Systematic reviews of diagnostic test accuracy studies

McGill Summer Institute, Montreal June 2017

Karen R Steingart, MD, MPH
Cochrane Infectious Diseases Group
Liverpool School of Tropical Medicine

karen.steingart@gmail.com

Declarations of Interest

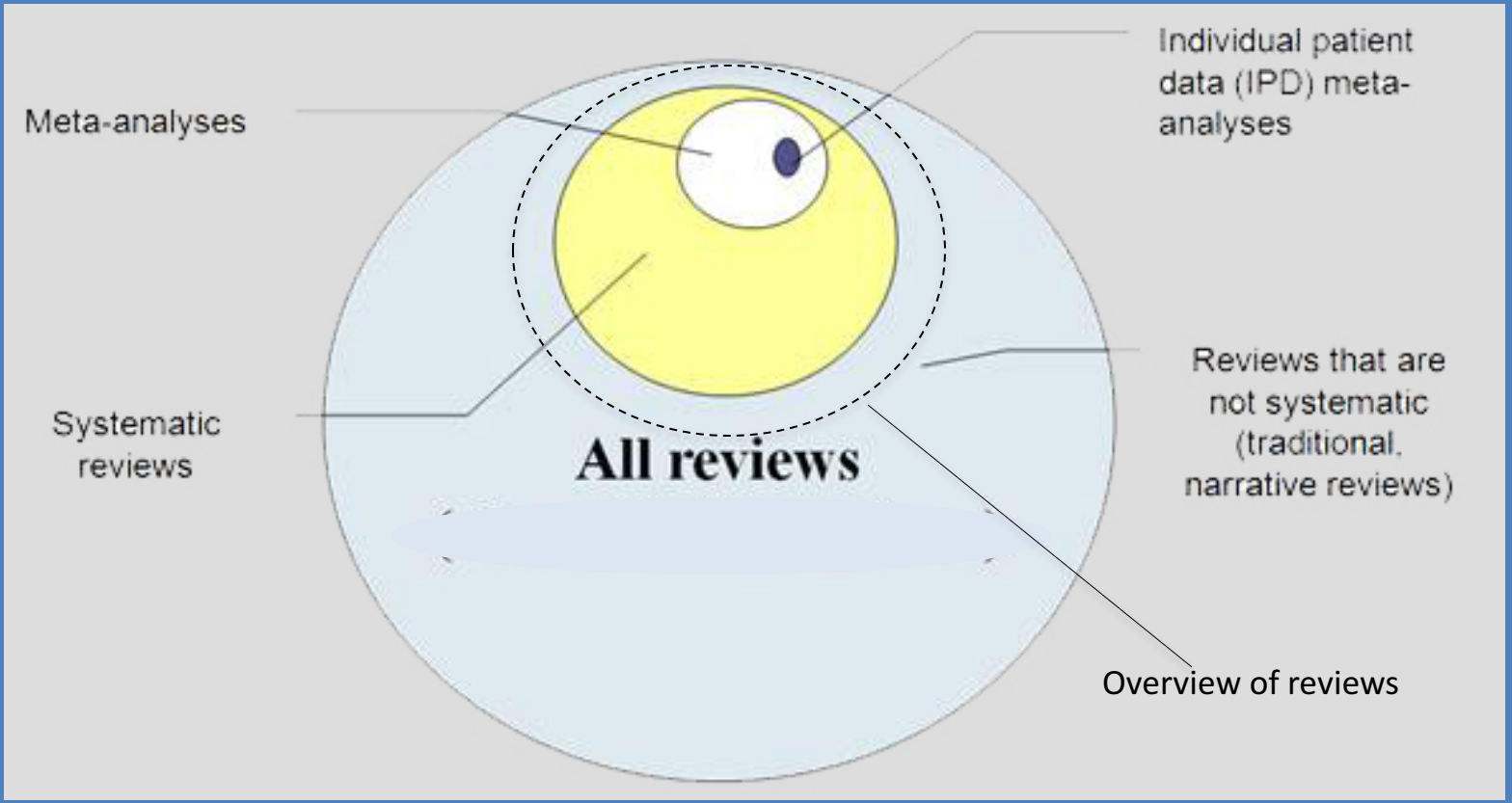
- Editor, Cochrane Infectious Diseases Group and Cochrane Diagnostic Test Accuracy Editorial Team

- No financial disclosures

Overview

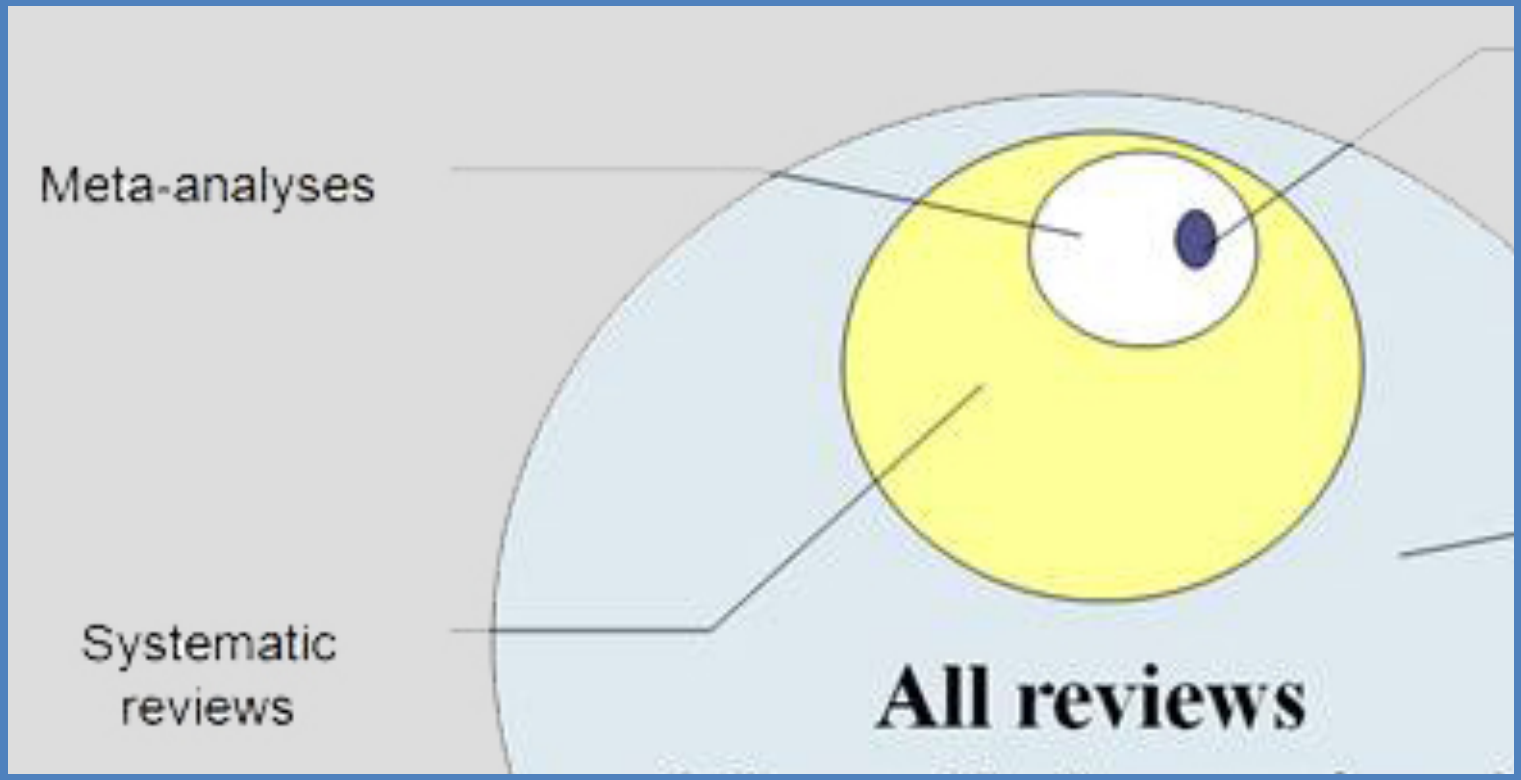
- Systematic reviews (DTA systematic reviews): what, why, and how
- Formulate a review question, PICO
- Describe quality assessment of DTA studies
- Describe methods for analyzing data, including meta-analysis
- Interpret results

- DTA = diagnostic test accuracy



A systematic review starts with a **clearly formulated question** and uses **systematic** and **explicit methods** to identify, select, and **critically appraise relevant research**, and to **collect** and **analyse** data from the studies that are included in the review

Egg slide adapted from Madhu Pai

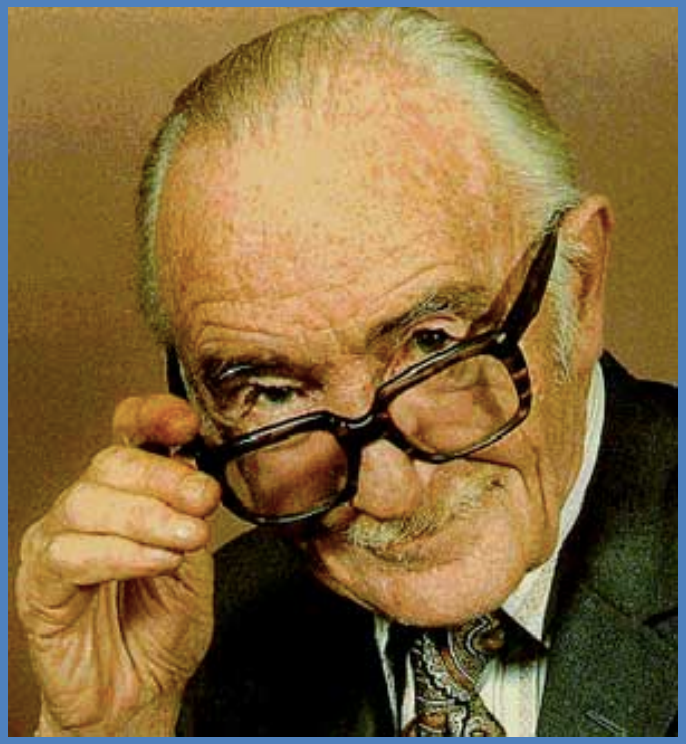


A meta-analysis is the use of statistical techniques in a systematic review to integrate the results of included studies and analyse data from the studies that are included in the review.

