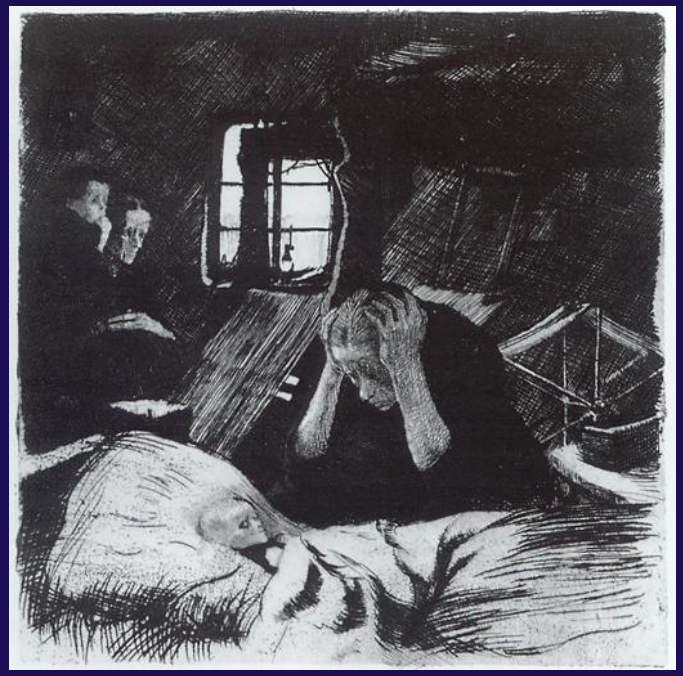
Beth Israel Deaconess Medical Center

A teaching hospital of Harvard Medical School



Käthe Kollwitz, Poverty, 1894

Promise versus reality -

Optimism bias in package inserts of TB diagnostics

Claudia Denkinger, MD PhD
Beth Israel Deaconess Medical Center, Boston
McGill University, Montreal
cdenking@bidmc.harvard.edu

No conflict of interest

JCM
Journals.ASM.org

Promise versus Reality: Optimism Bias in Package Inserts for Tuberculosis Diagnostics

Claudia M. Denkinger,[a] Jasmine Grenier,[b] Jessica Minion,[c] and Madhukar Pai[d,e]

Department of Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA[a]; Faculty of Medicine, McGill University, Montreal, Quebec, Canada[b]; Department of Medical Microbiology & Immunology, University of Alberta, Edmonton, Alberta, Canada[c]; Department of Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada[d]; and Respiratory Epidemiology & Clinical Research Unit, Montreal Chest Institute, Montreal, Quebec, Canada[e]

Laboratorians and clinicians often rely on package inserts of diagnostic tests to assess their accuracy. We compared test accuracy for tuberculosis diagnostics reported in 19 package inserts against estimates in published meta-analyses and found that package inserts generally report overoptimistic accuracy estimates. However, package inserts of most tests approved by the U.S. Food and Drug Administration (FDA) or endorsed by the World Health Organization provide more realistic estimates that agree with meta-analyses.

Acknowledgement

Co-authors:

- Jasmine Grenier, BSc, McGill University
- Jessica Minion, MD, MSc, University of Alberta
- Madhukar Pai, MD, PhD, McGill University

Acknowledgement:

- Daphne Ling, MPH, McGill University

Rose-colored Glasses

- Optimism bias is the unwarranted belief in the efficacy of new tools or interventions
- Optimism bias is everywhere in medicine

JAMA. 2005 Jul 13;294(2):218-28.

Contradicted and initially stronger effects in highly cited clinical research.

Ioannidis JP.

Department of Hygiene and

Epidemiology. 2008 Sep;19(5):640-8.

Why most discovered true associations are inflated.

Ioannidis JP.

N Engl J Med. 2008 Jan 17;358(3):252-60.

Selective publication of antidepressant trials and its influence on apparent efficacy.

Turner EH. Matthews AM. Linardatos E. Tell RA. Rosenthal R.

JAMA. 2011 Jun 1;305(21):2200-10.

Comparison of effect sizes associated with biomarkers reported in highly cited individual articles and in subsequent meta-analyses.

Ioannidis JP, Panagiotou OA.

Prevention Research Center, Department of Medicine, Stanford University School of Medicine, Stanford, California 94305, USA. jioannid@stanford.edu

What about bias in diagnostics research?

