Reporting of diagnostic accuracy studies

Madhukar Pai, MD, PhD
Assistant Professor
Dept. of Epidemiology, Biostatistics & Occupational Health
McGill University
Montreal, Canada H3A 1A2
Email: madhukar.pai@mcgill.ca

McGill

Poor reporting in diagnostic studies

Diagnostic studies in 4 general medical journals

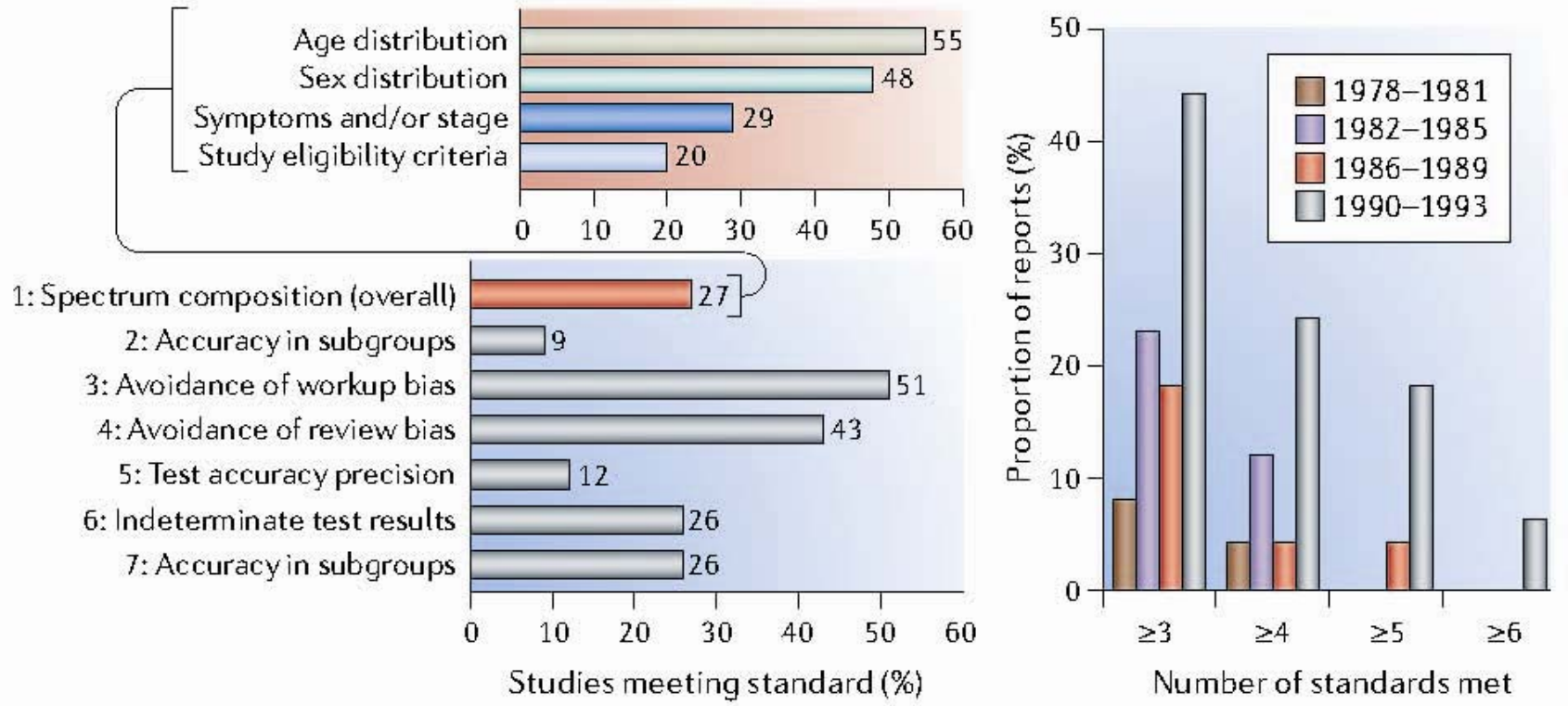


Figure 4 | **Proportion of diagnostic evaluations meeting accepted standards.** The seven standards are shown on the left. The data are taken from REF. 10.