Not all systematic reviews include a meta-analysis.

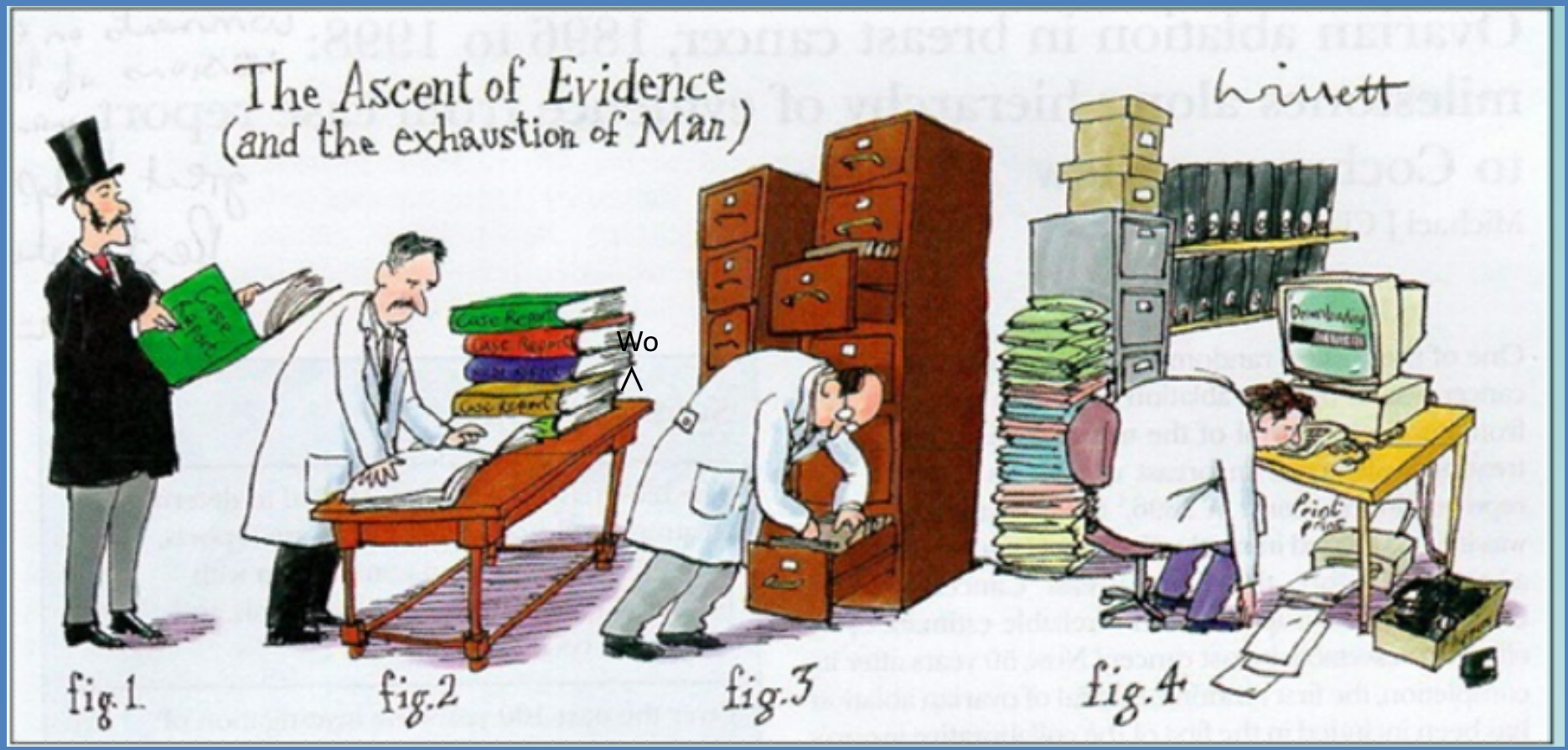
Why undertake a systematic review?

- **It is surely a great criticism of our profession that we have not organized a critical summary, by specialty and subspecialty, adapted periodically, of all relevant randomised controlled trials”**



Professor Archibald Leman Cochrane, CBE FRCP FFCM, (1909-1988)

- Archie L Cochrane, 1979 ‘Medicines for the year 2000’ London. Office for Health Economics. p1-11



- **The growing amount of information is challenging to manage**
- **Decision makers need to integrate critical pieces of information**
- **Systematic reviews are an efficient scientific technique**

Mulrow BMJ 1994

Systematic reviews of diagnostic test accuracy studies - terminology

- Index test(s) - new test
- Target condition - is a particular disease that the index test is intended to identify
- Reference standard (gold standard) - an agreed-upon, accurate method for identifying patients who have the target condition

Diagnostic test accuracy - definition

- Diagnosis asks, “does this person have this disease (more generally, this target condition) at this point in time?”

- Diagnostic test accuracy refers to the ability of a test to distinguish between patients with disease and those without disease

Reference standard - a few comments

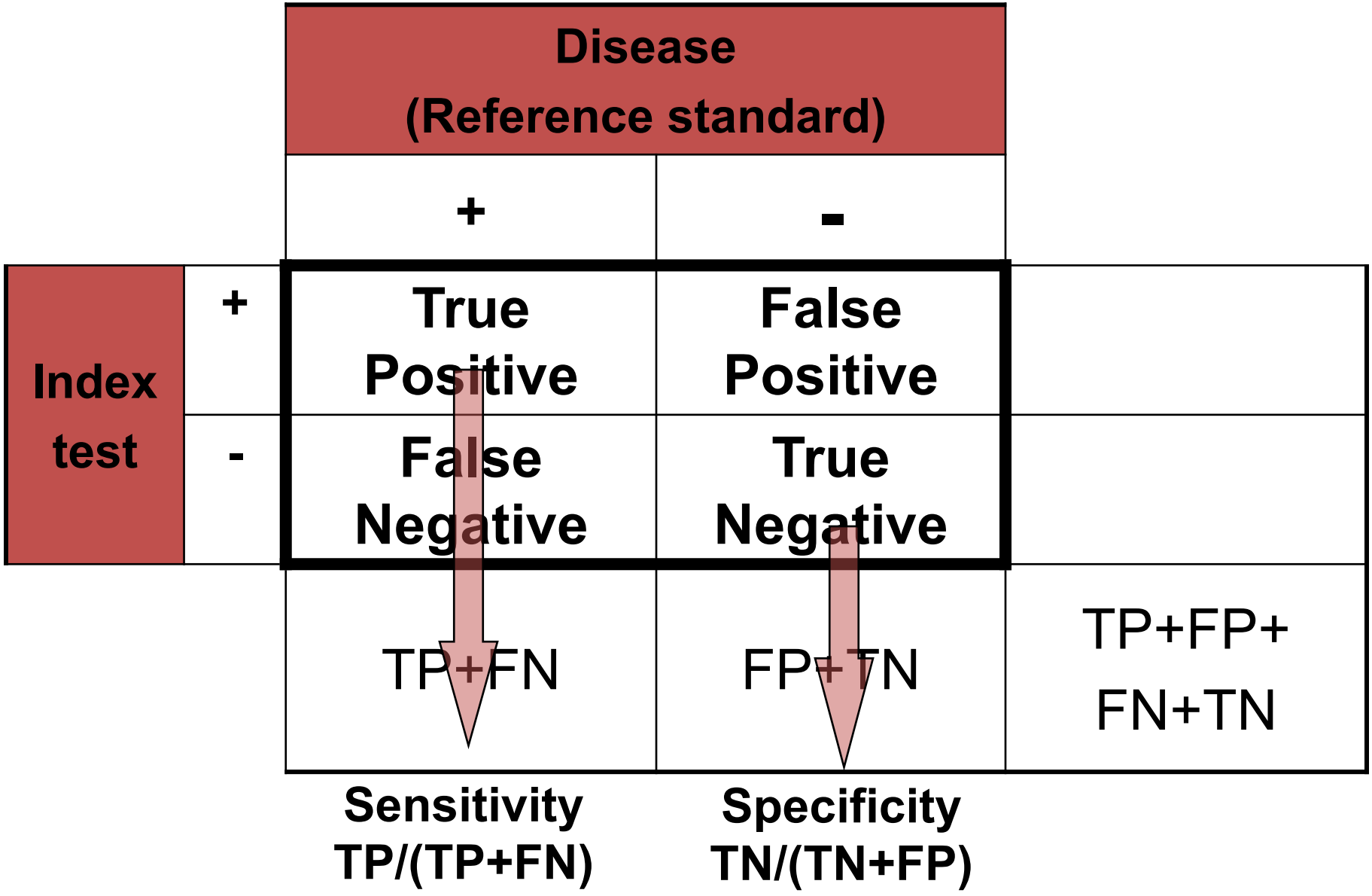
- The accuracy of an index test cannot be measured without a reference standard
- There should be general agreement that the reference standard is more accurate than the index test
- There may be more than one reference standard
- The reference standard may comprise several pieces of information
- The most accurate reference standard may not be feasible or ethical

Diagnostic test accuracy – what are we measuring?