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

Source	Bias or Variation	Description
Population		
Demographic features	Variation	Tests may perfo
Disease severity		performance.
Disease prevalence	Variation	
		more frequen performance
Distorted selection of participants	Variatio	
		used in pract
Test protocol: materials and methods		
Test execution		
Test technology	Variation	When the char improvement be affected.
Treatment paradox and disease progression bias	Bias	Disease progres the reference is performed. knowledge of the results of the index test, and the reference standard is applied after treatment has started.
Reference standard and verification procedure		
Inappropriate reference standard	Bias	
Differential verification bias	Bias	Part of the inde
Partial verification bias	Bias	Only a selected standard.
Interpretation (reading process)		
Review bias		Interpretation o the other test.
Clinical review bias	Bias	The avail interpretation
Incorporation bias	Bias	The result of th
Observer variability	Variation	The reproducib test. Because the same res intraobserver variability oc
Analysis		
Handling of indeterminate results	Bias	A diagn 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.

HIV +, children

Advanced disease: more likely +

If interpretation necessary, high

Choosing healthy controls rather than patients suspected of having TB is likely to overestimate accuracy

expertise and skill as test developers which may result in lower accuracy

Imperfect gold standard (low sensitivity and/or specificity) will affect the results of the new test

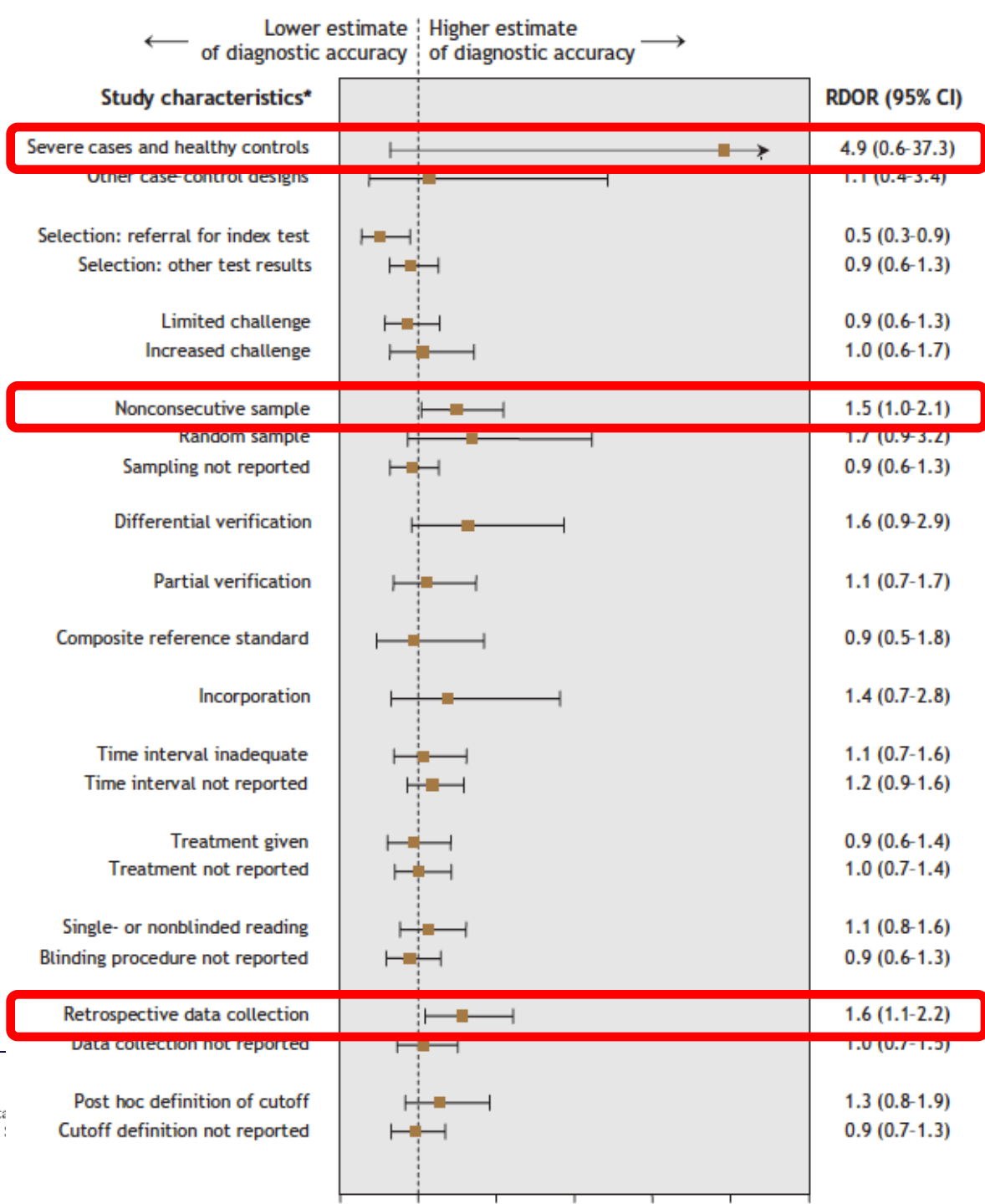
Interpretation of the reference standard needs to occur without knowledge of the index test or the clinical history of the patient

Excluding indeterminate results may lead to inflation of test accuracy

What about bias in diagnostics research?

