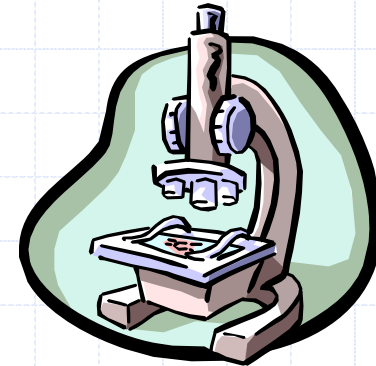
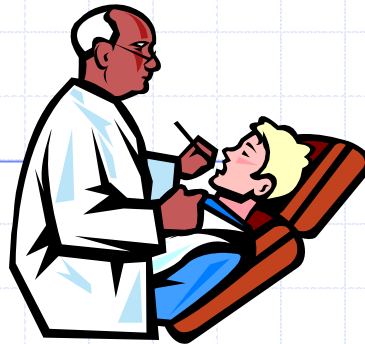
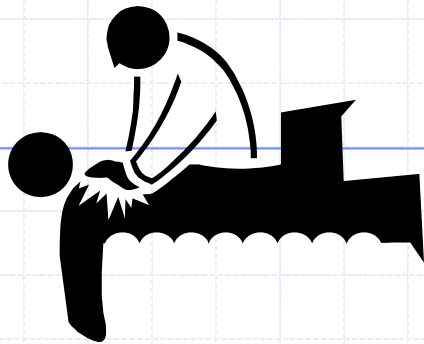


# Bias in diagnostic research



Madhukar Pai, MD, PhD

Assistant Professor of Epidemiology, McGill University

Montreal, Canada

Professor Extraordinary, Stellenbosch University, S Africa

Email: [madhukar.pai@mcgill.ca](mailto:madhukar.pai@mcgill.ca)

# Diagnostic trials lack methodologic rigor

## Diagnostic studies in 4 general medical journals

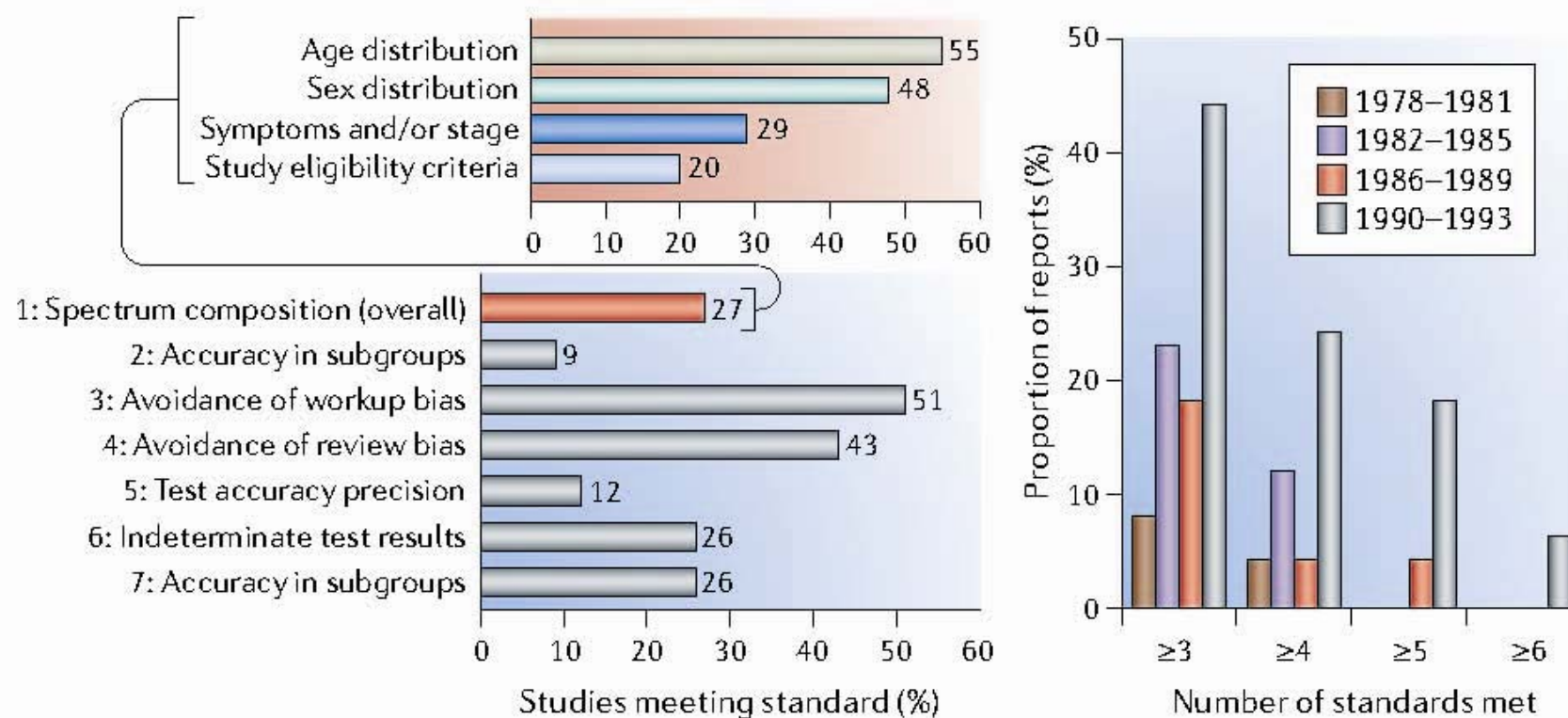


Figure 4 | **Proportion of diagnostic evaluations meeting accepted standards.** The seven standards are shown on the left. The data are taken from REF. 10.

# Lack of rigor: example from TB literature

12 meta-analysis with  
over 500 diagnostic  
studies

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

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- $\gamma$ 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]

\*At least single blind. †By reference standard.

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

## Performance of Purified Antigens for Serodiagnosis of Pulmonary Tuberculosis: a Meta-Analysis<sup>▽†</sup>

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

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) <sup>a</sup>
Index test results blinded to reference standard?	
Yes.....	65 (26)
No .....	1 (0)
Not reported.....	188 (74)

<sup>a</sup> The description of the test execution was deemed insufficient in one study.

STATISTICS IN MEDICINE, VOL. 6, 411-423 (1987)

## BIASES IN THE ASSESSMENT OF DIAGNOSTIC TESTS

COLIN B. BEGG

*Dana-Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115, U.S.A.*

### SUMMARY

Diagnostic tests are traditionally characterized by simple measures of efficacy such as the sensitivity and the specificity. These measures, though widely recognized and easy to understand, are subject to definitional arbitrariness. Moreover, studies constructed to estimate the sensitivity and specificity are susceptible to a variety of biases. In this paper the various potential problems are described with reference to examples from the diagnostic literature. These difficulties have implications for the design of diagnostic test evaluations, and the choice of suitable measures of test efficacy.

**KEY WORDS** Diagnostic test Bias Sensitivity Specificity



# 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 bias or variation in the results of diagnostic accuracy studies.

**Data Sources:** MEDLINE, EMBASE, and BIOSIS, and the methodologic databases of the Centre for Reviews and Dissemination and the Cochrane Collaboration. Methodologic experts in diagnostic tests were contacted.

**Study Selection:** Studies that investigated the effects of bias and variation on measures of test performance were eligible for inclusion, which was assessed by one reviewer and checked by a second reviewer. Discrepancies were resolved through discussion.

**Data Extraction:** Data extraction was conducted by one reviewer and checked by a second reviewer.

**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 reference standard, differential verification bias, and review bias, the amount of evidence was limited. Evidence was lacking for other features, including incorporation bias, treatment paradox, arbitrary choice of threshold value, and dropouts.

**Conclusions:** Many issues in the design and conduct of diagnostic accuracy studies can lead to bias or variation; however, the empirical evidence about the size and effect of these issues is limited.

*Ann Intern Med.* 2004;140:189-202.

For author affiliations, see end of text.