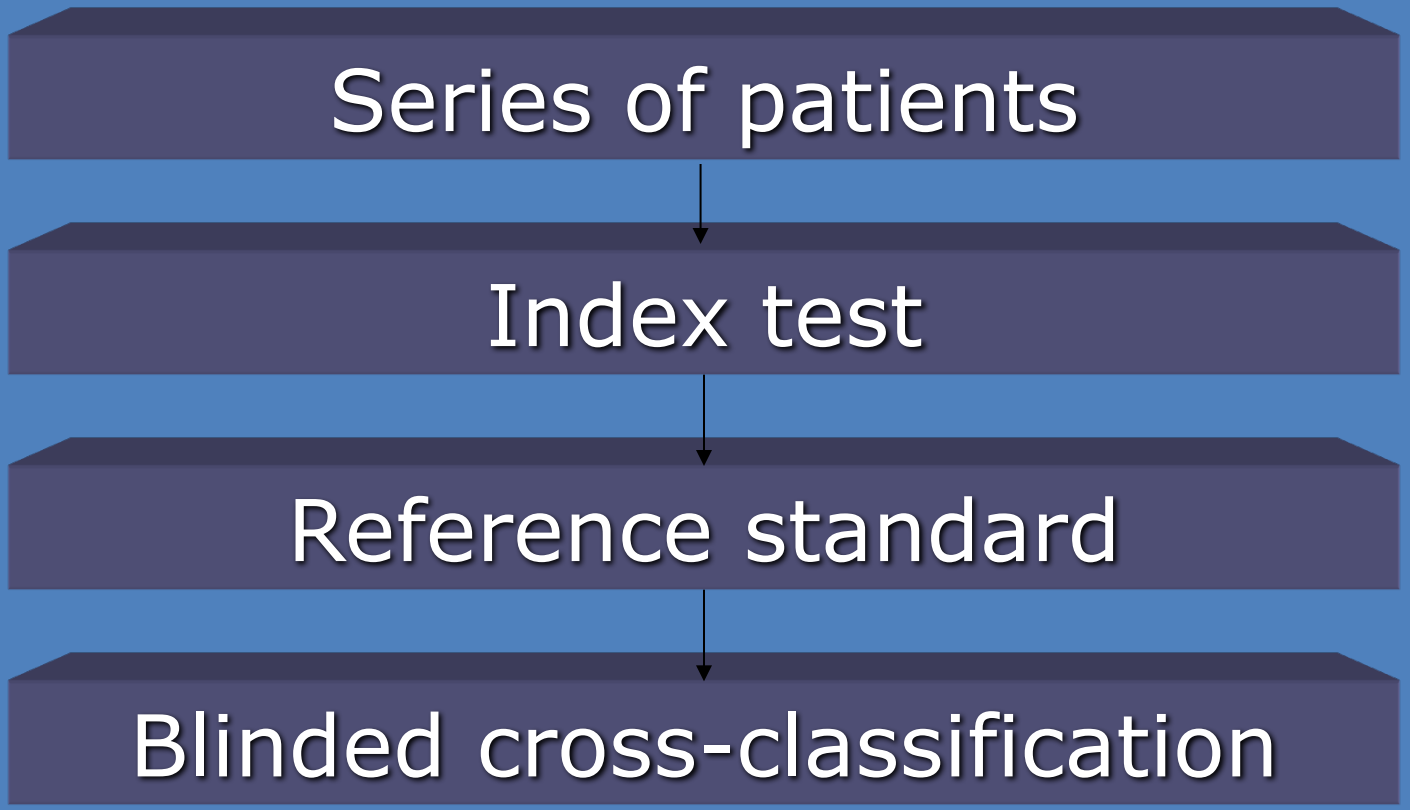
- Determine agreement between the results of the index test and the reference standard
  - Usually estimate sensitivity and specificity
  - Requires a 2×2 table

