

## *Mapping the landscape and quality of TB diagnostic research*



Madhukar Pai, MD, PhD  
Laurence Brunet, MSc  
Jessica Minion, MD  
Karen Steingart, MD, MPH  
Andrew Ramsay, MSc  
Christian Lienhardt, MD, PhD

### **PI contact information:**

Dr Madhukar Pai  
Dept. of Epidemiology, Biostatistics & Occupational Health  
McGill University  
1020 Pine Avenue West  
Montreal, Canada H3A 1A2  
Email: [madhukar.pai@mcgill.ca](mailto:madhukar.pai@mcgill.ca)

### **STP Research Movement focal point:**

Dr Christian Lienhardt  
Senior Scientific Advisor  
Stop TB Research Movement  
WHO/STB/TBP & Stop TB Partnership  
CH - 1211 Geneva 27  
Switzerland  
e-mail: [lienhardtc@who.int](mailto:lienhardtc@who.int)

**Report submitted to Stop TB Research Movement on Nov 12, 2009**



## BACKGROUND

The field of tuberculosis (TB) diagnostics is an active area of research and considerable work has been done to develop and evaluate new tools for TB diagnosis. High quality diagnostic studies are critical to evaluate new tools, to develop evidence-based policies on TB diagnostics, and, ultimately, for effective control of the global TB epidemic. The widespread application of new or improved diagnostic tests has the potential to save lives and greatly enhance global TB control efforts.

Despite the progress made and the high output of research on TB diagnostics, there is a concern that studies on the accuracy of TB diagnostics lack methodologic rigour [1-3]. Consequently, there is a perception that new tests that reportedly perform well in clinical trials may turn out to be less useful in routine clinical practice. Biased results from poorly designed studies can lead to premature adoption of tests that may have little or no clinical relevance, and result in adverse consequences for the patient and/or the healthcare service. Lack of methodologic rigour in TB trials is a cause for concern as it may prove to be a major hurdle for effective application of diagnostics in TB care and control.

There is also concern that existing TB diagnostic studies are almost entirely focused on Phase I/II questions - i.e. focused on sensitivity and specificity, with very little data on outcomes such as a) incremental or added value of new tests; b) impact on diagnostic thinking; c) impact on therapeutic choices, d) impact on patient-related outcomes such as mortality, treatment completion, and drop-out rates [4] and, ultimately, on programmatic aspects.

In this context, there is a need to map the landscape of TB diagnostic research and to also systematically study the quality of existing TB diagnostic research. This will enhance our ability to improve the quality and scope of TB diagnostic research, and add value to existing initiatives such as STARD [5], QUADAS [6] and DEEP [7]. It will also facilitate the development of a global research agenda on TB, one of the key goals of the Stop TB Research Movement [<http://www.stoptb.org/researchmovement/>].

## OBJECTIVES

Our project addressed two key questions:

**Question 1: What is the landscape of TB diagnostic research**, in terms of how many diagnostic studies are published annually, what types of diagnostic studies are published (e.g. study designs), which TB tests are most frequently evaluated (type of technology, purpose (e.g. active TB vs LTBI)), what types of diagnostic outcomes are reported (e.g. sensitivity & specificity versus other outcomes), which journals are these diagnostic studies published in, by whom, and from which countries?

**Question 2: What is the quality of TB diagnostic accuracy studies**, in terms of methodological quality (validity) and quality of reporting?

## METHODS

For question 1: To *map the landscape of TB diagnostic research*, we performed a bibliometric review of recent literature on TB diagnostic research. PubMed and EMBASE were selected to

search for all original TB publications in 2007-2008. Both databases were accessed online on May 7, 2009. For PubMed, the search strategy was: ("Mycobacterium tuberculosis"[Majr] OR "Tuberculosis"[Majr] OR "Tuberculosis/diagnosis"[Mesh] OR tuberculosis Field: Title) Limits: Publication Date from 2007/01/01 to 2008/12/31 NOT Field: Title, Editorial, Letter, Meta-Analysis, Practice Guideline, Review, Addresses, Bibliography, Biography, Comment, Dictionary, Directory, Interview, Newspaper Article). For EMBASE, the search strategy was: exp \*Mycobacterium Tuberculosis/ or exp \*Tuberculosis or exp Tuberculosis/di [Diagnosis] or tuberculosis.m\_titl. limit to yr="2007 - 2008" not (book or book series or editorial or letter or "review"). Database searches were done by a librarian.

A total of 8249 abstracts were retrieved and screened by a single reviewer. After exclusion of duplicates, of studies which primary focus was not tuberculosis and of non-original studies, 6459 citations were analysed. Information on the presence of an abstract, the language of publication and the journal were collected for every study included. The impact factors of the journals were obtained from the Journal Citation Report (JCR). The UK Clinical Research Collaboration Health Research Classification System (HRCS) was used to retrieve details on the type of research of each study [8]. The HRCS was adapted to encompass every type of study encountered during the pilot study. One broad category was added to the classification system in order to cover case reports and case series. A sub-category was also added to the Detection, Screening and Diagnosis category to include cost and cost-effectiveness studies.

Additional information was collected for the diagnosis studies on: study design and type of outcome reported (feasibility, reliability, sensitivity/specificity, concordance with reference standard, predictive values, ROC curve and AUC, LR, incremental value, cost, cost-effectiveness and impact on clinical outcomes), purpose of the test, technology platform, study participants, study population, reporting of HIV status, use of commercial vs. in-house test, country where study was done and whether or not it is a multicenter study.

A pilot study was conducted in order to optimize the questionnaire. PubMed was searched for articles on tuberculosis published in 2009. Two reviewers analysed the last 100 abstracts published. There was a 90% agreement between the two reviewers before the optimization of the questionnaire. When needed, the two reviewers consulted each other to categorize studies adequately. Stata 10.1 was used to analyse the data and produce the frequency tables.

## Overview of the Research Activity Codes

<b>1</b>	<b>Underpinning Research</b>
1.1	Normal biological development and functioning
1.2	Psychological and socioeconomic processes
1.3	Chemical and physical sciences
1.4	Methodologies and measurements
1.5	Resources and infrastructure (underpinning)
<b>2</b>	<b>Aetiology</b>
2.1	Biological and endogenous factors
2.2	Factors relating to physical environment
2.3	Psychological, social and economic factors
2.4	Surveillance and distribution
2.5	Research design and methodologies (aetiology)
2.6	Resources and infrastructure (aetiology)
<b>3</b>	<b>Prevention of Disease and Conditions, and Promotion of Well-Being</b>
3.1	Primary prevention interventions to modify behaviours or promote well-being
3.2	Interventions to alter physical and biological environmental risks
3.3	Nutrition and chemoprevention
3.4	Vaccines
3.5	Resources and infrastructure (prevention)
<b>4</b>	<b>Detection, Screening and Diagnosis</b>
4.1	Discovery and preclinical testing of markers and technologies
4.2	Evaluation of markers and technologies
4.3	Influences and impact
4.4	Population screening
4.5	Resources and infrastructure (detection)
<b>5</b>	<b>Development of Treatments and Therapeutic Interventions</b>
5.1	Pharmaceuticals
5.2	Cellular and gene therapies
5.3	Medical devices
5.4	Surgery
5.5	Radiotherapy
5.6	Psychological and behavioural
5.7	Physical
5.8	Complementary
5.9	Resources and infrastructure (development of treatments)
<b>6</b>	<b>Evaluation of Treatments and Therapeutic Interventions</b>
6.1	Pharmaceuticals
6.2	Cellular and gene therapies
6.3	Medical devices
6.4	Surgery
6.5	Radiotherapy
6.6	Psychological and behavioural
6.7	Physical
6.8	Complementary
6.9	Resources and infrastructure (evaluation of treatments)
<b>7</b>	<b>Management of Diseases and Conditions</b>
7.1	Individual care needs
7.2	End of life care
7.3	Management and decision making
7.4	Resources and infrastructure (disease management)
<b>8</b>	<b>Health and Social Care Services Research</b>
8.1	Organisation and delivery of services
8.2	Health and welfare economics
8.3	Policy, ethics and research governance
8.4	Research design and methodologies
8.5	Resources and infrastructure (health services)

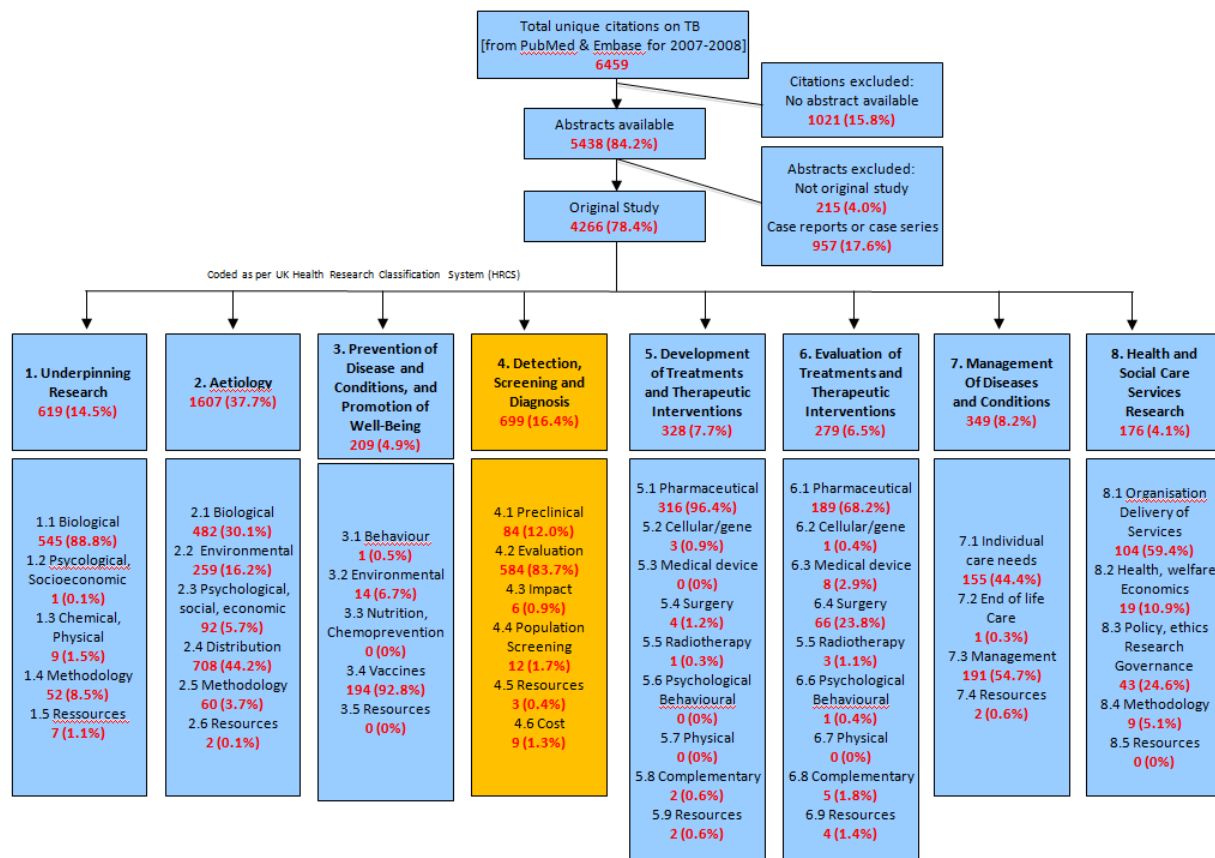
© UK Clinical Research Collaboration 2008