[www.annals.org](http://www.annals.org)

**Table 1. Description of Sources of Bias and Variation**

Source	Bias or Variation	Description
<b>Population</b>		
Demographic features	Variation	Tests may perform differently in various samples. Therefore, demographic features may lead to variations in estimates of test performance.
Disease severity	Variation	Differences in disease severity among studies may lead to differences in estimates of test performance.
Disease prevalence	Variation	The prevalence of the target condition varies according to setting and may affect estimates of test performance. Context bias, the tendency of interpreters to consider test results to be positive more frequently in settings with higher disease prevalence, may also affect estimates of test performance.
Distorted selection of participants	Variation	The selection process determines the composition of the study sample. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used in practice, the results of the study may have limited applicability.
<b>Test protocol: materials and methods</b>		
Test execution	Variation	A sufficient description of the execution of index and reference standards is important because variation in measures of diagnostic accuracy can be the result of differences in test execution.
Test technology	Variation	When the characteristics of a diagnostic test change over time as a result of technological improvement or the experience of the operator of the test, estimates of test performance may be affected.
Treatment paradox and disease progression bias	Bias	Disease progression bias occurs when the index test is performed an unusually long time before the reference standard, so the disease is at a more advanced stage when the reference standard is performed. Treatment paradox occurs when treatment is started on the basis of the knowledge of the results of the index test, and the reference standard is applied after treatment has started.
<b>Reference standard and verification procedure</b>		
Inappropriate reference standard	Bias	Errors of imperfect reference standard or standards bias the measurement of diagnostic accuracy of the index test.
Differential verification bias	Bias	Part of the index test results is verified by a different reference standard.
Partial verification bias	Bias	Only a selected sample of patients who underwent the index test is verified by the reference standard.
<b>Interpretation (reading process)</b>		
Review bias	Bias	Interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known when the reference standard is interpreted. Test review bias occurs when results of the reference standard are known while the index test is interpreted.
Clinical review bias	Bias	The availability of information on clinical data, such as age, sex, and symptoms, during interpretation of test results may affect estimates of test performance.
Incorporation bias	Bias	The result of the index test is used to establish the final diagnosis.
Observer variability	Variation	The reproducibility of test results is one of the determinants of diagnostic accuracy of an index test. Because of variation in laboratory procedures or observers, a test may not consistently yield the same result when repeated. In 2 or more observations of the same diagnostic study, intraobserver variability occurs when the same person obtains different results, and interobserver variability occurs when 2 or more people disagree.
<b>Analysis</b>		
Handling of indeterminate results	Bias	A diagnostic test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies; the uninterpretable results are simply removed from the analysis. This may lead to biased assessment of the test characteristics.
Arbitrary choice of threshold value	Variation	The selection of the threshold value for the index test that maximizes the sensitivity and specificity of the test may lead to overoptimistic measures of test performance. The performance of this cutoff in an independent set of patients may not be the same as in the original study.

# Sources of bias in diagnostic studies

- ◆ Bias due to an inappropriate reference standard
- ◆ Spectrum bias
- ◆ Verification (work-up) bias
  - Partial verification bias
  - Differential verification bias
- ◆ Review bias (lack of blinding)
- ◆ Incorporation bias
- ◆ Bias due to exclusions, indeterminates, etc



# Bias due to inappropriate or imperfect reference standard

- ◆ There is no such thing as a “gold” standard
- ◆ Imperfect reference standards are commonly used in diagnostic studies
  - Can lead to underestimation of test accuracy (under certain conditions)



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](#)

# Misclassification of disease status

◆ How accurately can the following be measured?

- Depression
- Tuberculosis in children
- Latent TB infection
- Appendicitis
- Dementia
- Migraine
- Attention deficit disorder
- Cause of death
- Irritable bowel syndrome
- Chronic fatigue syndrome
- Angina



Very rarely, you get tests that are nearly perfect (i.e. 100% sensitive and 100% specific)

OPEN ACCESS Freely available online



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

Nitika Pant Pai<sup>1\*</sup>, Rajnish Joshi<sup>2</sup>, Sandeep Dogra<sup>3</sup>, Bharati Taksande<sup>2</sup>, S. P. Kalantri<sup>2</sup>, Madhukar Pai<sup>4</sup>, Pratibha Narang<sup>2</sup>, Jacqueline P. Tulsy<sup>5</sup>, Arthur L. Reingold<sup>6</sup>

**1** Immunodeficiency Service, Montreal Chest Institute, McGill University Health Center, Montreal, Canada, **2** Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India, **3** Acharya Shri Chander College of Medical Sciences, Jammu, India, **4** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, **5** Department of Internal Medicine, University of California at San Francisco, San Francisco, California, United States of America, **6** Division of Epidemiology, University of California at Berkeley, Berkeley, California, United States of America

**Background.** Oral fluid-based rapid tests are promising for improving HIV diagnosis and screening. However, recent reports from the United States of false-positive results with the oral OraQuick® ADVANCE HIV1/2 test have raised concerns about their performance in routine practice. We report a field evaluation of the diagnostic accuracy, client preference, and feasibility for the oral fluid-based OraQuick® Rapid HIV1/2 test in a rural hospital in India. **Methodology/Principal Findings.** A cross-sectional, hospital-based study was conducted in 450 consenting participants with suspected HIV infection in rural India. The objectives were to evaluate performance, client preference and feasibility of the OraQuick® Rapid HIV-1/2 tests. Two Oraquick® Rapid HIV1/2 tests (oral fluid and finger stick) were administered in parallel with confirmatory ELISA/Western Blot (reference standard). Pre- and post-test counseling and face to face interviews were conducted to determine client preference. Of the 450 participants, 146 were deemed to be HIV sero-positive using the reference standard (seropositivity rate of 32% (95% confidence interval [CI] 28%, 37%)). 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). The OraQuick oral fluid-based test was preferred by 87% of the participants for first time testing and 60% of the participants for repeat testing. **Conclusion/Significance.** In a rural Indian hospital setting, the OraQuick® Rapid- HIV1/2 test was found to be highly accurate. The oral fluid-based test performed marginally better than the finger stick test. The oral OraQuick test was highly preferred by participants. In the context of global efforts to scale-up HIV testing, our data suggest that oral fluid-based rapid HIV testing may work well in rural, resource-limited settings.



# But even accurate tests run into problems!

OPEN ACCESS Freely available online



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

Krishna Jafa<sup>1,4\*</sup>, Pragna Patel<sup>1</sup>, Duncan A. MacKellar<sup>1</sup>, Patrick S. Sullivan<sup>1</sup>, Kevin P. Delaney<sup>1</sup>, Tracy L. Sides<sup>2a</sup>, Alexandra P. Newman<sup>3,4</sup>, Cindy M. Paul<sup>5</sup>, Evan M. Cadoff<sup>6</sup>, Eugene G. Martin<sup>6</sup>, Patrick A. Keenan<sup>7</sup>, Bernard M. Branson<sup>1</sup>, 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<sup>®</sup> 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<sup>®</sup> 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<sup>®</sup> 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

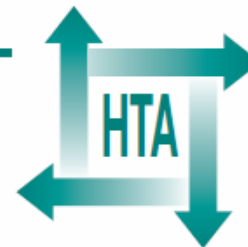
## Evaluation of diagnostic tests when there is no gold standard. A review of methods

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](http://www.hta.ac.uk)





# Four approaches

- ◆ Impute or adjust for missing data on reference standard; needs careful attention to the pattern and fraction of missing values.
- ◆ Correct imperfect reference standard; can be useful if there is reliable information about the degree of imperfection of the reference standard and about the correlation of the errors between the index test and the reference standard.
- ◆ Construct reference standard; combine multiple test results to construct a reference standard outcome including deterministic predefined rules, consensus procedures and statistical modelling (latent class analysis).
- ◆ Diagnostic test accuracy paradigm is abandoned and research examines, using a number of different methods, whether the results of an index test are meaningful in practice, for example by relating index test results to relevant other clinical characteristics and future clinical events.

