for some patients RDOR 1.6 (95% CI 0.9,2.9)
	Gold standard not used for some patients RDOR 1.0 (95% CI 0.8,1.3)	Gold standard not used for some patients (strong empirical support)	Gold standard not used for some patients RDOR 1.1 (95% CI 0.7,1.7)

Adapted from Furukawa and Guyatt. CMAJ 2006; 174(4):481-2

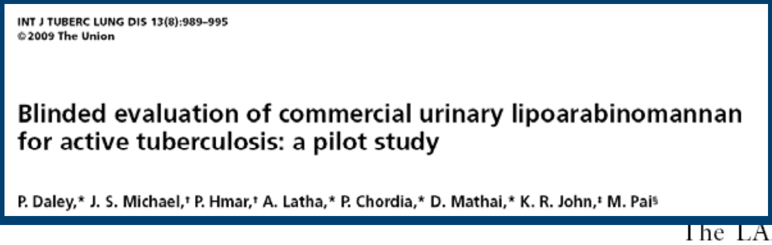
Lack of blinding (also called review bias)

- Diagnostic studies may be:
  - Unblinded
  - Single blind (study test *or* ref. standard result is blinded)
  - Double blind (study test *and* ref. std results are blinded)
- Lack of blinding can lead to overestimation of test accuracy

Lack of blinding

- Blinding is more important when the interpretation of test results is subjective (e.g., pain)
- Blinding is less important when study test and gold standard are produced by an automated system with little or no ambiguity in the reading of results (e.g. CD4 count)
- Lab tests can be easily blinded by coding specimens

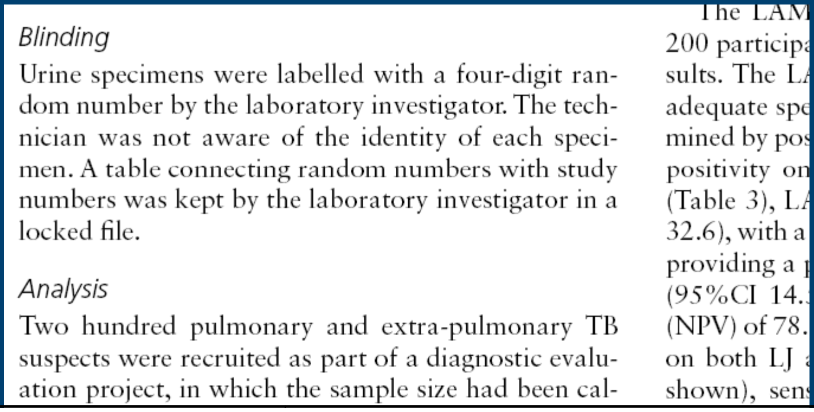
Example: blinding



INT J TUBERC LUNG DIS 13(8):989–995
©2009 The Union

Blinded evaluation of commercial urinary lipoarabinomannan for active tuberculosis: a pilot study