For question 2: To determine the *quality of TB diagnostic* accuracy research, we used two sources: 1) a recently published assessment of quality of TB, HIV and malaria diagnostic studies using QUADAS and STARD criteria [9]; and 2) quality assessments done as part of systematic reviews published on various TB diagnostic accuracy studies. Because systematic reviews and meta-analyses often include quality assessment as a key component of the systematic review process, and because several systematic reviews have been recently published on various TB diagnostic tests, there is now considerable evidence on quality of published TB diagnostic accuracy trials. From each systematic review, we extracted data on quality of studies that were included in the review. Some preliminary work has been done using this approach [3,4] and we updated the prior work for this analysis.

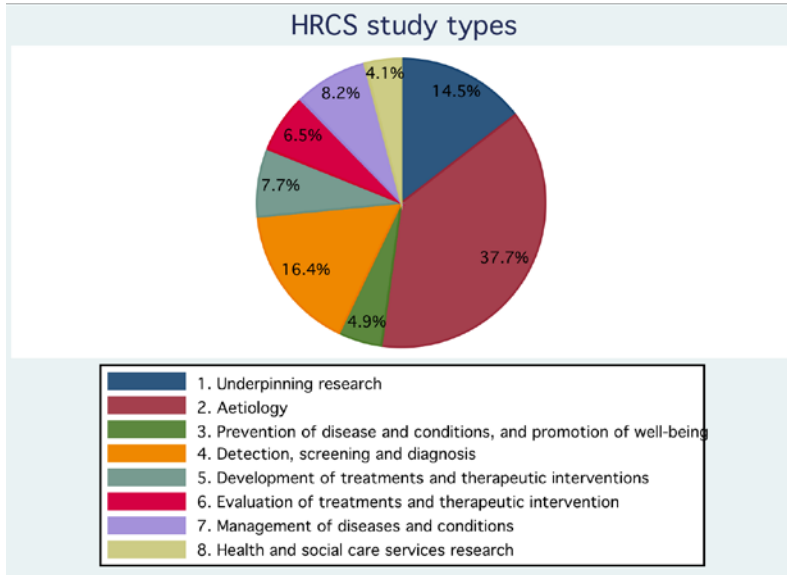
## RESULTS

### Question 1: What is the landscape of TB diagnostic research?

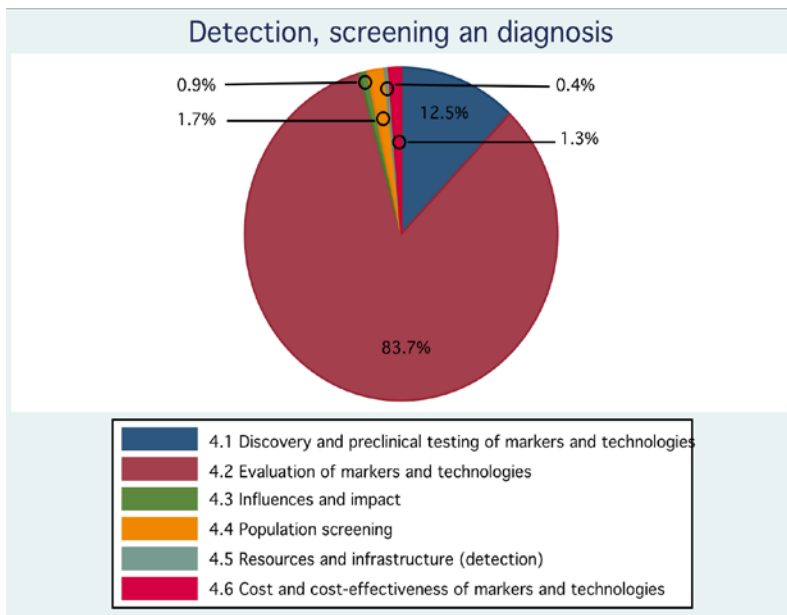
The flow chart below shows the main results, based on analysis of 4266 citations of original studies (after excluding citations of papers that were not original studies and case reports/series).



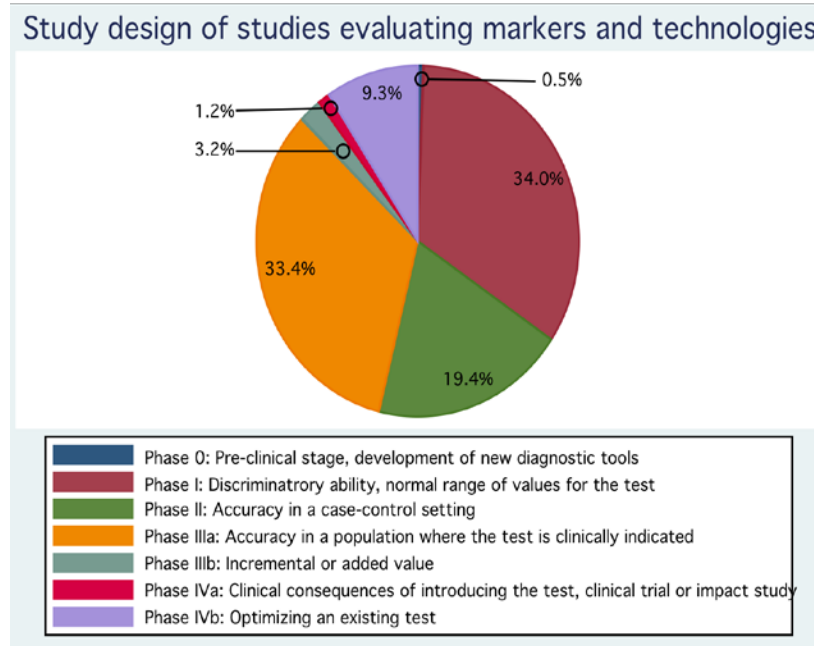
The appendix provides a detailed breakdown of the various codes and sub-categories. The distribution of main study types shows that diagnostic studies accounted for 16.4% of the 4266 citations (shown in the pie chart). Aetiology and underpinning research [basic science] accounted for about half of all citations.



Within the diagnostic category [N=699], about 84% of all the citations were focused on evaluation of markers and technologies.



Within the category of evaluation studies [N=584], >80% were early phase studies on accuracy.

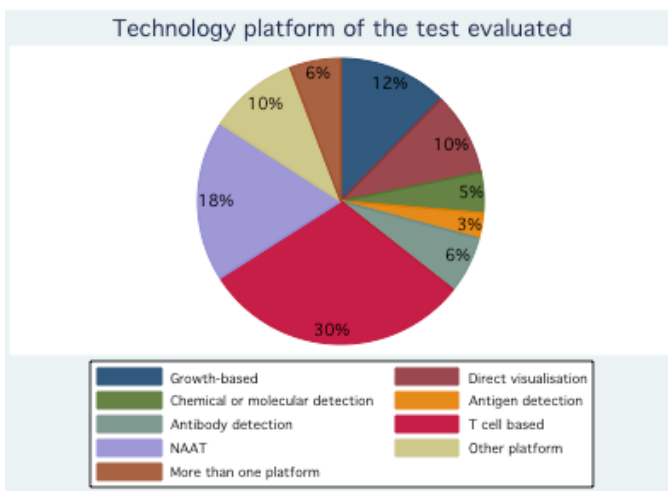
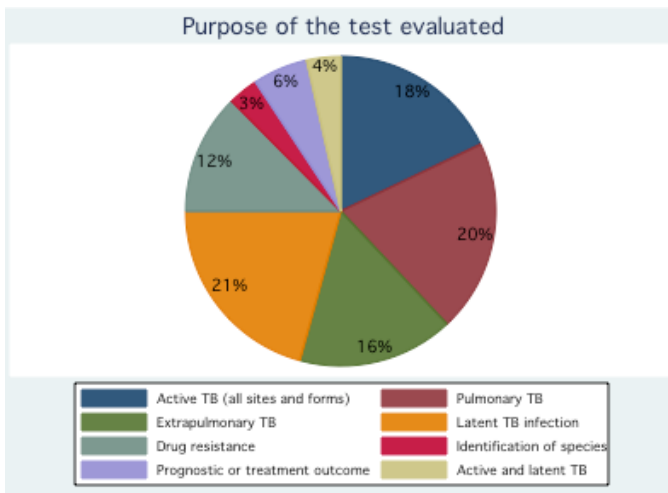
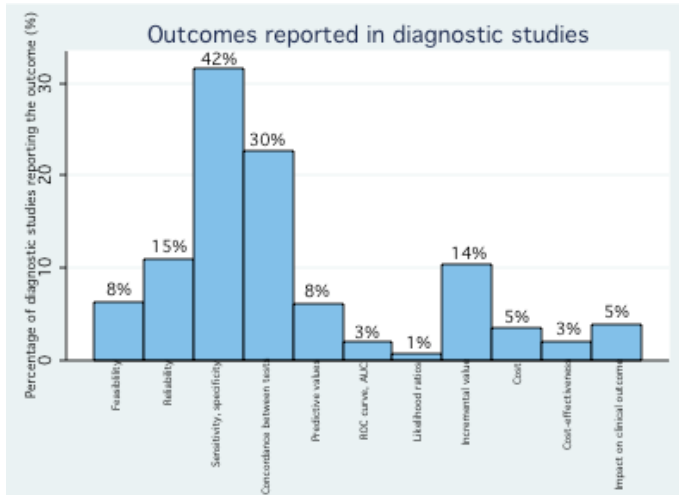


India, China, USA, Japan and Brazil were the top 5 countries that accounted for most of the diagnostic publications.