True positive	False positive
False negative	True negative

2x2 Table - sensitivity and specificity



Diagnostic accuracy cross-sectional study design



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

Tests do not make patients better

Diagnostic accuracy ≠ patient outcome

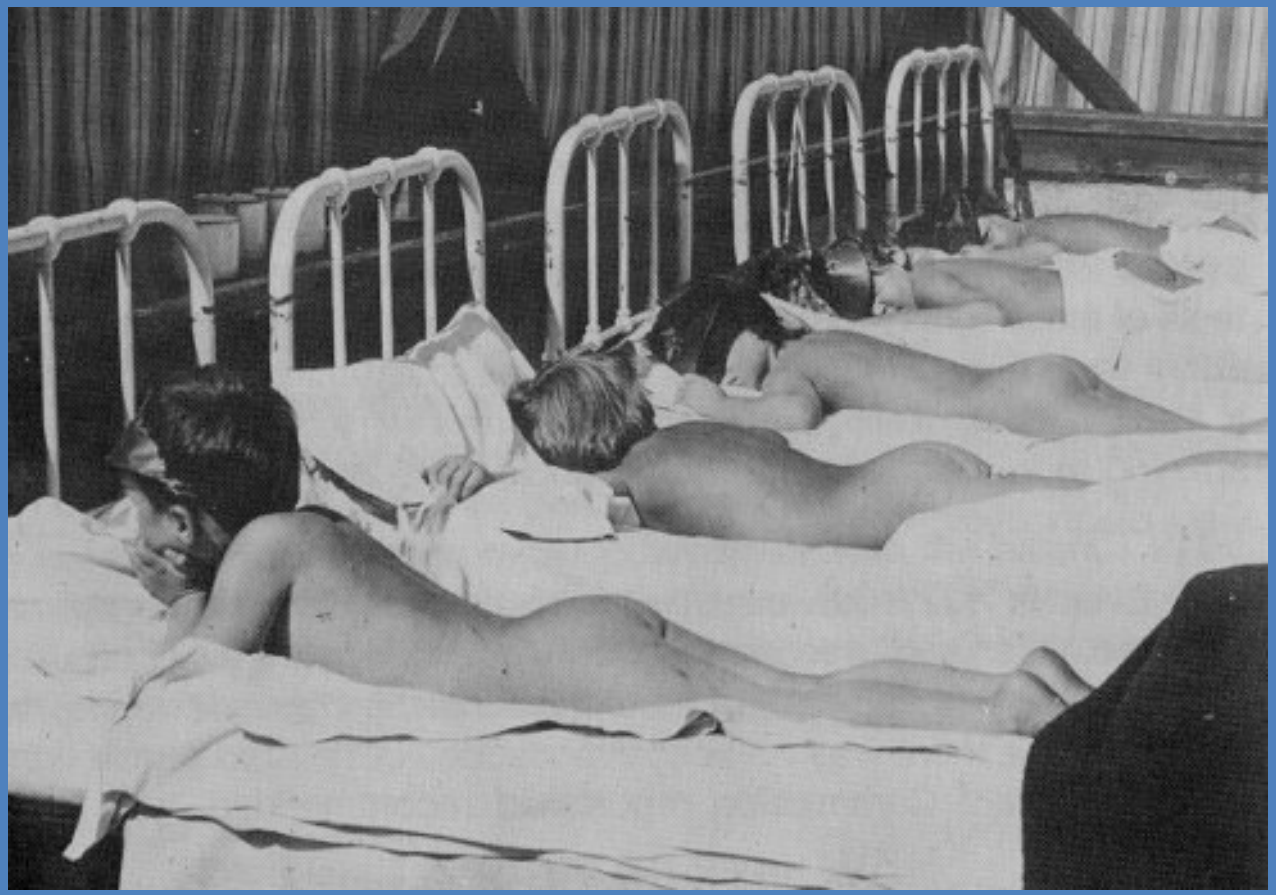


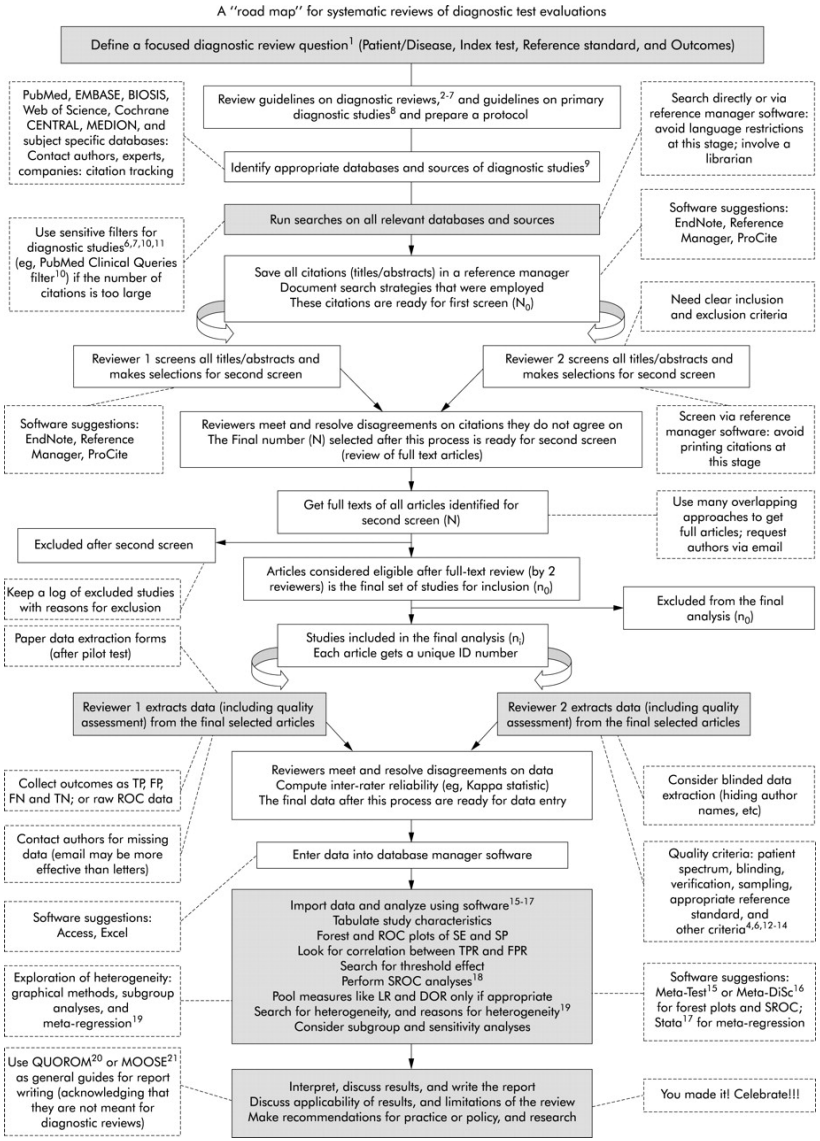
Image:
http://www.lung.ca/tb/images/full_archiv

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

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

di Ruffano, BMJ 2012

Road map for diagnostic accuracy reviews



Pai M et al. Systematic reviews of diagnostic test evaluations: what's behind the scenes? Evid Based Med 2004;9:101-103

©2004 by BMJ Publishing Group Ltd

EBM

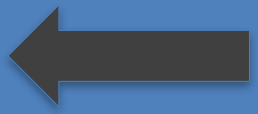
10 steps in performing a systematic review of diagnostic test accuracy studies

1. Define the review question and selection criteria
2. Write a protocol (plan the methods)
3. Search for studies
4. Apply selection criteria
5. Collect the data
6. Assess methodological quality
7. Analyze the data
8. Interpret results
9. Draw conclusions
10. Improve and update the review

Define the review question and selection (inclusion and exclusion) criteria

Begin with a well-framed question - 1

- **P**articipants
- **I**ndex test(s)



PI(C,O)

The index test was Alere Determine™ TB LAM Ag test. We evaluated the test at two different cut-off values for positivity (grade 1 and grade 2) based on the original manufacturer reference card.

- (**C**omparator test)
- **O**utcomes are usually sensitivity (true positives and false negatives) and specificity (false positives and true negatives)

Begin with a well-framed question - 2

- How do the participants present?
- What is the purpose of the test (such as screening, diagnosis, prognosis, monitoring)?
- What is the role of the new test (replacement, triage, add-on, combination with other tests?)

Role of tests

- Replacement: new test replaces existing test
- Triage: new test goes before existing test and only patients with a particular result go on to receive further testing
- Add-on: new test follows existing test
- Parallel: new test is used in combination with existing test(s)

Bossuyt et al. BMJ 2006

Most situations are replacement

Why would you want to replace an existing test with a new test?

- More accurate
- More rapid
- Less invasive
- Technically easier to do
- Easier to interpret
- Other: storage, type of specimen, environment, infection control

PICO for systematic review of diagnostic test accuracy

- Participants
- Index test(s)
- Purpose of test
- Role of testing
- Target condition
- Reference standard
- Comparator (may or may not be included)
- Outcomes (usually sensitivity and specificity)

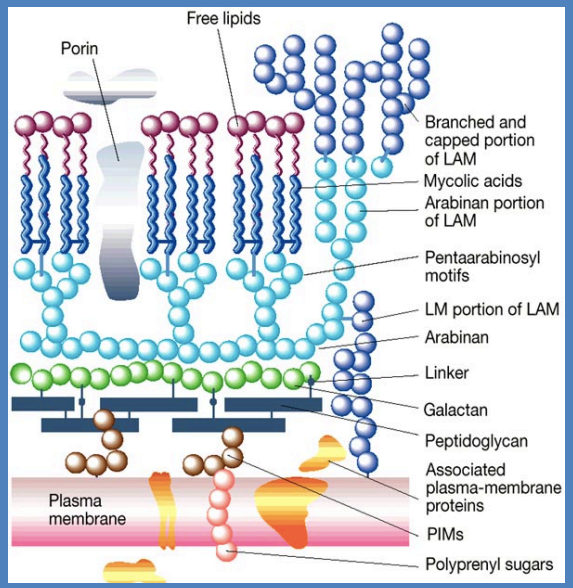
Lateral flow urine lipoarabinomannan assay for detecting active tuberculosis in HIV-positive adults

Maunank Shah et al
The Cochrane Library
2016
Open access

Professor Stephen Lawn

Urine lateral-flow lipoarabinomannan (LF-LAM) Background - 1

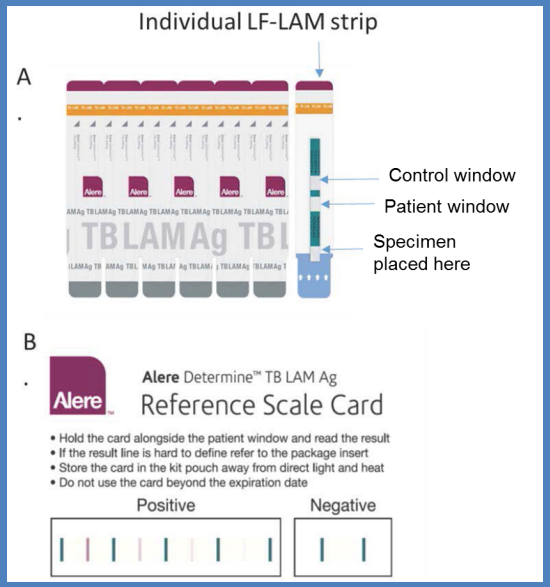
- LAM: 17.5 kilodalton structural component of mycobacterial cell walls
- Detectable in urine of patients with TB
- Meta-analysis of ELISA-LAM*
    - Increased sensitivity in HIV-positive compared with HIV-negative patients
    - Pooled sensitivity 56% (40-71%)
    - Pooled specificity 95% (77-99%)
    - Higher sensitivity with more severe immunosuppression



