Optimism bias, inflated accuracy estimates, and contradicted findings in TB diagnostic research

Madhukar Pai, MD, PhD [madhukar.pai@mcgill.ca]
Jessica Minion, MD

McGill University, Montreal

L'Institut de recherche du Centre universitaire de santé McGill
The Research Institute of the McGill University Health Centre
Les meilleurs soins pour la vie
The Best Care for Life

1

Context

There is some evidence that:

- Initially stronger effects and subsequent contradictions are not infrequent in highly cited research of clinical interventions and their outcomes.
- Claims from highly cited observational studies persist and continue to be supported in the medical literature despite strong contradictory evidence from randomized trials.
- Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes.
- Publication bias is a major concern and may be more widespread than we think; some have also challenged the conventional publishing model
- Even within published studies, selective reporting of positive outcomes in randomized trials as well as observational studies appears to be frequent
- Lack of replication of research findings and over-interpretation of findings are other concerns, especially in some fields

2

Context

- All of these likely result in "optimism bias" —unwarranted belief in the efficacy of new therapies, and overinterpretation of the applicability of findings
- Optimism bias is more likely in industry-supported research
- Optimism bias and conflicting study findings appear to be eroding the public's faith in research
- Even among some researchers, there is concern that most published research findings may be false!

3

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

CLINICAL RESEARCH ON IMPORTANT questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to address whether specific types of studies are more likely to be contradicted and to explain observed controversies. For example, evidence exists that small studies may sometimes be refuted by larger ones.[1,2]

Similarly, there is some evidence on disagreements between epidemiological studies and randomized trials.[3,4] Prior investigations have focused on a variety of studies without any particular attention to their relative importance and scientific impact. Yet, most research publications have little impact while a small minority receives

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discern whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.

Results Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials ($P = .008$). Among randomized trials, studies with contradicted or stronger effects were smaller ($P = .009$) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with "negative" results.

Conclusions Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. The extent to which high citations may provoke contradictions and vice versa needs more study. Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

JAMA. 2005;294:218-228 www.jama.com

Persistence of Contradicted Claims in the Literature

Athina Tatsioni, MD
Nikolaos G. Bonitsis, MD
John P. A. Ioannidis, MD

SOME RESEARCH FINDINGS THAT have received wide attention in the scientific community, as proven by the high citation counts of the respective articles, are eventually contradicted by subsequent evidence.[1] A number of such high-profile contradictions pertain to differences between nonrandomized and randomized studies. For example, the effect of vitamin E on cardiovascular disease prevention has been in the center of a major debate in clinical research over the last 2 decades. Vitamin E is known to have antioxidant activity, and a long list of citations in the preclinical literature on antioxidants[2-4] suggested that these agents may be beneficial for cancer and cardiovascular disease. Two highly cited publications suggested in the 1990s that vitamin E could decrease cardiovascular disease risk by almost half in men and in women.[5,6] However, subsequent randomized trials showed no benefit or even suggested increased harm.[7-9] Several other highly prominent contradictions have also been recorded pertaining to the effects of other dietary components and hormones.[9-12] The prominent refutation of the epidemiological studies has spurred considerable controversy for observational epidemiology in general.[9-11]

Such debate offers opportunities to

Context Some research findings based on observational epidemiology are contradicted by randomized trials, but may nevertheless still be supported in some scientific circles.

Objectives To evaluate the change over time in the content of citations for 2 highly cited epidemiological studies that proposed major cardiovascular benefits associated with vitamin E in 1993; and to understand how these benefits continued being defended in the literature, despite strong contradicting evidence from large randomized clinical trials (RCTs). To examine the generalizability of these findings, we also examined the extent of persistence of supporting citations for the highly cited and contradicted protective effects of beta-carotene on cancer and of estrogen on Alzheimer disease.

Data Sources For vitamin E, we sampled articles published in 1997, 2001, and 2005 (before, early, and late after publication of refuting evidence) that referenced the highly cited epidemiological studies and separately sampled articles published in 2005 and referencing the major contradicting RCT (HOPE trial). We also sampled articles published in 2006 that referenced highly cited articles proposing benefits associated with beta-carotene for cancer (published in 1981 and contradicted long ago by RCTs in 1994-1996) and estrogen for Alzheimer disease (published in 1996 and contradicted recently by RCTs in 2004).

Data Extraction The stance of the citing articles was rated as favorable, equivocal, and unfavorable to the intervention. We also recorded the range of counterarguments raised to defend effectiveness against contradicting evidence.

Results For the 2 vitamin E epidemiological studies, even in 2005, 50% of citing articles remained favorable. A favorable stance was independently less likely in more recent articles, specifically in articles that also cited the HOPE trial (odds ratio for 2001, 0.06 [95% confidence interval, 0.01-0.19; $P < .001$] and the odds ratio for 2005, 0.06 [95% confidence interval, 0.02-0.24; $P < .001$], as compared with 1997), and in general/internal medicine vs specialty journals. Among articles citing the HOPE trial in 2005, 41.4% were unfavorable. In 2006, 62.5% of articles referencing the highly cited article that had proposed beta-carotene and 61.7% of those referencing the highly cited article on estrogen effectiveness were still favorable; 100% and 96%, respectively, of the citations appeared in specialty journals; and citations were significantly less favorable ($P = .009$ and $P = .009$, respectively) when the major contradicting trials were also mentioned. Counterarguments defending vitamin E or estrogen included diverse selection and information biases and genuine differences across studies in participants, interventions, contraventions, and outcomes. Favorable citations to beta-carotene, long after evidence contradicted its effectiveness, did not consider the contradicting evidence.

Conclusion Claims from highly cited observational studies persist and continue to be supported in the medical literature despite strong contradictory evidence from randomized trials.

JAMA. 2007;298(21):2517-2526 www.jama.com

Why Most Discovered True Associations Are Inflated

John P. A. Ioannidis

Abstract: Newly discovered true (non-null) associations often have inflated effects compared with the true effect sizes. I discuss here the main reasons for this inflation. First, theoretical considerations prove that when true discovery is claimed based on crossing a threshold of statistical significance and the discovery study is underpowered, the observed effects are expected to be inflated. This has been demonstrated in various fields ranging from early stopped clinical trials to genome-wide associations. Second, flexible analyses coupled with selective reporting may inflate the published discovered effects. The vibration ratio (the ratio of the largest vs. smallest effect on the same association approached with different analytic choices) can be very large. Third, effects may be inflated at the stage of interpretation due to diverse conflicts of interest. Discovered effects are not always inflated, and under some circumstances may be deflated—for example, in the setting of late discovery of associations in sequentially accumulated overpowered evidence, in some types of misclassification from measurement error, and in conflicts causing reverse biases. Finally, I discuss potential approaches to this problem. These include being cautious about newly discovered effect sizes, considering some rational down-adjustment, using analytical methods that correct for the anticipated inflation, ignoring the magnitude of the effect (if not necessary), conducting large studies in the discovery phase, using strict protocols for analyses, pursuing complete and transparent reporting of all results, placing emphasis on replication, and being fair with interpretation of results.

(Epidemiology 2008;19:640–648)

prognostic studies, and so forth. I start here with the assumption that a research finding is indeed true (non-null), ie, it reflects a genuine association that is not entirely due to chance or biases (confounding, misclassification, selection biases, selective reporting, or other). The question is: do the effect sizes for such associations, at the time they are first discovered and published in the scientific literature, accurately reflect the true effect sizes?

The article has the following sections: a brief literature review on inflated early-effect sizes based on theoretical and empirical considerations; a description of the major reasons why early discovered effects are inflated and the major countering forces that may occasionally lead to deflated effects (underestimates); and suggestions on how to deal with these problems.

Evidence About Inflated Early-Effect Sizes

Table 1 cites articles suggesting that early studies give (on average) inflated estimates of effect.[2-34] I list here only selected evaluations that cover either many different articles/effects or a whole research domain or method. This list is nowhere close to exhaustive. For some topics, such as the inflation of regression coefficients for variables selected through stepwise statistical-significance-based processes, the literature is

4

Publication bias and selective publication

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy

Erick H. Turner, M.D., Annette M. Matthews, M.D., Eftihia Linardatos, B.S., Robert A. Tell, L.C.S.W., and Robert Rosenthal, Ph.D.