## Using latent TB as an example, there are several approaches to the gold standard problem

- ◆ a) use the tuberculin skin test as the gold standard, but TST itself is an imperfect test;
- ◆ b) use both TST and IGRA and then use latent class analysis or mixture models
- ◆ c) use active TB as a surrogate for LTBI, but active disease can lead to depressed immunity;
- ◆ d) use a gradient of exposure among contacts of active cases, and examine if IGRA correlates more closely with exposure than the TST;
- ◆ e) use future progression from latency to active disease as the gold standard

# Active TB as gold standard

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

**Background:** Interferon- $\gamma$ -release assays (IGRAs) are alternatives to the tuberculin skin test (TST). A recent meta-analysis showed that IGRAs have high specificity, even among populations that have received bacille Calmette–Guérin (BCG) vaccination. Sensitivity was suboptimal for TST and IGRAs.

**Purpose:** To incorporate new evidence into an updated meta-analysis on the sensitivity and specificity of IGRAs.

**Data Sources:** PubMed was searched through 31 March 2008, and citations of all original articles, guidelines, and reviews for studies published in English were reviewed.

**Study Selection:** Studies that evaluated QuantiFERON-TB Gold, QuantiFERON-TB Gold In-Tube (both from Cellestis, Victoria, Australia), and T-SPOT.TB (Oxford Immunotec, Oxford, United Kingdom) or its precommercial ELISpot version, when data on the commercial version were lacking. For assessing sensitivity, the study sample had to have microbiologically confirmed active tuberculosis. For assessing specificity, the sample had to comprise healthy, low-risk individuals without known exposure to tuberculosis. Studies with fewer than 10 participants and those that included only immunocompromised participants were excluded.

**Data Extraction:** One reviewer abstracted data on participant characteristics, test characteristics, and test performance from 38 studies; these data were double-checked by a second reviewer. The original investigators were contacted for additional information when necessary.

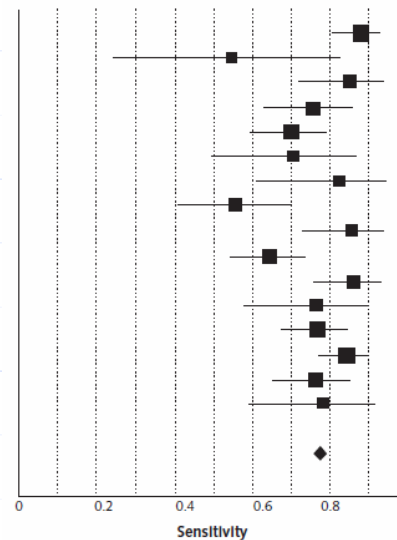
**Data Synthesis:** A fixed-effects meta-analysis with correction for overdispersion was done to pool data within prespecified subgroups. The pooled sensitivity was 78% (95% CI, 73% to 82%) for QuantiFERON-TB Gold, 70% (CI, 63% to 78%) for QuantiFERON-TB Gold In-Tube, and 90% (CI, 86% to 93%) for T-SPOT.TB. The pooled specificity for both QuantiFERON tests was 99% among non-BCG-vaccinated participants (CI, 98% to 100%) and 96% (CI, 94% to 98%) among BCG-vaccinated participants. The pooled specificity of T-SPOT.TB (including its precommercial ELISpot version) was 93% (CI, 86% to 100%). Tuberculin skin test results were heterogeneous, but specificity in non-BCG-vaccinated participants was consistently high (97% [CI, 95% to 99%]).

**Limitation:** Most studies were small and had limitations, including no gold standard for diagnosing latent tuberculosis and variable TST methods and cutoff values. Data on the specificity of the commercial T-SPOT.TB assay were limited.

**Conclusion:** The IGRAs, especially QuantiFERON-TB Gold and QuantiFERON-TB Gold In-Tube, have excellent specificity that is unaffected by BCG vaccination. Tuberculin skin test specificity is high in non-BCG-vaccinated populations but low and variable in BCG-vaccinated populations. Sensitivity of IGRAs and TST is not consistent across tests and populations, but T-SPOT.TB appears to be more sensitive than both QuantiFERON tests and TST.

Ann Intern Med. 2008;149.  
For author affiliations, see end of text.

www.annals.org



Study, Year (Reference)

Sensitivity (95% CI)

Patients, n/n


Mori et al., 2004 (7)	0.88 (0.81–0.93)	105/119
Ferrara et al., 2005 (8)	0.55 (0.23–0.83)	6/11
Ravn et al., 2005 (9)	0.85 (0.72–0.94)	41/48
Kang et al., 2005 (10)	0.76 (0.63–0.86)	44/58
Lee et al., 2006 (11)	0.70 (0.59–0.79)	61/87
Ferrara et al., 2006 (12)	0.71 (0.49–0.87)	17/24
Goletti et al., 2006 (13)	0.83 (0.61–0.95)	19/23
Dewan et al., 2007 (14)	0.56 (0.40–0.70)	25/45
Kobashi et al., 2006 (15)	0.86 (0.73–0.94)	43/50
Mazurek et al., 2007 (16)	0.65 (0.54–0.74)	62/96
Kang et al., 2007 (17)	0.87 (0.76–0.94)	58/67
Bua et al., 2007 (18)	0.77 (0.58–0.90)	23/30
Soysal et al., 2008 (19)	0.77 (0.68–0.85)	77/100
Kobashi et al., 2008 (26)	0.85 (0.77–0.90)	110/130
Nishimura et al., 2008 (27)	0.77 (0.66–0.86)	59/77
Kobashi et al., 2008 (28)	0.79 (0.59–0.92)	22/28

Pooled sensitivity = 0.78 (0.73–0.82)

Chi-square = 46.23;  $P < 0.001$

Inconsistency  $I^2 = 67.6\%$

# Latent class analysis

  
INT J TUBERC LUNG DIS 12(8):895–902  
© 2008 The Union

## Improving the estimation of tuberculosis infection prevalence using T-cell-based assay and mixture models

M. Pai,<sup>\*\*</sup> N. Dendukuri,<sup>\*\*</sup> L. Wang,<sup>§</sup> R. Joshi,<sup>¶</sup> S. Kalantri,<sup>¶</sup> H. L. Rieder<sup>#</sup>

<sup>\*</sup> Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, <sup>†</sup> Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, Montreal, <sup>‡</sup> Technology Assessment Unit, McGill University Health Center, Montreal, Quebec, <sup>§</sup> Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada; <sup>¶</sup> Mahatma Gandhi Institute of Medical Sciences, Sevagram, Maharashtra, India;

<sup>#</sup> International Union Against Tuberculosis and Lung Disease, Paris, France

### SUMMARY

**BACKGROUND:** The prevalence of latent tuberculosis infection (LTBI) is traditionally estimated using the tuberculin skin test (TST). Highly specific blood-based interferon-gamma release assays (IGRAs) are now available and could enhance the estimation of LTBI prevalence in combination with model-based methods.

**DESIGN:** We compared conventional and model-based methods for estimating LTBI prevalence among 719 Indian health care workers who underwent both TST and QuantiFERON-TB Gold In-Tube (QFT-G). In addition to using standard cut-off points on TST and QFT-G, Bayesian mixture model analyses were performed with: 1) continuous TST data and 2) categorical data using both TST and QFT-G results in a latent class analysis (LCA), accounting for prior information on sensitivity and specificity.

**RESULTS:** Estimates of LTBI prevalence varied from 33.8% to 60.7%, depending on the method used. The mixture model based on TST alone estimated the prevalence at 36.5% (95%CI 28.5–47.0). When results from both tests were combined using LCA, the prevalence was 45.4% (95%CI 39.5–51.1). The LCA provided additional results on the sensitivity, specificity and predictive values of joint results.

**CONCLUSION:** The availability of novel, specific IGRAs and development of methods such as mixture analyses allow a more realistic and informative approach to prevalence estimation.