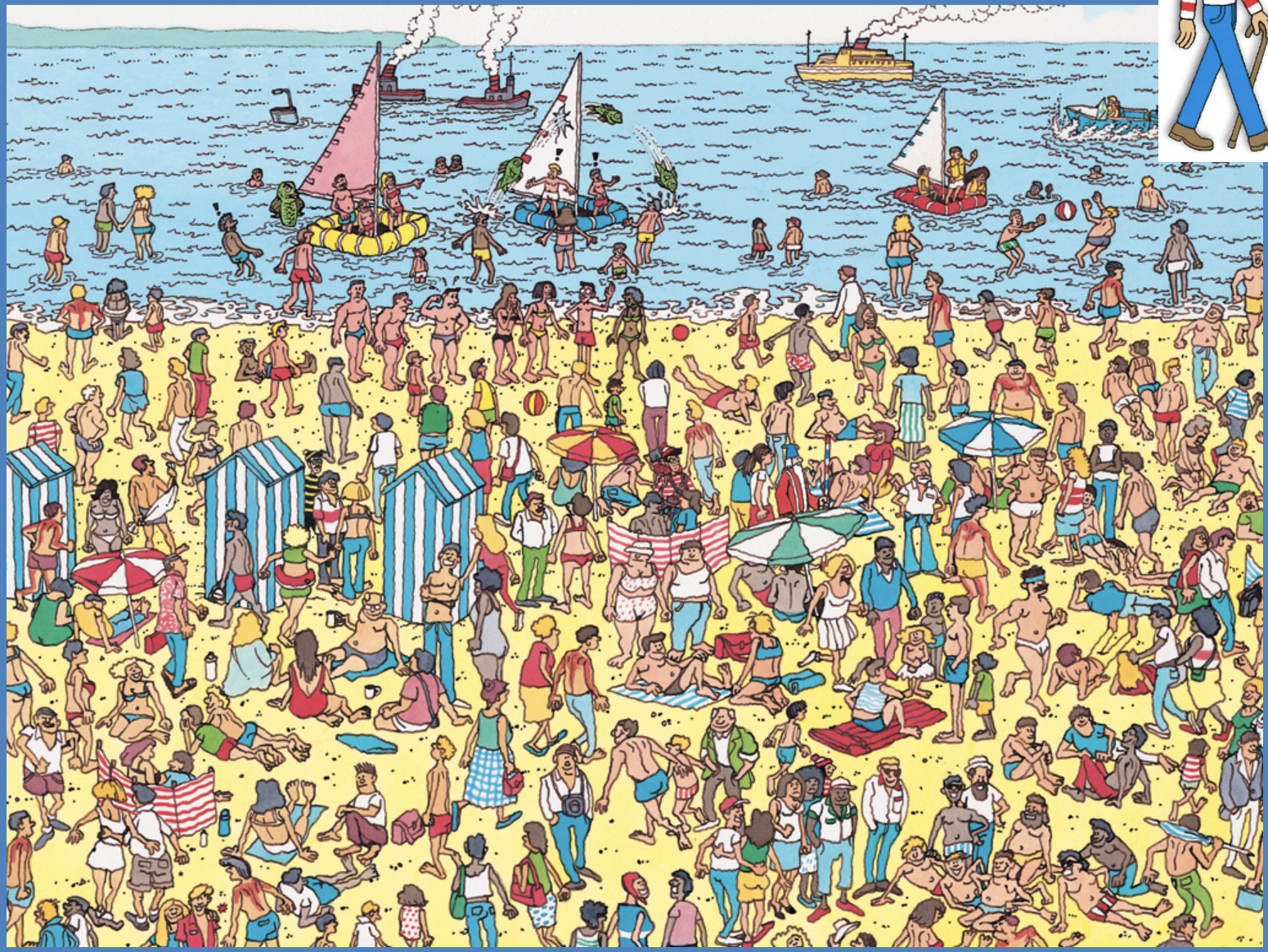
*Minion et al. ERJ 2011; Image: Se-Ho Park and Albert Bendelac. Nature 406, 788-792

LF-LAM Background – 2

- Urine LF-LAM is a new diagnostic test that may overcome limitations of other approaches
  - Point of care, lateral flow format
  - Fast (< 20 minutes)
  - No equipment needed
  - Low cost (~$3.50)
  - Accuracy has varied across published studies



Where's PICO?



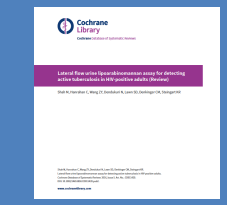
Objectives

- To assess the accuracy of LF-LAM for the diagnosis of active TB disease in HIV-positive adults who have signs and symptoms suggestive of TB (TB diagnosis)

- To assess the accuracy of LF-LAM as a screening test for active TB disease in HIV-positive adults irrespective of signs and symptoms suggestive of TB (TB screening)

Secondary objectives

- To compare diagnostic accuracy of LF-LAM and existing tests, sputum smear microscopy or sputum Xpert® MTB/RIF, as well as determine the diagnostic accuracy of LF-LAM when added to existing tests
- To investigate heterogeneity of test accuracy in the included studies (CD4 count and clinical setting)



Selection criteria - 1

- We included randomized controlled trials, cross-sectional studies, and cohort studies that determined LF-LAM accuracy for TB against a microbiological reference standard (culture or nucleic acid amplification test from any body site)
- We included abstracts with sufficient data
- We excluded case-control studies

Selection criteria - 2

- Participants were HIV positive adults
- We required studies to diagnose TB using at least one of the following two reference standards
  - Microbiological reference standard
  - Composite reference standard with (1) a positive culture, or (2) a positive NAAT, or (3) a positive smear, or (4) a clinical decision to start TB treatment

PICO for LF-LAM ?

- Participants =
- Presentation =
- Index test(s) =
- Purpose of test =
- Role of testing =
- Target condition =
- Reference standard =
- Comparator =
- Outcomes =

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

Assess methodological quality (bias and applicability)

NATURE | COLUMN: WORLD VIEW

Beware the creeping cracks of bias

Evidence is mounting that research is riddled with systematic errors. Left unchecked, this could erode public trust, warns Daniel Sarewitz.

09 May 2012

Alarming cracks are starting to penetrate deep into the scientific edifice. They threaten the status of science and its value to society. And they cannot be blamed on the usual suspects — inadequate funding, misconduct, political interference, an illiterate public. Their cause is bias, and the threat they pose goes to the heart of research.

print
email
download pdf
rights and permissions

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

E-alert RSS Facebook Twitter

Enjoy an exclusive 40% discount! nature

Recent Read Commented Emailed

Nothing will corrode public trust more than a creeping awareness that scientists are unable to live up to the standards that they have set for themselves.

What is bias?

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

- This is the tendency of some (poor) study designs **systematically** to produce results that are **better** (rarely if ever worse) than those with a robust design. Bias for diagnostic tests works in different ways to bias in trials of treatment. *Bandolier*

Three key sources of bias in diagnostic studies

1. Inclusion of right spectrum of patients
2. Verification of patients
    - choice of reference standard
    - complete verification
3. Independent interpretation of index test and reference standard results (blinding)

Quality Assessment of Diagnostic Accuracy Studies, QUADAS-2

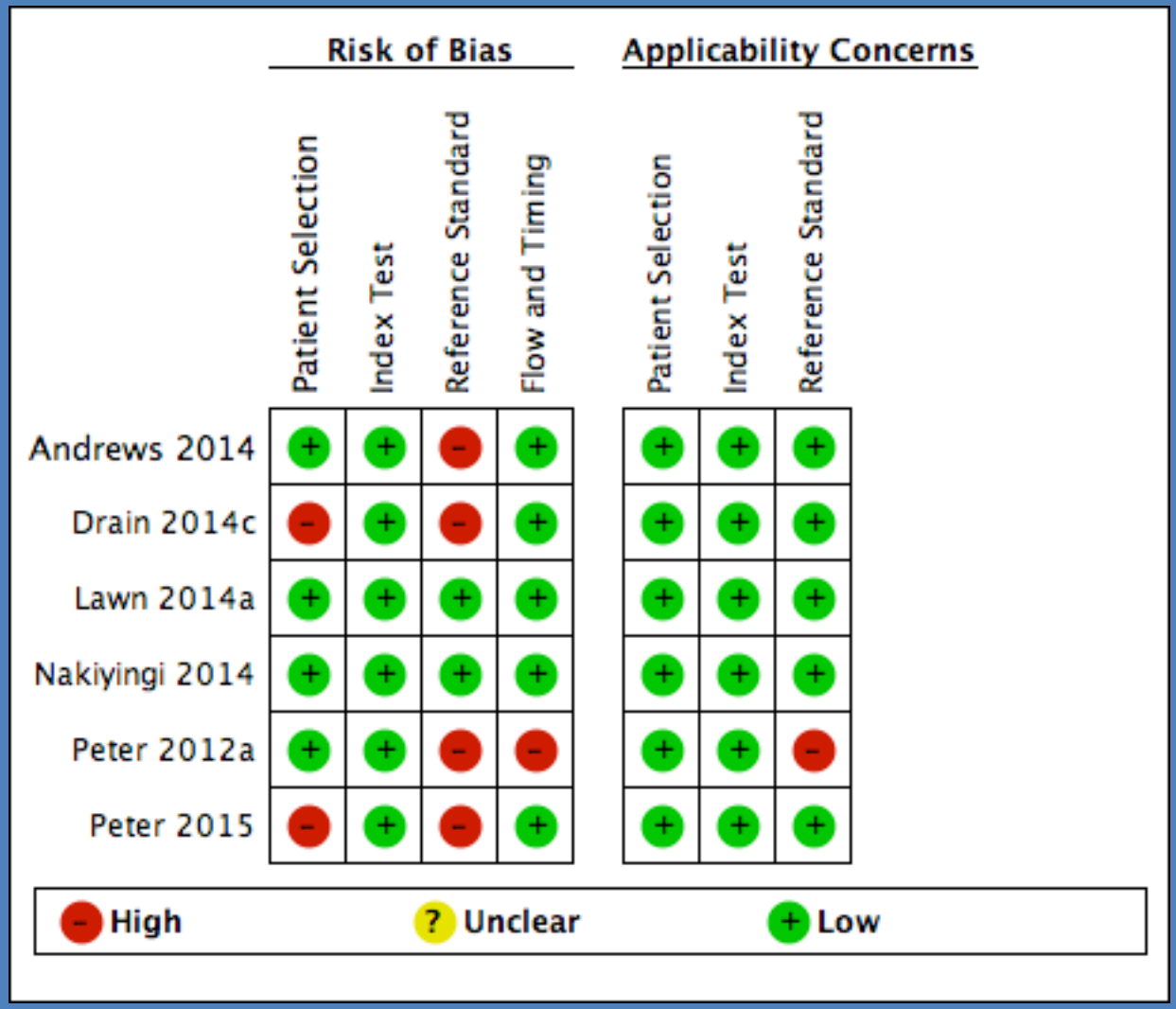
- Domain list
  - patient selection
  - index test
  - reference standard
  - flow and timing
- Signalling questions are used for judgments of risk of bias
- First three domains are also assessed for applicability

QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies

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

Whiting et al. Annals Internal Medicine. 2011

Are you concerned about risk of bias?