Peeling et al. Nature Rev Micro 2006 [data from Reid et al.JAMA 1995]

McGill

Poor reporting: example from TB literature

12 meta-analysis with over 500 diagnostic studies

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

Table 2. Methodological quality of studies on tuberculosis diagnostics in recently published meta-analyses.

Meta-analysis	No. of studies	Diagnostic test	Average size of each study	Prospective data collection (%)	Consecutive or random sampling of subjects (%)	Cross-sectional design (%)	Blinded interpretation of test results* (%)	Complete verification of index test results† (%)	Ref.
Sarmiento et al. (2003)	16	PCR on respiratory specimens for smear-negative pulmonary TB	NR	50	NR	NR	63	100	[12]
Goto et al. (2003)	40	ADA for TB pleural effusion	137	NR	NR	NR	0	NR	[13]
Pai et al. (2003)	49	NAT for TB meningitis	42	61	49	61	59	94	[14]
Greco et al. (2003)	44	ADA and IFN-γ tests for TB pleural effusion	135	NR	NR	NR	9	NR	[15]
Pai et al. (2004)	40	NAT for TB pleural effusion	60	63	53	70	55	100	[16]
Flores et al. (2005)	84	In-house PCR for pulmonary TB	149	NR	NR	71	34	NR	[17]
Kalantri et al. (2005)	13	Phage amplification tests for pulmonary TB	448	NR	NR	85	23	100	[18]
Pai et al. (2005)	21	Phage-based tests for rifampin resistance	85	NR	38	NR	57	100	[19]
Morgan et al. (2005)	15	Line probe assay for rifampin resistance	91	NR	0	NR	13	100	[20]
Greco et al. (2006)	63	Commercial NAT for pulmonary TB	410	16	32	NR	16	NR	[21]
Steingart et al. (2006)	45	Fluorescence versus conventional sputum smear microscopy for pulmonary TB	493	100	36	NR	49	NR	[22]
Steingart et al. (2006)	83	Direct versus concentrated sputum smear microscopy for pulmonary TB	256	100	21	NR	31	NR	[23]

*At least single blind. †By reference standard.
ADA: Adenosine deaminase; IFN: Interferon; NAT: Nucleic acid amplification test; NR: Not reported; TB: Tuberculosis.

Pai M, et al. Exp Rev Mol Diagn 2006.

McGill

CLINICAL AND VACCINE IMMUNOLOGY, Feb. 2009, p. 260–276 Vol. 16, No. 2
1556-6811/09/$08.00+0 doi:10.1128/CVI.00355-08

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]

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.

Steingart et al. CVI 2009

McGill

Study quality vs. study reporting

Data from a meta-analysis of NAAT for TB meningitis (Pai et al. *Lancet Infect Dis* 2003)

Characteristic	Before contact % [N = 49]	After contact % [N = 49]
Blinding		
Double blind	12	35
Single blind	14	24
Unblinded	0	10
Not reported	74	31
Sampling		
Consecutive/random	18	49
Not consecutive/random	6	20
Not reported	76	31
Data collection		
Prospective	51	61
Retrospective	0	4
Both	2	10
Not reported	47	25

Pai M, et al. Quality assessment in meta-analyses of diagnostic studies: what difference does email contact with authors make? Abstract: XI Cochrane Colloquium, Barcelona, Spain, 26 – 31, October 2003.

McGill

What can be done to improve quality and reporting of diagnostic studies?