ABSTRACT

BACKGROUND

Evidence-based medicine is valuable to the extent that the evidence base is complete and unbiased. Selective publication of clinical trials — and the outcomes within those trials — can lead to unrealistic estimates of drug effectiveness and alter the apparent risk–benefit ratio.

METHODS

We obtained reviews from the Food and Drug Administration (FDA) for studies of 12 antidepressant agents involving 12,564 patients. We conducted a systematic literature search to identify matching publications. For trials that were reported in the literature, we compared the published outcomes with the FDA outcomes. We also compared the effect size derived from the published reports with the effect size derived from the entire FDA data set.

RESULTS

Among 74 FDA-registered studies, 31%, accounting for 3449 study participants, were not published. Whether and how the studies were published were associated with the study outcome. A total of 37 studies viewed by the FDA as having positive results were published; 1 study viewed as positive was not published. Studies viewed by the FDA as having negative or questionable results were, with 3 exceptions, either not published (22 studies) or published in a way that, in our opinion, conveyed a positive outcome (11 studies). According to the published literature, it appeared that 94% of the trials conducted were positive. By contrast, the FDA analysis showed that 51% were positive. Separate meta-analyses of the FDA and journal data sets showed that the increase in effect size ranged from 11 to 69% for individual drugs and was 32% overall.

PLoS MEDICINE

Initial Severity and Antidepressant Benefits: A Meta-Analysis of Data Submitted to the Food and Drug Administration

ABSTRACT

Background

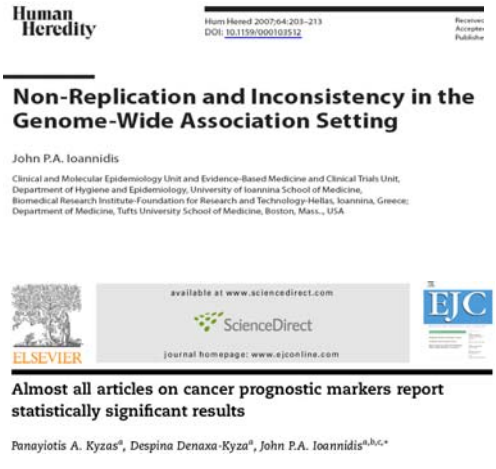
Methods and Findings

Conclusions

While almost all trials with "positive" results on antidepressants had been published, trials with "negative" results submitted to the US Food and Drug Administration, with few exceptions, remained either unpublished or were published with the results presented so that they would appear "positive"

5

Non-replicated studies and publication bias – especially in genetic and biomarker studies



Human Heredity

Hum Hered 2007;64:203–213
DOI: 10.1159/000103512

Non-Replication and Inconsistency in the Genome-Wide Association Setting

John P.A. Ioannidis

Clinical and Molecular Epidemiology Unit and Evidence-Based Medicine and Clinical Trials Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Biomedical Research Institute-Foundation for Research and Technology-Hellas, Ioannina, Greece; Department of Medicine, Tufts University School of Medicine, Boston, Mass., USA

ELSEVIER

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejconline.com

EJC

Almost all articles on cancer prognostic markers report statistically significant results

Panayiotis A. Kyzas[a], Despina Denaxa-Kyza[a], John P.A. Ioannidis[a,b,c,*]

6

Even within published studies, selective reporting of outcomes

Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials

Comparison of Protocols to Published Articles

An-Wen Chan, MD, DPhil
Asbjørn Hróbjartsson, MD, PhD
Mette T. Haahr, BSc
Peter C. Gøtzsche, MD, DrMedSci
Douglas G. Altman, DSc

Context Selective reporting of outcomes within published studies based on the nature or direction of their results has been widely suspected, but direct evidence of such bias is currently limited to case reports.

Objective To study empirically the extent and nature of outcome reporting bias in a cohort of randomized trials.

Design Cohort study using protocols and published reports of randomized trials approved by the Scientific-Ethical Committees for Copenhagen and Frederiksberg, Denmark, in 1994-1995. The number and characteristics of reported and unreported trial outcomes were recorded from protocols, journal articles, and a survey of trialists. An outcome was considered incompletely reported if insufficient data were presented in the published articles for meta-analysis. Odds ratios relating the completeness of outcome reporting to statistical significance were calculated for each trial and then pooled to provide an overall estimate of bias. Protocols and published articles were also compared to identify discrepancies in primary outcomes.

Main Outcome Measures Completeness of reporting of efficacy and harm outcomes and of statistically significant vs nonsignificant outcomes; consistency between primary outcomes defined in the most recent protocols and those defined in published articles.

Results One hundred two trials with 122 published journal articles and 3736 outcomes were identified. Overall, 50% of efficacy and 65% of harm outcomes per trial were incompletely reported. Statistically significant outcomes had a higher odds of being fully reported compared with nonsignificant outcomes for both efficacy (pooled odds ratio, 2.4; 95% confidence interval [CI], 1.4-4.0) and harm (pooled odds ratio, 4.7; 95% CI, 1.8-12.0) data. In comparing published articles with protocols, 62% of trials had at least 1 primary outcome that was changed, introduced, or omitted. Eighty-six percent of survey responders (42/49) denied the existence of unreported outcomes despite clear evidence to the contrary.

Conclusions The reporting of trial outcomes is not only frequently incomplete but also biased and inconsistent with protocols. Published articles, as well as reviews that incorporate them, may therefore be unreliable and overestimate the benefits of an intervention. To ensure transparency, planned trials should be registered and protocols should be made publicly available prior to trial completion.

METHODS

In RCTs

Selection in Reported Epidemiological Risks: An Empirical Assessment

Fotini K. Kavvoura[1], George Liberopoulos[1], John P. A. Ioannidis[1,2*]

ABSTRACT

Background

Epidemiological studies may be subject to selective reporting, but empirical evidence thereof is limited. We empirically evaluated the extent of selection of significant results and large effect sizes in a large sample of recent articles.

Methods and Findings

We evaluated 389 articles of epidemiological studies that reported, in their respective abstracts, at least one relative risk for a continuous risk factor in contrasts based on median, tertile, quartile, or quintile categorizations. We examined the proportion and correlates of reporting statistically significant and nonsignificant results in the abstract and whether the magnitude of the relative risks presented (coined to be consistently ≥1.00) differs depending on the type of contrast used for the risk factor. In 342 articles (87.9%), ≥1 statistically significant relative risk was reported in the abstract, while only 169 articles (43.4%) reported ≥1 statistically nonsignificant relative risk in the abstract. Reporting of statistically significant results was more common with structured abstracts, and was less common in US-based studies and in cancer outcomes. Among 50 randomly selected articles in which the full text was examined, a median of nine (interquartile range 5-16) statistically significant and six (interquartile range 3-16) statistically nonsignificant relative risks were presented ($p = 0.25$). Paradoxically, the smallest presented relative risks were based on the contrasts of extreme quintiles; on average, the relative risk magnitude was 1.41-, 1.42-, and 1.34-fold larger in contrasts of extreme quartiles, extreme tertiles, and above-versus-below median values, respectively ($p < 0.001$).

Conclusions

Published epidemiological investigations almost universally highlight significant associations between risk factors and outcomes. For continuous risk factors, investigators selectively present contrasts between more extreme groups, when relative risks are inherently lower.