Shah Cochrane 2016

Domain: reference standard

- **Could the reference standard, its conduct, or its interpretation have introduced bias?**
- **Signalling question 1: is the reference standard likely to correctly classify the target condition?**

- Due to the difficulties in diagnosing HIV-associated TB, it is recommended that multiple cultures from sputum and other specimens be evaluated.

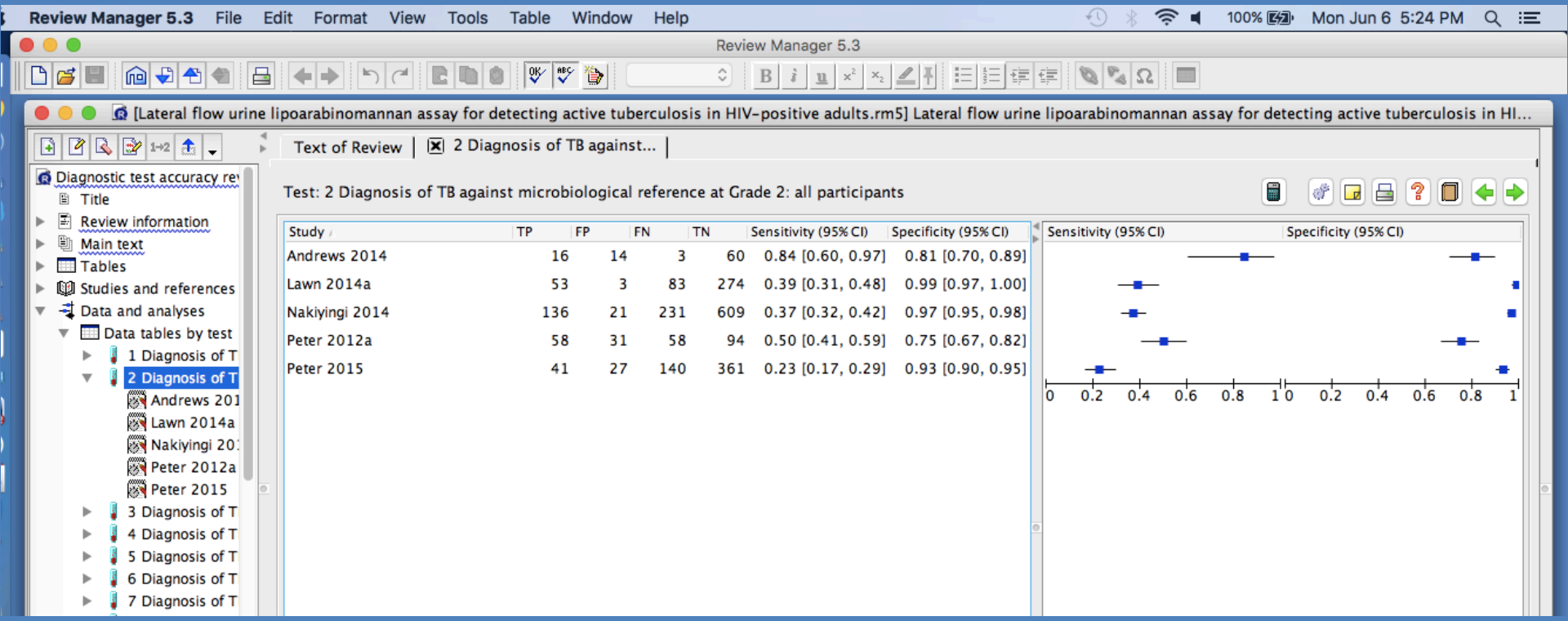
STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies

http://www.equator-network.org/reporting-guidelines/stard/

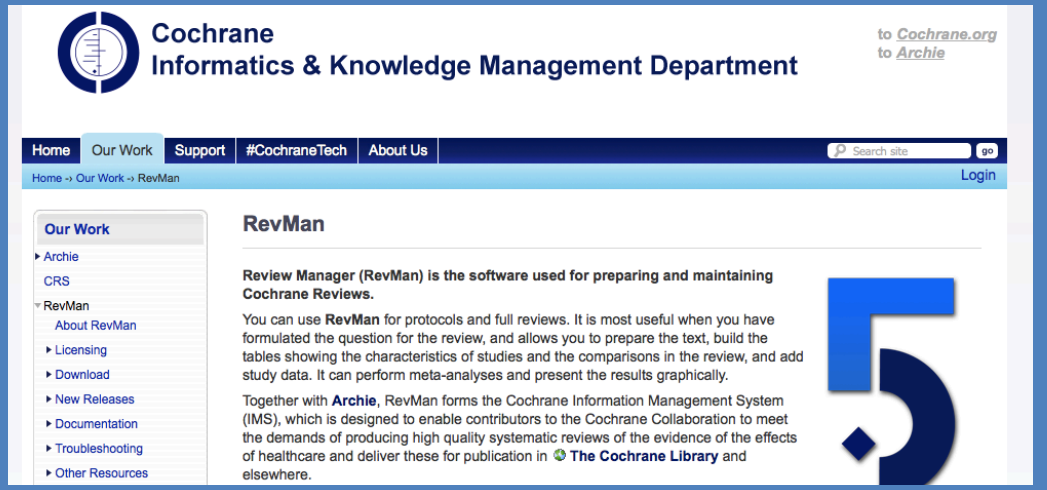
Analyze the data

Statistical analysis and data synthesis

- Calculate estimates of sensitivity and specificity and 95% confidence intervals for individual studies
- Visually examine results of individual studies
- Calculate summary (pooled) accuracy estimates using recommended methods for meta-analysis
- Investigate possible reasons for heterogeneity

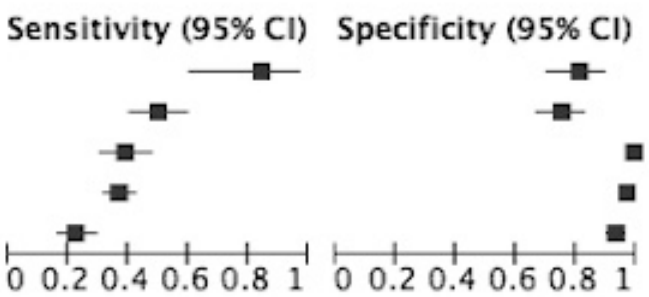


http://tech.cochrane.org/revman



Forest plots of urine LAM sensitivity and specificity for TB diagnosis measured against a microbiological reference standard. The studies are ordered by decreasing sensitivity

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Andrews 2014	16	14	3	60	0.84 [0.60, 0.97]	0.81 [0.70, 0.89]
Peter 2012a	58	31	58	94	0.50 [0.41, 0.59]	0.75 [0.67, 0.82]
Lawn 2014a	53	3	83	274	0.39 [0.31, 0.48]	0.99 [0.97, 1.00]
Nakiyingi 2014	136	21	231	609	0.37 [0.32, 0.42]	0.97 [0.95, 0.98]
Peter 2015	41	27	140	361	0.23 [0.17, 0.29]	0.93 [0.90, 0.95]



TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative

Meta-analysis of diagnostic test accuracy

- Calculate the diagnostic accuracy of a test
- Compare the diagnostic accuracy of two or more tests
- Investigate the variability of results between studies (heterogeneity is to be expected)

How is a meta-analysis performed?

- The bivariate model
  - gives average sensitivity and specificity
  - use when studies report a *common threshold* for a positive result
- Hierarchical summary ROC curve
  - gives a summary ROC curve
  - use when studies report *several different thresholds*
- Both models use random effects

Macaskill P et al. Chapter 10. Cochrane Handbook, Diagnostic Test Accuracy, Version 1.0.2010.