### Countries accounting for the majority of diagnostic studies

Country	N	%	Country	N	%
India	86	12.3	Germany	19	2.7
China	50	7.1	Italy	19	2.7
USA	47	6.7	Peru	17	2.4
Japan	44	6.3	UK	15	2.1
Brazil	36	5.1	Taiwan	14	2.0
Russia	36	5.1	Netherlands	13	1.8
South Africa	30	4.3	Spain	12	1.7
Turkey	29	4.1	Iran	10	1.4
Republic of Korea	23	3.3			

Among the 699 citations coded as diagnostic studies, sensitivity and specificity were the most frequently reported outcome. Cost, cost effectiveness and impact on patient outcomes were rarely reported. Active TB was the purposes of testing in more than half the diagnostic studies. T-cell assays accounted for nearly a third of all the technology platforms evaluated.



## Question 2: What is the quality of TB diagnostic accuracy studies?

In a recent study [9], we evaluated the quality of recently published diagnostic accuracy studies in TB, HIV and malaria. We identified 90 diagnostic studies of commercial tests for TB, malaria and HIV, through a systematic search of the literature published over a 3 year period. We then critically evaluated the study quality and completeness of reporting by using standardized, already published checklists called QUADAS and STARD. The results showed that diagnostic studies on TB, malaria and HIV commercial tests had moderate to low methodological quality and were often poorly reported. Sources of bias and variation were present in all the studies, and important criteria for determining the presence of bias were often either not mentioned or unclearly reported. For example, essential methodological elements, such as selection of a representative population and blinding, were not used and/or not reported by many researchers. Furthermore, only a small proportion of the studies adequately described how exactly the tests were performed and whether the tests were reproducible. The table below summarizes the results for the 45 studies on TB diagnostics.

Quality item	45 studies n (%)
Adequate spectrum composition	26 (58)
Clear description of selection criteria	21 (47)
Adequate reference standard	44 (98)
Absence of disease progression bias	42 (93)
Absence of partial verification bias	44 (98)
Absence of differential verification bias	42 (93)
Absence of incorporation bias	45 (100)
Absence of index test review bias	6 (13)
Absence of reference test review bias	7 (16)
Absence of clinical review bias	14 (31)
Report of uninterpretable results	9 (20)
Description of withdrawals	3 (7)

In another analysis [updated of reference 3], we extracted data on quality of studies that were included in several systematic reviews and meta-analyses (Table below).

Author, year [reference]	Studies included in the meta-analysis	Diagnostic test	Average size of each study included in the meta-analysis	Prospective data collection (%)	Consecutive or random sampling of subjects (%)	Cross-sectional design (%)	Blinded interpretation of test results [at least single blind] (%)	Complete verification of index test results by reference standard (%)
<b>Sarmiento et al. 2003</b>	16	PCR on respiratory specimens for smear-negative pulmonary TB	NR	50	NR	NR	63	<b>100</b>
<b>Goto et al. 2003</b>	40	ADA for TB pleural effusion	137	NR	NR	NR	0	<b>NR</b>
<b>Pai et al. 2003</b>	49	NAAT for TB meningitis	42	61	49	61	59	<b>94</b>
<b>Greco et al. 2003</b>	40	ADA and IFN- $\gamma$ tests for TB pleural effusion	135	23	15	NR	10	<b>NR</b>
<b>Pai et al. 2004</b>	40	NAAT for TB pleural effusion	60	63	53	70	55	<b>100</b>
<b>Flores et al. 2005</b>	84	In-house PCR for pulmonary TB	149	NR	NR	71	34	<b>NR</b>
<b>Kalantri et al. 2005</b>	13	Phage amplification tests for pulmonary TB	448	NR	NR	85	23	<b>100</b>
<b>Pai et al. 2005</b>	21	Phage based tests for rifampin resistance	85	NR	38	NR	57	<b>100</b>
<b>Morgan et al. 2005</b>	15	Line probe assay for rifampin resistance	91	NR	0	NR	13	<b>100</b>
<b>Greco et al. 2006</b>	63	Commercial NAAT for pulmonary TB	410	16	32	90	16	<b>NR</b>
<b>Steingart et al. 2006</b>	45	Fluorescence versus conventional sputum smear microscopy for pulmonary TB	493	100	36	NR	49	<b>NR</b>
<b>Steingart et al. 2006</b>	83	Direct versus concentrated sputum smear microscopy for pulmonary TB	256	100	21	NR	31	<b>NR</b>

Table (contd.)

Author, year [reference]	Studies included in the meta-analysis	Diagnostic test	Average size of each study included in the meta-analysis	Prospective data collection (%)	Consecutive or random sampling of subjects (%)	Cross-sectional design (%)	Blinded interpretation of test results [at least single blind] (%)	Complete verification of index test results by reference standard (%)
<b>Martin et al., 2007</b>	RIF: 18 INH: 16	Colorimetric redox-indicator methods for the rapid detection of multidrug resistance	RIF: 90 INH: 99	RIF: 28 INH: 31	RIF: 6 INH: 6	RIF: 100 INH: 100	RIF: 78 INH: 75	<b>RIF: 100</b> <b>INH: 100</b>
<b>Mase et al. 2007</b>	37	Yield of serial sputum smears for pulmonary TB	1134	49	95	NA	3	<b>NA</b>
<b>Steingart et al. 2007</b>	21	Commercial serological antibody detection tests for extra-pulmonary TB	45	43	0	43	5	<b>10</b>
<b>Steingart et al. 2007</b>	68	Commercial serological antibody detection tests for pulmonary TB	64	47	35	47	46	<b>57</b>
<b>Daley et al. 2007</b>	49	Nucleic acid amplification tests for TB lymphadenitis	57	69	18	80	12	<b>92</b>
<b>Ling et al. 2008</b>	10	Line probe assay for drug resistance (Genotype MTBDR assay)	116	NR	20	70	20	<b>100</b>
<b>Steingart et al. 2009</b>	254	Non-commercial serological tests	108	NR	8	15	26	<b>42</b>

ADA: adenosine deaminase; IFN- $\gamma$ : interferon- $\gamma$ ; INH: isoniazid; NAAT: nucleic acid amplification test; NA: not applicable; NR: not reported; PCR: polymerase chain reaction; RIF: rifampin; TB: tuberculosis

Table adapted from Pai & O'Brien (2006) [reference 3]

The table shows the quality elements that were most frequently reported in many of the meta-analyses: prospective data collection, consecutive or random sampling, cross-sectional design (as opposed to case-control), blinded interpretation of test results, and complete verification of index test results by reference standard. Only blinding was uniformly reported in all meta-analyses. On average, about 52% (range 16 – 100%) of the trials used a prospective data collection design. However, only 30% (range 0 – 95%) of the trials used a consecutive or random sampling method to recruit subjects. About 75% (range 43 – 100%) of the trials used a cross-sectional design, and

the case-control approach was used in about 25% of the studies. Any form of blinding was used in only 35% (range 0 – 78%) of the trials. In most studies (87%; range 10 – 100%), the index test results were verified by a reference standard test.

## **Conclusions**

Our extensive analysis of the landscape and quality of TB diagnostic research shows that:

- About 15% of all TB papers are mainly focused on TB diagnosis.
- Of these, about 85% are evaluation studies of tests and markers.
- Of these evaluation studies, about 85% are early phase studies of test accuracy
- There are very little data on patient outcomes, cost-effectiveness and impact in real world settings.
- Most test accuracy studies are of moderate to low quality and are poorly reported.
- Essential methodological and design elements are often either not reported or poorly reported.

These results have important implications for evidence-based policy making. High quality diagnostic studies are critical to evaluate new tools, to develop evidence-based policies on TB diagnostics. Based on the results of our analysis, it is evident that TB diagnostic trials are poorly conducted and poorly reported. Lack of methodologic rigour in TB trials is a cause for concern as it may prove to be a major hurdle for effective application of diagnostics in TB care and control.

Furthermore, it's evident that a majority of TB diagnostic studies are focused on test accuracy. There are limited data on outcomes such as accuracy of diagnostic algorithms (rather than single tests) and their relative contributions to the health care system, incremental value of new tests, impact of new tests on clinical decision-making and therapeutic choices, cost-effectiveness in routine programmatic settings, and impact on patient-important outcomes. This poses problems because research on test accuracy, while necessary, is not sufficient for policy and guideline development. Test accuracy data are surrogates for patient-important outcomes and cannot provide high quality evidence for policy making [10]. Therefore, accuracy studies must be considered along with impact of the test on patient-important outcomes, and other factors such as quality of the evidence, the uncertainty about values and preferences associated with the tests and presumed patient-important outcomes, and cost and feasibility, especially in resource-limited settings.

## **Acknowledgments**

This work was supported in part by a contract (APW Contract #200097280) from the Stop TB Partnership and World Health Organization.