In observational studies

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

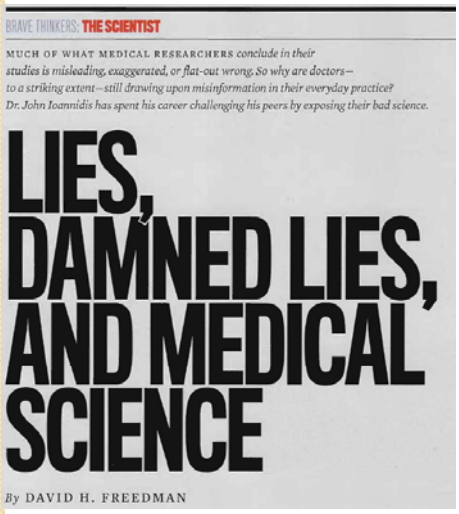
Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on p-values. Research findings are defined here as any relationship reaching formal statistical significance, e.g., effective interventions, informative predictors, risk factors, or associations. "Negative" research is also very useful

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R+1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV. The PPV is also the complementary probability of what Wacholder et al. have called the false positive report probability [10]. According to the 2

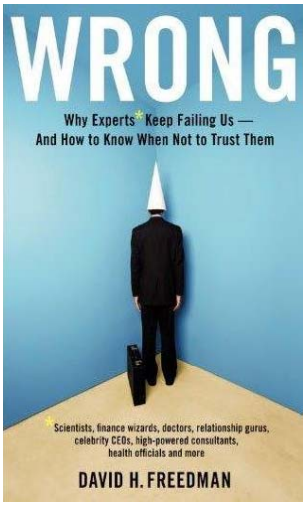
PLoS Med 2005


BRAVE THINKERS: THE SCIENTIST

MUCH OF WHAT MEDICAL RESEARCHERS *conclude in their studies is misleading, exaggerated, or flat-out wrong. So why are doctors— to a striking extent—still drawing upon misinformation in their everyday practice? Dr. John Ioannidis has spent his career challenging his peers by exposing their bad science.*

LIES, DAMNED LIES, AND MEDICAL SCIENCE

By DAVID H. FREEDMAN



As researchers in tuberculosis we asked the question:

"is there evidence for 'optimism bias' in TB diagnostic research?"

We present several case studies to answer this question

Several new diagnostics are in the pipeline
But do they work? Will optimism bias prove to be a big issue?

10

Case study 1:

How much evidence is sufficient for commercialization?

11

Promising new Point of Care test: LAM antigen detection

ELSEVIER

Journal of Microbiological Methods 45 (2001) 41–52

Journal of Microbiological Methods

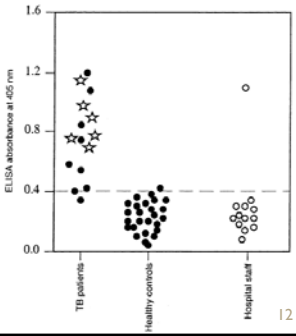
www.elsevier.com/locate/jmicmeth

Rapid diagnosis of tuberculosis by detection of mycobacterial lipoarabinomannan in urine

Beston Hamasur [a], Judith Bruchfeld [b], Melles Haile [a], Andrzej Pawlowski [a], Bjarne Bjorvatn [c], Gunilla Källenius [a,d], Stefan B. Svenson [a,e,*]

Sensitivity 93%

Specificity 95%



12

Early data lead to rapid commercialization and marketing of a urine LAM assay

Transactions of the Royal Society of Tropical Medicine and Hygiene (2005) 99, 893–900

ELSEVIER

www.elsevierhealth.com/journals/trst

Detection of mycobacterial lipoarabinomannan with an antigen-capture ELISA in unprocessed urine of Tanzanian patients with suspected tuberculosis

C. Boehme[a,*], E. Molokova[b], F. Minja[c], S. Geis[a], T. Loscher[a], L. Maboko[d], V. Koulchin[b], M. Hoelscher[a]

231 patients with suspected pulmonary TB and 103 healthy volunteers were screened with standard TB tests and with the new LAM-ELISA.

Of 132 patients with positive sputum culture, 106 were positive using the LAM-ELISA (sensitivity 80.3%)

To define the specificity of the assay, urine samples from 103 healthy volunteers were also screened using LAM-ELISA. All but one had an optical density below the cut-off (specificity 99%)

13

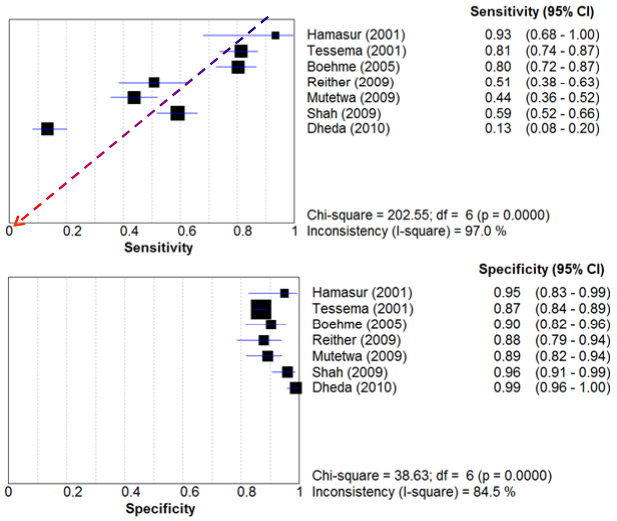
Marketed in 2007/08 by Inverness Medical Innovations

Clearview® TB ELISA

LAM Specific Direct Urinary Antigen Test

14

Subsequent evidence from field studies in India, S Africa, Zimbabwe, Tanzania



Minion et al. Under review

15

Lessons

- Rapid commercialization on the basis of early data may be problematic (especially case-control studies that can exaggerate accuracy estimates)
- Thorough field evaluation in diverse settings (e.g. varying HIV prevalence) should have been done
- This case study raises an interesting question: at what point in time after a test is introduced should meta-analyses be done?

16

Case study 2:

How should we design and analyze diagnostic studies?

17

Serologic (antibody) tests for TB

- Attractive … especially if point of care (POC) option
- >80 antigenic targets evaluated and several commercial assays developed
- All existing serologic tests have failed to demonstrate adequate accuracy

A systematic review of commercial serological antibody detection tests for the diagnosis of extrapulmonary tuberculosis

Karen R Steingart, Megan Henry, Suman Laal, Philip C Hopewell, Andrew Ramsay, Dick Menzies, Jane Cunningham, Karin Weldingh, Madhukar Pai

Thorax 2007

PLoS Medicine 2007

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[3], Karin Weldingh[10], Madhukar Pai[8,9]*

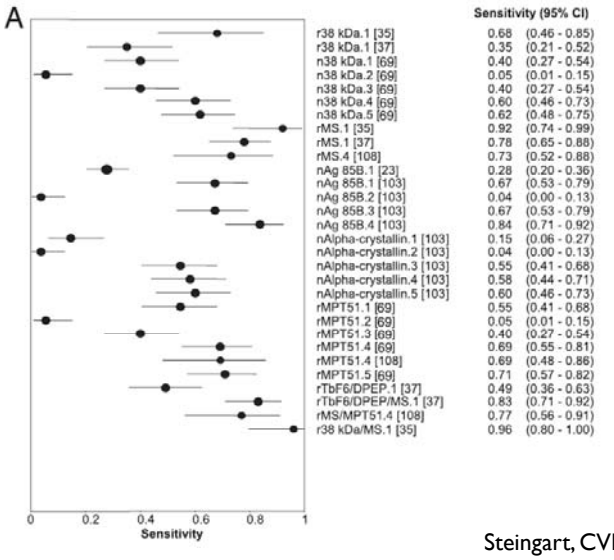
Performance of Purified Antigens for Serodiagnosis of Pulmonary Tuberculosis: a Meta-Analysis[▽†]

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

Clin Vaccine Immunol 2009

18

Sensitivity varies from 0 – 100%

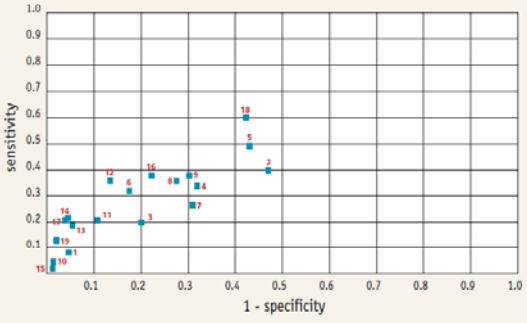


Steingart, CVI 2009

19

WHO evaluation of 19 commercial serological tests for TB

Figure 4. ROC curve of commercial rapid tests for the diagnosis of pulmonary tuberculosis (all patients, n=355)



1. ABP Diagnostics 2. Advanced Diagnostics 3. Products 6. Chembio Diagnostic Systems 7. CTK Biotech American Bionostica 4. Ameritek USA 5. Bio-Medical 8. Hema Diagnostic Systems 9. Laboratorios Silanes

TDR/WHO Report 2008

20

Why do these tests fail in field studies?