- Spectrum bias appears to be the most important bias

- Quality assessment tools for diagnostic accuracy studies have been developed - QUADAS



Rutjes CMAJ 2006
Bossuyt JAMA 1999
Fontela PLOSone 2009
Whiting BMC Med Res Meth 2003

Industry Involvement – a problem?

- Well documented in drug trials
- Conclusions of studies favor industry

JAMA. 2003 Jan 22-29;289(4):454-65.

Scope and impact of financial conflicts of interest in biomedical research: a systematic review.

Bekelman JE, Li Y, Gross CP.

Department of Medicine, Yale University School of Medicine, New Haven, Conn 06520, USA.

BMJ. 2002 Aug 3;325(7358):249.

Association between competing interests and authors' conclusions: epidemiological study of randomised clinical trials published in the BMJ.

Kjaergard LL, Als-Nielsen B.

Cochrane Hepatobiliary Group, Copenhagen Trial Unit, Centre for Clinical Intervention Research, Copenhagen University Hospital, Department 7102, H:S Rigshospitalet, DK-2100 Copenhagen, Denmark. Kjaergard@ctu.rh.dk

- Likely a problem in diagnostic research as well

- Publication bias and issues with analysis of data (e.g. discrepant analysis) are also an issue

Many of these biases, variations, interests may contribute to an over optimistic assessment of a novel test

Approval of Diagnostics

- In the USA, diagnostic companies seek approval from the Food and Drug Administration (FDA)
- In Europe, diagnostics need to obtain CE marking
- On a global level, the World Health Organization (WHO)
  - Aims to assure quality for diagnostics in resource limited settings through the pre-qualifications program for HIV, malaria, HBV, HCV
  - Can endorse diagnostic tests in TB (e.g. Xpert MTB/RIF)
- However, aside from these regulatory bodies, the market regulation for diagnostic devices is **limited** and many diagnostic products are brought to the market **without undergoing any independent evaluation**

Is there bias in TB diagnostics?

- Many new diagnostics for TB have been put on the market
- Laboratory professionals and clinicians often rely on package inserts of diagnostic tests to assess their accuracy
- It is often anticipated that test performance when applied to patient care may be less impressive than what is reported in package inserts, however this has not been examined in a systematic way
- How biased are package inserts in TB diagnostics?
  - Investigated by comparing package insert performance data with data reported in meta-analyses
  - What are the sources of bias in the package insert performance data?

Methods

- Systematic search for systematic reviews on the accuracy of diagnostics for TB published through March 2012 in
  - Pubmed
  - Cochrane Library
  - Evidence-based TB Diagnosis website (www.tbevidence.org)
- Excluded meta-analyses that reported performance characteristics other than sensitivity and specificity (e.g. LR), > not comparable to information provided in package inserts
- Searched company websites and contacted test manufacturers for package inserts
- Assessed only commercially available diagnostics tests
- Diagnostic tests for LTBI were not considered because no good reference-standard is available to determine accuracy

Diagnostics assessed

Not FDA/WHO approved

- Interferon-gamma release assays (IGRAs) for active TB
- Lipoarabinomannan urine antigen test
- Serological antibody-detection assays
- Bacteriophage-based tests

FDA approved

- Amplified MTD Gen-Probe assay
- BD Probe Tec ET

WHO endorsed

- Xpert MTB/RIF
- GenoType MTBDRplus
- Inno LIPA RIF/TB
- MODS

IGRAs in active TB

- Not an FDA-approved indication
- Number of samples used to derive PI estimates for active TB:
  - For TB-Spot: 189 for sensitivity, 311 for specificity
  - For QFT: 54 for sensitivity, 581 for specificity
- Type of analysis: case-control, controls were healthy subjects
- Unpublished
- Not stratified by clinically relevant subgroups for active TB (i.e. HIVpositive) or by TB prevalence

Comparison PI vs MA

QuantiFERON®-TB Gold (In-Tube Method)

	Sensitivity	Specificity
PI	89	99
MA	69-81	52-99

T-SPOT.TB

	Sensitivity	Specificity
PI	96	97
MA	83-92	59-88

- Discrepancy
  - In specificity: due to population used: Patients suspected of having TB vs healthy controls > spectrum bias
  - In sensitivity:
    - Due to study set-up: case-control
    - Lack of inclusion of clinically relevant groups (i.e. HIV+)
    - Better performance in skilled hands