**KEY WORDS:** tuberculosis; prevalence; tuberculin skin test; interferon-gamma release assay; mixture model; latent class analysis

**Table 3** Results on positive predictive values, sensitivity and specificity from latent class analysis model

Variable	Posterior distribution	
	Median %	95%CrI
<i>P</i> (LTBI+ TST+, QFT-G+)	99.2	99.0–100.0
<i>P</i> (LTBI+ TST+, QFT-G–)	46.0	29.0–65.0
<i>P</i> (LTBI+ TST–, QFT-G+)	85.0	69.0–94.0
<i>P</i> (LTBI+ TST–, QFT-G–)	2.0	1.0–4.0
Sensitivity of TST	79.5	74.9–84.4
Specificity of TST	87.4	82.3–91.8
Sensitivity of QFT-G	89.9	86.1–93.7
Specificity of QFT-G	97.4	94.2–98.9

CrI = credible interval; LTBI = latent tuberculosis infection; TST = tuberculin skin test; QFT-G = QuantiFERON-TB Gold In-Tube assay.

# Exposure gradient

## Comparison of T-cell-based assay with tuberculin skin test for diagnosis of *Mycobacterium tuberculosis* infection in a school tuberculosis outbreak

Katie Ewer, Jonathan Deeks, Lydia Alvarez, Gerry Bryant, Sue Waller, Peter Andersen, Phillip Monk, Ajit Lalvani

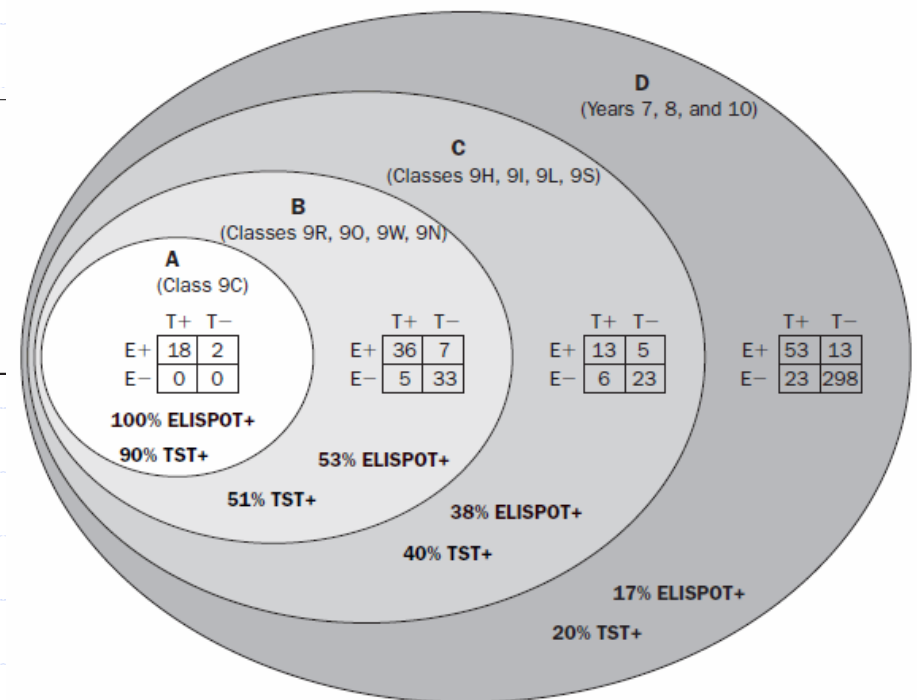


Figure 1: TST and ELISPOT results for students stratified by decreasing proximity to index case based on school year and class

T+=TST positive. T-=TST negative. E+=ELISPOT positive. ELISPOT-=ELISPOT negative.

A: students in same class as index case. B: students in classes in same year who regularly shared lessons with index case. C: students in the four remaining classes in same year who shared only weekly school events but no lessons with index case. D: students in different years who shared no school events with index case.



# Predictive value of IGRAs: longitudinal studies

JOURNAL OF CLINICAL MICROBIOLOGY, Feb. 2002, p. 704–706  
0095-1137/02/\$04.00+0 DOI: 10.1128/JCM.40.2.704–706.2002  
Copyright © 2002, American Society for Microbiology. All Rights Reserved.

Vol. 40, No. 2

## Immune Responses to the *Mycobacterium tuberculosis*-Specific Antigen ESAT-6 Signal Subclinical Infection among Contacts of Tuberculosis Patients

T. Mark Doherty,<sup>1\*</sup> Abebech Demissie,<sup>2</sup> Joseph Olobo,<sup>2</sup> Dawit Wolday,<sup>3</sup> Sven Britton,<sup>4</sup> Tewodros Eguale,<sup>5</sup> Pernille Ravn,<sup>6</sup> and Peter Andersen<sup>1</sup>

Department of Tuberculosis Immunology, Statens Serum Institute,<sup>1</sup> and Hvidovre Hospital,<sup>6</sup> Copenhagen, Denmark; Armauer Hansen Research Institute,<sup>2</sup> Black Lion Hospital,<sup>3</sup> and Hossana Regional Hospital, Ministry of Health,<sup>5</sup> Hossana, Ethiopia; and Karolinska Institute, Stockholm, Sweden<sup>4</sup>

High Incidence

OPEN ACCESS Freely available online



## Incidence of Tuberculosis and the Predictive Value of ELISPOT and Mantoux Tests in Gambian Case Contacts

Philip C. Hill<sup>1</sup>, Dolly J. Jackson-Sillah, Annette Fox, Roger H. Brookes, Bouke C. de Jong, Moses D. Lugos, Ifedayo M. Adetifa, Simon A. Donkor, Alex M. Aiken, Stephen R. Howie, Tumani Corrah, Keith P. McAdam, Richard A. Adegbola

Bacterial Diseases Programme, Medical Research Council (MRC) Laboratories, Banjul, The Gambia

## Predictive Value of a Whole Blood IFN- $\gamma$ Assay for the Development of Active Tuberculosis Disease after Recent Infection with *Mycobacterium tuberculosis*

Roland Diel<sup>1</sup>, Robert Loddenkemper<sup>2</sup>, Karen Meywald-Walter<sup>3</sup>, Stefan Niemann<sup>4</sup>, and Albert Nienhaus<sup>5</sup>

<sup>1</sup>School of Public Health, University of Düsseldorf, Düsseldorf, Germany; <sup>2</sup>German Central Committee against Tuberculosis, Lungenklinik Heckeshorn, HELIOS, Klinikum Emil von Behring, Berlin, Germany; <sup>3</sup>Public Health Department Hamburg-Mitte, Hamburg, Germany;

<sup>4</sup>National Reference Center for Mycobacteria, Research Center Borstel, Borstel, Germany; and <sup>5</sup>Institution for Statutory Accident Insurance and Prevention in the Health and Welfare Services, Hamburg, Germany

Low/Intermediate Incidence

Annals of Internal Medicine

ARTICLE

## Prognostic Value of a T-Cell–Based, Interferon- $\gamma$ Biomarker in Children with Tuberculosis Contact

Mustafa Bakir, MD; Kerry A. Millington, DPhil; Ahmet Soysal, MD; Jonathan J. Deeks, PhD; Serpil Efe; Yasemin Aslan, SRN; Davinder P.S. Dosanjh, DPhil; and Ajit Lalvani, DM

## Detection and Prediction of Active Tuberculosis Disease by a Whole-Blood Interferon- $\gamma$ Release Assay in HIV-1–Infected Individuals

Maximilian C. Aichelburg,<sup>1</sup> Armin Rieger,<sup>1</sup> Florian Breiteneker,<sup>1</sup> Katharina Pfistershammer,<sup>1</sup> Julia Tittes,<sup>1</sup> Stephanie Eltz,<sup>1</sup> Alexander C. Aichelburg,<sup>3</sup> Georg Stingl,<sup>1</sup> Athanasios Makristathis,<sup>2</sup> and Norbert Kohrgruber<sup>1,4</sup>