TABLE 3. Characteristics of study quality

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

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

A large % were case-control studies

Confirmed TB cases
Vs.
Healthy controls (often from low-incidence countries)

21

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

Clinical Chemistry 51:8
1335–1341 (2005)

Minireview

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

Anne W.S. Rutjes,[1*] Johannes B. Reitsma,[1] Jan P. Vandenbroucke,[2] Afina S. Glas,[3] and Patrick M.M. Bossuyt[1]

22

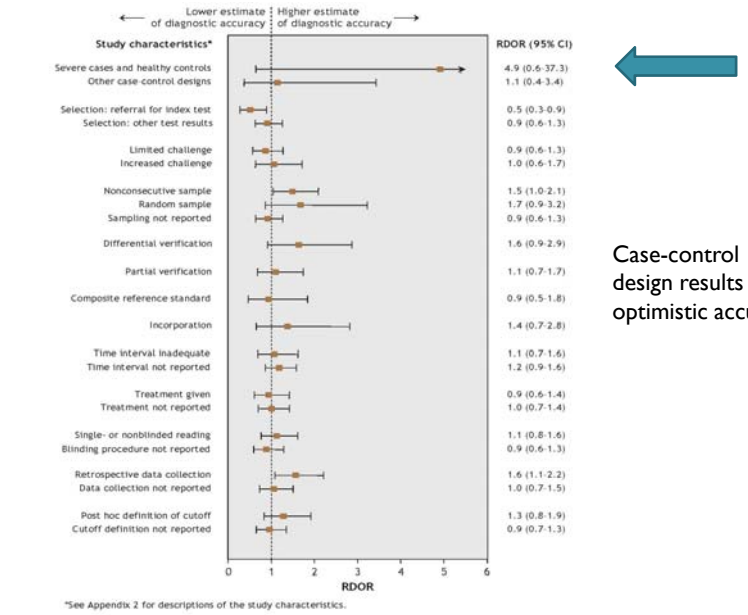


Fig. 2: Effects of study design characteristics on estimates of diagnostic accuracy. RDOR = relative diagnostic odds ratio (adjusted RDORs were estimated in a multivariable random-effects meta-epidemiologic regression model).

Case-control design results in optimistic accuracy

Rutjes et al. CMAJ 2006

23

We find this in TB as well: Example: PCR tests for TB meningitis

Diagnostic accuracy of nucleic acid amplification tests for tuberculous meningitis: a systematic review and meta-analysis

Madhukar Pai, Laura L Flores, Nitika Pai, Alan Hubbard, Lee W Riley, and John M Colford Jr

Case-control studies had a two-fold higher diagnostic odds ratios 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
Study design			
Case-control	19	86·5 (39·3, 190·2)	0·03
Cross-sectional	16	43·3 (22·5, 83·3)	0·94
Blinded interpretation of test and/or reference standard results			
Yes	21	46·9 (24·9, 88·6)	0·16
No	14	82·3 (39·8, 170·2)	0·70
Consecutive or random sampling of participants			
Yes	18	63·3 (32·8, 122·4)	0·20
No	17	46·8 (23·6, 92·8)	0·42
Prospective data collection			
Yes	18	59·9 (28·1, 127·6)	0·12
No	17	55·2 (29·9, 101·6)	0·59

*Random effects model. †χ^2 test for heterogeneity. CI=confidence interval.

Pai et al. Lancet Infect Dis 2003

24

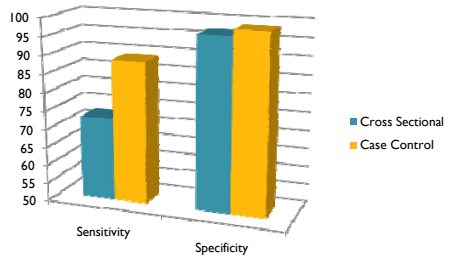
LED microscopy for sputum examination

- Cross Sectional Studies

Sensitivity 72.6%
(69.2, 75.8)
Specificity 96.9
(92.1, 98.8)

- Case Control Studies

Sensitivity 88.7%
(81.4, 93.4)
Specificity: 98.6%
(97.3, 99.3)



Minion et al. Unpublished

25

Analysis of diagnostic studies

- It is not uncommon to see researchers:
  - Excluding patients or controls with no definitive diagnoses ("diagnostic myopia bias")
  - Excluding indeterminate or inconclusive results
  - Perform post-hoc "discrepant" analysis to move numbers within 2 x 2 tables
- Such analyses often result in spuriously inflated accuracy estimates

26

Example: exclusion of indeterminates can inflate accuracy estimates

OPEN ACCESS Freely available online PLoS one

The Impact of HIV Infection and CD4 Cell Count on the Performance of an Interferon Gamma Release Assay in Patients with Pulmonary Tuberculosis

Martine G. Aabye[1]*, Pernille Ravn[2], George PrayGod[3], Kidola Jeremiah[3], Apolinary Mugomela[4], Maria Jepsen[5], Daniel Faurholt[5], Nyagosya Range[6], Henrik Friis[5], John Changalucha[3], Aase B. Andersen[1]

1 Department of Infectious Diseases, University of Copenhagen, Rigshospitalet, Copenhagen, Denmark, 2 Unit for Infectious Diseases Q, University of Copenhagen, Herlev Hospital, Herlev, Denmark, 3 National Institute for Medical Research, Mwanza Medical Research Centre, Mwanza, Tanzania, 4 Zonal Tuberculosis Reference Laboratory, Bugando Medical Centre, Mwanza, Tanzania, 5 Department of Human Nutrition, Faculty of Life Sciences, University of Copenhagen, Frederiksberg, Denmark, 6 National Institute for Medical Research, Muhimbili Medical Research Centre, Dar Es Salaam, Tanzania

Abstract

Background: The performance of the tuberculosis specific Interferon Gamma Release Assays (IGRAs) has not been sufficiently documented in tuberculosis- and HIV-endemic settings. This study evaluated the sensitivity of the QuantiFERON TB-Gold In-Tube (QFT-IT) in patients with culture confirmed pulmonary tuberculosis (PTB) in a TB- and HIV-endemic population and the effect of HIV-infection and CD4 cell count on test performance.

Methodology/Principal Findings: 161 patients with sputum culture confirmed PTB were subjected to HIV- and QFT-IT testing and measurement of CD4 cell count. The QFT-IT was positive in 74% (119/161; 95% CI: 67–81%). Sensitivity was higher in HIV-negative (75/93) than in HIV-positive (44/68) patients (81% vs. 65%, p = 0.02) and increased with CD4 cell count in HIV-positive patients (test for trend p = 0.03). 23 patients (14%) had an indeterminate result and this proportion decreased with increasing CD4 cell count in HIV-positive patients (test for trend p = 0.03). Low CD4 cell count (<300 cells/µl) did not account for all QFT-IT indeterminate nor all negative results. Sensitivity when excluding indeterminate results was 86% (95% CI: 81–92%) and did not differ between HIV-negative and HIV-positive patients (88 vs. 83%, p = 0.39).

Conclusions/Significance: Sensitivity of the QFT-IT for diagnosing active PTB infection was reasonable when excluding indeterminate results and in HIV-negative patients. However, since the test missed more than 10% of patients, its potential as a rule-out test for active TB disease is limited. Furthermore, test performance is impaired by low CD4 cell count in HIV-positive patients and possibly by other factors as well in both HIV-positive and HIV-negative patients. This might limit the potential of the test in populations where HIV-infection is prevalent.

- If indeterminates are included:
  - Sens = 74%
- If indeterminates are excluded:
  - Sens = 86%

J Clin Epidemiol Vol. 52, No. 12, pp. 1231–1237, 1999
Published by Elsevier Science Inc.