Sester ERJ 2011
Metcalfe JID 2011
Chang Thorax 2010
Diel Chest 2010

Serological tests

- Many tests, evaluated in the published studies, are no longer on the market
- Most PIs only state numbers for sensitivity or specificity but do not provide data on how these numbers were obtained
- If further details are provided, PIs almost always report a retro-spective evaluation on in-house samples with healthy controls
- Sample sizes often do not exceed 100 patients and the reference standard is rarely mentioned

IBL INTERNATIONAL *Instructions for Use*

Mycobacterium tuberculosis IgG ELISA

Clinical Specificity	99 %	99 %	100 %
Clinical Sensitivity	100 %	100 %	100 %

PERFORMANCE CHARACTERISTICS

In an in-house evaluation, thirty known positive and one hundred and ten known negative specimens were tested with SEROCHECK-MTB and compared with a licensed commercially available test. The results obtained are as follows:

Specimen Data	Number	Licensed Test	SEROCHECK-MTB
Negative for Ab. to M. tuberculosis	110	110	110
Positive for Ab. to M. tuberculosis	30	30	30

Based on the above study, the specificity and sensitivity of SEROCHECK-MTB is 100%.

SEROCHECK-MTB

Rapid test for detection of antibodies to *Mycobacterium tuberculosis*

Serological Testing

- Attractive option for a point of care test

Grenier ERJ 2011

TABLE 1 Data from 22 high-tuberculosis (TB) burden countries on use of serological tests for the diagnosis of TB

Country	Availability and use of serological tests	Local policy advising use of serological tests	Use in NTP	Use in the public sector#	Use in the private sector	Local, imported tests or both¶	Types of tests used+	Use of serological tests to initiate active TB therapy	Rating of regulatory agency by respondent	Do regulations permit import or use of inaccurate diagnostics?	Do serological tests exceed microbiological methods on the market?	Crude estimate of serological test volume per year
Afghanistan	Yes	No	No	No	Yes	Imported	Rapid	Sometimes	Weak	Yes	Yes, in the private sector	<1000
Bangladesh	Yes	No	No	No	Yes	Imported	Rapid	Sometimes	Weak	Unsure	No	1000–10000
Brazil	Yes	No	No	No	Yes	Both	Both	Unsure	Strong	Yes	No	1000–10000
Cambodia	Yes	No	No	No	Yes	Imported	Both	Rarely	Weak	Yes	No	<1000
China	Yes	No	No	Yes	No	Both	Both	Sometimes	Strong	No	No	>10000–50000
DRC	No											
Ethiopia	No											
India	Yes	No	No	Yes	Yes	Both	Both	Often (in the private sector)	Weak	Yes	Yes, in the private sector	1.5 million
Indonesia	Yes	No	No	Yes	Yes	Imported	Both	Often (in the private sector)	Weak	Yes	Yes, in the private sector	>10000–50000
Kenya	Yes	No	No	Yes	Yes	Imported	Both	Rarely	Weak	Yes	No	1000–10000
Mozambique	No											
Myanmar	Yes	No	No	No	Yes	Both	Both	Rarely	Weak	Yes	No	1000–10000
Nigeria	Yes	No	No	No	Yes	Imported	Both	Sometimes	Weak	Yes	No	1000–10000
Pakistan	Yes	No	No	Yes	Yes	Imported	Both	Sometimes	Weak	Yes	No	>10000–50000
Philippines	Yes	No	No	No	Yes	Imported	Both	Sometimes	Strong	Unsure	No	<1000
Russia	Yes	No	No	Yes	No	Both	Both	Sometimes	Strong	No	No	1000–10000
South Africa	Yes	No	No	No	Yes	Imported	Both	Often (in the private sector)	Weak	Yes	No	1000–10000
Tanzania	No											
Thailand	Yes	No	No	Yes	Yes	Both	Rapid	Unsure	Strong	Yes	Unsure	>10000–50000
Uganda	Yes	No	No	No	Yes	Imported	Rapid	Rarely	Weak	Yes	No	>10000–50000
Vietnam	Yes	No	No	Yes	Yes	Imported	Both	Sometimes	Weak	Yes	No	>50000–100000
Zimbabwe	No											

NTP: National Tuberculosis Control Program; DRC: Democratic Republic of Congo. #: includes public sector establishments (*e.g.* public or government hospitals) other than the NTP. ¶: "local" implies tests that have been manufactured in that country, while "imported" is used to designate the tests that have been manufactured outside the country; +: ELISA, rapid test kits or both