<sup>1</sup>Department of Dermatology, Division of Immunology, Allergy, and Infectious Diseases, and <sup>2</sup>Department of Hygiene and Medical Microbiology, Division of Clinical Microbiology, Vienna General Hospital, Medical University of Vienna, <sup>3</sup>Medical Department Pulmological Centre SMZ Baumgartner Höhe, Otto Wagner Hospital, and <sup>4</sup>Department of Dermatology and Venerology, Wilhelminen Hospital, Vienna, Austria

# Spectrum bias (a form of selection bias)

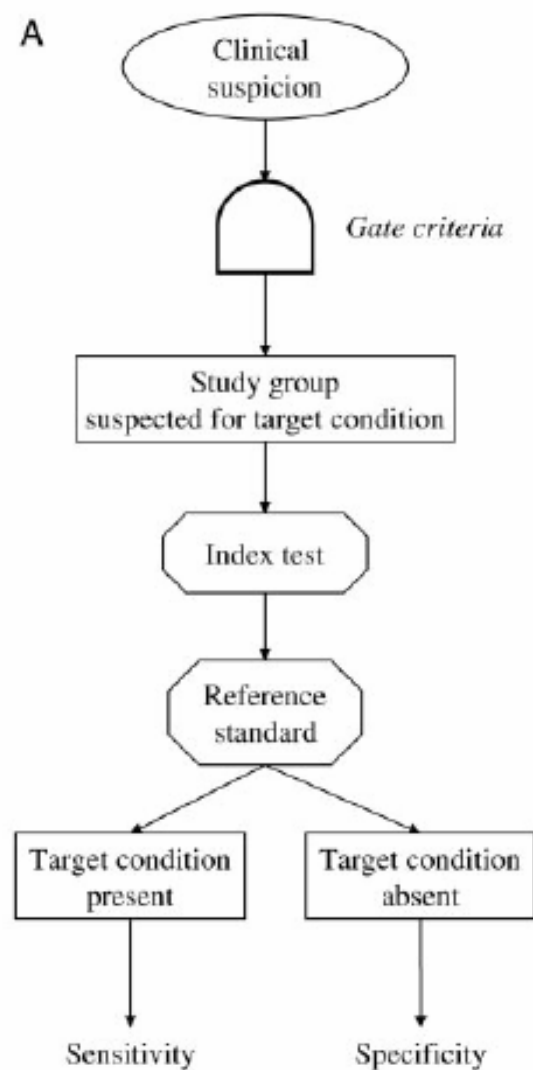
## ◆ Population used for evaluating the test:

- Extreme contrast
  - ◆ Case-control design
- Normal contrast (Indicated population)
  - ◆ Consecutively recruited patients in whom the disease is suspected

## ◆ Extreme contrast (spectrum bias) can result in overestimation of test accuracy

The selection process determines the composition of the study sample. If the selection process does not aim to include a patient spectrum similar to the population in which the test will be used in practice, the results of the study may have limited applicability.

A

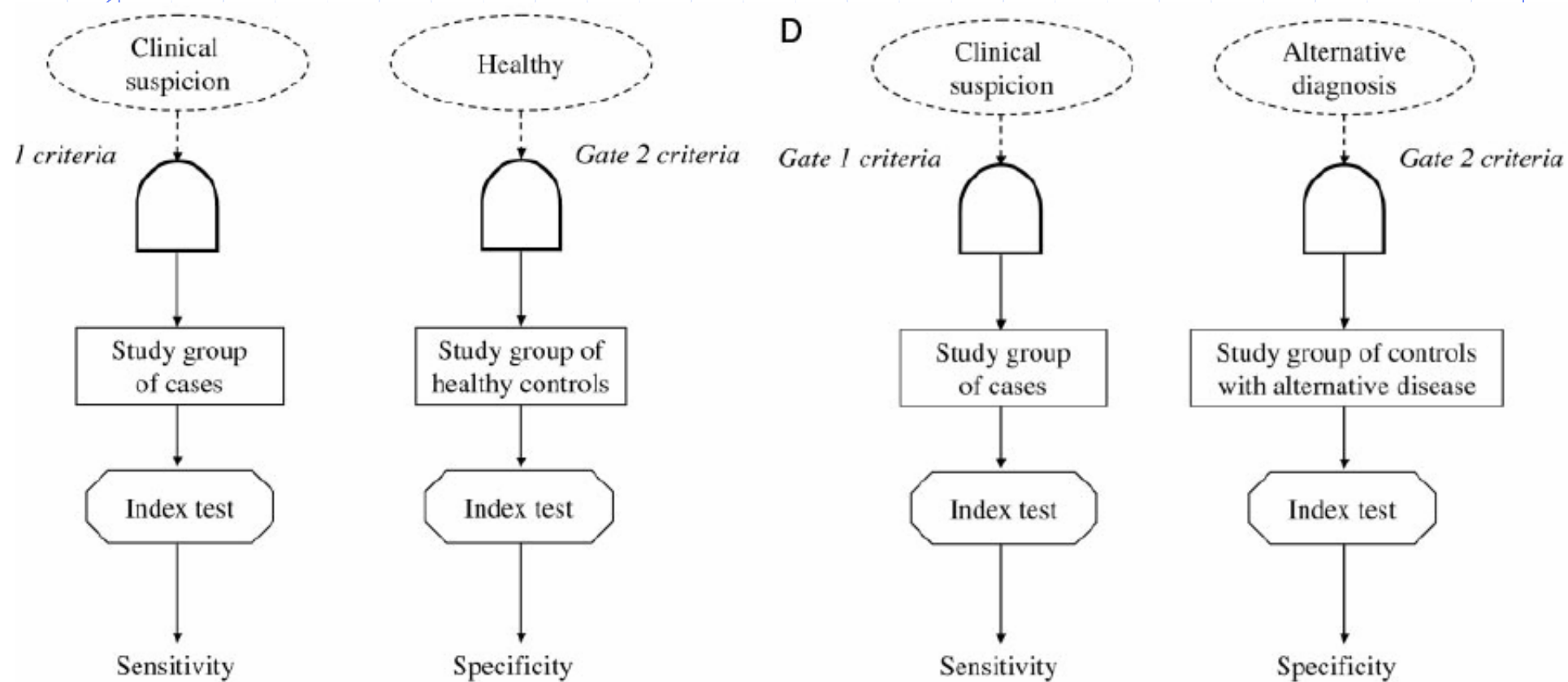


*Clinical Chemistry* 51:8  
1335–1341 (2005)

Minireview

## Case-Control and Two-Gate Designs in Diagnostic Accuracy Studies

ANNE W.S. RUTJES,<sup>1\*</sup> JOHANNES B. REITSMA,<sup>1</sup> JAN P. VANDENBROUCKE,<sup>2</sup> AFINA S. GLAS,<sup>3</sup> and  
PATRICK M.M. BOSSUYT<sup>1</sup>



# Spectrum bias example

- ◆ Story of carcinoembryonic antigen (CEA) for colorectal cancer:
  - Initial case-control showed high sens and spec; in advanced cancer vs normal people
  - In subsequent studies with less advanced cancer and patients with other disorders, the accuracy was significantly less
  - Clinicians were forced to abandon CEA



# Spectrum bias example

- ◆ Lachs et al. (1992) studied the leukocyte esterase and nitrite on a urine dipstick as predictors of a urinary tract infection (UTI), defined as a urine culture with greater than 105 bacteria/mL.
- ◆ They divided the 366 adults subjects in the study into those with high ( $>50\%$ ) and low ( $\leq 50\%$ ) prior probability of UTI, based on the signs and symptoms recorded by clinicians before obtaining the urine dipstick result, which was classified as positive if either the leukocyte esterase or nitrite was positive.
- ◆ They found marked differences in both sensitivity and specificity in 2 groups defined by prior probability:

**Table 5.3.** Differences in test characteristics of the urine dipstick in women at high and low prior probability of UTI, based on signs and symptoms (from Lachs et al. 1992)

	Sensitivity	Specificity	LR+	LR–
High Prior Prob.	92%	42%	1.6	0.19
Low Prior Prob.	56%	78%	2.5	0.56

# NAAT for TBM

## 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

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

**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
<b>Study design</b>			
Case-control	19	86.5 (39.3, 190.2)	0.03
Cross-sectional	16	43.3 (22.5, 83.3)	0.94
<b>Blinded interpretation of test and/or reference standard results</b>			
Yes	21	46.9 (24.9, 88.6)	0.16
No	14	82.3 (39.8, 170.2)	0.70
<b>Consecutive or random sampling of participants</b>			
Yes	18	63.3 (32.8, 122.4)	0.20
No	17	46.8 (23.6, 92.8)	0.42
<b>Prospective data collection</b>			
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.