ELSEVIER

0895-4356/99/$–see front matter
PII S0895-4356(99)00101-8

Discrepant Analysis: A Biased and an Unscientific Method for Estimating Test Sensitivity and Specificity

*Alula Hadgu**

CENTERS FOR DISEASE CONTROL AND PREVENTION, DIVISION OF STD PREVENTION, ATLANTA, GEORGIA

ABSTRACT. Discrepant analysis is a widely used technique for estimating test performance indices (sensitivity, specificity, etc.) of DNA-amplification tests for detecting infectious diseases. It has recently been claimed that the discrepant analysis–based estimates of specificity are typically less biased than those based on culture and that the discrepant analysis–based specificity shows little appreciable bias. In this article, I show that those conclusions are incorrect. Using a typical example from the published literature, I show that the discrepant analysis–based estimates of sensitivity and specificity can generate a significant and clinically important overestimation of the true sensitivity and specificity values. Moreover, I demonstrate that the concept of discrepant analysis is profoundly flawed and unscientific. It violates a fundamental principle of diagnostic testing—the principle that the new test should not be used to determine the true disease status. Thus, the major problem with discrepant analysis is not only that it is biased but that it is unscientific. Therefore, discrepant analysis should not be adopted for the evaluation of any diagnostic or screening test. J CLIN EPIDEMIOL 52;12:1231–1237, 1999. Published by Elsevier Science Inc.

KEY WORDS. Discrepant analysis, sensitivity, specificity, DNA-amplification tests, *Chlamydia trachomatis*

Lessons

- Early case-control studies are often used to promote and market tests
- But a large proportion of tests fail, once they are used in real world settings (e.g. large number of failed commercial serological tests)
- Case-control studies exaggerate accuracy estimates, especially if the two-gate approach is used
- Certain data analytic approaches can also inflate accuracy estimates
- Diagnostic studies can begin as case-control studies, but need to move beyond that to prospective studies in clinically indicated populations
- Even accuracy data may be insufficient to decide on clinical impact
- Regulatory agencies should demand prospective data and not just rely on case-control accuracy studies

29

Case study 3:

Where should TB tests be evaluated and which populations are appropriate?

30

It is not uncommon to see TB test evaluations where:

- Cases come from a high-incidence country and controls from a low-incidence country
- Tests work well in a low-incidence country and fall apart in a high-incidence country
- Tests that work well in immunocompetent persons fail in populations with high HIV prevalence

31

Example: cases from Zambia and controls from England

Table 1. Response rates in the ex-vivo enzyme-linked immunospot assay to ESAT-6- and CFP-10-derived peptides, recombinant ESAT-6 antigen and purified protein derivative in Zambian pulmonary tuberculosis patients and healthy Zambian and British adults, and tuberculin skin test results in healthy Zambian adults.

	Tuberculosis patients		Healthy Zambian adults		Healthy British adults (n = 40)
	HIV− (n = 11)	HIV+ (n = 39)	HIV− (n = 54)	HIV+ (n = 21)	
Response rates (%)					
ESAT-6 peptides	11 (100)	34 (87)	28 (52)	6 (29)	0 (0)
CFP-10 peptides	8 (73)	25[a] (66)	34 (63)	7 (33)	0 (0)
Combined ESAT-6/CFP-10 peptides	11 (100)	35 (90)	37* (69)*	9* (43)*	0 (0)
ESAT-6 antigen	9 (82)	18 (46)	23 (43)	4 (19)	0 (0)
PPD	11 (100)	28 (72)	45 (83)	6 (29)	33 (83)
TST	–	–	28/35** (80)	5/14** (36)	–

PPD, Purified protein derivative; TST, tuberculin skin test.
[a]n = 38.
*P value for difference 0.064.
**P value for difference 0.0057.

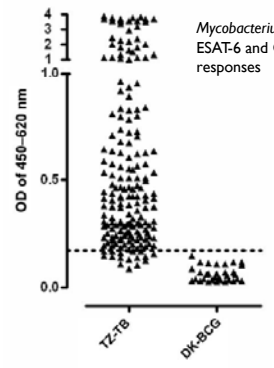
Rapid detection of active and latent tuberculosis infection in HIV-positive individuals by enumeration of *Mycobacterium tuberculosis*-specific T cells

Ann L.N. Chapman[a], Mwansa Munkanta[b], Katalin A. Wilkinson[a], Ansar A. Pathan[a], Katie Ewer[a], Helen Ayles[b,c], William H. Reece[a], Alwyn Mwinga[b], Peter Godfrey-Faussett[b,c] and Ajit Lalvani[a]

32

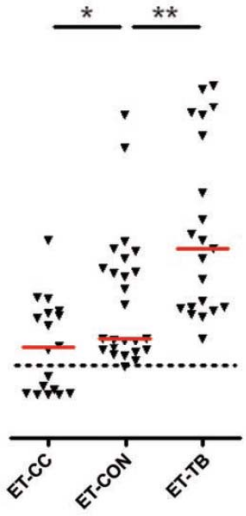
Lack of discrimination in TB endemic settings: example

Tanzanian cases



Mycobacterium tuberculosis–specific ESAT-6 and CFP-10 antibody responses

Danish controls



Ethiopia

Figure 3. Dot-plot showing the optical density (OD) values obtained from 184 patients with active tuberculosis disease who resided in northern Tanzania (TZ-TB) and 32 healthy, bacille Calmette-Guérin–vaccinated, Danish resident volunteer donors with no known risk factors for tuberculosis (DK-BCG). The dotted line indicates the cutoff value, calculated as the mean OD + 3 SDs for the 32 healthy Danish resident volunteers.

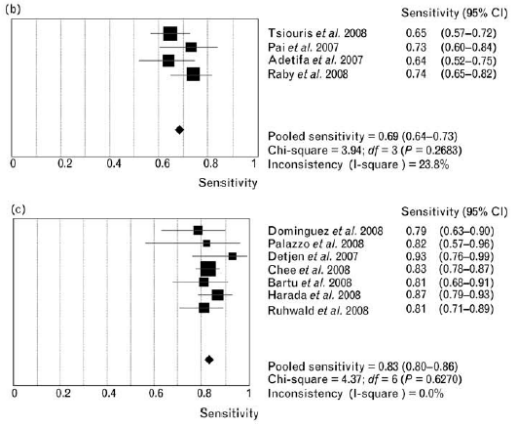
Hoff et al. Clinical Infectious Diseases 2007; 45:575–82

33

Variation in performance in high vs low endemic countries: example

T-cell interferon-γ release assays for the rapid immunodiagnosis of tuberculosis: clinical utility in high-burden vs. low-burden settings

Keertan Dheda[a,b,c], Richard van Zyl Smit[a], Motasim Badri[a] and Madhukar Pai[d]



High incidence countries

Low incidence countries

34

HIV can prove to be the acid test for any test! Example of MycoDot

MycoDot was hailed to be a breakthrough because it was a simple dipstick test

Commercialized and marketed by Mossman Associates (with support of PATH)

Package insert: sensitivity of 70% and specificity of 95%

35

But when the test was evaluated in countries with high HIV prevalence, the performance was disastrous

Evaluation of the MycoDot™ test in patients with suspected tuberculosis in a field setting in Tanzania

G. R. Somi,* R. J. O'Brien,† G. S. Mfinanga,* Y. A. Ipuge*

*National Institute for Medical Research, Dar Es Salaam, Tanzania, †WHO Global Tuberculosis Programme, Geneva, Switzerland, ‡National Tuberculosis and Leprosy Programme, Dar Es Salaam, Tanzania

SUMMARY

SETTING: Rapid, simple and inexpensive methods are needed to improve the diagnosis of tuberculosis in low-income countries. The MycoDot™ test has these characteristics.

OBJECTIVE: To assess the utility of the MycoDot™ test in screening patients with suspected tuberculosis.

DESIGN: Ambulatory patients presenting with symptoms of pulmonary tuberculosis were evaluated by physical examination and sputum acid-fast bacilli (AFB) microscopy. Separately, the MycoDot™ test was performed on whole blood. Patients with AFB-negative smears were treated with a 10-day course of erythromycin. Those remaining symptomatic had a chest radiograph. All sputum specimens were cultured for mycobacteria. Patients with culture-negative tuberculosis and those without a tuberculosis diagnosis were reassessed at 2 months.