Serological Testing – a Useless Tool

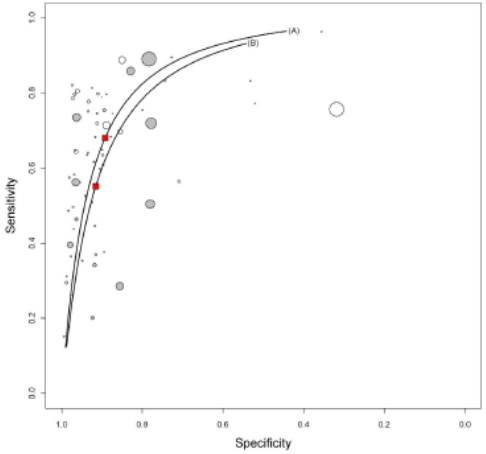
OPEN ACCESS Freely available online PLOS MEDICINE

Commercial Serological Antibody Detection Tests for the Diagnosis of Pulmonary Tuberculosis: A Systematic Review

Karen R. Steingart[1,2], Megan Henry[3], Suman Laal[4,5,6], Philip C. Hopewell[1,2], Andrew Ramsay[7], Dick Menzies[8,9], Jane Cunningham[7], Karin Weldingh[10], Madhukar Pai[8,9*]

PLOS Medicine
2007 and 2011

- For all tests, sensitivity was estimated at 0% to 100%, specificity at 59% to 100%
- Considering added value of serology to smear
  - Additional 57% of the smear-negative cases detected with serology
  - However specificity decreased to 58%



"None of the commercial tests evaluated perform well enough to replace [..] smear microscopy. Thus, these tests have little or no role in the diagnosis of pulmonary tuberculosis."

World Health Organization

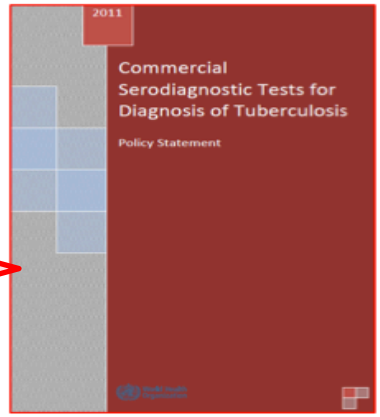
www.who.int/tb

TUBERCULOSIS Serodiagnostic Tests Policy Statement 2011

COMMERCIAL SERODIAGNOSTIC TESTS FOR DIAGNOSIS OF ACTIVE TUBERCULOSIS

CONCLUSION

- It is strongly recommended that **these commercial tests not be used** for the diagnosis of pulmonary and extra-pulmonary TB.
- Currently available commercial serodiagnostic tests (also referred to as serological tests) provide inconsistent and imprecise findings.
- There is no evidence that existing commercial serological assays improve patient outcomes, and high proportions of false-positive and false-negative results may have an adverse impact on the health of patients.



Steingart PLOS one 2011
WHO Statement 2011

Antigen-based tests

Clearview® TB ELISA

- Lipoarabinomannan is a cell wall lipopolysaccharide antigen of MTB that can be detected in urine
- Early evaluation suggested 93% sensitivity and 95% specificity for diagnosis of TB in HIV+
- PI (469 samples in total)
  - 73-81% sensitivity for HIV+ only
  - 93-98% specificity in healthy controls (70-88% in TB suspects)
- Two meta-analyses with data from > 3,000 patients subsequently estimated a sensitivity of about 47-51% in HIV+ and a specificity of 94-96% in healthy controls
- Only in patients with very advanced HIV do sensitivity estimates in meta-analyses approach those of PI > disease severity bias