- Report better using standardized reporting formats (e.g. STARD)
- Improve study design using guidelines specific for diagnostic trials
  - QUADAS
  - DEEP
- Use GCP, GLP and GCLP to upgrade overall research standards
- Strengthen lab capacity and research capacity in developing countries

McGill

Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative

Patrick M. Bossuyt, Johannes B. Reitsma, David E. Bruns, Constantine A. Gatsonis, Paul P. Glasziou, Les M. Irwig, Jeroen G. Lijmer, David Moher, Drummond Rennie, and Henrica C.W. de Vet, for the STARD Group*

Background: To comprehend the results of diagnostic accuracy studies, readers must understand the design, conduct, analysis, and results of such studies. That goal can be achieved only through complete transparency from authors.

Objective: To improve the accuracy and completeness of reporting of studies of diagnostic accuracy in order to allow readers to assess the potential for bias in the study and to evaluate its generalizability.

Methods: The Standards for Reporting of Diagnostic Accuracy (STARD) steering committee searched the literature to identify publications on the appropriate conduct and reporting of diagnostic studies and extracted potential items into an extensive list. Researchers, editors, methodologists and statisticians, and members of professional organizations shortened this list during a 2-day consensus meeting with the goal of developing a checklist and a generic flow diagram for studies of diagnostic accuracy.

Results: The search for published guidelines on diagnostic research yielded 33 previously published checklists, from which we extracted a list of 75 potential items. The consensus meeting shortened the list to 25 items, using evidence on bias whenever available. A prototypical flow diagram provides information about the method of patient recruitment, the order of test execution, and the numbers of patients undergoing the test under evaluation, the reference standard, or both.

Conclusions: Evaluation of research depends on complete and accurate reporting. If medical journals adopt the checklist and the flow diagram, the quality of reporting of studies of diagnostic accuracy should improve to the advantage of the clinicians, researchers, reviewers, journals, and the public.

Ann Intern Med. 2003;138:40-44. www.annals.org

For author affiliations, see end of text.

*For members of the STARD Group, see Appendix.

See related article, available only at www.annals.org.

McGill

STARD reporting standards

STARD checklist

APPENDIX 1 | STANDARDS FOR REPORTING OF DIAGNOSTIC ACCURACY (STARD) CHECKLIST

Section and topic	Item #		On page #
Title/abstract/ keywords	1	Identify the article as a study of diagnostic accuracy (recommended MeSH heading 'sensitivity and specificity').	❑
Introduction	2	State the research questions or study aims, such as estimating the diagnostic accuracy or comparing accuracy between tests or across participant groups.	❑
Methods		Describe:	
Participants	3	The study population: the inclusion and exclusion criteria, the setting and the locations where the data were collected.	❑
	4	Participant recruitment: was the recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	❑
	5	Participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.	❑
	6	Data collection: was data collection planned before (prospective study) or after (retrospective study) the index test and reference standard were performed?	❑
Test methods	7	The reference standard and its rationale.	❑
	8	Technical specifications of the material and methods involved, including how and when the measurements were taken, and/or cite references for the index tests and reference standard.	❑
	9	Definition of, and rationale for, the units, cut offs and/or categories of the results of the index tests and the reference standard.	❑
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.	❑
	11	Whether or not the readers of the index tests and reference standard were blind to the results of the other test and describe any other clinical information available to the readers.	❑
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).	❑
	13	Methods for calculating test reproducibility, if done.	❑
Results		Report:	❑
Participants	14	When the study was done, including the start and end dates of recruitment.	❑
	15	The clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co-morbidity, current treatments and recruitment centres).	❑
	16	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).	❑
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered inbetween.	❑
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	❑
	19*	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard.	❑
	20	Any adverse events from performing the index tests or the reference standard.	❑
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).	❑
	22	How indeterminate results, missing responses and outliers of the index tests were handled.	❑
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done.	❑
	24	Estimates of test reproducibility, if done.	❑
	25	Discuss the clinical applicability of the study findings.	❑

* This entry has been modified from the original.