P. Daley,* J. S. Michael,† P. Hmar,† A. Latha,* P. Chordia,* D. Mathai,* K. R. John,‡ M. Pai§



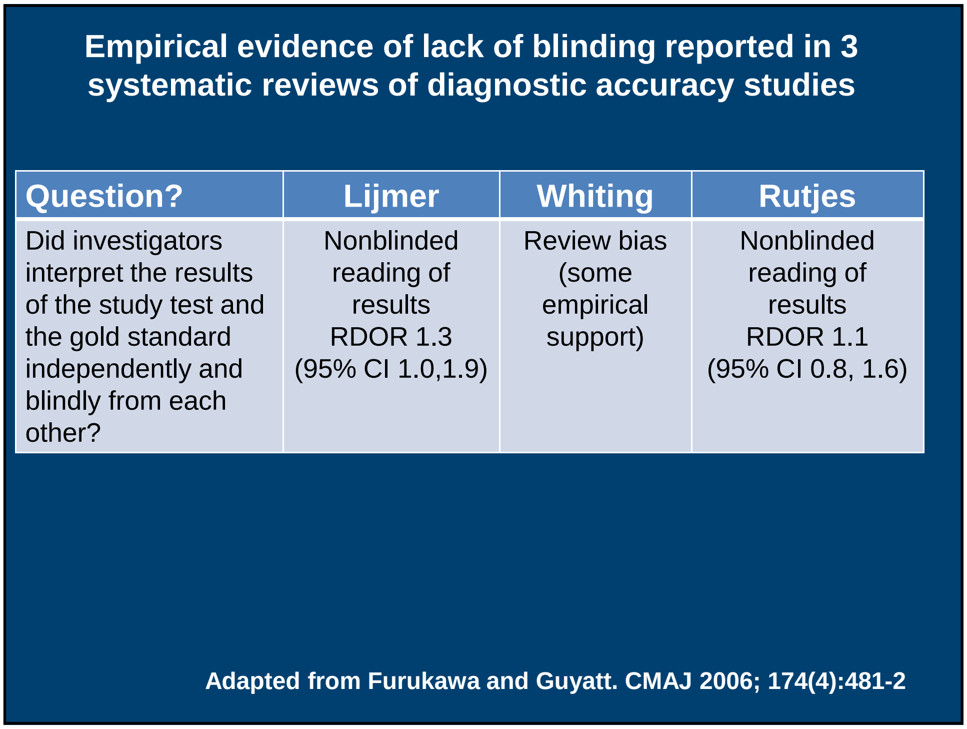
Blinding

Urine specimens were labelled with a four-digit random number by the laboratory investigator. The technician was not aware of the identity of each specimen. A table connecting random numbers with study numbers was kept by the laboratory investigator in a locked file.

Analysis

Two hundred pulmonary and extra-pulmonary TB suspects were recruited as part of a diagnostic evaluation project, in which the sample size had been cal-

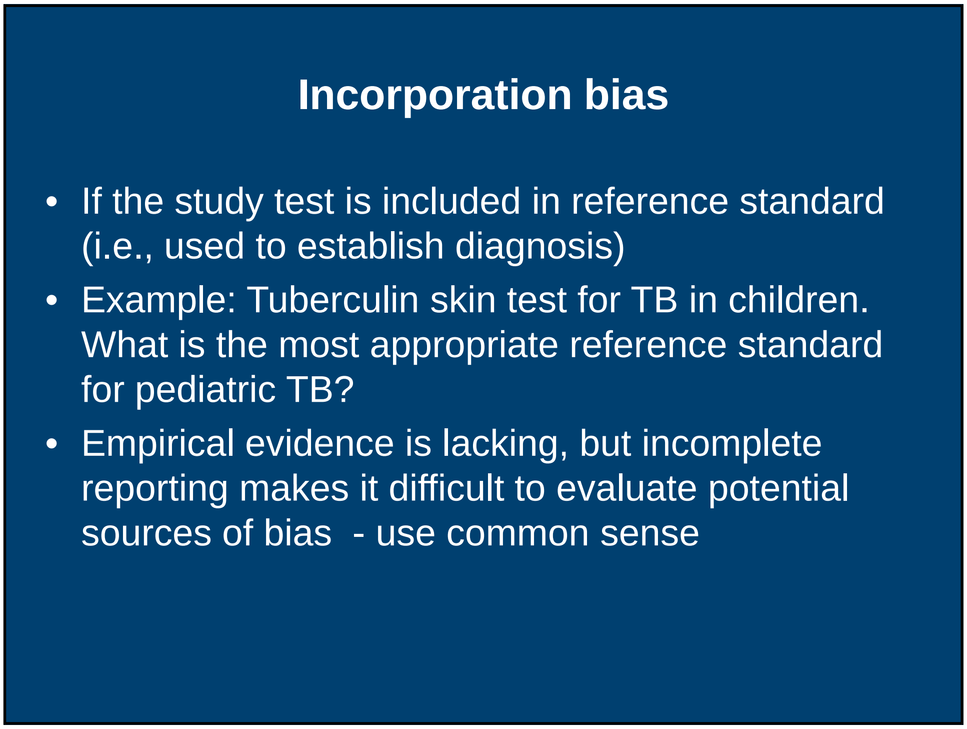
The LAM
200 particip
sults. The L
adequate spe
mined by pos
positivity on
(Table 3), L
32.6), with a
providing a
(95%CI 14.
(NPV) of 78.
on both LJ
shown), sen



Empirical evidence of lack of blinding reported in 3 systematic reviews of diagnostic accuracy studies

Question?	Lijmer	Whiting	Rutjes
Did investigators interpret the results of the study test and the gold standard independently and blindly from each other?	Nonblinded reading of results RDOR 1.3 (95% CI 1.0,1.9)	Review bias (some empirical support)	Nonblinded reading of results RDOR 1.1 (95% CI 0.8, 1.6)