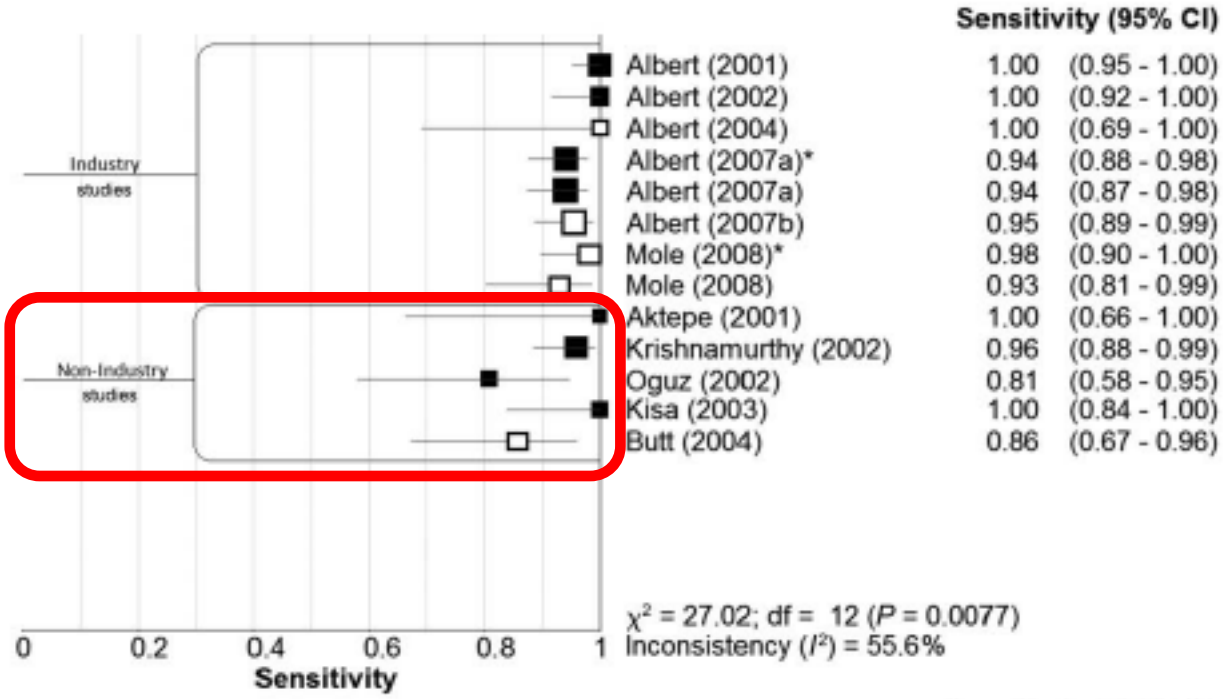
Hamasur J Micro Meth 2001
Minion ERJ 2011
Flores Clin Vacc Immu 2011

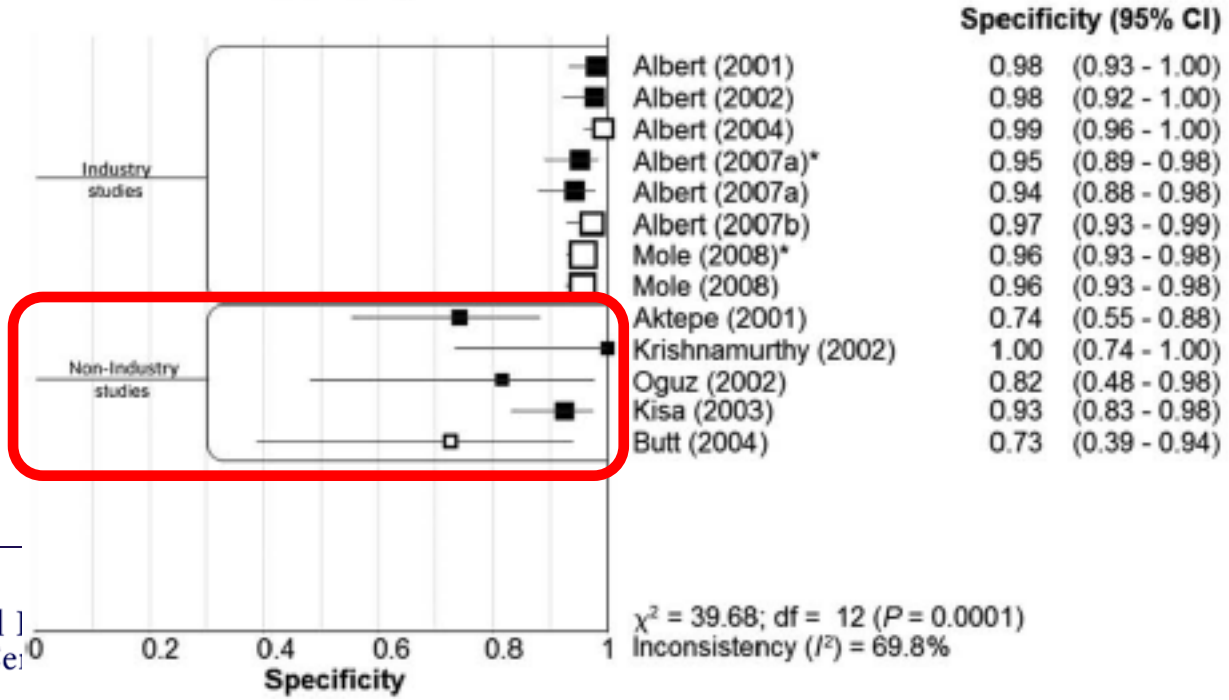
Phage-based assays

- Bacteriophage-based assays were also initially hailed for their high performance
- PIs report sensitivity of 73-82% (S- 49-67, S+ 87) and specificity of 98-99% based on prospective data from over 2000 patients
- Subsequent meta-analysis based on almost 6000 patients suggested sensitivity of 21-94% (S-13-78, S+ 75-87) and specificity of 83-100%
- Data in meta-analysis was not pooled due to heterogeneity
- Up to about 20% of results were uninterpretable > exclusion results in an inflation of test result