McGill

STARD reporting standards

Table. STARD Checklist for the Reporting of Studies of Diagnostic Accuracy*

Section and Topic	Item #		On page #
TITLE/ABSTRACT/KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	

McGill

STARD reporting standards

METHODS		Describe	
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	
	4	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?	
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	The reference standard and its rationale.	
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	
	9	Definition of and rationale for the units, cutoffs, and/or categories of the results of the index tests and the reference standard.	
	10	The number, training, and expertise of the persons executing and reading the index tests and the reference standard.	
	11	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.	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% confidence intervals).	
	13	Methods for calculating test reproducibility, if done.	

McGill

STARD reporting standards

RESULTS		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment.	
	15	Clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).	
	16	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).	
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.	
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	
	19	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.	
	20	Any adverse events from performing the index tests or the reference standard.	
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).	
	22	How indeterminate results, missing responses, and outliers of the index tests were handled.	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.	
	24	Estimates of test reproducibility, if done.	
DISCUSSION	25	Discuss the clinical applicability of the study findings.	

* MeSH = Medical Subject Heading; STARD = Standards for Reporting of Diagnostic Accuracy.

McGill

STARD Explanation and elaboration

ACADEMIA AND CLINIC

The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and Elaboration

Patrick M. Bossuyt, Johannes B. Reitsma, David E. Bruns, Constantine A. Gatsonis, Paul P. Glasziou, Les M. Irwig, David Moher, Drummond Rennie, Henrica C.W. de Vet, and Jeroen G. Lijmer

The quality of reporting of studies of diagnostic accuracy is less than optimal. Complete and accurate reporting is necessary to enable readers to assess the potential for bias in the study and to evaluate the generalizability of the results.

A group of scientists and editors has developed the STARD (Standards for Reporting of Diagnostic Accuracy) statement to improve the reporting the quality of reporting of studies of diagnostic accuracy. The statement consists of a checklist of 25 items and flow diagram that authors can use to ensure that all relevant information is present.

This explanatory document aims to facilitate the use, understanding, and dissemination of the checklist. The document contains a clarification of the meaning, rationale, and optimal use of each item on the checklist, as well as a short summary of the available evidence on bias and applicability.

The STARD statement, checklist, flowchart, and this explanation and elaboration document should be useful resources to improve reporting of diagnostic accuracy studies. Complete and informative reporting can only lead to better decisions in health care.

Ann Intern Med. 2003;138:W1-W12. www.annals.org

For author affiliations, see end of text.

See related article on pp 40-44.

McGill

STARD reporting standards

Item 1. Identify the Article as a Study of Diagnostic Accuracy (Recommend MeSH Heading "Sensitivity and Specificity")

Example (an Excerpt from a Structured Abstract)

Purpose: To determine the sensitivity and specificity of computed tomographic colonography for colorectal polyp and cancer detection by using colonoscopy as the reference standard (14).

Electronic databases have become indispensable tools to identify studies. To facilitate retrieval of their study, authors should explicitly identify it as a report of a study of diagnostic accuracy. We recommend the use of the term "diagnostic accuracy" in the title or abstract of a report that compares the results of one or more index tests with the results of a reference standard. In 1991 the National Library of Medicine's MEDLINE database introduced a specific keyword (MeSH heading) for diagnostic studies: "Sensitivity and Specificity." Using this keyword to search for studies of diagnostic accuracy remains problematic (15–19). In a selected set of MEDLINE journals covering publications between 1992 through 1995, the use of the MeSH heading "Sensitivity and Specificity" identified only 51% of all studies of diagnostic accuracy and incorrectly identified many articles that were not reports of studies on diagnostic accuracy (18).

In the example, the authors used the more general term "Performance Characteristics of CT Colonography" in the title. The purpose section of the structured abstract explicitly mentions sensitivity and specificity. The MEDLINE record for this paper contains the MeSH "Sensitivity and Specificity."

McGill

STARD reporting standards

Item 2. State the Research Questions or Study Aims, Such as Estimating Diagnostic Accuracy or Comparing Accuracy between Tests or across Participant Groups

Example

Invasive x-ray coronary angiography remains the gold standard for the identification of clinically significant coronary artery disease. . . . A noninvasive test would be desirable. Coronary magnetic resonance angiography performed while the patient is breathing freely has reached sufficient technical maturity to allow more widespread application with a standardized protocol. Therefore, we conducted a study to determine the [accuracy] *of coronary magnetic resonance angiography in the diagnosis of native-vessel coronary artery disease* (20).