RESULTS: Among the 241 patients who were evaluated, the MycoDot™ test was positive in 26% of patients with AFB-positive/culture-positive tuberculosis, 7% with AFB-negative/culture-positive tuberculosis, 7% with culture-negative tuberculosis, 19% treated for tuberculosis who did not meet study case definitions, and 16% without tuberculosis. Twenty four patients did not complete the assessment. Test sensitivity was 16%, specificity 94% and positive predictive value 45%. Sensitivity was highest (41%) in AFB-positive/HIV-negative patients and lowest (3%) in AFB-negative/HIV-positive patients.

CONCLUSION: The MycoDot™ test is not useful for the diagnosis of tuberculosis in sub-Saharan African countries, especially where HIV infection is prevalent.

KEY WORDS: tuberculosis; diagnosis; HIV; serology

Sens in HIV+ = 26%

Evaluation of a commercial immunodiagnostic kit incorporating lipoarabinomannan in the serodiagnosis of pulmonary tuberculosis in Ghana

S. D. Lawn*, E. H. Frimpong† and E. Nyarko‡

*Department of Medicine, School of Medical Sciences, University of Science and Technology, Kumasi, Ghana
†Department of Microbiology, School of Medical Sciences, University of Science and Technology, Kumasi, Ghana
‡National Tuberculosis Control Programme, Ministry of Health, Accra, Ghana

Summary

We evaluated 'Mycodot', a commercially marketed immunodiagnostic test for tuberculosis which detects antibodies to lipoarabinomannan antigen. Serum was tested from 32 patients with newly diagnosed smear-positive pulmonary tuberculosis, of whom 20 were HIV-positive and 32 HIV-negative. Control sera were taken from 40 patients of whom 20 had acute non-tuberculous lobar pneumonia and 20 patients had no respiratory disease. The test was found to have a very high specificity of 97.5% (95%CI 86.5–100%). However, the sensitivity in HIV-negative patients was 56% (95%CI 39–73%), and was substantially lower at 25% (95%CI=8–44%) in HIV-positive patients. In conclusion 'Mycodot' was found to be a highly specific and easily performed assay. However, the poor sensitivity, especially in HIV- infected patients, renders it unlikely to be useful either as a primary or adjunctive diagnostic test for tuberculosis, particularly in countries with a high prevalence of HIV. A larger trial of this assay in Ghana was not deemed necessary.

Sens in HIV+ = 25%

Despite these results, the test is still available on the market!

36

Lessons

- TB evaluation studies must be done in high TB incidence countries, especially in high HIV prevalent settings
- Performance outcomes from low incidence countries may be deceptive and not reflect the performance in high incidence settings where the challenges include:
  - HIV
  - Severe TB
  - High background prevalence of TB infection
  - Widespread BCG vaccination
  - Malnutrition
  - Other diseases that can affect performance (e.g. worm infestations)
- If tests perform well in TB/HIV endemic countries, then they are likely to hold up well!

37

Case study 4:

Who should conduct TB diagnostic studies?

38

Industry involvement in drug trials and its impact on study outcomes and conclusions

Scope and Impact of Financial Conflicts of Interest in Biomedical Research
A Systematic Review

JAMA 2003

Pharmaceutical industry sponsorship and research outcome and quality: systematic review

BMJ 2003

Association between industry funding and statistically significant pro-industry findings in medical and surgical randomized trials

CMAJ 2004

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

BMJ 2002

39

Industry involvement in diagnostic studies?

OPEN ACCESS Freely available online

PLoS one

Quality and Reporting of Diagnostic Accuracy Studies in TB, HIV and Malaria: Evaluation Using QUADAS and STARD Standards

Patricia Scolari Fontela[1], Nitika Pant Pai[2], Ian Schiller[2], Nandini Dendukuri[2], Andrew Ramsay[3], Madhukar Pai[1,4]*

Table 2. Characteristics of the studies included (N = 90).

Characteristic	Frequency (%)
Disease	
Tuberculosis	45 (50)
Malaria	18 (20)
HIV	27 (30)
Studies' origin*	
Africa	16
Asia	29
Australia and Oceania	01
Europe	27
North America	11
South America	06
Number of patients per study	
Median (interquartile range)	209 (110–555)
Number of studies with industry involvement	39 (43)
Number of studies with conflict of interest	38 (42)
Year of publication	
2004	42 (47)
2005	21 (23)
2006	27 (30)
Number of journals where included studies were published	46

About 40% of TB, HIV, Malaria diagnostic studies had industry involvement or known conflict of interest

Fontela P et al. PLoS One 2009

40

Industry involvement in TB diagnostic studies: example from IGRA literature

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

Of the 38 studies in the meta-analysis, 21 (55%) had some sort of industry involvement or support, such as sponsorship, donation of test kits, participation in advisory boards, involvement of test developers, or ownership of patents.

41

Industry involvement in TB, HIV, Malaria studies and likely impact: McGill-TDR/WHO study

Table 10 Multivariate logistic regression results using authors' conclusion (dependent variable) and industry involvement, disease of interest and quality assessment variables (independent variables)
[n = 153]

Variable	OR	95% CI
Industry involvement		
• No	1.0	Reference group
• Yes	4.28	1.83 - 10.02
• NR	5.11	1.77 - 14.74

Pai, Fontela, Dendukuri, Ramsay, et al.

42

Industry involvement in TB studies and likely impact: commercial IGRAs

- We searched for cost-effectiveness studies on commercial IFN-gamma release assays
- We found a total of 10 studies
- Of these 6 studies had industry involvement of some sort
  - 2 of 6 had CEO of a test making company as author!
- Of the 6 studies with industry involvement: ALL concluded in favor of the commercial test and claimed superior cost-effectiveness
- Of the 4 independent studies, two were in favor of the test, and two were cautious and recommended a more selective use of the test

43

Industry involvement in TB studies and likely impact: commercial IGRAs

Studies with industry involvement

Direct costs of three models for the screening of latent tuberculosis infection

P. Wrighton-Smith* and J-P. Zellweger#

Cost-optimisation of screening for latent tuberculosis in close contacts

R. Diel*, A. Nienhaus#, C. Lange† and T. Schaberg†

Cost-effectiveness of interferon-γ release assay testing for the treatment of latent tuberculosis

R. Diel*, P. Wrighton-Smith# and J-P. Zellweger†

Cost-effectiveness of Interferon-γ Release Assay Screening for Latent Tuberculosis Infection Treatment in Germany*

Roland Diel, MD, MPH, Albert Nienhaus, MD, MPH, and Robert Loddenkemper, MD, FCCP

Enhanced cost-benefit analysis of strategies for LTBI screening and INH chemoprevention in Germany

R. Diel [a,*], T. Schaberg [b], R. Loddenkemper [c], T. Welte [a], A. Nienhaus [d]

Targeted screening and treatment for latent tuberculosis infection using QuantiFERON®-TB Gold is cost-effective in Mexico

J. L. Burgos,* J. G. Kahn,† S. A. Strathdee,* A. Valencia-Mendoza,‡ S. Bautista-Arredondo,‡ R. Laniado-Laborín,§ R. Castañeda,‖ K. Deiss,* R. S. Garfein*

Independent studies

Interferon-gamma release assays and TB screening in high-income countries: a cost-effectiveness analysis

O. Oxlade, K. Schwartzman, D. Menzies

Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, McGill University, Montreal, Canada

Cost-effectiveness of Interferon Gamma Release Assays vs Tuberculin Skin Tests in Health Care Workers

Marie A. de Perio, MD; Joel Tsevat, MD, MPH; Gary A. Roselle, MD; Stephen M. Kralovic, MD, MPH; Mark H. Eckman, MD, MS

Cost-effectiveness of a new interferon-based blood assay, QuantiFERON®-TB Gold, in screening tuberculosis contacts