>> For both bacteriophage-based assays and LAM, a strong bias due to industry sponsorship for early studies was invoked

Kalantri BMC Infect Dis 2005
Minion Int J Tuberc Lung Dis 2010


Sensitivity (95% CI)
Industry studies
Non-Industry studies
Albert (2001) 1.00 (0.95 - 1.00)
Albert (2002) 1.00 (0.92 - 1.00)
Albert (2004) 1.00 (0.69 - 1.00)
Albert (2007a)* 0.94 (0.88 - 0.98)
Albert (2007a) 0.94 (0.87 - 0.98)
Albert (2007b) 0.95 (0.89 - 0.99)
Mole (2008)* 0.98 (0.90 - 1.00)
Mole (2008) 0.93 (0.81 - 0.99)
Aktepe (2001) 1.00 (0.66 - 1.00)
Krishnamurthy (2002) 0.96 (0.88 - 0.99)
Oguz (2002) 0.81 (0.58 - 0.95)
Kisa (2003) 1.00 (0.84 - 1.00)
Butt (2004) 0.86 (0.67 - 0.96)
0 0.2 0.4 0.6 0.8 1
Sensitivity
χ² = 27.02; df = 12 (P = 0.0077)
Inconsistency (I²) = 55.6%


Specificity (95% CI)
Industry studies
Non-Industry studies
Albert (2001) 0.98 (0.93 - 1.00)
Albert (2002) 0.98 (0.92 - 1.00)
Albert (2004) 0.99 (0.96 - 1.00)
Albert (2007a)* 0.95 (0.89 - 0.98)
Albert (2007a) 0.94 (0.88 - 0.98)
Albert (2007b) 0.97 (0.93 - 0.99)
Mole (2008)* 0.96 (0.93 - 0.98)
Mole (2008) 0.96 (0.93 - 0.98)
Aktepe (2001) 0.74 (0.55 - 0.88)
Krishnamurthy (2002) 1.00 (0.74 - 1.00)
Oguz (2002) 0.82 (0.48 - 0.98)
Kisa (2003) 0.93 (0.83 - 0.98)
Butt (2004) 0.73 (0.39 - 0.94)
0 0.2 0.4 0.6 0.8 1
Specificity
χ² = 39.68; df = 12 (P = 0.0001)
Inconsistency (I²) = 69.8%

Nucleic-acid based tests

- For Gen-Probe amplified MTD, BD ProbeTec ET MTB and Xpert MTB/RIF the discrepancy between the PI-reported data and the estimates in the meta-analysis is small (<5%)
- These tests are either FDA-approved or WHO-endorsed
- For Line-Probe-Assays, a comparison of PI and meta-analysis data is not possible because
  - No information on performance characteristics is reported in PI of GenoType MTBDRplus
  - The data in PI for InnoLIPA reports only indirect testing while the meta-analysis includes studies that tested directly on sputum
- Discrepant analysis is used in some of the analyses for NAAT which might contribute to an overestimate of accuracy

Chang J Infect 2012
Dinnes Health Technol Assess 2007
Greco Thorax 2006
Ling PLoS One 2008