The Helsinki Declaration states that biomedical research involving people should be based on a thorough knowledge of the scientific literature (21). In the introduction of scientific reports authors describe the scientific background, previous work on the subject, the remaining uncertainty, and, hence, the rationale for their study.

Clearly specified research questions help the readers to judge the appropriateness of the study design and data analysis. A single general description, such as "diagnostic value" or "clinical usefulness," is usually not very helpful to the readers.

In the example, the authors use the introduction section of their paper to describe the potential of coronary magnetic resonance angiography as a non-invasive alternative to conventional x-ray angiography in the diagnosis of clinically significant coronary stenosis. This description helps the reader to judge the appropriateness of the selection criteria, the choice of the reference standard, and the statistical methods used to summarize and analyze the data.

McGill

STARD reporting standards

Item 3. Describe the Study Population: The Inclusion and Exclusion Criteria, Setting and Locations Where Data Were Collected

Example

Patient population. Female patients attending participating family planning clinics in the states of Washington and Oregon during 1992 and 1993 were considered for enrollment in the study. The previously published screening criteria of the Region X Chlamydia Project were used to establish eligibility for enrollment.[ref] These criteria included any of the following: (i) mucopurulent cervicitis, pelvic inflammatory disease, friable cervix, or abnormal bleeding; (ii) a partner with signs and/or symptoms suggestive of urethritis; (iii) client request; (iv) rape within the previous 60 days; (v) candidacy for intrauterine device insertion; and (vi) a positive pregnancy test and a bimanual pelvic examination. Alternatively, the criteria included two or more of the following: (i) age under 24 years and being sexually active; (ii) new sex partner in the previous 60 days; (iii) sex partner with multiple partners in the previous 30 days; (iv) multiple sex partners in the previous 30 days; and (v) use of nonbarrier birth control method or no birth control method (nonbarrier birth control methods include oral contraceptives, the intrauterine device, sterilization, and all natural family planning methods) (22).