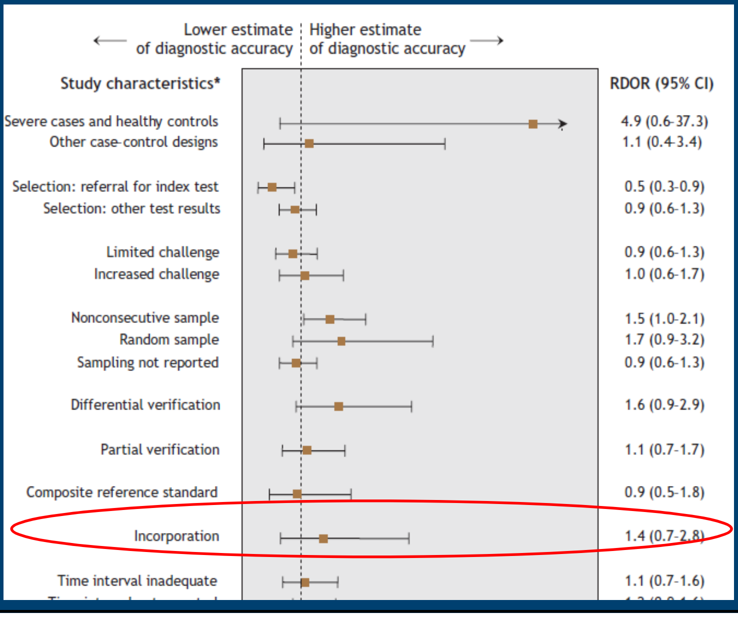
Adapted from Furukawa and Guyatt. CMAJ 2006; 174(4):481-2



Incorporation bias

- If the study test is included in reference standard (i.e., used to establish diagnosis)
- Example: Tuberculin skin test for TB in children. What is the most appropriate reference standard for pediatric TB?
- Empirical evidence is lacking, but incomplete reporting makes it difficult to evaluate potential sources of bias - use common sense

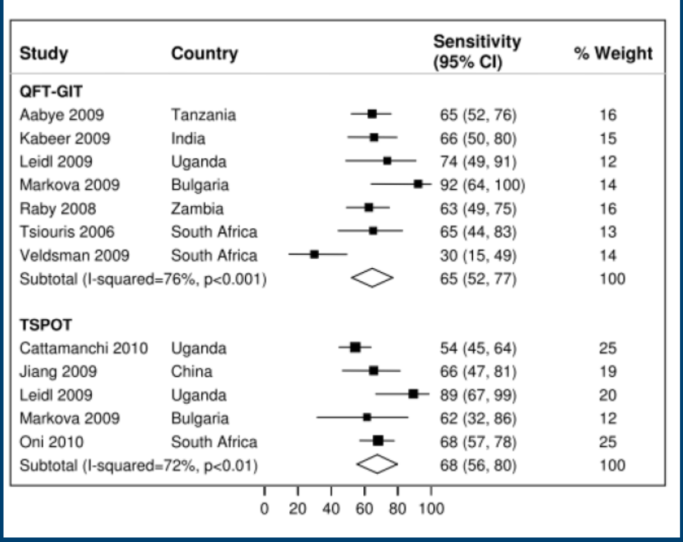
Evidence of bias and variation in diagnostic accuracy studies. Rutjes. CMAJ.2006



Bias due to withdrawals, indeterminates, missing data

- Example: “High sensitivity of IGRA in HIV+ TB patients”
  - Sensitivity of IGRA ~90%
    - But nearly 30% of all patients had indeterminate IGRA results!
    - These results were excluded for computation of sensitivity

Sensitivity of QuantiFERON-TB Gold In-Tube and T-SPOT.TB in HIV-infected persons with confirmed active tuberculosis (low/middle-income countries)