Culture-based tests

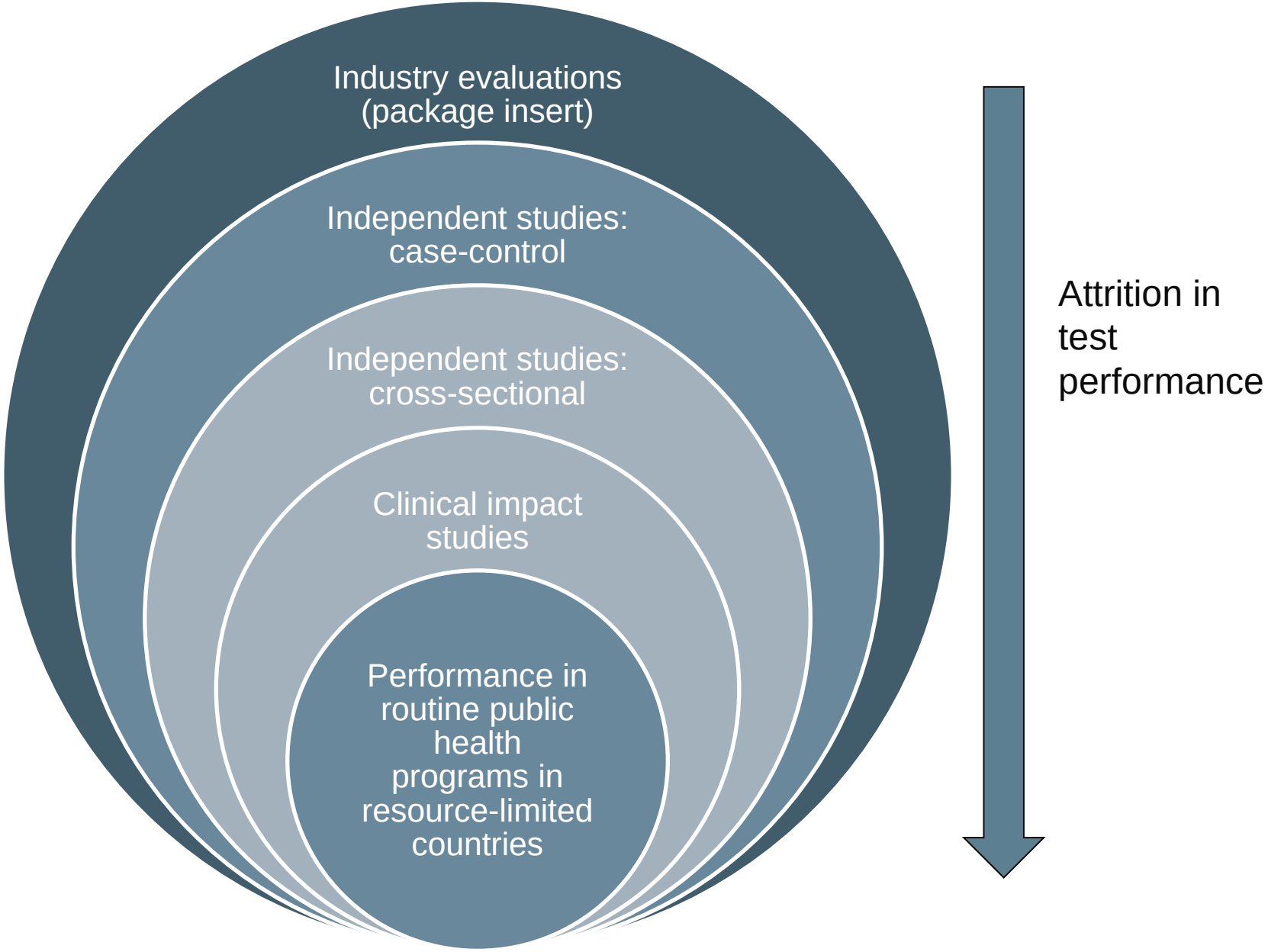
- For the WHO-endorsed MODS (Microscopic-observation drug susceptibility) the PIs claims a sensitivity of 98% and specificity of 100% but only refers to one major published study without further elaborating on the results
- The meta-analysis on over 10,000 patients concludes with a sensitivity of 92% and a specificity of 96%

- Unable to assess liquid culture as PIs do not report sensitivity or specificity but relative false negative rates and contamination rates

Bwanga BMC Infect Dis 2009
Minion Lancet Infect Dis 2010
Leung Int J Tuberc Lung Dis 2011

Limitations of Study

- Unable to compute numeric differences in the estimates of meta-analyses versus PIs because
  - pooling of data were often not possible due to heterogeneity between studies
  - presence of several overlapping meta-analyses
- The performance gap might be even wider, because:
  - A proportion of studies included in the meta-analyses also is industry-supported and might be biased
  - There might be publication bias for studies included in meta-analyses
  - Real-world performance of tests, especially when tests are scaled-up in public health programs, may be worse than those reported in research studies including meta-analyses


Industry evaluations (package insert)
Independent studies: case-control
Independent studies: cross-sectional
Clinical impact studies
Performance in routine public health programs in resource-limited countries
Attrition in test performance

Conclusion

- PIs often present biased data and conclude with over-optimistic estimates of test accuracy
- Particularly pronounced in tests that are not FDA- or WHO-endorsed
- What could be done?
  - Creation of in-country validation of all TB tests, guided by their NTPs
  - Expansion of the WHO prequalification program to TB diagnostics could be considered
- Crucial aspects for evaluation of diagnostics:
  - Evaluation of new tests under clinical and programmatic conditions
  - Independent of industry sponsorship or test developers
- Post-marketing surveillance should be done
- Companies should consider revising the package insert data based on post-marketing data and independent studies

Thank you

Merci

Danke!