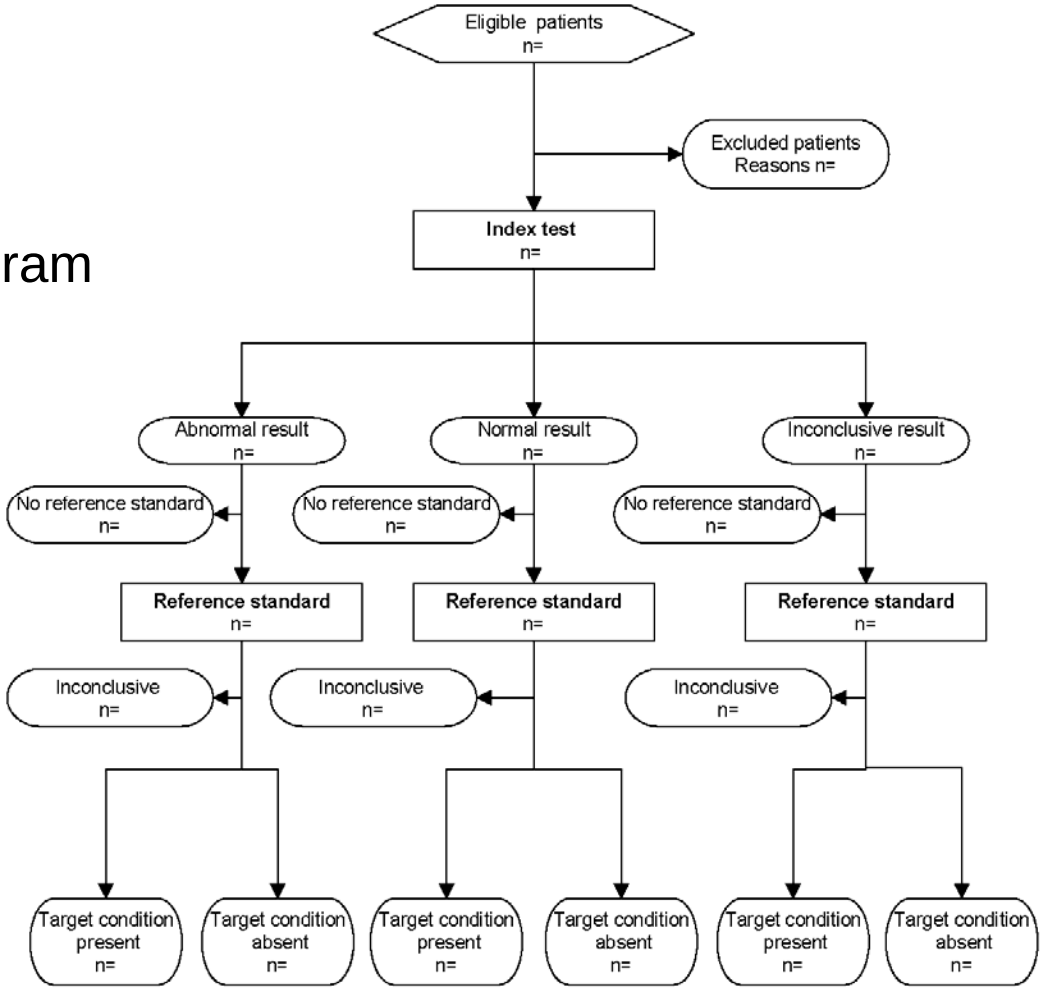
Since diagnostic accuracy describes the behavior of a test under particular circumstances, a report of the study must also include a helpful description of the targeted population. The eligibility criteria describe the targeted patient population, including additional exclusion criteria used for reasons of safety or feasibility.

Readers must know whether or not the study excluded patients with a specific condition known to adversely affect the way the test works, which would inflate diagnostic accuracy (limited challenge bias) (23). Examples are the exclusion of patients using beta-blockers in studies of exercise electrocardiography and the exclusion of patients with preexisting pulmonary diseases in studies of ventilation-perfusion scintigraphy (24, 25).