Meta-analysis with STATA

Data Editor (Edit)

var9[10]

	studyid	tp	fp	fn	tn
1	Andrews 2014	16	14	3	60
2	Lawn 2014a	53	3	83	274
3	Nakiyingi 2014	136	21	231	609
4	Peter 2012a	58	31	58	94
5	Peter 2015	41	27	140	361

Stata 12.1

Review

	Command	_rc
1	metandi tp fp fn tn	

Results

	Lambda	1.822691	.7692647			.3149664	3.330422
	Theta	-1.183008	.4236904			-2.013426	-.3525899
	beta	.4108552	.4851453	0.85	0.397	-.540012	1.361722
	s2alpha	.8228375	.6413983			.178569	3.791596
	s2theta	.6978667	.4917079			.1753989	2.776631
Summary pt.							
	Se	.4449208	.0913434			.2796638	.6233288
	Sp	.9290358	.035616			.8195279	.9741886
	DOR	10.49351	5.051922			4.084361	26.95984
	LR+	6.269648	2.750944			2.653139	14.81584
	LR-	.5974788	.0892902			.4457739	.8008116
	1/LR-	1.6737	.2501261			1.248733	2.24329

Covariance between estimates of E(logitSe) & E(logitSp) -.0984308

Stata command: metandi

Practical Stata tutorial:

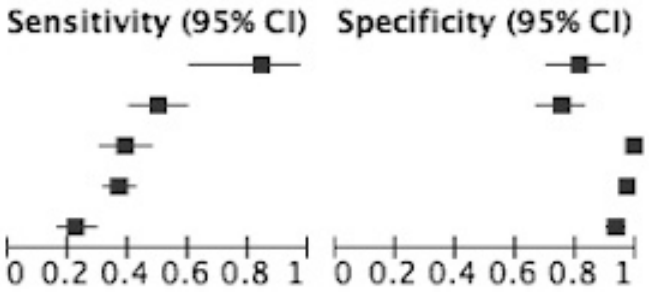
http://methods.cochrane.org/sdt/software-meta-analysis-dta-studies

Heterogeneity

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

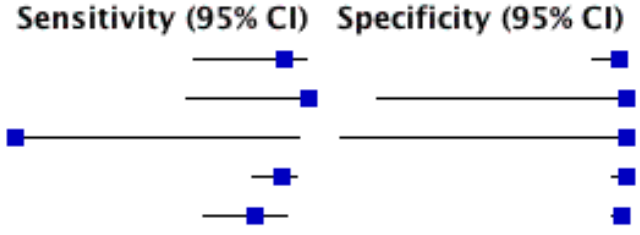
Visual inspection, do the confidence intervals overlap?

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Andrews 2014	16	14	3	60	0.84 [0.60, 0.97]	0.81 [0.70, 0.89]
Peter 2012a	58	31	58	94	0.50 [0.41, 0.59]	0.75 [0.67, 0.82]
Lawn 2014a	53	3	83	274	0.39 [0.31, 0.48]	0.99 [0.97, 1.00]
Nakiyingi 2014	136	21	231	609	0.37 [0.32, 0.42]	0.97 [0.95, 0.98]
Peter 2015	41	27	140	361	0.23 [0.17, 0.29]	0.93 [0.90, 0.95]

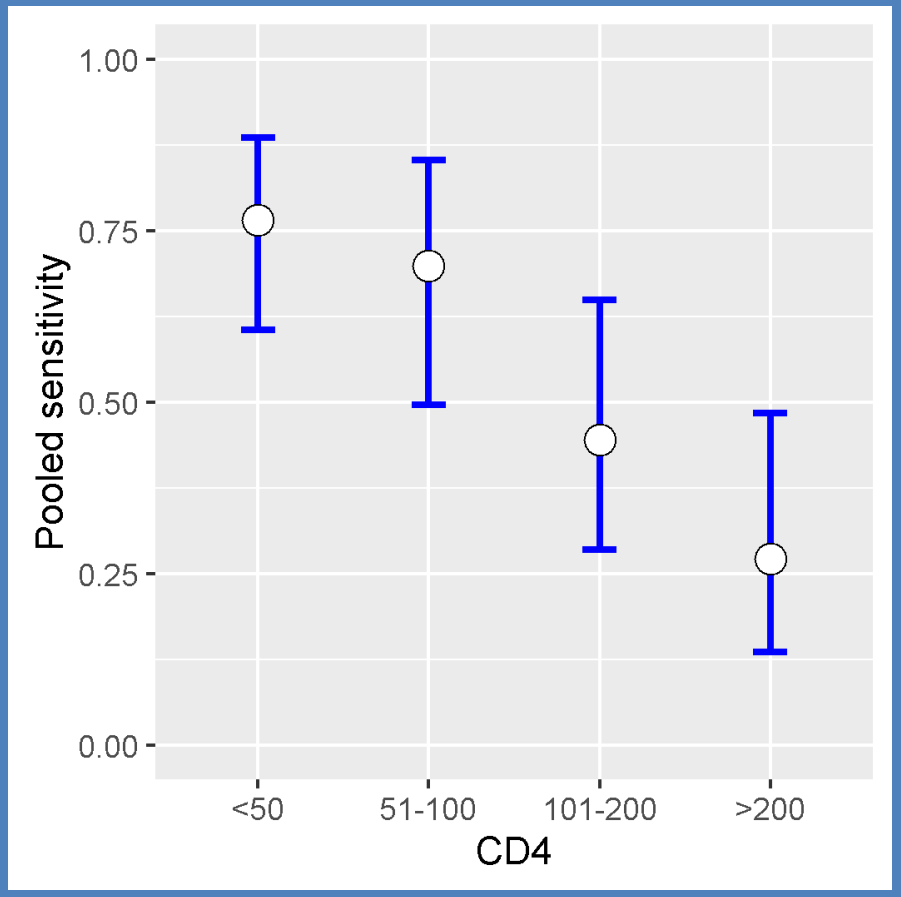


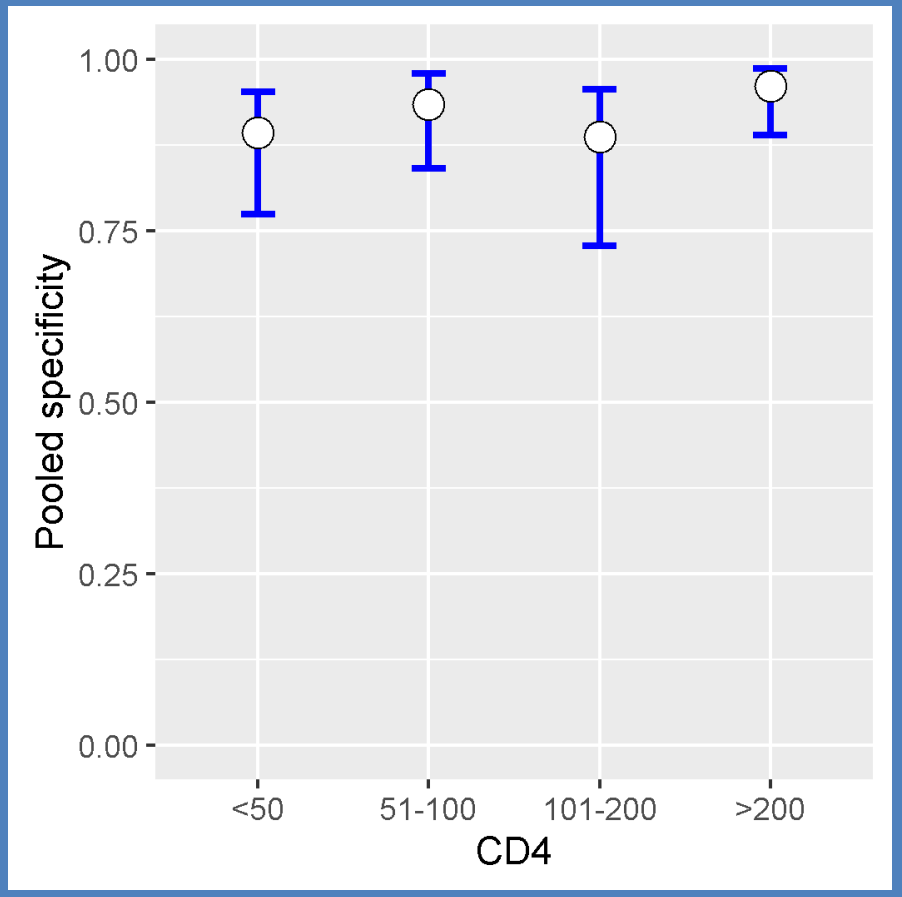
HIV positive

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Balcells 2012	11	1	1	47	0.92 [0.62, 1.00]	0.98 [0.89, 1.00]
Boehme 2010a	7	0	0	2	1.00 [0.59, 1.00]	1.00 [0.16, 1.00]
Boehme 2010b	0	0	1	1	0.00 [0.00, 0.97]	1.00 [0.03, 1.00]
Boehme 2010c	60	0	6	81	0.91 [0.81, 0.97]	1.00 [0.96, 1.00]
Boehme 2010d	27	2	6	141	0.82 [0.65, 0.93]	0.99 [0.95, 1.00]



Heterogeneity, sensitivity and specificity of urine LAM for TB diagnosis stratified by CD4 count

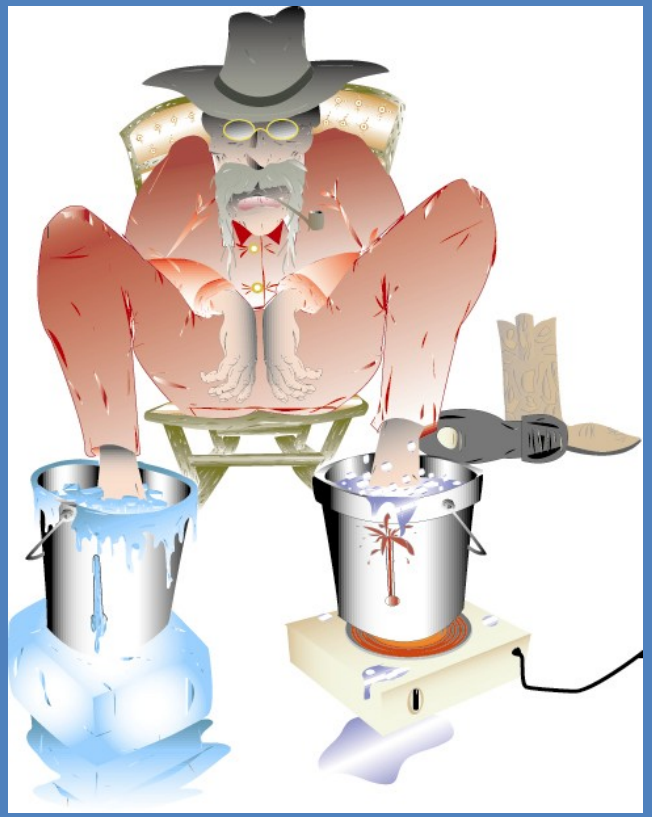




When would you not do a meta-analysis?