## Performance of Purified Antigens for Serodiagnosis of Pulmonary Tuberculosis: a Meta-Analysis<sup>†</sup>

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

**TABLE 8. Specificity estimates by type of comparison**

Antigen name	Specificity (%) <sup>a</sup>	
	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)

<sup>a</sup> The data represent the posterior means (95% credible intervals) (number of studies).

# What is the right population for a diagnostic accuracy study?

- ◆ Those in whom we are uncertain of the diagnosis
- ◆ Those in whom we will use the test in clinical practice to resolve our uncertainty
- ◆ Patients with the disease who suffer from a wide spectrum of severity and patients without the disease who have other conditions that are commonly confused with the target disease

# Verification bias (selection bias)

## ◆ Verification bias in general:

- When the decision to perform the reference standard depends on the result of the index test
- When the type of reference standard used depends on the result of the index test

## ◆ Partial verification:

- Reference standard performed on test-positives, but not test-negatives

## ◆ Differential verification:

- Reference standard used for test-positives is different from that used for test-negatives

Part of the index test results is verified by a different reference standard.

Only a selected sample of patients who underwent the index test is verified by the reference standard.



# Verification bias: example

- ◆ PIOPED study on ventilation perfusion scan for pulmonary embolism:
  - Pulm angiography was the gold standard
  - Angio was more commonly done in patients with abnormal VQ scan results
  - Clinicians were reluctant to order angio in patients with low risk of pulmonary embolism
  - Researchers got around this problem by doing a 1 year follow up on patients who did not undergo angio - to make sure they were really negative

# Verification bias

In routine care, not every patient suspected of a particular disease undergoes the entire diagnostic work-up. Referral to subsequent testing is always based on previous test results. Hence, only a selected sample undergoes further testing or disease verification, including the reference test. In diagnostic research, ideally *all* patients undergo the entire diagnostic work-up, including the reference test, to determine the final diagnosis. It has been shown extensively that selective referral or work-up or disease verification leads to biased estimates of the accuracy of the tests under study, and still occurs in many (up to about 25%) of the published diagnostic studies [Lijmer et al., 1999; Rutjes et al., 2005; Whiting et al., 2004]. However, we prefer to discourage all of this different terminology for the same bias or problem, because it adds to the confusion surrounding diagnostic research. To prevent this bias, outcome assessment (“verification”) should be ensured in all patients in the design of data collection. This means that each study patient undergoes the reference test(s). If this is not feasible or deemed unethical, a clinical follow-up period and/or outcome panel to ultimately determine the presence or absence of the target disease in all patients could offer a solution (see above).

# How does verification bias work?

- ◆ Consider a study evaluating the usefulness of ankle swelling to predict a fracture on x-ray in patients with ankle injuries. X-rays are less likely to be ordered in patients with no swelling, and the study includes only those with x-rays.
- ◆ This design decreases the numbers of subjects with negative tests (no swelling), both with and without disease (fracture), as represented in cells C and D (table below):

	Fracture	No Fracture
Ankle Swelling	a	b
No Ankle Swelling	c↓	d↓

**Figure 5.1** How verification bias leads to overestimation of sensitivity and underestimation of specificity by lowering numbers in cells (c) and (d).

# Review bias

- ◆ Diagnostic studies may be:
  - Unblinded
  - Single blind (test or reference standard result is blinded)
  - Double blind (both test and ref. std results are blinded)
- ◆ Lack of blinding can lead to overestimation of test accuracy
- ◆ Examples: physical examination for ascitis and ultrasound, echo and cardiac murmur

Interpretation of the index test or reference standard is influenced by knowledge of the results of the other test. Diagnostic review bias occurs when the results of the index test are known when the reference standard is interpreted. Test review bias occurs when results of the reference standard are known while the index test is interpreted.

# Review bias

- ◆ Blinding is really important with “soft” outcomes (e.g. touch, physical signs, etc)
- ◆ Blinding is less relevant for a “hard” outcome (e.g. CD4 count, thyroxine levels)
- ◆ Lab tests can be easily blinded by coding specimens



# Incorporation bias

- ◆ If the test that is being evaluated is included in the reference standard
- ◆ Can lead to overestimation of test accuracy
- ◆ Can happen if final diagnosis is made on the basis of all clinical data (which might include the index test)
- ◆ Examples: PCR for tuberculosis, Mantoux for TB among kids, screening for depression

The result of the index test is used to establish the final diagnosis.

# Incorporation bias: example

- ◆ A study was done on screening instruments for depression in terminally ill people
- ◆ The authors reported 100% sens and 100% spec for a single question: ‘are you depressed?’ to detect depression
- ◆ Their diagnostic test included 9 questions, of which 1 was “Are you depressed”?

# Bias due to exclusions, indeterminates, missing data

- ◆ In real life studies, several problems can occur:
  - Drop-out of patients who don't complete all the tests
  - Invalid results
  - Indeterminate results
  - Insufficient specimen volume
- ◆ Should these results be excluded for computation of accuracy measures?

A diagnostic test can produce an uninterpretable result with varying frequency depending on the test. These problems are often not reported in test efficacy studies; the uninterpretable results are simply removed from the analysis. This may lead to biased assessment of the test characteristics.

# Bias due to exclusions, indeterminates, missing data

## ◆ Example:

- Manuscript entitled “High sensitivity of IGRA in HIV+ TB patients”
- ~90% sensitivity of IGRA
  - ◆ But nearly 30% of all patients had indeterminate IGRA results!
  - ◆ These results were excluded for computation of sensitivity
- How should the authors have addressed this problem? Is their title justified??

In reality, the 2 x 2 table, should be a 3 x 3 table:

		Reference standard		
		Pos	Neg	Invalid/Missing
Index test	Pos	a	b	c
	Neg	d	e	f
	Invalid/missing	g	h	i

If the invalid/missing rows and columns are excluded then we get the standard 2 x 2 table



# Do design flaws affect study results?

## 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 number of available diagnostic tests has been rapidly increasing. As for all new medical technologies, new diagnostic tests should be thoroughly evaluated prior to their introduction into daily practice. The number of test evaluations in the literature is increasing but the methodological quality of these studies is on average poor. A survey of the diagnostic literature (1990-1993) showed that only 18% of the studies satisfied 5 of the 7 methodological standards examined.<sup>1</sup> Different guidelines have been written to help physicians with the critical appraisal of the diagnostic literature consisting of lists of criteria for the assessment of study quality.<sup>2-4</sup> Criteria enable

**Context** The literature contains a large number of potential biases in the evaluation of diagnostic tests. Strict application of appropriate methodological criteria would invalidate the clinical application of most study results.

**Objective** To empirically determine the quantitative effect of study design shortcomings on estimates of diagnostic accuracy.

**Design and Setting** Observational study of the methodological features of 184 original studies evaluating 218 diagnostic tests. Meta-analyses on diagnostic tests were identified through a systematic search of the literature using MEDLINE, EMBASE, and DARE databases and the Cochrane Library (1996-1997). Associations between study characteristics and estimates of diagnostic accuracy were evaluated with a regression model.