McGill

STARD reporting standards

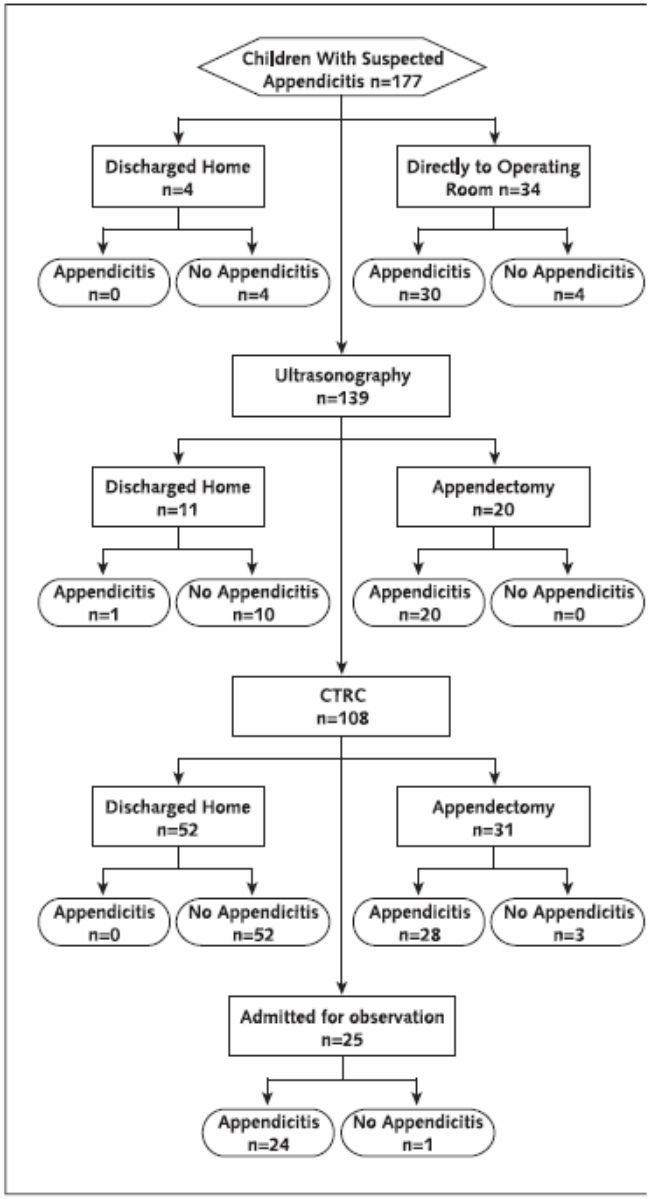
STARD flow diagram



Bossuyt PM, Reitsma JB, Bruns DE, et al. *Clin Chem* 2003;**49**(1)**:**1-6.

McGill

Figure 1. **Example of a flow diagram of a diagnostic accuracy study (55).**



McGill

QUADAS tool for quality assessment of diagnostic studies

BMC Medical Research Methodology

BioMed Central

Research article

Open Access

The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews

Penny Whiting*[1], Anne WS Rutjes[2], Johannes B Reitsma[2], Patrick MM Bossuyt[2] and Jos Kleijnen[1]

Although designed for quality assessment in systematic reviews, it can be used to improve study design