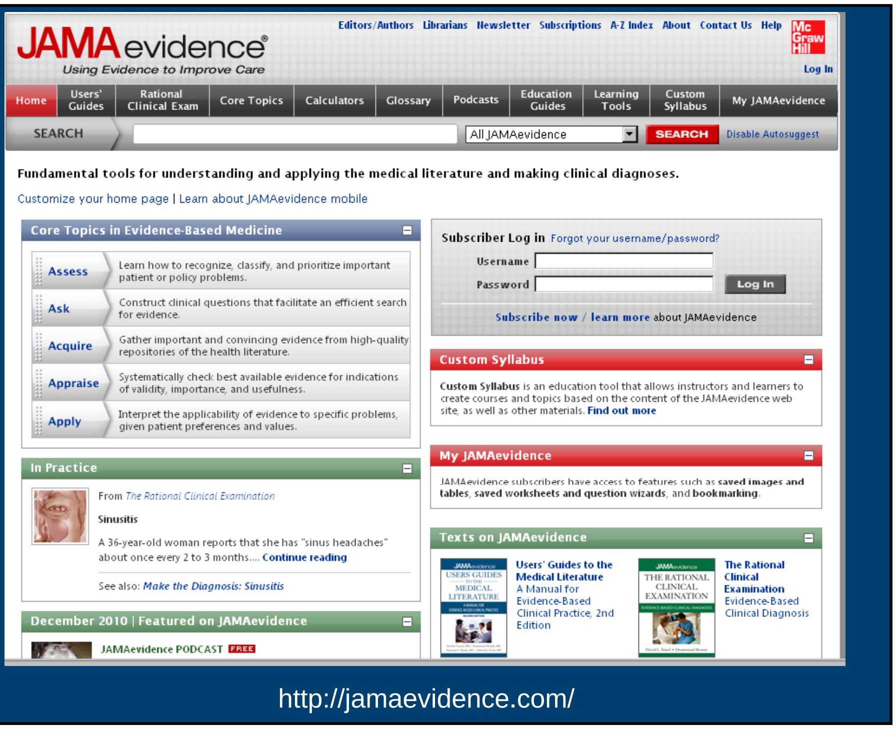


Study	Country	Sensitivity (95% CI)	% Weight
QFT-GIT			
Aabye 2009	Tanzania	65 (52, 76)	16
Kabeer 2009	India	66 (50, 80)	15
Leidl 2009	Uganda	74 (49, 91)	12
Markova 2009	Bulgaria	92 (64, 100)	14
Raby 2008	Zambia	63 (49, 75)	16
Tsiouris 2006	South Africa	65 (44, 83)	13
Veldsman 2009	South Africa	30 (15, 49)	14
Subtotal (I-squared=76%, p<0.001)		65 (52, 77)	100
TSPOT			
Cattamanchi 2010	Uganda	54 (45, 64)	25
Jiang 2009	China	66 (47, 81)	19
Leidl 2009	Uganda	89 (67, 99)	20
Markova 2009	Bulgaria	62 (32, 86)	12
Oni 2010	South Africa	68 (57, 78)	25
Subtotal (I-squared=72%, p<0.01)		68 (56, 80)	100

Metcalfe et al.- Methods

- We used the following definitions for primary outcomes
- (1) Sensitivity - the proportion of individuals with a positive IGRA result among those with culture-positive TB (we included indeterminate IGRA results in the denominator if they occurred in individuals with culture positive TB)

Assessment of methodological quality of diagnostic accuracy studies



Fundamental tools for understanding and applying the medical literature and making clinical diagnoses.

Customize your home page | Learn about JAMAevidence mobile

Core Topics in Evidence-Based Medicine

Assess	Learn how to recognize, classify, and prioritize important patient or policy problems.
Ask	Construct clinical questions that facilitate an efficient search for evidence.
Acquire	Gather important and convincing evidence from high-quality repositories of the health literature.
Appraise	Systematically check best available evidence for indications of validity, importance, and usefulness.
Apply	Interpret the applicability of evidence to specific problems, given patient preferences and values.

In Practice

From *The Rational Clinical Examination*

Sinusitis

A 36-year-old woman reports that she has "sinus headaches" about once every 2 to 3 months.... **Continue reading**

See also: ***Make the Diagnosis: Sinusitis***

December 2010 | Featured on JAMAevidence

JAMAevidence PODCAST FREE

Subscriber Log in Forgot your username/password?

Username

Password

Log In

Subscribe now / learn more about JAMAevidence

Custom Syllabus

Custom Syllabus is an education tool that allows instructors and learners to create courses and topics based on the content of the JAMAevidence web site, as well as other materials. **Find out more**

My JAMAevidence

JAMAevidence subscribers have access to features such as **saved images and tables**, **saved worksheets and question wizards**, and **bookmarking**.

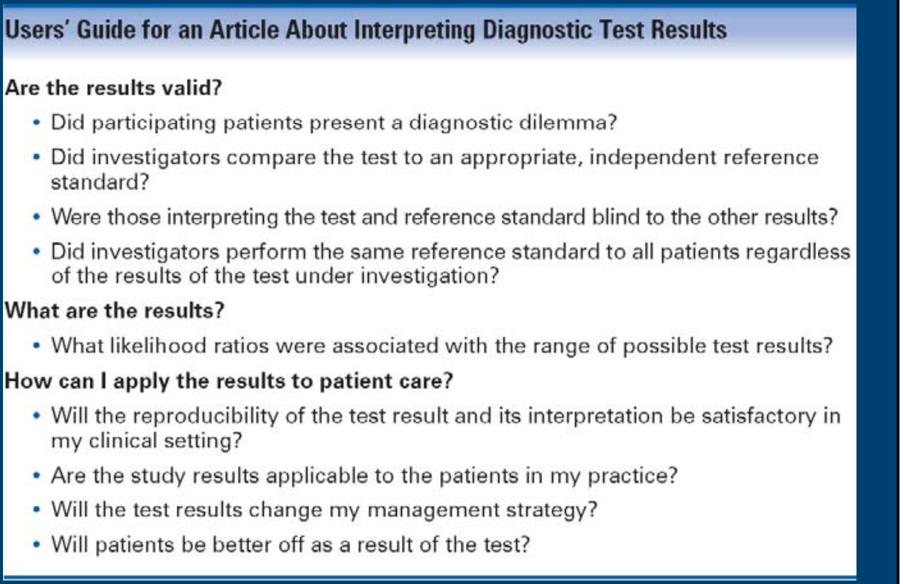
Texts on JAMAevidence

Users' Guides to the Medical Literature A Manual for Evidence-Based Clinical Practice, 2nd Edition

The Rational Clinical Examination Evidence-Based Clinical Diagnosis

http://jamaevidence.com/

Users' guide for a diagnostic study



Users' Guide for an Article About Interpreting Diagnostic Test Results

Are the results valid?