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

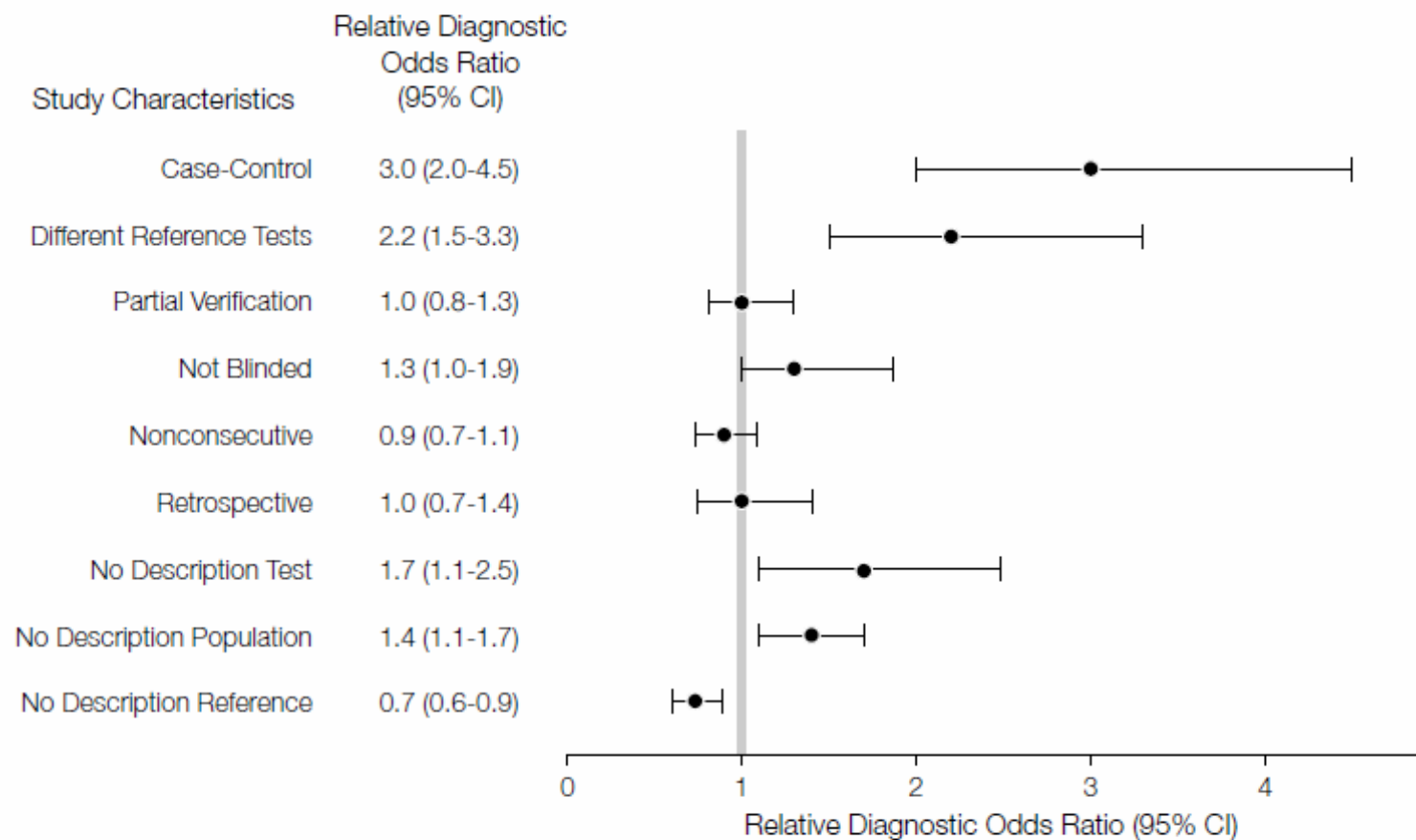
**Results** Fifteen (6.8%) of 218 evaluations met all 8 criteria; 64 (30%) met 6 or more. Studies evaluating tests in a diseased population and a separate control group overestimated the diagnostic performance compared with studies that used a clinical population (RDOR, 3.0; 95% confidence interval [CI], 2.0-4.5). Studies in which different reference tests were used for positive and negative results of the test under study overestimated the diagnostic performance compared with studies using a single reference test for all patients (RDOR, 2.2; 95% CI, 1.5-3.3). Diagnostic performance was also overestimated when the reference test was interpreted with knowledge of the test result (RDOR, 1.3; 95% CI, 1.0-1.9), when no criteria for the test were described (RDOR, 1.7; 95% CI, 1.1-2.5), and when no description of the population under study was provided (RDOR, 1.4; 95% CI, 1.1-1.7).

**Conclusion** These data provide empirical evidence that diagnostic studies with methodological shortcomings may overestimate the accuracy of a diagnostic test, particularly those including nonrepresentative patients or applying different reference standards.

*JAMA.* 1999;282:1061-1066

[www.jama.com](http://www.jama.com)

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



# Do design flaws affect study results?

## 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*.

#### ABSTRACT

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

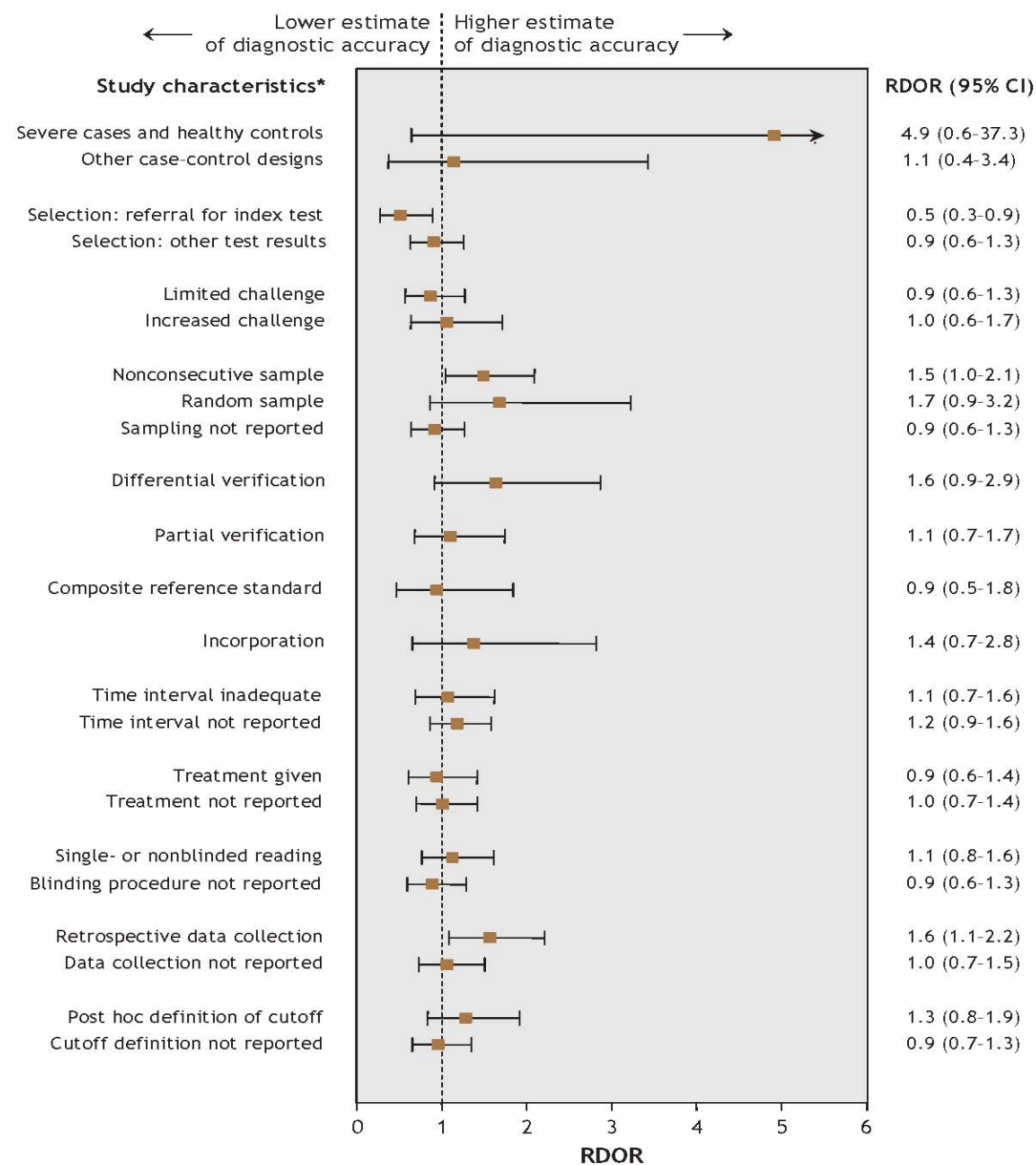
**Methods:** We identified meta-analyses of the diagnostic accuracy of tests through an electronic search of the databases MEDLINE, EMBASE, DARE and MEDION (1999–2002). We included meta-analyses with at least 10 primary studies without preselection based on design features. Pairs of reviewers independently extracted study characteristics and original data from the primary studies. We used a multivariable meta-epidemiologic regression model to investigate the direction and strength of the association between 15 study features on estimates of diagnostic accuracy.