## **Conflicts of interest**

None

## REFERENCES

1. Small PM, Perkins MD. More rigour needed in trials of new diagnostic agents for tuberculosis. *Lancet* 2000;356:1048-9
2. Walsh A, McNerney R. Guidelines for establishing trials of new tests to diagnose tuberculosis in endemic countries. *Int J Tuberc Lung Dis* 2004;8:609-13
3. Pai M, O'Brien R. Tuberculosis diagnostics trials: do they lack methodological rigor? *Expert Rev Mol Diagn* 2006;6:509-14
4. Pai M, Ramsay A, O'Brien. Evidence-based tuberculosis diagnosis. *PLoS Med*. 2008 Jul 22;5(7):e156.
5. Bossuyt PM et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med*. 2003 Jan 7;138(1):40-4.
6. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol*. 2003 Nov 10;3:25.
7. Banoo et al. Evaluation of diagnostic tests for infectious diseases: general principles. *Nat Rev Microbiol*. 2006 Dec;4(12 Suppl):S20-32.
8. UK Clinical Research Collaboration. Health Research Classification System (HRCS). UK. URL: <http://www.hrcsonline.net/pages/front> .
9. Fontela PS, Pai NP, Schiller I, Dendukuri N, Ramsay A, Pai M. Quality and Reporting of Diagnostic Accuracy Studies in TB, HIV and Malaria: Evaluation Using QUADAS and STARD Standards. *PLoS ONE* 2009;4(11): e7753.
10. Schunemann HJ, Oxman AD, Brozek J *et al*. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008; 336(7653), 1106-1110.

**Appendix: Breakdown of results in citations coded as: 4. Detection, Screening and Diagnosis (N=698)**

		4.1 Detection, Screening and Diagnosis 84 (12.0%)	4.2 Evaluation of markers and technologies 584 (83.7%)	4.3 Influences an d impact 6 (0.9%)	4.4 Populationscre ening 12 (1.7%)	4.5 Ressources and infrastructure(det ection) 3 (0.4%)	4.6 Cost and cost- effectiveness 9 (1.3%)
Study design	Phase 0	68 (80.9)	3 (0.5)	1 (16.7)	0	0	0
	Phase I	9 (10.7)	198 (34.0)	0	1 (8.3)	0	0
	Phase II	2 (2.4)	113 (19.4)	0	0	0	0
	Phase IIIa	2 (2.4)	195 (33.4)	0	1 (8.3)	0	6 (66.7)
	Phase IIIb	0	13 (2.2)	0	3 (25.0)	1 (33.3)	3 (33.3)
	Phase IVa	0	7 (1.2)	5 (83.3)	0	0	0
	Phase IVb	3 (3.6)	54 (9.3)	0	7 (58.4)	2 (66.7)	0
Feasibility		3 (3.6)	49 (8.4)	1 (16.7)	4 (33.3)	1 (33.3)	0
Agreement		1 (1.2)	99 (17.0)	0	2 (16.7)	1 (33.3)	0
Sensitivity/ Specificity		23 (27.4)	265 (45.5)	1 (16.7)	5 (41.7)	0	0
Concordance between Index and GS		8 (9.5)	201 (34.6)	1 (16.7)	3 (25)	0	0
Predicted values		5 (6.0)	51 (8.8)	1 (16.7)	1 (8.3)	0	0
ROC/AUC		2 (2.4)	16 (2.8)	0	0	0	0
Likelihood ratio		0	6 (1.0)	0	0	0	0
Incremental value		3 (3.6)	82 (14.1)	2 (33.3)	9 (75)	1 (33.3)	0
Cost		1 (1.2)	20 (3.4)	1 (16.7)	1 (8.3)	1 (33.3)	8 (88.9)
Cost effectiveness		1 (1.2)	7 (1.2)	0	1 (8.3)	1 (33.3)	9 (100)
Impact on clinical outcome		1 (1.2)	32 (5.6)	3 (50)	0	0	0
Purpose of test	Active all forms	30 (35.7)	87 (14.9)	1 (16.7)	2 (16.6)	1 (33.3)	3 (33.3)
	Pulmonary	10 (11.9)	130 (22.3)	0	0	0	1 (11.1)
	Extrapulmonary	9 (10.7)	103 (17.7)	1 (16.7)	0	0	0
	LTBI	10 (11.9)	122 (21.0)	2 (33.3)	5 (41.7)	1 (33.3)	5 (55.6)
	Drug resistance	11 (13.1)	75 (12.9)	0	0	1 (33.3)	0
	Identify species	3 (3.6)	19 (3.3)	0	0	0	0
	Prognostic/Treat ment outcome	8 (9.5)	32 (5.5)	0	0	0	0
Active+Latent	3 (3.6)	14 (2.4)	2 (33.3)	5 (41.7)	0	0	
Technology Platform	Growth-based	8 (9.5)	75 (12.9)	0	0	0	2 (22.2)
	Direct	6 (7.1)	55 (9.5)	0	4 (33.3)	1 (33.3)	0
	visualisation	10 (11.9)	22 (3.8)	0	0	0	0
	Chemical/mole cular detection	8 (9.5)	12 (2.1)	0	0	0	0
	Antigen detection	14 (16.7)	31 (5.3)	0	0	0	0
	Antigen detection	15 (17.9)	182 (31.3)	3 (50.0)	5 (41.7)	1 (33.3)	4 (44.5)
	Antibody detection	17 (20.2)	108 (18.6)	0	0	0	1 (11.1)
	Antibody detection	4 (4.8)	62 (10.7)	2 (33.3)	2 (16.7))	0	1 (11.1)
	T cell based NAAT	2 (2.4)	34 (5.8)	1 (16.7)	1 (8.3)	1 (33.3)	1 (11.1)

Other >1 platform							
Multi centre		2 (2.4)	16 (2.8)	0	2 (16.7)	1 (33.3)	0
Commercial test		0	217 (37.4)	2 (33.3)	2 (16.7)	1 (33.3)	5 (55.6)
Study participants	Adults	0	30 (5.2)	0	5 (41.7)	0	1 (11.1)
	Children	2 (2.4)	44 (7.6)	1 (16.7)	3 (25.0)	0	1 (11.1)
	Both	0	10 (1.7)	0	0	0	0
	Not specified	56 (66.7)	448 (77.1)	5 (83.3)	4 (33.3)	2 (66.7)	56 (66.7)
	No human involved	26 (30.9)	49 (8.4)	0	0	1 (33.3)	1 (11.1)
Population	Suspects	3 (3.6)	137 (23.6)	0	1 (8.3)	0	0
	Diseased	25 (29.8)	180 (30.1)	1 (16.7)	0	0	0
	Contacts	0	35 (6.0)	1 (16.7)	2 (16.7)	0	4 (44.5)
	General population	0	21 (3.6)	0	1 (8.3)	0	1 (11.1)
	High risk group	2 (2.4)	47 (8.1)	2 (33.2)	8 (66.7)	0	0
	Not specified	20 (23.8)	93 (16.0)	1 (16.7)	0	2 (66.7)	3 (33.3)
	>1	7 (8.3)	19 (3.3)	1 (16.7)	0	0	0
	No human involved	27 (32.1)	49 (8.4)	0	0	1 (33.3)	1 (11.1)
HIV		5 (6)	57 (9.8)	1 (16.7)	0	0	1 (11.1)

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

Patricia Scolari Fontela<sup>1</sup>, Nitika Pant Pai<sup>2</sup>, Ian Schiller<sup>2</sup>, Nandini Dendukuri<sup>2</sup>, Andrew Ramsay<sup>3</sup>, Madhukar Pai<sup>1,4\*</sup>

**1** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Canada, **2** Department of Medicine, Division of Clinical Epidemiology, McGill University, Montreal, Canada, **3** Special Programme for Research and Training in Tropical Diseases, World Health Organization, Geneva, Switzerland, **4** Respiratory Epidemiology and Clinical Research Unit, Montreal Chest Institute, Montreal, Canada

## Abstract

**Background:** Poor methodological quality and reporting are known concerns with diagnostic accuracy studies. In 2003, the QUADAS tool and the STARD standards were published for evaluating the quality and improving the reporting of diagnostic studies, respectively. However, it is unclear whether these tools have been applied to diagnostic studies of infectious diseases. We performed a systematic review on the methodological and reporting quality of diagnostic studies in TB, malaria and HIV.

**Methods:** We identified diagnostic accuracy studies of commercial tests for TB, malaria and HIV through a systematic search of the literature using PubMed and EMBASE (2004–2006). Original studies that reported sensitivity and specificity data were included. Two reviewers independently extracted data on study characteristics and diagnostic accuracy, and used QUADAS and STARD to evaluate the quality of methods and reporting, respectively.

**Findings:** Ninety (38%) of 238 articles met inclusion criteria. All studies had design deficiencies. Study quality indicators that were met in less than 25% of the studies included adequate description of withdrawals (6%) and reference test execution (10%), absence of index test review bias (19%) and reference test review bias (24%), and report of uninterpretable results (22%). In terms of quality of reporting, 9 STARD indicators were reported in less than 25% of the studies: methods for calculation and estimates of reproducibility (0%), adverse effects of the diagnostic tests (1%), estimates of diagnostic accuracy between subgroups (10%), distribution of severity of disease/other diagnoses (11%), number of eligible patients who did not participate in the study (14%), blinding of the test readers (16%), and description of the team executing the test and management of indeterminate/outlier results (both 17%). The use of STARD was not explicitly mentioned in any study. Only 22% of 46 journals that published the studies included in this review required authors to use STARD.

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

**Citation:** Fontela PS, Pant Pai N, Schiller I, Dendukuri N, Ramsay A, et al. (2009) Quality and Reporting of Diagnostic Accuracy Studies in TB, HIV and Malaria: Evaluation Using QUADAS and STARD Standards. PLoS ONE 4(11): e7753. doi:10.1371/journal.pone.0007753

**Editor:** Ben Marais, University of Stellenbosch, South Africa

**Received:** August 21, 2009; **Accepted:** October 16, 2009; **Published:** November 13, 2009

**Copyright:** © 2009 Fontela et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This study was funded in part by the United States Agency for International Development (USAID) through a grant awarded to the UNICEF/UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases [TDR] (Grant AAGG-00-99-00005-31) and by the Canadian Institutes of Health Research (CIHR grant MOP-89918) and European Commission (TBSusgent grant, EU-FP7). MP is a recipient of a CIHR New Investigator Award and a FRSQ grant (subvention d'établissement jeune chercheur). PF is a recipient of a Montreal Children's Hospital Research Institute (MCH-RI) fellowship for doctoral training. ND is a recipient of a Chercheur Boursier Junior 2 award from the Fonds de Recherche en Santé du Québec (FRSQ). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** Madhukar Pai is an editorial board member of PLoS Med and PLoS ONE.