- Did participating patients present a diagnostic dilemma?
- Did investigators compare the test to an appropriate, independent reference standard?
- Were those interpreting the test and reference standard blind to the other results?
- Did investigators perform the same reference standard to all patients regardless of the results of the test under investigation?

What are the results?

- What likelihood ratios were associated with the range of possible test results?

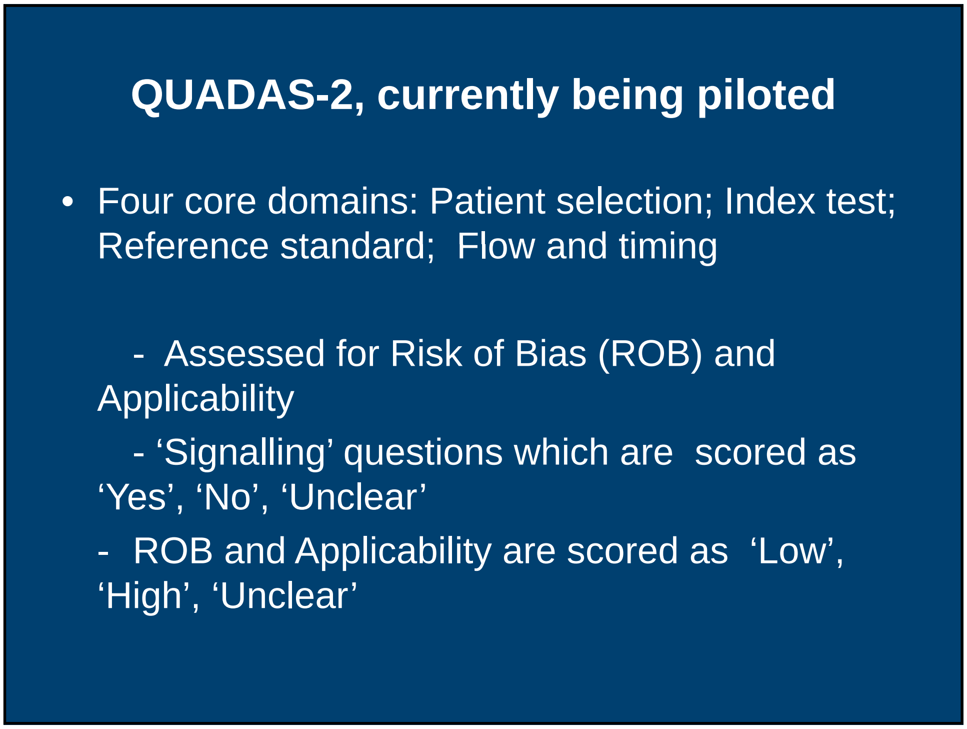
How can I apply the results to patient care?

- Will the reproducibility of the test result and its interpretation be satisfactory in my clinical setting?
- Are the study results applicable to the patients in my practice?
- Will the test results change my management strategy?
- Will patients be better off as a result of the test?

QUality Assessment of Diagnostic Accuracy Studies (QUADAS)

- Systematically developed based on empirical evidence and a formal consensus method (modified Delphi)
- Recommended tool by Cochrane Collaboration

Whiting et al. The Development of QUADAS… BMC Med Res Methodol 2003; 3:25.