**Results:** We selected 31 meta-analyses with 487 primary studies of test evaluations. Only 1 study had no design deficiencies.

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

Shortcomings in the design of clinical trials are known to affect results. The biasing effects of inadequate randomization procedures and differential dropout have been discussed and demonstrated in several publications.<sup>4–6</sup> A growing understanding of the potential sources of bias and variation has led to the development of guidelines to help researchers and readers in the reporting and appraisal of results from randomized trials.<sup>7,8</sup> More recently, similar guidelines have been published to assess the quality of reporting and design of studies evaluating the diagnostic accuracy of tests. For many of the items in these guidelines, there is no or limited empirical evidence available on their potential for bias.<sup>9</sup>

In principle, such evidence can be collected by comparing studies that have design deficiencies with studies of the same type that have no such imperfections. Several large meta-



\*See Appendix 2 for descriptions of the study characteristics.

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

Table 16–2 Empirical Evidence of Sources of Bias in Diagnostic Accuracy Studies<sup>a</sup>

	Lijmer et al <sup>3</sup> (RDOR; 95% CI)	Whiting et al <sup>5</sup>	Rutjes et al <sup>4</sup> (RDOR; 95% CI)
Did participating patients present a diagnostic dilemma?	Case-control design (3.0; 2.0–4.5)	Distorted selection of participants (some empirical support)	Case-control design (4.9; 0.6–37.3)
	Nonconsecutive patient selection (0.9; 0.7–1.1)		Nonconsecutive sampling (1.5; 1.0–2.1)
	Retrospective data collection (1.0; 0.7–1.4)		Retrospective data collection (1.6; 1.1–2.2)
Did investigators compare the test to an appropriate, independent reference standard?		Inappropriate reference standard (some empirical support)	
		Incorporation bias (using test as part of reference standard) (no empirical support)	Incorporation (1.4; 0.7–2.8)
Were those interpreting the test and reference standard blind to the other result?	Not blinded (1.3; 1.0–1.9)	Review bias (some empirical support)	Single or nonblinded reading (1.1; 0.8–1.6)
Did investigators perform the same reference standard to all patients regardless of the results of the test under investigation?	Different reference tests (2.2; 1.5–3.3)	Differential verification bias (some empirical support)	Differential verification (1.6; 0.9–2.9)
	Partial verification (1.0; 0.8–1.3)	Partial verification bias (strong empirical support)	Partial verification (1.1; 0.7–1.7)

Abbreviations: CI, confidence interval; RDOR, relative diagnostic odds ratio.

<sup>a</sup>RDOR, point estimates, and 95% CIs are shown.



Bias Type	General description	Specific situations	Sensitivity is falsely...	Specificity is falsely...
Incorporation bias	Classification of disease status partly depends on the results of the index test. Gold standard incorporates the index test.		↑	↑
Verification bias	Patients with positive index tests are more likely to get the gold standard, and only patients who get the gold standard are included in the study.		↑	↓
Double gold standard bias	Patients with a positive index test are more likely to receive one (often invasive) gold standard, whereas patients with a negative index test are more likely to receive a different gold standard (often clinical follow-up). Bias occurs only if there is a subgroup where the two gold standards give different answers.	For disease that can resolve spontaneously.	↑	↑
		For disease that becomes detectable during the follow-up period.	↓	↓
Spectrum bias	Spectrum of disease and nondisease differs from clinical practice. Sensitivity depends on spectrum of disease. Specificity depends on spectrum of nondisease or of diseases that might mimic the disease of interest.	When disease is skewed toward higher severity than in clinical practice – “sickest of the sick.”	↑	NA
		When nondisease is skewed toward greater health – “weldest of the well.”	NA	↑



# Critical appraisal of diagnostic studies

# How to critically appraise diagnostic studies?

- ◆ Users' Guides to the Medical Literature
- ◆ QUADAS
- ◆ Several others

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

[Customize your home page](#)

## 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*

### Penicillin Allergy

A 12-year-old boy with pharyngitis has a positive rapid streptococcal test result for strep... [Continue reading](#)

See also: [Make the Diagnosis: Penicillin Allergy](#)

## My JAMAevidence

Get free access to personalized features such as **saved images and tables**, **saved worksheets and question wizards**, and **bookmarking**.

### Log in to your personal profile

Username

Password

Log In

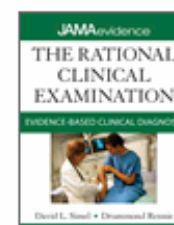
[Forgot your username/password?](#)

### Create a Personal Profile

## 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  
*Includes online-only content*

# Users' Guides 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?
-



# QUADAS tool for quality assessment of diagnostic studies

**BMC Medical Research  
Methodology**



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\*<sup>1</sup>, Anne WS Rutjes<sup>2</sup>, Johannes B Reitsma<sup>2</sup>,  
Patrick MM Bossuyt<sup>2</sup> and Jos Kleijnen<sup>1</sup>

# QUADAS tool for quality assessment of diagnostic studies

Table 1: QUADAS

Item #	Description
1.	Was the spectrum of patients representative of the patients who will receive the test in practice?
2.	Were selection criteria clearly described?
3.	Is the reference standard likely to correctly classify the target condition?
4.	Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (disease progression bias)
5.	Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis? (partial verification bias)
6.	Did patients receive the same reference standard regardless of the index test result? (differential verification bias)
7.	Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (incorporation bias)
8.	Was the execution of the index test described in sufficient detail to permit replication of the test?
9.	Was the execution of the reference standard described in sufficient detail to permit its replication?
10.	Were the index test results interpreted without knowledge of the results of the reference standard? (test review bias)
11.	Were the reference standard results interpreted without knowledge of the results of the index test? (diagnostic review bias)
12.	Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (clinical review bias)
13.	Were uninterpretable/ intermediate test results reported?
14.	Were withdrawals from the study explained?

# Quality of diagnostic accuracy studies: evaluation using QUADAS and STARD standards

Fontela PS, Pai NP, Schiller I, Dendukuri N, Ramsay A, Pai M

**METHODS:** We identified diagnostic studies of tests for tuberculosis, malaria and HIV through a systematic search of the literature using MEDLINE and EMBASE databases (2004-2006). Original studies about commercial tests that presented a cross tabulation of the results were included. Two reviewers independently extracted data on study characteristics and diagnostic accuracy. We used QUADAS and STARD criteria to evaluate the quality of reporting.

**RESULTS:** Ninety (38%) of 238 papers were selected. All the studies presented design deficiencies. Quality items that were present in less than 25% of the studies included description of withdrawals (6%), adequate description of the reference test execution (10%), absence of index test review bias (19%), report of uninterpretable results (22%), and absence of reference test review bias (24%). In terms of quality of report, nine STARD items 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 QUADAS or STARD was not mentioned in any study. Only 22% of the journals studied required authors to use STARD.