\* E-mail: madhukar.pai@mcgill.ca

## Introduction

Tuberculosis (TB), malaria and human immunodeficiency virus (HIV), the 'big three' among infectious diseases, are major global causes of morbidity and mortality. Together, they cause more than 3.5 million deaths per year.[1,2,3] Consequently, considerable financial and other investments have been directed towards the control of these diseases in recent years, which includes the development of diagnostic and treatment services that are

accessible to patients. For example, the Global Fund to Fight AIDS, TB and Malaria has committed US\$ 15.6 billion in 140 countries to support large-scale prevention, treatment and care programs against these three diseases.[4]

Recently, simple and robust technological platforms that allow rapid diagnostic testing at the primary health care level have greatly increased diagnostic capability, particularly in developing countries. The use of such tests for HIV is well-established, and the use of rapid diagnostic tests (RDT) in malaria control programmes

is increasing.[5,6] Although point-of-care (POC) tests for TB have not been successful, the WHO has recently endorsed the use of two new diagnostic technologies for TB and drug-resistance, and several other new TB diagnostics are in the pipeline. [7,8,9,10]

The increasing number of diagnostic tests for TB, malaria and HIV leaves regulatory authorities, policy makers and health care professionals with the difficult task of choosing the tests that would best fit their patient populations and health-care delivery systems. In order to make evidence-based decisions, they often use published diagnostic accuracy studies as a way of gathering evidence about their options. [8] Also, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to guideline development requires a careful assessment of evidence on diagnostic accuracy, as well as other considerations, such as patient-important outcomes, the overall quality of evidence across these outcomes and the balance between benefits and harms and the strength of recommendations. [11,12] However, systematic reviews have revealed that the value of diagnostic accuracy studies is frequently compromised by poor methodological quality and/or poor reporting.[13,14,15] There is also a growing realization that design flaws can systematically bias estimates of diagnostic accuracy.[16,17,18] Furthermore, even diagnostic test accuracy data may not be sufficient for policy making, because they are surrogates for patient-important outcomes.[12]

In 2003, two tools were developed with the objective of providing researchers with a standardized and validated format for assessing quality of diagnostic studies and a template for improving reporting: QUADAS (Quality Assessment of Studies of Diagnostic Accuracy) and STARD (STAndards for the Reporting of Diagnostic accuracy studies).[19,20,21] QUADAS was designed to be used in systematic reviews to evaluate the quality of primary diagnostic accuracy studies, while STARD was developed to improve the quality of reporting of diagnostic accuracy studies in general.

Both tools are slowly gaining acceptance in the diagnostic literature. In April 2008, it was estimated that more than 200 biomedical journals encouraged the use of the STARD statement in their instructions for authors.[22] The QUADAS tool is increasingly being used in diagnostic accuracy meta-analyses. However, it is unclear if these tools have been widely accepted and applied to diagnostic accuracy studies of major infectious diseases. We performed a systematic review with the objective to describe the methodological and reporting quality of recently published diagnostic accuracy studies on commercial tests for TB, malaria and HIV.

## Methods

### Search Strategy

We searched PubMed and EMBASE (OVID interface) for primary diagnostic accuracy studies published between January 2004 and December 2006. We chose these databases because together they have a wide coverage of the health literature and would therefore enable us to obtain a fairly representative sample of indexed diagnostic studies published in the time period of interest. We limited the search to the period between 2004 and 2006 because we wanted to determine the methodological and reporting quality of diagnostic studies following the publication and dissemination of QUADAS and STARD.

The keywords and search terms that were used included {'tuberculosis' (explode) OR '*Mycobacterium tuberculosis*' (explode) OR '(tuberculosis or tuberculous).ti'} OR {'malaria' (explode) OR 'Plasmodium' (explode) OR 'malaria.ti'} OR {'HIV' (explode) OR 'HIV seropositivity' (explode) OR 'HIV infections' (explode)

OR 'acquired immunodeficiency syndrome' (explode) OR 'HIV.ti'}} AND {'sensitivity and specificity' (explode) OR 'specificity.ti' OR 'specificity.ab' OR 'accuracy.ti' OR 'diagn\$.ti'}}. The search was limited to studies in humans.

### Study Eligibility

We included diagnostic accuracy studies on commercial tests for TB, malaria and HIV that aimed to determine sensitivity and specificity of a given diagnostic test for one of these three infections. To be eligible, the studies had to be original, describe their methods, report sensitivity and specificity data and be published between January 2004 and December 2006. Languages were restricted to English, French, Spanish and Portuguese (languages that our study team was able to cover). Because commercial tests are standardized and usually test methods are well reported and easily defined, we restricted the study to commercial kits. In addition, commercial tests are more likely to be used in routine clinical practice than exclusively for research.

### Study Selection

Initially, one reviewer (PSF) screened the titles and abstracts of the citations retrieved by the electronic search (first screen). Citations that were identified as diagnostic accuracy studies were classified according to the disease (TB, malaria or HIV).

One researcher (PSF) reviewed the full text of all potentially eligible studies. A second researcher (NPP) independently reviewed 20% of all full text articles considered relevant in the first screen. Disagreements among reviewers were resolved by consensus. Figure 1 describes the study selection process.

### Data Abstraction

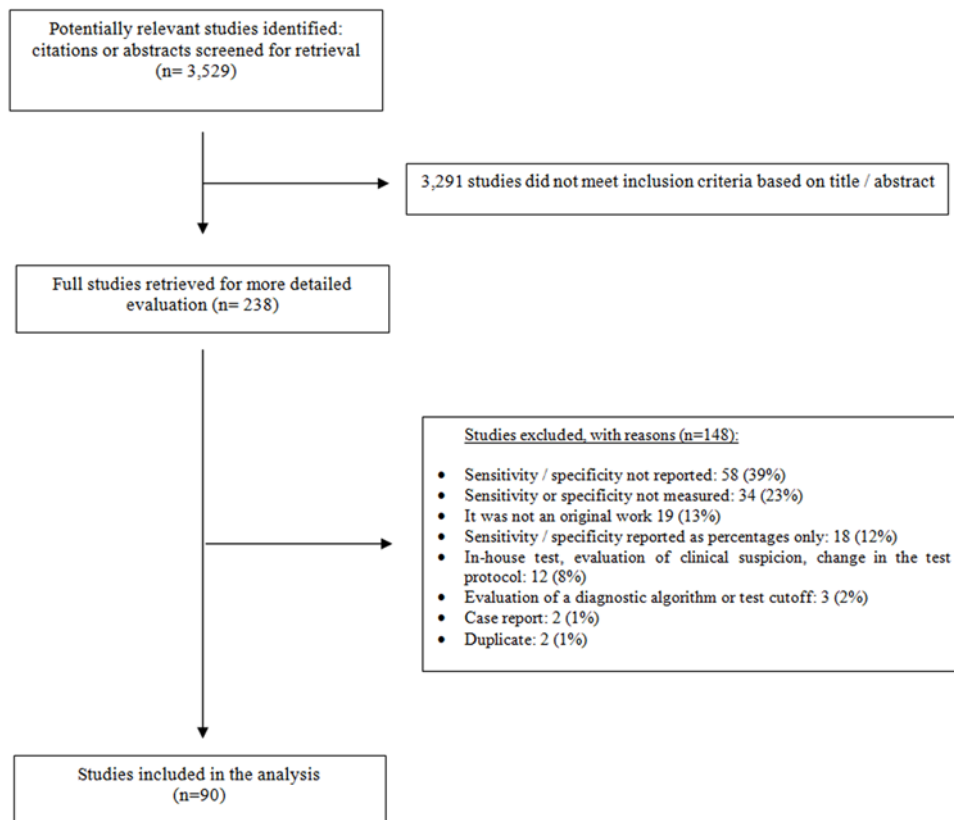
Two researchers (MP and PSF) created a data extraction form to be used in this review. The initial form was piloted by two reviewers (PSF and NPP) with 5% of the included publications. Based upon experience gained in the pilot, we modified and finalized the data extraction form.

Data extracted only included information explicitly stated in the text. Data retrieved included the following: year of publication, journal, disease of interest, type of commercial diagnostic test, reference standard employed, and data on quality of methods and reporting (listed below). When data were unavailable or not stated explicitly, the reviewers coded the information as "not reported". Any remaining disagreements were resolved by consensus before finalizing the data extraction.

### Assessment of Methodological Quality

We assessed the methodological quality of studies using QUADAS.[20,21] QUADAS is a validated quality checklist composed of 14 items, which encompass the most important sources of bias and variation observed in diagnostic accuracy studies. It was developed using a Delphi procedure which was used to reduce an initial list of 28 quality items.

The quality assessment items included in QUADAS are: spectrum composition, description of selection criteria and reference standard, disease progression bias, partial and differential verification, incorporation bias, description of index and reference test execution, test and reference standard review bias, clinical review bias, and description of uninterpretable test results. The definition of the items listed above can be found in Table 1. All the researchers involved in data extraction (PSF and NPP) were trained in the use of QUADAS checklist. Each item in the QUADAS checklist was scored as "Yes", "No", or "Unclear", as per the recommendations of the authors of the QUADAS checklist.



**Figure 1. Flow diagram for study selection.**  
doi:10.1371/journal.pone.0007753.g001

### Assessment of Quality of Reporting

The quality of the reporting was evaluated using the STARD criteria.[19] STARD, developed by a group of scientists and editors, consists of a checklist of 25 items that assess the completeness of reporting in diagnostic studies, potential sources of bias and generalizability. The checklist is subdivided in 5 sections: title/abstract/keywords, introduction, methods, results, and discussion. The majority of items in the STARD checklist were scored as “Not reported” or “Reported”. The “Reported” category included both “Fully reported” and “Partially reported” sub-categories. A “Partially reported” item means that the authors

mentioned the item, but did not provide all the information required by the STARD checklist about it.