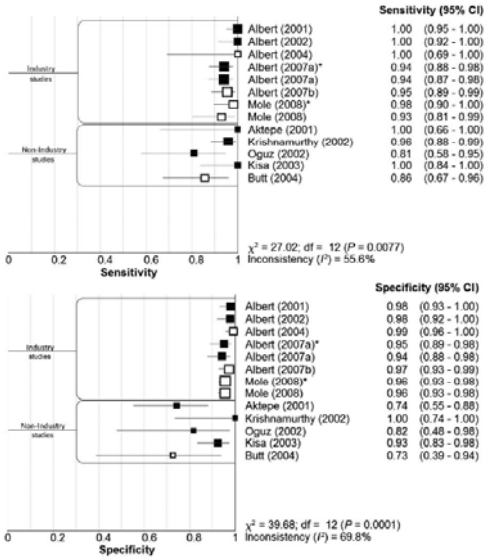
F. Marra,*† C. A. Marra,*† M. Sadatsafavi,* O. Morán-Mendoza,‡§ V. Cook,*† R. K. Elwood,*† M. Morshed,*† R. C. Brunham,*† J. M. FitzGerald*†

Cost Effectiveness of Interferon-γ Release Assay for Tuberculosis Contact Screening in Japan

Akiko Kowada,[1] Osamu Takahashi,[2] Takuro Shimbo,[3] Sachiko Ohde,[4] Yasuharu Tokuda[2] and Tsuguya Fukui[2]

44

FASTPlaque tests for drug-resistant TB



Minion J et al. ITJLD 2010

45

Lessons

- When test developers do the studies, test performance is always good; performance is less optimal when others try to replicate the results
  - may be suppression of unfavourable data
  - may just be a learning curve issue (test developers, by definition, understand the test better and know how to make it work!)
- While industry is critical for test development and commercialization, test evaluations should, ideally, be done independent of industry support
- At the very least, industry involvement should be clearly disclosed in all publications and presentations
- Industry and test developers should definitely not be involved in guideline and policy development
  - At least 17 countries have guidelines and statements on IGRAs
  - Vast majority of these guidelines had no disclosures on conflicts of interest

46

Case study 5:

Can we trust the package insert?

47

Commercial package inserts always provide data on test accuracy: can we trust them?

2009 package insert

TABLE 8. QuantiFERON®-TB Gold IT: Summary of results from clinical studies of subjects with culture-confirmed *M. tuberculosis* infection.

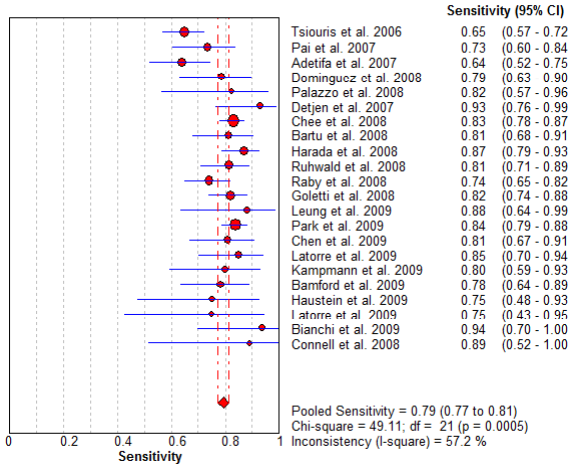
STUDY	QuantiFERON®-TB Gold IT			QuantiFERON®-TB Gold (liquid antigen)			TST (5mm)*	
	Pos	Neg	Ind	Pos	Neg	Ind	Pos	Neg
Australian	24	3	0	20	7	0	–	–
USA	47	11	3	34	10	6	60	19
Japanese	86	6	8	78	14	8	–	–
Overall Sensitivity	89% (157/177)			81% (132/163)			76% (60/79)	

Pos – Positive; Neg – Negative; Ind – Indeterminate
* In the U.S. study of 86 *M. tuberculosis* patients, TST results were missing for 4 and invalid for 3.

According to the company, this test has 89% sensitivity in active TB

48

Updated meta-analyses on sensitivity of QuantiFERON-TB Gold In Tube



Pooled estimate was about 79%

49

More examples...

Test	Package insert sens	Package insert spec	Meta-analysis sens	Meta-analysis spec
FASTPlaque-Response	96 – 100%	99 – 100%	95%	95%
Anda-TB IgG	85 - 90%	85 - 100 %	60 - 75%	~90%
MycoDot	70%	95%	26% - 76%	84% - 97%
Clearview TB ELISA	81% (HIV+)	93 – 98%	56% (HIV+)	95%
GenoType MDTBDrplus	99%	99%	98%	99%
Gen-Probe MTD	97% (S+) 72 (S-)	100% (S+) 99% (S-)	97% (S+) 76% (S-)	96% (S+) 95% (S-)

50

Lessons

- Company package inserts often present optimistic estimates based on small in-house evaluations that are usually sponsored by the companies
- Lab professionals and clinicians must be critical of advertised estimates of accuracy and performance
- Even when contradictory data are published, companies may not revise their package inserts or advertisements
- There is very little post-marketing surveillance of diagnostics and devices
- Regulatory agencies may not require companies to revise their package inserts
- Poorly performing tests may, in fact, never get pulled off the market

51

Case study 6:

Should we expect tests to be transferable and replicable?

Transferability: technologies that work well in the hands of developers will not necessarily work well everywhere

Eg. MODS, phage assays

52

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Microscopic-Observation Drug-Susceptibility Assay for the Diagnosis of TB

David A.J. Moore, M.D., Carlton A.W. Evans, M.D., Ph.D., Robert H. Gilman, M.D., Luz Caviedes, B.Sc., Jorge Coronel, B.Sc., Aldo Vivar, M.D., Eduardo Sanchez, M.D., Yvette Piñedo, M.D., Juan Carlos Saravia, M.D., Cayo Salazar, M.D., Richard Oberhelman, M.D., Maria-Graciela Hollm-Delgado, M.Sc., Doris LaChira, M.D., A. Roderick Escombe, M.D., Ph.D., and Jon S. Friedland, M.D., Ph.D.

MODS: developed in Peru – performs excellent

Sensitivity better than LJ (98 vs. 84%)

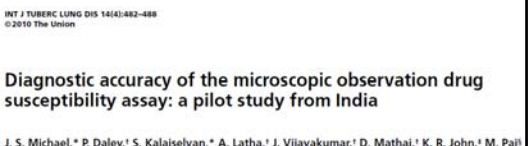
Fast turnaround time (1 week vs. 6 weeks+)

Implemented in India – performs poorly

Sensitivity 80%

Issues with contamination

Issues with reliability

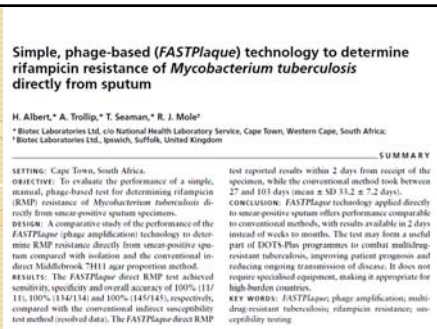


INT J TUBERC LUNG DIS 14(4):482–488

Diagnostic accuracy of the microscopic observation drug susceptibility assay: a pilot study from India

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

53



Simple, phage-based (*FASTPlaque*) technology to determine rifampicin resistance of *Mycobacterium tuberculosis* directly from sputum

H. Albert,* A. Trollip,* T. Seaman,* R. J. Mole†

* Biotec Laboratories Ltd, c/o National Health Laboratory Service, Cape Town, Western Cape, South Africa; † Biotec Laboratories Ltd., Ipswich, Suffolk, United Kingdom

SUMMARY

SETTING: Cape Town, South Africa.
OBJECTIVE: To evaluate the performance of a simple, manual, phage-based test for determining rifampicin (RMP) resistance of *Mycobacterium tuberculosis* directly from smear-positive sputum specimens.
DESIGN: A comparative study of the performance of the *FASTPlaque* (phage amplification) technology to determine RMP resistance directly from smear-positive sputum compared with isolation and the conventional indirect Middlebrook 7H11 agar proportion method.
RESULTS: The *FASTPlaque* direct RMP test achieved sensitivity, specificity and overall accuracy of 100% (11/11), 100% (134/134) and 100% (145/145), respectively, compared with the conventional indirect susceptibility test method (resolved data). The *FASTPlaque* direct RMP test reported results within 2 days from receipt of the specimen, while the conventional method took between 27 and 103 days (mean ± SD 33.2 ± 7.2 days).
CONCLUSION: *FASTPlaque* technology applied directly to smear-positive sputum offers performance comparable to conventional methods, with results available in 2 days instead of weeks to months. The test may form a useful part of DOTS-Plus programmes to combat multidrug-resistant tuberculosis, improving patient prognosis and reducing ongoing transmission of disease. It does not require specialised equipment, making it appropriate for high-burden countries.
KEY WORDS: *FASTPlaque*; phage amplification; multidrug-resistant tuberculosis; rifampicin resistance; susceptibility testing