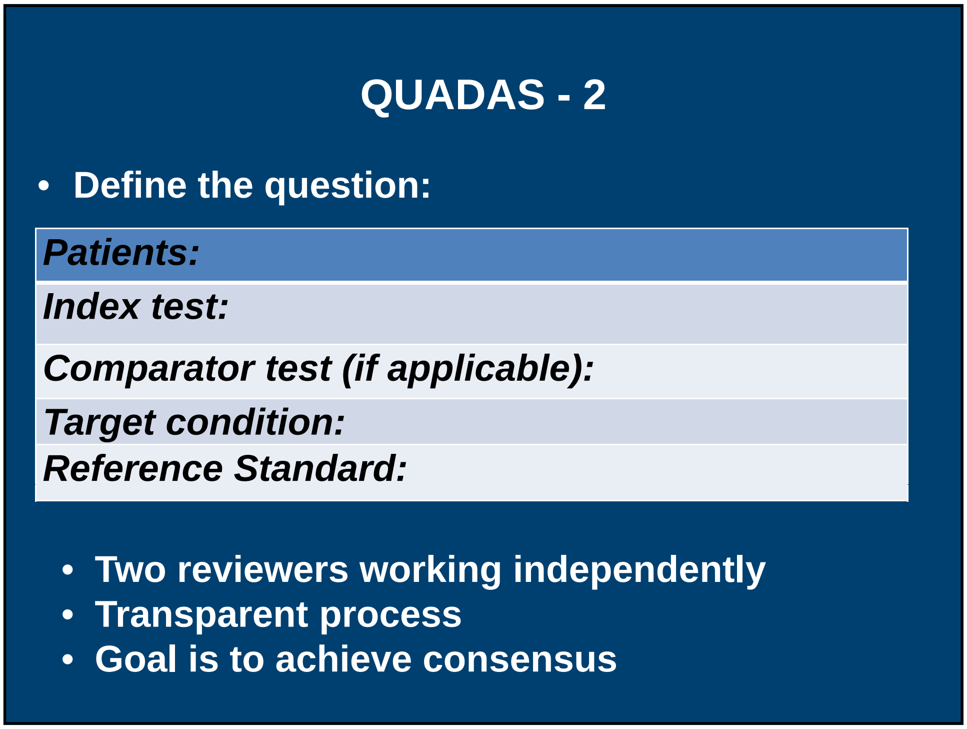
QUADAS-2, currently being piloted

- Four core domains: Patient selection; Index test; Reference standard; Flow and timing

  - Assessed for Risk of Bias (ROB) and Applicability

  - ‘Signalling’ questions which are scored as ‘Yes’, ‘No’, ‘Unclear’

  - ROB and Applicability are scored as ‘Low’, ‘High’, ‘Unclear’



QUADAS - 2

- **Define the question:**

Patients:
Index test:
Comparator test (if applicable):
Target condition:
Reference Standard:

- **Two reviewers working independently**
- **Transparent process**
- **Goal is to achieve consensus**

Domain 1: Patient Selection

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

- ***Signalling Question 1: Were eligibility criteria defined?***
- ***Signalling Question 2: Was an unselected sample of patients enrolled?***

Whiting P, QUADAS2, DRAFT

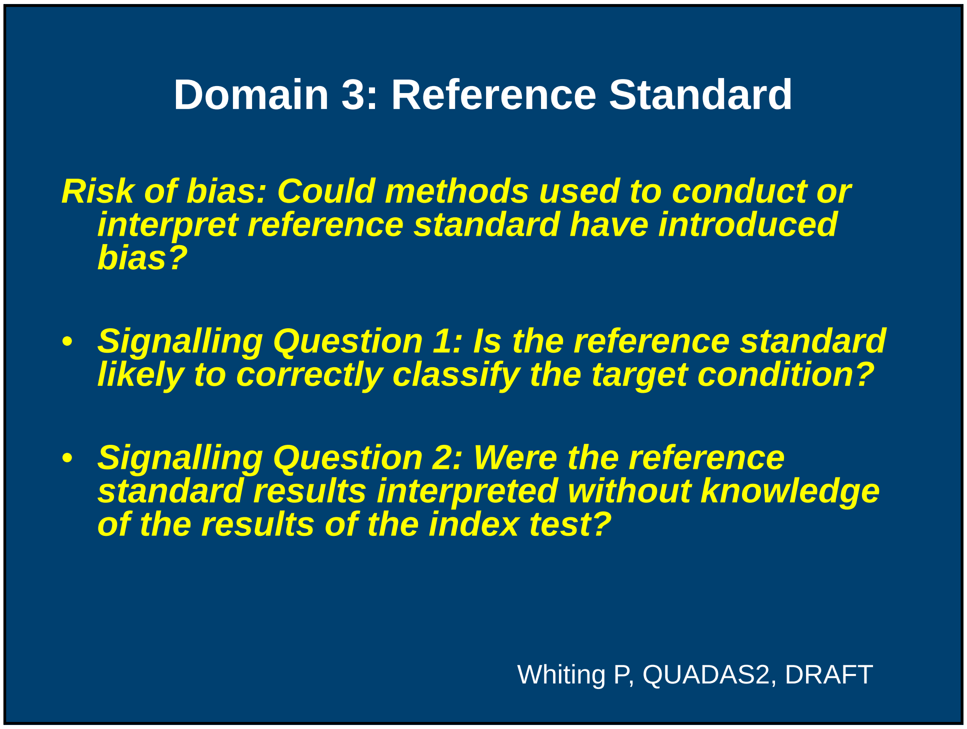
Domain 2: Index Test - DRAFT

Risk of bias: Could methods used to interpret or conduct the index test have introduced bias?