Three STARD items were scored using other criteria: the item “participant recruitment” was scored as “recruitment based on symptoms” or “other recruitment/unclear”, while the item “participant sampling” was classified as “consecutive sampling” or “other sampling strategy/unclear”. Finally, the item “data collection” was scored as “prospective” and “retrospective”.

Eight out of the 25 STARD reporting items were considered essential by our group for the purposes of our project: reporting of the sampling strategy used, reference standard test, data collection

**Table 1. Biases in diagnostic accuracy test studies.**

Bias	Definition
Spectrum composition bias	When the spectrum of patients is not representative of the patients who will receive the test in practice
Disease progression bias	When the time period between reference standard and index test is not short enough to be reasonably sure that the target condition did not change between the two tests
Partial verification bias	When the whole sample or a random selection of the sample does not receive verification using a reference standard of diagnosis
Differential verification bias	When patients receive different reference standard depending on the index test result
Incorporation bias	When the reference standard is not independent of the index test, i.e., when the index test forms part of the reference standard
Test review bias	When the index test results are interpreted with knowledge of the results of the reference standard
Reference standard review bias	When the reference standard test results are interpreted with knowledge of the results of the index test
Clinical review bias	When test results are interpreted in the light of the clinical data that would not be available when the test is used in practice

Adapted from: Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003;3:25.  
doi:10.1371/journal.pone.0007753.t001

methods, blinding, proportion of eligible patients that did not participate in the study, inclusion and exclusion criteria, participant recruitment and description of clinic and demographic characteristics of the study population. These items were used to compare the quality of reporting of studies after stratifying them by disease (TB, Malaria and HIV).

### Use of STARD

In order to determine the frequency of use of STARD in diagnostic accuracy studies, we searched the full-text of all the included papers for any explicit mention of their use by the authors. Furthermore, in September 2008, we accessed the sections containing “information for the authors” (author guidelines) on the websites of all the journals (46 in all) in which the included papers were published. In doing so, we wanted to determine if the use of STARD was required when submitting a diagnostic accuracy manuscript to these journals.

### Data Synthesis and Statistical Analysis

Descriptive statistics were used to summarize the number and proportion of included studies that met the QUADAS and STARD criteria. We carried out a qualitative synthesis of the study characteristics, and quality of the methodology and reporting. Since the studies were heterogeneous with respect to diseases (TB, malaria and HIV), we decided to present overall results, as well as results stratified by disease subgroup. We also stratified the results by year of study publication in order to capture any temporal change since the publication of the STARD and QUADAS guidelines.

## Results

### Study Selection

We identified a total of 3,529 potentially relevant citations from the database searches. After the first and second screens, a total of 90 full-text studies were eligible for inclusion in this systematic review (Figure 1).

### Description of Included Studies

The characteristics of the included studies are shown in Table 2. Most papers were published in 2004 (47%). The 90 studies included were published in 46 different medical journals, Fifty percent evaluated TB diagnostic tests, 21% malaria diagnostic tests, and 29% HIV diagnostic tests.

### Use of STARD

No study explicitly mentioned using STARD for preparing the manuscript (this, however, does not mean that this tool was not actually used). When the journal websites of the 46 journals that published the included papers were searched in September 2008, only 10 of them (22%) required the authors to use STARD when submitting diagnostic accuracy study manuscripts.

### Assessment of the Methodological Quality Using QUADAS

The overall results of the quality assessment using QUADAS, as well as the results after stratification by disease and year of publication are presented in Tables 3 and 4.

The majority of studies used an adequate reference standard test (96%), and did not suffer from incorporation and partial or differential verification biases (98 and 92%, respectively). Reference standard tests considered “adequate” for TB, malaria and HIV were, respectively, sputum culture, blood smear examination and

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

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

\*The total number of countries is not 90 because there were some studies that were performed in more than one country.  
doi:10.1371/journal.pone.0007753.t002

ELISA and/or Western Blot. Nevertheless, all 90 studies included in this systematic review had at least one design flaw. The most commonly noted problems were associated with poor description of test execution, withdrawal of patients, and interpretation and reporting of test results.

Quality items that were reported in less than 25% of the studies included description of withdrawals (6%), adequate description of the reference test execution (10%), absence of index test review bias (19%), report of uninterpretable results (22%), and absence of reference test review bias (24%). Two other quality items were clearly described in less than 50% of the papers: index test execution (28%) and absence of clinical review bias (38%). Finally, a clear description of selection criteria and adequacy of spectrum composition, which are essential quality items for diagnostic accuracy studies, were reported in only 51 and 62% of studies, respectively.

Specific problems with some quality items were detected after we stratified the studies by disease (TB, malaria and HIV) and year of publication. In TB and HIV diagnostic accuracy studies, a clear description of selection criteria was present in less than 50% of time (47 and 48%, respectively). Moreover, the same item was reported in only 48% of the study sample published in 2006.

Furthermore, the results stratified by disease showed that HIV diagnostic accuracy studies met fewer of the methodological quality criteria when compared to those of TB and malaria. HIV studies were affected by higher prevalence of important biases such as partial (19%) and differential (37%) verification, incorporation (7%) and clinical review (70%) biases.

Finally, when the results were analyzed according to year of publication, we observed that in 2006, compared to previous years, a greater number of studies adequately described the index (37%)

**Table 3.** Assessment of methodological quality using QUADAS\* stratified by disease.

QUADAS item(scored as "Yes")	Disease			Total
	Tuberculosis (N = 45)	Malaria (N = 18)	HIV (N = 27)	(N = 90)
	n (%)	n (%)	n (%)	n (%)
<b>QUADAS 1</b>				
Adequate spectrum composition	26 (58)	13 (72)	17 (63)	56 (62)
<b>QUADAS 2</b>				
Clear description of selection criteria	21 (47)	12 (67)	13 (48)	46 (51)
<b>QUADAS 3</b>				
Adequate reference standard	44 (98)	18 (100)	24 (89)	86 (96)
<b>QUADAS 4</b>				
Absence of disease progression bias	42 (93)	15 (83)	21 (78)	78 (87)
<b>QUADAS 5</b>				
Absence of partial verification bias	44 (98)	17 (94)	22 (81)	83 (92)
<b>QUADAS 6</b>				
Absence of differential verification bias	42 (93)	17 (94)	17 (63)	76 (84)
<b>QUADAS 7</b>				
Absence of incorporation bias	45 (100)	18 (100)	25 (93)	88 (98)
<b>QUADAS 8</b>				
Adequate description of the index test execution	15 (33)	3 (17)	7 (26)	25 (28)
<b>QUADAS 9</b>				
Adequate description of the reference test execution	6 (13)	2 (11)	1 (4)	9 (10)
<b>QUADAS 10</b>				
Absence of index test review bias	6 (13)	5 (28)	6 (22)	17 (19)
<b>QUADAS 11</b>				
Absence of reference test review bias	7 (16)	8 (44)	7 (26)	22 (24)
<b>QUADAS 12</b>				
Absence of clinical review bias	14 (31)	12 (67)	8 (30)	34 (38)
<b>QUADAS 13</b>				
Report of uninterpretable results	9 (20)	1 (6)	10 (37)	20 (22)
<b>QUADAS 14</b>				
Description of withdrawals	3 (7)	1 (6)	1 (4)	5 (6)

\*Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003;3:25. doi:10.1371/journal.pone.0007753.t003

and reference standard (22%) tests used, as well as withdrawals (11%). These numbers, however, can still be considered very low.

### Assessment of the Quality of Report Using STARD

Tables 5 and 6 present the overall and stratified results (by disease and year of publication) in detail. No study fulfilled all the 25 items of STARD checklist. Overall, the major reporting problems encountered were in the sections about description of participants, test and statistical methods, and reporting of results.

Nine STARD items were reported in less than 25% of the studies: methods for calculation and estimates of test reproducibility (0%), adverse effects of the diagnostic tests (1%), estimates of diagnostic accuracy between subgroups (10%), distribution of severity of disease/other diagnoses in study participants (11%), number of eligible patients who did not participate in the study (14%), blinding of the test readers (16%), and description of the team executing the test and management of indeterminate, invalid/outlier results (both 17%).

Two other STARD items were poorly reported (less than 50% of time): participant sampling method (31%) and statistical methods to calculate diagnostic accuracy and uncertainty/precision (47%). When specifically analyzing the reporting of results' uncertainty, we observed that only 22 of the studies (24%) presented 95% confidence intervals.

When stratifying the studies by disease, HIV diagnostic accuracy studies met fewer of the reporting standards compared to those of TB and malaria diagnostics. Reports of HIV diagnostic accuracy studies failed, more frequently, to describe 5 out of 8 reporting items considered essential by our group: sampling strategies used (reported in 22% of studies), reference standard test (reported in 93% of HIV studies compared to 100% in TB and malaria studies), data collection methods (reported in 78% of studies), blinding (reported in 11% of studies – same as malaria) and proportion of eligible patients that did not participate in the study (reported in only 7% of studies). The 3 other reporting items considered essential were inclusion and exclusion criteria, participant recruitment and

**Table 4.** Assessment of methodological quality using QUADAS\* stratified by year of publication.

QUADAS item (scored as "Yes")	Year			Total
	2004 (N = 42)	2005 (N = 21)	2006 (N = 27)	(N = 90)
	n (%)	n (%)	n (%)	n (%)
<b>QUADAS 1</b>				
Adequate spectrum composition	26 (62)	14 (67)	16 (59)	56 (62)
<b>QUADAS 2</b>				
Clear description of selection criteria	21 (50)	12 (57)	13 (48)	46 (51)
<b>QUADAS 3</b>				
Adequate reference standard	41 (98)	20 (95)	25 (93)	86 (96)
<b>QUADAS 4</b>				
Absence of disease progression bias	38 (91)	16 (76)	24 (89)	78 (87)
<b>QUADAS 5</b>				
Absence of partial verification	40 (95)	17 (81)	26 (96)	83 (92)
<b>QUADAS 6</b>				
Absence of differential verification bias	36 (86)	16 (76)	24 (89)	76 (84)
<b>QUADAS 7</b>				
Absence of incorporation bias	42 (100)	19 (91)	27 (100)	88 (98)
<b>QUADAS 8</b>				
Adequate description of the index test execution	11 (26)	4 (19)	10 (37)	25 (28)
<b>QUADAS 9</b>				
Adequate description of the reference test execution	3 (7)	0 (0)	6 (22)	9 (10)
<b>QUADAS 10</b>				
Absence of index test review bias	10 (24)	4 (19)	3 (11)	17 (19)
<b>QUADAS 11</b>				
Absence of reference test review bias	10 (24)	3 (14)	9 (33)	22 (24)
<b>QUADAS 12</b>				
Absence of clinical review bias	17 (41)	7 (33)	10 (37)	34 (38)
<b>QUADAS 13</b>				
Report of uninterpretable results	10 (24)	4 (19)	6 (22)	20 (22)
<b>QUADAS 14</b>				
Description of withdrawals	2 (5)	0 (0)	3 (11)	5 (6)

\*Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 2003;3:25. doi:10.1371/journal.pone.0007753.t004

description of clinic and demographic characteristics of the study population.