FASTPlaque phage assay – performed well when done by industry

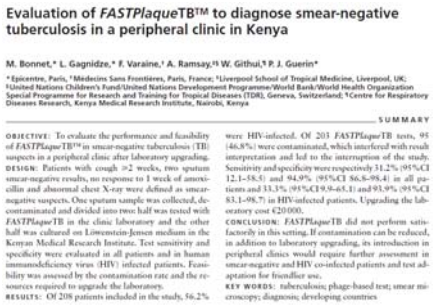
100% sens
100% spec

Implemented in Kenya – performs poorly

Despite upgrading the lab:

Low accuracy (31% sens; 95% spec)

Issues with contamination (nearly have were not interpretable)



Evaluation of *FASTPlaque*TB™ to diagnose smear-negative tuberculosis in a peripheral clinic in Kenya

M. Bonnet,* L. Gagnidze,* F. Varaine,† A. Ramsay,‡§ W. Githui,¶ P. J. Guerin*

* Epicentre, Paris, † Medecins Sans Frontières, Paris, France; ‡ Liverpool School of Tropical Medicine, Liverpool, UK; § United Nations Children's Fund/United Nations Development Programme/World Bank/World Health Organization Special Programme for Research and Training for Tropical Diseases (TDR), Geneva, Switzerland; ¶ Centre for Respiratory Diseases Research, Kenya Medical Research Institute, Nairobi, Kenya

SUMMARY

OBJECTIVE: To evaluate the performance and feasibility of *FASTPlaque*TB™ in smear-negative tuberculosis (TB) suspects in a peripheral clinic after laboratory upgrading.
DESIGN: Patients with cough ≥2 weeks, two sputum smear-negative results, no response to 1 week of amoxicillin and abnormal chest X-ray were defined as smear-negative suspects. One sputum sample was collected, decontaminated and divided into two; half was tested with *FASTPlaque*TB in the clinic laboratory and the other half was cultured on Löwenstein-Jensen medium in the Kenyan Medical Research Institute. Test sensitivity and specificity were evaluated in all patients and in human immunodeficiency virus (HIV) infected patients. Feasibility was assessed by the contamination rate and the resources required to upgrade the laboratory.
RESULTS: Of 208 patients included in the study, 56.2% were HIV-infected. Of 203 *FASTPlaque*TB tests, 95 (46.8%) were contaminated, which interfered with result interpretation and led to the interruption of the study. Sensitivity and specificity were respectively 31.2% (95%CI 12.1–58.5) and 94.9% (95%CI 86.8–98.4) in all patients and 33.3% (95%CI 9.9–65.1) and 93.9% (95%CI 83.1–98.7) in HIV-infected patients. Upgrading the laboratory cost €20 000.
CONCLUSION: *FASTPlaque*TB did not perform satisfactorily in this setting. If contamination can be reduced, in addition to laboratory upgrading, its introduction in peripheral clinics would require further assessment in smear-negative and HIV co-infected patients and test adaptation for friendlier use.
KEY WORDS: tuberculosis; phage-based test; smear microscopy; diagnosis; developing countries

54

Replication

- There are many examples of novel tests for TB that show great promise, but do not get replicated
- Or subsequent results are disappointing and commercialization is abandoned
- Or test may be quite good, but impossible to develop and manufacture in a cost-effective way
- Results in a graveyard of inexplicably abandoned diagnostics

55

Example: MPB64 skin patch test (Sequella Inc.)

INT J TUBERC LUNG DIS 2(7):543-546
© 1998 IUATLD

MPB64 mycobacterial antigen: a new skin-test reagent through patch method for rapid diagnosis of active tuberculosis

R. M. Nakamura,* M. A. Velmonte,† K. Kawajiri,* C. F. Ang,† R. A. Frias,† M. T. Mendoza,† J. C. Montoya,† I. Honda,* S. Haga,‡ I. Toida*

*Japan BCG Laboratory, Kiyose-shi, Tokyo, Japan, †Infectious Disease Section, Philippine General Hospital, Manila, Philippines, ‡National Institute of Infectious Diseases, Toyama, Shinjuku-ku, Tokyo, Japan

SUMMARY

SETTING: A collaborative study between the Japan BCG Laboratory, Tokyo, Japan, and the Infectious Disease Section, Philippine General Hospital, Manila, the Philippines. Tuberculosis patients from four clinics in the vicinity of Manila, Our Lady of Grace Parish, Sto. Niño de Tondo Parish, the Canossa Health and Social Center, and the Health Care Development Center, were examined.

OBJECTIVE: To develop a new, simple and rapid diagnostic method for active tuberculosis. Subjects were tested for skin reaction to a special antigen, MPB64, by the patch test method instead of intradermal injection of purified protein derivative (PPD).

DESIGN: Fifty-three active tuberculosis patients and 43 healthy PPD-positive controls were tested to determine whether or not the reaction to MPB64 was positive only in active tuberculosis patients.

RESULTS: Fifty-two of the 53 active tuberculosis patients showed a positive reaction to MPB64, while none of the 43 PPD-positive controls did. The specificity of MPB64 to active tuberculosis was 100%, and the sensitivity was 98.1%. The efficacy of the test was 98.9%.

CONCLUSION: The patch test with MPB64 is a promising method for the diagnosis of active tuberculosis, distinguishing tuberculous patients from those who are infected but have not developed the disease, and also from BCG-vaccinated individuals. This new skin test is a subject for further evaluation and it is important to compare the results with PPD Mantoux.

KEY WORDS: MPB64; patch skin test; rapid diagnosis; active TB

Early data in 1998:

Sensitivity: 98%
Specificity: 100%

In 2010, still not commercially available – plans have been abandoned

Pai M et al. Exp Rev Mol Diagn 2006;6(3):423-432; Image courtesy Sequella Inc.

56

Lessons

- Many novel tests and tools are bound to fail
- We need to appreciate the "failure rate" of new tests and interventions
- Replication, in diverse settings, is required, before proceeding with commercialization and clinical use
- Transferability of technologies must receive attention; tests need to be robust if they have to work well in all settings
- Tests that work well in the hands of developers may not work well in field settings, especially in resource-limited countries
- Single studies are never sufficient for policy and guideline development; we need more extensive evidence
- Even accuracy data are not sufficient for evidence-based policies

57

Roadmap for rolling out Xpert MTB/RIF for rapid diagnosis of TB and MDR-TB

6 December 2010

WHO THE STOP TB DEPARTMENT

World Health Organization

58

Will Xpert MTB/RIF survive "optimism bias"?

- Validation data and early demonstration data look very good
- Not much "real world" experience in resource-limited and routine programmatic settings
- Impact of "point-of-treatment" use is not demonstrated

Rapid Molecular Detection of Tuberculosis and Rifampin Resistance

Catharina C. Boehme, M.D., Pamela Nabeta, M.D., Doris Hillemann, Ph.D., Mark P. Nicol, Ph.D., Shubhada Shenai, Ph.D., Fiorella Krapp, M.D., Jenny Allen, B.Tech., Rasim Tahirli, M.D., Robert Blakemore, B.S., Roxana Rustomjee, M.D., Ph.D., Ana Milovic, M.S., Martin Jones, Ph.D., Sean M. O'Brien, Ph.D., David H. Persing, M.D., Ph.D., Sabine Ruesch-Gerdes, M.D., Eduardo Gotuzzo, M.D., Camilla Rodrigues, M.D., David Alland, M.D., and Mark D. Perkins, M.D.

59