It has been said that a fellow with one leg frozen in ice and the other leg in boiling water is comfortable - on average.

JM Yancey. Am J Surg. 1990;159:553-559.



Interpret results and draw conclusions

Summary of Findings Table - 1

- **Question:** what is the diagnostic accuracy of LF-LAM for diagnosing TB in adults living with HIV?
- **Participants:** HIV-positive adults with symptoms of TB
- **Index test:** LF-LAM
- **Role:** a replacement test or test in combination with sputum smear microscopy or sputum Xpert® MTB/RIF
- **Reference standard:** microbiological (mainly mycobacterial culture)

Summary of Findings Table - 2

- **Studies:** cross-sectional
- **Setting:** inpatient and outpatient
- **Limitations:** the main limitations of the review were the use of a lower quality reference standard in most included studies, and the small number of studies and participants included in the analyses
- **Pooled sensitivity:** 45% (95% CrI: 29 to 63); pooled specificity: 92% (95% CrI: 80 to 97)

Summary of Findings Table - 3

Test result	Number of results per 1000 patients tested (95% CrI)		Number of participants (studies)	Quality of the Evidence (GRADE)
	Prevalence 10%	Prevalence 30%		
True positives (patients with tuberculosis)	45 (29 to 63)	135 (87 to 189)	819 (5)	⊕⊕⊕◯ Moderate[1,2,3]
False negatives (patients incorrectly classified as not having tuberculosis)	55 (37 to 71)	165 (111 to 213)		
True negatives (patients without tuberculosis)	828 (720 to 873)	644 (560 to 679)	1494 (5)	⊕⊕◯◯ Low[2,4,5]
False positives (patients incorrectly classified as having tuberculosis)	72 (27 to 180)	56 (21 to 140)		

Pooled sensitivity: 45% (29, 63)
Pooled specificity: 92% (80, 97)

Do you have confidence in the results, why or why not?

Pooled sensitivity: 45% (29, 63)
Pooled specificity: 92% (80, 97)

- Completeness of evidence?
- Accuracy of the reference standard?
- Quality and quality of reporting of included studies?
- Applicability of studies?

Completeness of evidence?

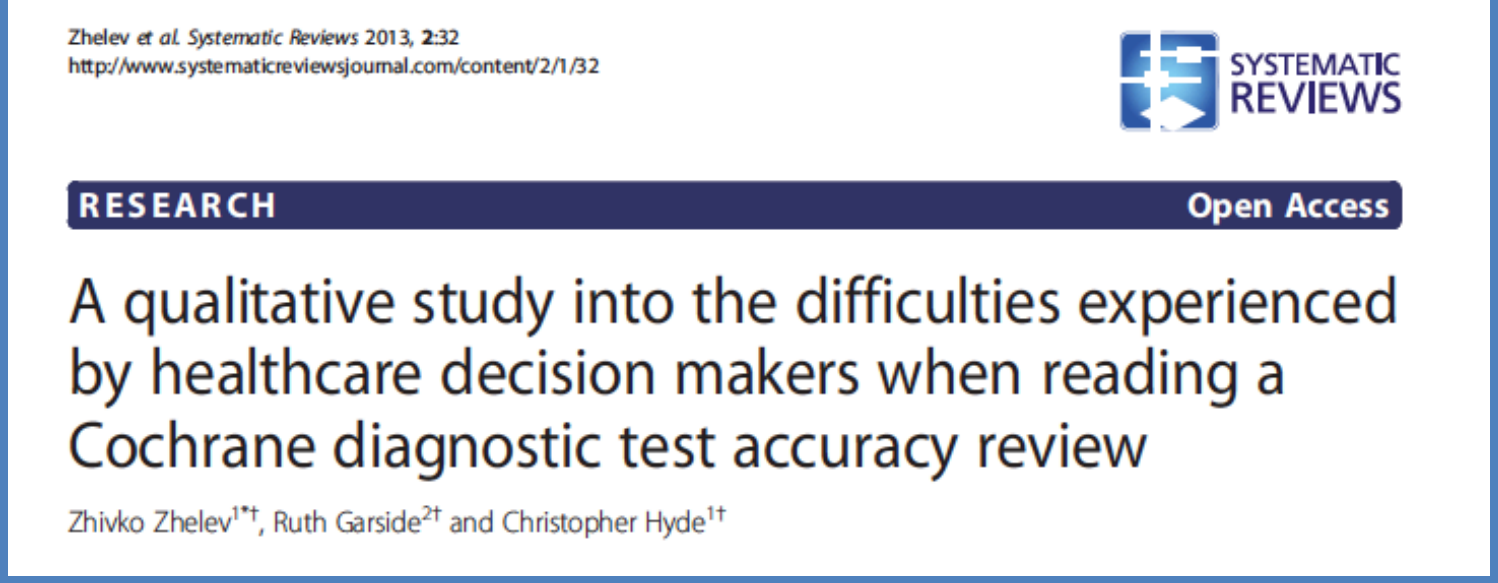
- This data set involved comprehensive searching and correspondence with experts in the field and the test manufacturer to identify additional studies, as well as repeated correspondence with study authors to obtain additional and unpublished data.
- The search strategy included studies published in all languages.
- We acknowledge that we may have missed some studies despite the comprehensive search.

Accuracy of the reference standard?

- HIV-positive TB patients may have pulmonary TB, extrapulmonary TB, or both pulmonary and extrapulmonary TB.
- Due to the difficulties in diagnosing HIV-associated TB, it is recommended that multiple cultures from sputum and other specimen types be evaluated.
- We therefore considered a reference standard using two or more specimen types to be of higher quality than a reference standard using one specimen type.

Quality, reporting, applicability

- Using GRADE approach, we judged the certainty of evidence for accuracy of LF-LAM for TB diagnosis to be low. This means that our confidence in the effect estimate is limited: The true effect may be substantially different rom the estimate of the effect
- Studies were fairly well reported, though we corresponded with almost all study authors for additional data
- We had low concern about the applicability of the included studies to our review question



Zhelev et al. Systematic Reviews 2013, 2:32
http://www.systematicreviewsjournal.com/content/2/1/32

SYSTEMATIC REVIEWS

RESEARCH Open Access

A qualitative study into the difficulties experienced by healthcare decision makers when reading a Cochrane diagnostic test accuracy review

Zhivko Zhelev[1*†], Ruth Garside[2†] and Christopher Hyde[1†]

- The most common error was to misinterpret the role of the reference standard as a comparator
- Participants could not recall definitions for sensitivity and specificity
- The Summary of Findings format was understood

In summary

- Express all components of the review question (PICO and more)
- Heterogeneity is to be expected in meta-analyses of diagnostic test accuracy; investigate the reasons for heterogeneity
- The Summary of Findings Table is useful for interpretation because it brings together the key elements of a review's findings

References

- Cochrane Diagnostic Test Accuracy srdta.cochrane.org/
- Leeflang. Clinical Microbiology and Infection, Volume 20 Number 2, February 2014
- Whiting PF et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011 Oct 18;155(8):529-36.

Initiatives to improve quality and reporting

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

www.equator-network.org/

Acknowledgements

- Clare Davenport
- Mariska Leeflang
- Hans Reitsma
- Maunank Shah
- Penny Whiting

Analysis	Participants (number of studies)	Pooled estimates (95% CrI)	
		Sensitivity	Specificity
LF-LAM, ALL	2313 (5 studies)	45% (29, 63)	92% (80, 97)
LAM alone	1876 (4 studies)	38% (34, 42)	98% (93,100)
Microscopy alone	1876 (4 studies)	40% (27, 54)	95% (94, 97)
LAM and microscopy*	1876 (4 studies)	59% (47, 70)	92% (73, 97)

* Either test positive

Analysis	Number of studies	Pooled estimates (95% CrI)	
		Sensitivity	Specificity
LAM alone	3	36% (31, 42)	96% (94, 98)
Xpert alone	3	61% (39, 77)	97% (94, 99)
LAM and Xpert*	3	75% (61, 87)	93% (81, 97)

* Either test positive