Whiting P et al. BMC Med Res Meth 2003

McGill

QUADAS tool for quality assessment of diagnostic studies

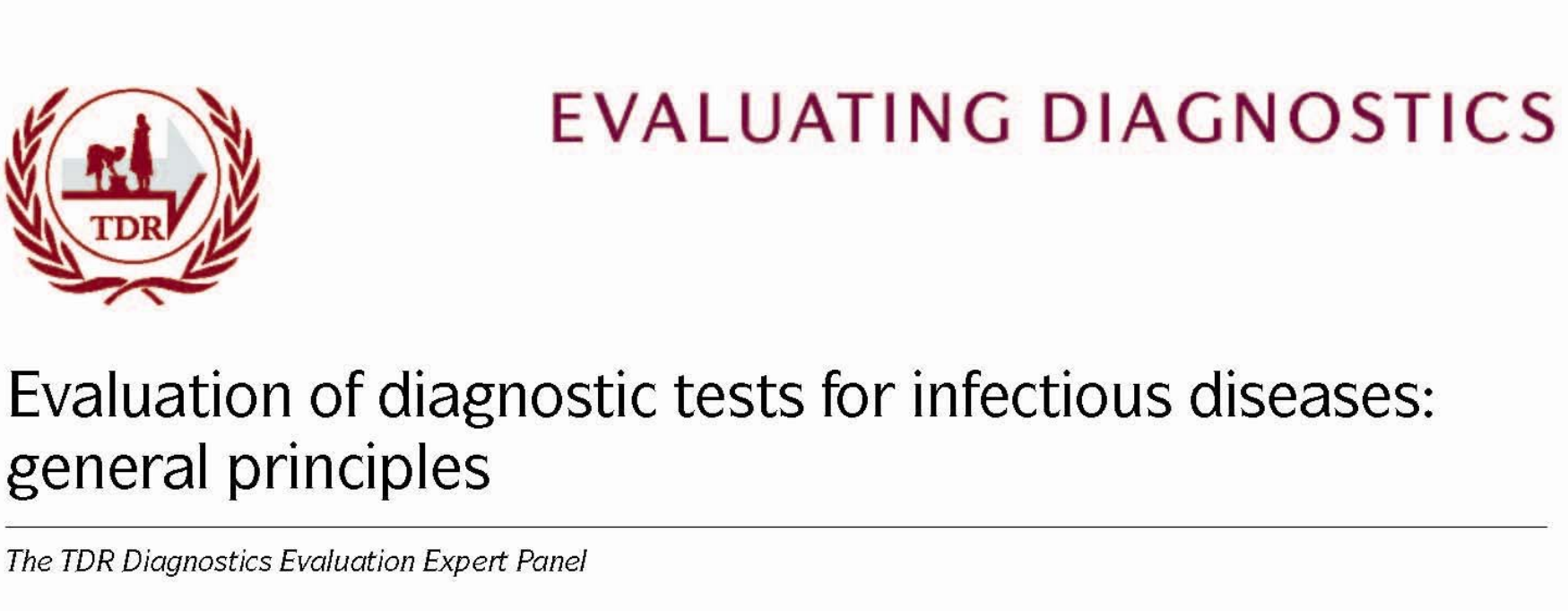
Table 1: QUADAS

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

Whiting P et al. BMC Med Res Meth 2003

McGill

TDR/WHO Diagnostics Evaluation Expert Panel (DEEP) guidelines



Nature Reviews Microbiology September 2006

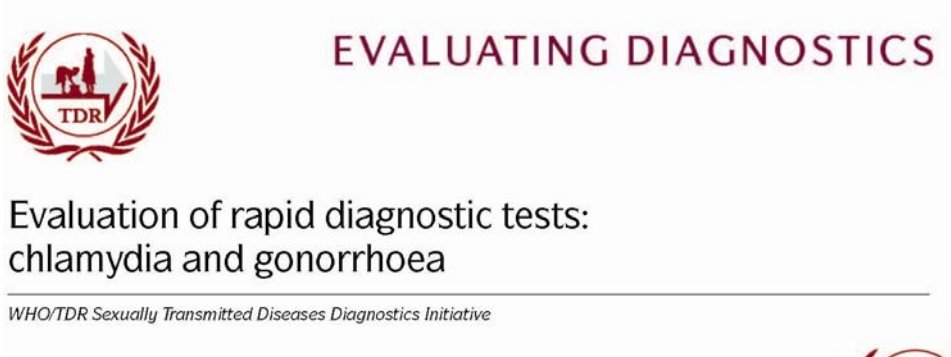
McGill

DEEP guidelines for specific infectious diseases

Nature Reviews Microbiology September 2006

McGill

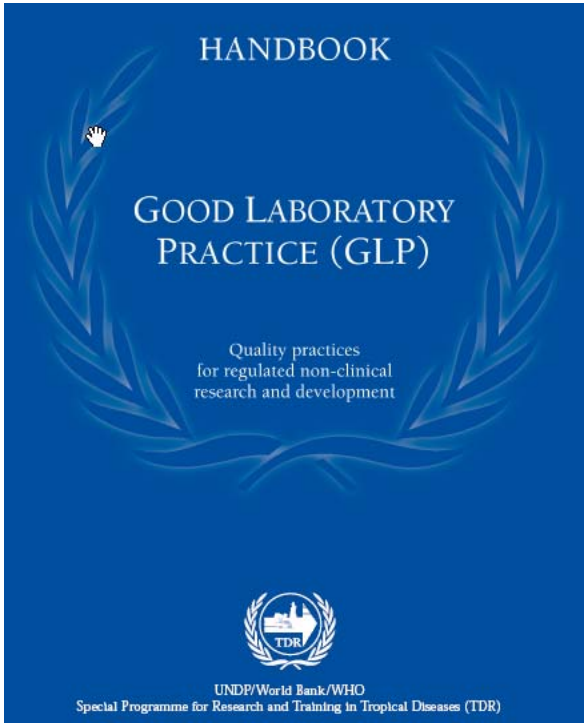
DEEP guidelines for specific infectious diseases



McGill

Improve overall research standards

Good Laboratory Practice (GLP) & Good Clinical Lab Practice (GCLP)



GCLP is increasingly being adopted as the laboratory standard of choice for clinical and diagnostic trials

McGill