- ***Signalling Question 1: Were the index test results interpreted without knowledge of the results of the reference standard?***
- ***Signalling Question 2: Did the study pre-specify the threshold?***

  - Selecting the threshold to maximise the sensitivity and/or specificity of the test may lead to overoptimistic measures of test performance

Whiting P, QUADAS2, DRAFT

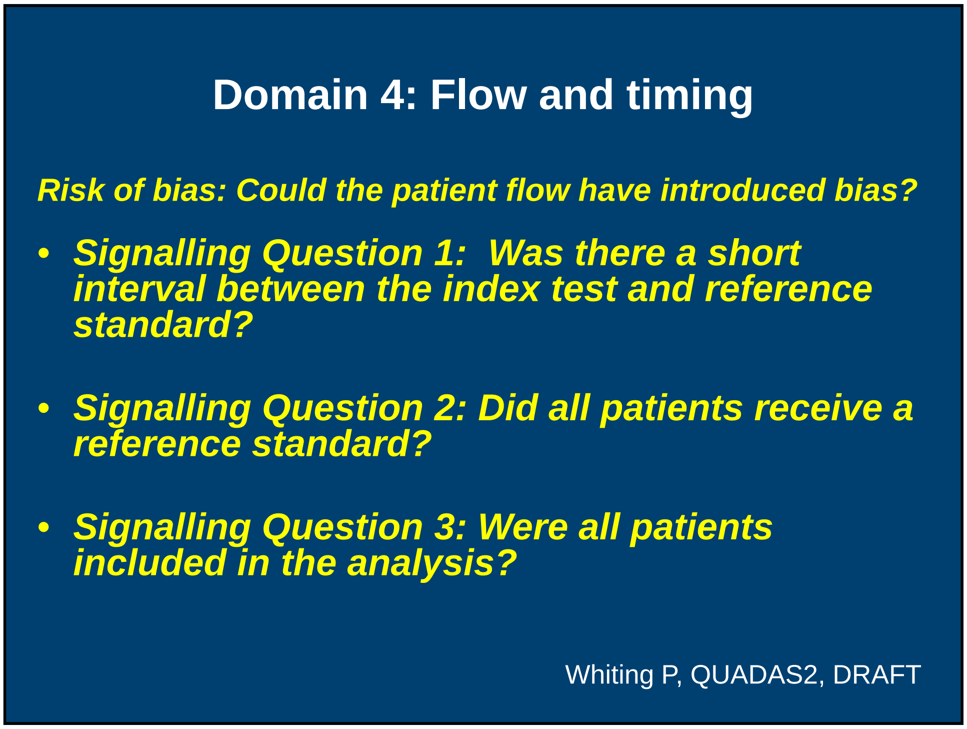


Domain 3: Reference Standard

Risk of bias: Could methods used to conduct or interpret reference standard have introduced bias?

- *Signalling Question 1: Is the reference standard likely to correctly classify the target condition?*
- *Signalling Question 2: Were the reference standard results interpreted without knowledge of the results of the index test?*

Whiting P, QUADAS2, DRAFT



Domain 4: Flow and timing

Risk of bias: Could the patient flow have introduced bias?

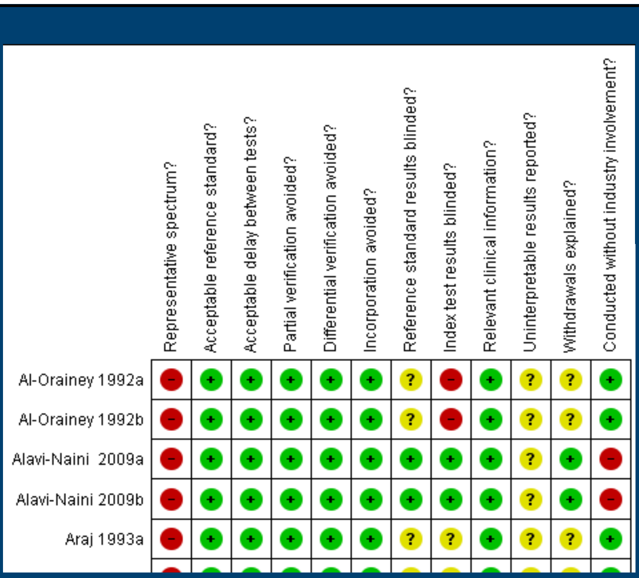
- *Signalling Question 1: Was there a short interval between the index test and reference standard?*
- *Signalling Question 2: Did all patients receive a reference standard?*
- *Signalling Question 3: Were all patients included in the analysis?*