Analysis by year of publication, revealed that in 2006, a greater number of studies reported the recruitment strategies used (63%), technical specifications of material and methods (100%), characteristics of study population (70%), number of eligible patients that did not undergo index/reference standard test (24%), distribution of severity of disease (24%) and estimate of diagnostic accuracy and 95% confidence intervals (100%) compared to previous years. However, it is important to highlight that the more frequent reporting of items such as description of material and methods does not mean that the quality of the report was adequate.

## Discussion

TB, malaria and HIV are major killers with enormous global burden. High-quality evidence on diagnostics is critical for the development of evidence-based policies on diagnosis, and, ultimately, for effective control of these global epidemics.[23] In

this study, we evaluated the methodological quality and reporting quality of recently published diagnostic accuracy studies in TB, HIV and malaria.

Our results show that diagnostic studies on TB, malaria and HIV commercial tests published between 2004 and 2006 had moderate to low methodological quality and were often poorly reported. Sources of bias and variation were present in all the studies, and important criteria for determining the presence of bias were often either not mentioned or unclearly reported. At least for TB and malaria, these results are consistent with previous observations made by several researchers.[8,24,25,26]

Most worrisome is the fact that essential methodological elements, such as selection of a representative population and blinding, were not used and/or not reported by many researchers. Furthermore, only a small proportion of the studies adequately described the execution of both reference (10%) and index (28%) tests, and no study reported on reproducibility. The implications of the under-reporting of these elements are several. For example, the value of sensitivity and specificity estimates are unclear in the

**Table 5.** Assessment of the quality of report using STARD\* stratified by disease.

Section and Topic in the STARD checklist (scored as "Reported")	Disease			Total
	TB (N = 45)	Malaria (N = 18)	HIV (N = 27)	(N = 90)
	n (%)	n (%)	n (%)	n (%)
<b>TITLE/ABSTRACT/KEYWORDS</b>				
Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	44 (98)	18 (100)	27 (100)	89 (99)
<b>INTRODUCTION</b>				
State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	44 (98)	17 (94)	25 (93)	86 (96)
<b>METHODS</b> (describe)				
<i>Participants</i>				
The study population: the inclusion and exclusion criteria, setting and locations where the data were collected.	30 (67)	17 (94)	23 (85)	70 (78)
Participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard? <sup>7</sup>	28 (62)	13 (29)	13 (48)	54 (60)
Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in the previous 2 items? If not, specify how participants were further selected. <sup>5</sup>	14 (31)	8 (44)	6 (22)	28 (31)
Data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)? <sup>9</sup>	38 (84)	16 (89)	21 (78)	75 (83)
<i>Test methods</i>				
The reference standard and its rationale.	45 (100)	18 (100)	25 (93)	88 (98)
Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	44 (98)	16 (89)	21 (78)	81 (90)
Definition of and rationale for the units, cutoffs, and/or categories of the results of the index tests and the reference standard.	41 (91)	16 (89)	18 (67)	75 (83)
The number, training, and expertise of the persons executing and reading the index tests and the reference standard.	3 (7)	7 (39)	5 (19)	15 (17)
Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	5 (11)	6 (33)	3 (11)	14 (16)
<i>Statistical methods</i>				
Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals).	16 (36)	12 (67)	14 (52)	42 (47)
Methods for calculating test reproducibility, if done.	0 (0)	0 (0)	0 (0)	0 (0)
<b>RESULTS</b> (report)				
<i>Participants</i>				
When study was done, including beginning and ending dates of recruitment.	34 (76)	16 (89)	16 (59)	66 (73)
Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	27 (60)	13 (29)	19 (70)	59 (66)
The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	7 (16)	4 (22)	2 (7)	13 (14)
<i>Test results</i>				
Time interval from the index tests to the reference standard, and any treatment administered between.	36 (80)	13 (29)	18 (67)	67 (74)
Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	6 (13)	1 (6)	3 (11)	10 (11)
A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	45 (100)	18 (100)	26 (96)	89 (99)
Any adverse events from performing the index tests or the reference standard.	1 (2)	0 (0)	0 (0)	1 (1)
<i>Estimates</i>				
Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).	43 (96)	17 (94)	27 (100)	87 (97)
How indeterminate results, missing responses, and outliers of the index tests were handled.	8 (18)	0 (0)	7 (26)	15 (17)
Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	1 (2)	2 (11)	6 (22)	9 (10)
Estimates of test reproducibility, if done.	0 (0)	0 (0)	0 (0)	0 (0)
<b>DISCUSSION</b>				
Discuss the clinical applicability of the study findings.	44 (98)	18 (100)	27 (100)	89 (99)

TB = tuberculosis MeSH = medical subject heading <sup>7</sup> = recruitment based on symptoms <sup>5</sup> = consecutive sampling <sup>9</sup> = prospective study.

\*Adapted from Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003;138:40-4.

doi:10.1371/journal.pone.0007753.t005

**Table 6.** Assessment of the quality of report using STARD\* stratified by year of publication.

Section and Topic in the STARD checklist (scored as "Reported")	Year			Total
	2004 (N = 42)	2005 (N = 21)	2006 (N = 27)	(N = 90)
	n (%)	n (%)	n (%)	n (%)
<b>TITLE/ABSTRACT/KEYWORDS</b>				
Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	42 (100)	21 (100)	26 (96)	89 (99)
<b>INTRODUCTION</b>				
State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	39 (93)	21 (100)	26 (96)	86 (96)
<b>METHODS</b> (describe)				
<i>Participants</i>				
The study population: the inclusion and exclusion criteria, setting and locations where the data were collected.	34 (81)	17 (81)	19 (70)	70 (78)
Participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard? <sup>7</sup>	24 (57)	13 (62)	17 (63)	54 (60)
Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in the previous 2 items? If not, specify how participants were further selected. <sup>5</sup>	14 (33)	8 (38)	6 (22)	28 (31)
Data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)? <sup>0</sup>	36 (86)	18 (86)	21 (78)	75 (83)
<i>Test methods</i>				
The reference standard and its rationale.	45 (100)	18 (100)	25 (93)	88 (98)
Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	41 (98)	20 (95)	27 (100)	88 (98)
Definition of and rationale for the units, cutoffs, and/or categories of the results of the index tests and the reference standard.	38 (91)	19 (91)	24 (89)	81 (90)
The number, training, and expertise of the persons executing and reading the index tests and the reference standard.	34 (81)	17 (81)	24 (89)	75 (83)
Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.	8 (19)	3 (14)	4 (15)	15 (17)
<i>Statistical methods</i>				
Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals).	17 (41)	11 (52)	14 (52)	42 (47)
Methods for calculating test reproducibility, if done.	0 (0)	0 (0)	0 (0)	0 (0)
<b>RESULTS</b> (report)				
<i>Participants</i>				
When study was done, including beginning and ending dates of recruitment.	34 (81)	13 (62)	19 (70)	66 (73)
Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	29 (69)	13 (62)	17 (70)	59 (66)
The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).	7 (17)	2 (10)	4 (24)	13 (14)
<i>Test results</i>				
Time interval from the index tests to the reference standard, and any treatment administered between.	33 (79)	14 (67)	20 (74)	67 (74)
Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	4 (10)	2 (10)	4 (24)	10 (11)
A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.	42 (100)	20 (95)	27 (100)	89 (99)
Any adverse events from performing the index tests or the reference standard.	1 (2)	0 (0)	0 (0)	1 (1)
<i>Estimates</i>				
Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).	40 (95)	20 (95)	27 (100)	87 (97)
How indeterminate results, missing responses, and outliers of the index tests were handled.	8 (18)	4 (19)	3 (11)	15 (17)
Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	5 (12)	2 (10)	2 (7)	9 (10)
Estimates of test reproducibility, if done.	0 (0)	0 (0)	0 (0)	0 (0)
<b>DISCUSSION</b>				
Discuss the clinical applicability of the study findings.	41 (98)	21 (100)	27 (100)	89 (99)

MeSH = Medical Subject Heading <sup>7</sup> = recruitment based on symptoms <sup>5</sup> = consecutive sampling <sup>0</sup> = prospective study.

\*Adapted from Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 2003;138:40-4.

doi:10.1371/journal.pone.0007753.t006

absence of clear information about test reproducibility. Moreover, if a reference standard is imperfect or poorly done, then this can potentially under-estimate or over-estimate the accuracy of a test. If the index test is poorly described, other researchers cannot replicate the study results (although this is less of an issue with standardized commercial kits).

### Strengths and Limitations

The major strength of this study is the systematic search for diagnostic accuracy studies via PubMed and EMBASE, two of the most widely used health literature databases. Furthermore, we used rigorous methods to select studies and abstract data, the latter independently conducted by two trained researchers.

The use of both QUADAS and STARD to evaluate diagnostic accuracy studies is also a strength of this systematic review. Both tools were developed by experts with the respective aims of assessing the quality of diagnostic studies included in systematic reviews and improve the quality of reporting of diagnostic studies in general. Furthermore, QUADAS and STARD are well standardized and easy to implement.[21,27] The complementary aspect of these tools also allowed us to have a deeper understanding of the current methodological and reporting quality of these studies. For example, for the item “reference test execution”, while more than 90% of the studies reported the reference test execution (STARD), only less than 25% of them did it in an adequate and clear manner (QUADAS).

An important limitation of our study is that we did not compare our results to a sample of studies published before the publication of QUADAS and STARD instruments (i.e., prior to 2003). Consequently, we can provide information about the current quality of methods and reporting of diagnostic studies, but not about changes in quality or reporting over time.

Wilczynski and colleagues compared the quality of report of papers published in journals that endorsed STARD versus those that did not (i.e., journals that published or not the STARD statement in 2003).[28] Studies were also compared according to year of publication (2001, 2002, 2004 and 2005). The results showed that the quality of report was not affected by the type of journal, and that it remained similar over time.

Another limitation of our study is the fact that we decided to only record information that was clearly stated in the paper, coding as “not reported” when data were not available. Thus, it may be possible that methodological quality items were met in the actual study, but not reported. Because we did not contact all the authors, we were unable to resolve this issue.

### Implications

Poor quality of diagnostic studies is a recognized problem. After the publication of QUADAS and STARD in 2003, the expectation was that the methodological quality of diagnostic studies, and the quality of their reporting, would improve over the years. Unfortunately, this objective seems to be far from being achieved, at least with respect to diagnostic studies on major infectious diseases.

Our results suggest that STARD is probably not used by researchers as often as expected or desired, at least in the field of infectious diseases. Furthermore, we have shown that, based on the results of a search performed in September 2008, only 22% of the journals in our study sample required authors to use STARD when submitting a diagnostic accuracy manuscript for publication. Consequently, we hypothesize that fact that not many journals require authors to use STARD may be one of the causes behind the lack of improvement of reporting of diagnostic studies over time. When we repeating this search in October 2009, we

observed that this number increased to 50%, probably due to the adoption of the *Uniform Requirements for Manuscripts Submitted to Biomedical Journal* (URM) created by the International Committee of Medical Journal Editors (ICMJE), which recommends authors to use “reporting guidelines relevant to their specific research design”, such as STARD.[29] Despite the substantial increase in the proportion of journals recommending the use of STARD, this proportion is still far from ideal.

Decreasing the burden of TB, malaria and HIV is a priority worldwide, and the provision of universal, high-quality and affordable diagnostic tests to affected populations is the first key step to achieve this goal. Regulatory authorities, policy makers and healthcare professionals frequently use diagnostic accuracy studies to decide which test should be implemented in a particular setting. However, choices based on biased study results may lead to detrimental consequences.

Lack of methodological rigour in diagnostic trials is a cause for concern as it may prove to be a major hurdle for effective application of diagnostics in controlling TB, malaria and HIV. Depending on how the presence of bias affects the estimates of diagnostic accuracy, a large number of patients could be harmed by not being properly diagnosed and consequently not receiving adequate care. [16,17] Furthermore, biased results from poorly designed studies can lead to premature or misguided adoption of tests that may have little or no clinical and public health relevance, and result in incorrect diagnosis and adverse consequences for the patient and/or the healthcare service. A good example of this is widespread use of serological, antibody tests for TB, when all the evidence suggests that they have poor accuracy and have no clinical role.[8] The situation is exacerbated by the fact that most developing countries have poor or nonexistent regulatory mechanisms for marketing and post-marketing surveillance of diagnostics.[30]

Thus, due to the negative implications that biased studies can present, efforts are urgently needed to improve quality of diagnostic research as well as quality of reporting. The more frequent use of tools such as QUADAS and STARD could aid in this process. While not designed with this intent, QUADAS, for example, could be used by researchers as a guideline when designing diagnostic studies, as it describes all the quality elements that should be present in this type of study. QUADAS can also be used as an educational tool, to help train researches in improving research design. STARD can be very useful at the manuscript development stage. However, because voluntary use of tools such as QUADAS and STARD is likely to be limited, their widespread use will probably only happen if more journals explicitly required and mandated authors to use these tools.

While improving diagnostic accuracy studies is a good starting point, efforts must also be made to go beyond test accuracy and generate evidence on patient-important outcomes that can inform policy and guideline development. For example, much of the existing evidence-base in TB is focused on test accuracy [8,31]. There are limited data on outcomes such as accuracy of diagnostic algorithms (rather than single tests) and their relative contributions to the health care system, incremental value of new tests, impact of new tests on clinical decision-making and therapeutic choices, cost-effectiveness in routine programmatic settings, and impact on patient-important outcomes. Future diagnostic studies must attempt to collect data on these outcomes and not merely focus on test accuracy.

In conclusion, our data suggests that recently published diagnostic studies on commercial tests for TB, malaria and HIV are of moderate to low quality and are poorly reported. Essential methodological and design elements were often either not reported

or poorly reported. The more frequent use of tools such as QUADAS and STARD may be necessary to improve methodological quality and reporting of future diagnostic accuracy studies in infectious diseases. This may happen only when more journals require authors to use instruments such as STARD.

## Acknowledgments

The authors thank Dr. Jesse Papenburg for his thoughtful review of this manuscript.

## References

- Aregawi M, Cibulskis R, Williams R, Dye C (2008) World malaria report 2008. Geneva, Switzerland: World Health Organization.
- Dye C, Floyd K, Uplekar M (2008) Global tuberculosis control: surveillance, planning, financing: WHO report 2008. Geneva, Switzerland: World Health Organization.
- Joint United Nations Programme on HIV/AIDS (2008) Report on the global AIDS epidemic. Geneva, Switzerland: UNAIDS.
- The Global Fund to Fight AIDS Tuberculosis and Malaria (2009) The Global Fund to Fight AIDS, Tuberculosis and Malaria. <http://www.theglobalfund.org/en/> (Access date: August 7, 2009).
- Peeling RW, Holmes KK, Mabey D, Ronald A (2006) Rapid tests for sexually transmitted infections (STIs): the way forward. *Sex Transm Inf* 82: v1–v6.
- Hopkins H, Asimwe C, Bell D (2009) Access to antimalarial therapy: accurate diagnosis is essential to achieving long term goals. *BMJ* 339: b2606.
- World Health Organization (2008) WHO policy statement: molecular line probe assays for rapid screening of patients at risk of multidrug-resistant tuberculosis. [http://www.who.int/tb/features\\_archive/policy\\_statement.pdf](http://www.who.int/tb/features_archive/policy_statement.pdf) (Access date: August 3, 2009).
- Pai M, Ramsay A, O'Brien R (2008) Evidence-Based Tuberculosis Diagnosis. *PLoS Medicine* 5: e156.
- Pai M, O'Brien R (2008) New diagnostics for latent and active tuberculosis: state of the art and future prospects. *Semin Respir Crit Care Med* 29: 560–568.
- World Health Organization (2007) New WHO policies: the use of liquid medium for culture and DST. <http://www.who.int/tb/dots/laboratory/policy/en/index3.htm> (Access date: August 3, 2009).
- Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, et al. (2008) Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 336: 1106–1110.
- Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, et al. (2004) Grading quality of evidence and strength of recommendations. *BMJ* 328: 1490–.
- Reid MC, Lachs MS, Feinstein AR (1995) Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 274: 645–651.
- Rama KRBS, Poovali S, Apsingi S (2006) Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. *Clinical Orthopaedics & Related Research* 447: 237–246.
- Siddiqui MAR, Azuara-Blanco A, Burr J (2005) The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *British Journal of Ophthalmology* 89: 261–265.
- Lijmer JG, Mol BW, Heisterkamp S, Bossel GJ, Prins MH, et al. (1999) Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests. *JAMA* 282: 1061–1066.
- Rutjes AWS, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, et al. (2006) Evidence of bias and variation in diagnostic accuracy studies. *CMAJ* 174: 469–476.
- Westwood M, Whiting P, Kleijnen J (2005) How does study quality affect the results of a diagnostic meta-analysis? *BMC Medical Research Methodology* 5: 20.
- Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, et al. (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD Initiative. *Ann Intern Med* 138: 40–44.
- Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology* 3: 25.
- Whiting P, Westwood M, Rutjes A, Reitsma J, Bossuyt P, et al. (2006) Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. *BMC Medical Research Methodology* 6: 9.
- Standards for the Reporting of Diagnostic Accuracy Studies Group (2008) STARD statement: news. <http://www.stard-statement.org/> (Access date: August 7, 2009).
- Mabey D, Peeling RW, Ustianowski A, Perkins MD (2004) Tropical infectious diseases: Diagnostics for the developing world. *Nat Rev Micro* 2: 231–240.
- Small PM, Perkins MD (2000) More rigour needed in trials of new diagnostic agents for tuberculosis. *The Lancet* 356: 1048–1049.
- Cot M (2005) Clinical research on malaria: what for the future? *Rev Epidemiol Sante Publique* 53: 291–297.
- Pai M, O'Brien R (2006) Tuberculosis diagnostics trials: do they lack methodological rigor? *Expert Review of Molecular Diagnostics* 6: 509–514.
- Smidt N, Rutjes A, van der Windt D, Ostelo R, Bossuyt P, et al. (2006) Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC Medical Research Methodology* 6: 12.
- Wilczynski NL (2008) Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication—before-and-after study. *Radiology* 248: 817–823.
- International Committee of Medical Journal Editors (2008) International Committee of Medical Journal Editors (ICMJE). Uniform Requirements for Manuscripts Submitted to Biomedical Journal.
- Peeling RW, Smith PG, Bossuyt PMM (2006) A guide for diagnostic evaluations. *Nat Rev Micro* 4: S2–S6.
- Pai M, Ramsay A, O'Brien R (2009) Comprehensive new resource for evidence-based TB diagnosis. *Expert Rev Mol Diagn* 9(7): 637–9.

## Author Contributions

Conceived and designed the experiments: ND AR MP. Performed the experiments: PF NPP. Analyzed the data: PF IS ND MP. Wrote the paper: PF NPP IS ND AR MP. Obtained funding: MP AR. Provided supervision: MP.