Whiting P, QUADAS2, DRAFT

Applicability

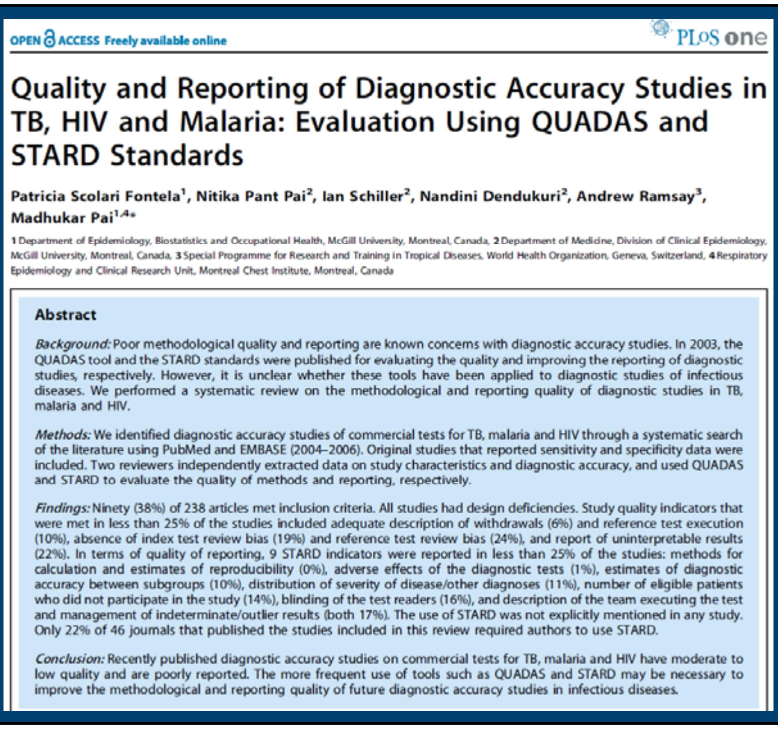
- ***Patient selection: Do the included patients and setting match the review question?***
- ***Index test: Does the test technology, execution and interpretation match the question?***
- ***Reference Standard: Does the target condition as defined by the reference standard match the question?***

Whiting P, QUADAS2, DRAFT



	Representative spectrum?	Acceptable reference standard?	Acceptable delay between tests?	Partial verification avoided?	Differential verification avoided?	Incorporation avoided?	Reference standard results blinded?	Index test results blinded?	Relevant clinical information?	Uninterpretable results reported?	Withdrawals explained?	Conducted without industry involvement?
Al-Orainey 1992a	−	+	+	+	+	+	?	−	+	?	?	+
Al-Orainey 1992b	−	+	+	+	+	+	?	−	+	?	?	+
Alavi-Naini 2009a	−	+	+	+	+	+	+	+	+	?	+	−
Alavi-Naini 2009b	−	+	+	+	+	+	+	+	+	?	+	−
Araj 1993a	−	+	+	+	+	+	?	?	+	?	?	+

Methodological quality summary: review authors' judgments about each methodological quality item for each included study, created in RevMan http://ims.cochrane.org/revman



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.

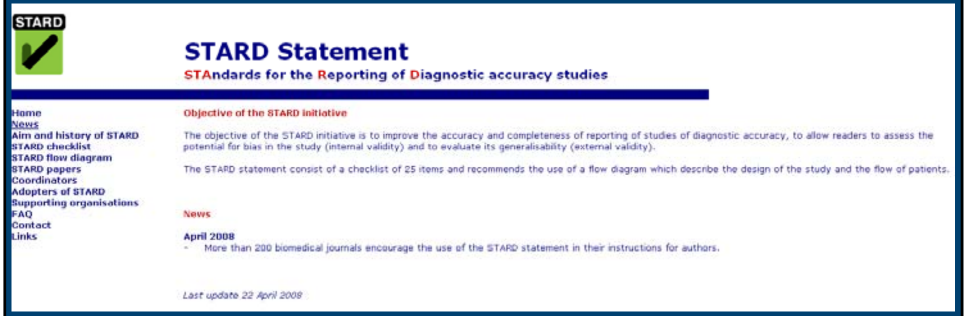
Conclusion: Recently published diagnostic accuracy studies on commercial tests for TB, malaria and HIV have moderate to low quality and are poorly reported. The more frequent use of tools such as QUADAS and STARD may be necessary to improve the methodological and reporting quality of future diagnostic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

Initiative to improve reporting of diagnostic accuracy studies



STARD Statement

STAndards for the Reporting of Diagnostic accuracy studies

Objective of the STARD initiative

The objective of the STARD initiative is to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study (internal validity) and to evaluate its generalisability (external validity).

The STARD statement consist of a checklist of 25 items and recommends the use of a flow diagram which describe the design of the study and the flow of patients.

News

April 2008

- More than 200 biomedical journals encourage the use of the STARD statement in their instructions for authors.

Last update 22 April 2008

http://www.stard-statement.org/

Thank you!

Questions?