Toward Complete and Accurate Reporting of Studies of Diagnostic Accuracy

The STARD Initiative

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

Key Words: Sensitivity and specificity; Diagnosis; Accuracy; Internal validity; Bias; External validity; Checklist; Guideline; Reporting

DOI: 10.1092/8EXCCM6YR1THUBAF

Abstract

Our objective was to improve the accuracy and completeness of reporting of studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study, and to evaluate its generalizability.

The Standards for Reporting of Diagnostic Accuracy 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, 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.

The search for published guidelines regarding diagnostic research yielded 33 previously published checklists, from which we extracted a list of 75 potential items. At the consensus meeting, participants shortened the list to a 25-item checklist, using evidence 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.

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 clinicians, researchers, reviewers, journals, and the public.

The world of diagnostic tests is highly dynamic. New tests are developed at a fast rate, and the technology of existing tests is continuously being improved. Exaggerated and biased results from poorly designed and reported diagnostic studies can trigger their premature dissemination and lead physicians into making incorrect treatment decisions. A rigorous evaluation process of diagnostic tests before introduction into clinical practice could not only reduce the number of unwanted clinical consequences related to misleading estimates of test accuracy but also limit health care costs by preventing unnecessary testing. Studies to determine the diagnostic accuracy of a test are a vital part of this evaluation process.[1-3]

In studies of diagnostic accuracy, the outcomes from one or more tests under evaluation are compared with outcomes from the reference standard, both measured in subjects who are suspected of having the condition of interest. The term *test* refers to any method for obtaining additional information on a patient's health status. It includes information from history and physical examination, laboratory tests, imaging tests, function tests, and histopathology. The condition of interest or target condition can refer to a particular disease or to any other identifiable condition that may prompt clinical actions, such as further diagnostic testing, or the initiation, modification, or termination of treatment. In this framework, the *reference standard* is considered to be the best available method for establishing the presence or absence of the condition of interest. The reference standard can be a single method or a combination of methods to establish the presence of the target condition. It can include laboratory tests, imaging tests, and pathology, but also dedicated clinical follow-up of subjects. The term *accuracy* refers to the amount of agreement between the information from the test under evaluation, referred to as

the *index test*, and the reference standard. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver operating characteristic curve.[4-6]

There are several potential threats to the internal and external validity of a study on diagnostic accuracy. A survey of studies of diagnostic accuracy published in 4 major medical journals between 1978 and 1993 revealed that the methodological quality was mediocre at best.[7] However, evaluations were hampered because many reports lacked information on key elements of design, conduct, and analysis of diagnostic studies.[7] The absence of critical information about the design and conduct of diagnostic studies has been confirmed by authors of meta-analyses.[8,9] As in any other type of research, flaws in study design can lead to biased results. One report showed that diagnostic studies with specific design features are associated with biased, optimistic estimates of diagnostic accuracy compared with studies without such deficiencies.[10]

At the 1999 Cochrane Colloquium meeting in Rome, the Cochrane Diagnostic and Screening Test Methods Working Group discussed the low methodological quality and substandard reporting of diagnostic test evaluations. The Working Group thought that the first step to correct these problems was to improve the quality of reporting of diagnostic studies. Following the successful CONSORT (Consolidated Standards of Reporting Trials) initiative,[11-13] the Working Group aimed at the development of a checklist of items that should be included in the report of a study on diagnostic accuracy.

The objective of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative is to improve the quality of reporting of studies of diagnostic accuracy. Complete and accurate reporting allows the reader to detect the potential for bias in the study (internal validity) and to assess the generalizability and applicability of the results (external validity).

Materials and Methods

The STARD Steering Committee **Appendix 1** started with an extensive search to identify publications on the conduct and reporting of diagnostic studies. This search included MEDLINE, Embase, BIOSIS, and the methodological database from the Cochrane Collaboration up to July 2000. In addition, the steering committee members examined reference lists of retrieved articles, searched personal files, and contacted other experts in the field of diagnostic research. They reviewed all relevant publications and extracted an extended list of potential checklist items.

Subsequently, the STARD Steering Committee convened a 2-day consensus meeting for invited experts from the following interest groups: researchers, editors, methodologists, and professional organizations. The aims of the conference were to reduce the extended list of potential items, where appropriate, and to discuss the optimal format and phrasing of the checklist. The selection of items to retain was based on evidence whenever possible.

The meeting format consisted of a mixture of small group and plenary sessions. Each small group focused on a group of related items on the list. The suggestions of the small groups were then discussed in plenary sessions. Overnight, a first draft of the STARD checklist was assembled based on the suggestions from the small groups and the additional remarks from the plenary sessions. All meeting attendees discussed this version the next day and made additional changes. The members of the STARD Group could suggest further changes through a later round of comments by electronic mail.

Potential users field-tested the conference version of the checklist and flow diagram, and additional comments were collected. This version was placed on the CONSORT Web site with a call for comments. The STARD Steering Committee discussed all comments and assembled the final checklist.

Results

The search for published guidelines for diagnostic research yielded 33 lists. Based on these published guidelines and on input from steering committee and STARD Group members, the steering committee assembled a list of 75 items. During the consensus meeting on September 16 and 17, 2000, participants consolidated and eliminated items to form the 25-item checklist. Conference members made major revisions to the phrasing and format of the checklist.

The STARD Group received valuable comments and remarks during the various stages of evaluation after the conference, which resulted in the version of the STARD checklist that appears in **Table 1**.

The flow diagram provides information about the method of patient recruitment (eg, based on a consecutive series of patients with specific symptoms, case-control), the order of test execution, and the number of patients undergoing the test under evaluation (index test) and the reference test **Figure 1**. We provide 1 prototypical flow diagram that reflects the most commonly used design in diagnostic research. Examples that reflect other designs are on the STARD Web site (www.consort-statement.org\stardstatement.htm).

Discussion

The purpose of the STARD initiative is to improve the quality of the reporting of diagnostic studies. The items in the checklist and the flow diagram can help authors describe essential elements of the design and conduct of the study, the execution of tests, and the results.

DOI: 10.1092/8EXCCM6YR1THUBAF

Table 1
STARD Checklist for Reporting Diagnostic Accuracy Studies

Section and Topic	Item No.	Comments	On Page No.*
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	
Methods		Describe	
Participants	3	The study population: the inclusion and exclusion criteria, setting and locations where 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 items 3 and 4? If not, specify how patients 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 materials and methods involved, including how and when measurements were taken, and/or cite references for index tests and the 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	Were the readers of the index tests and the reference standard blind (masked) to the results of the other test? Describe any other clinical information available to readers	
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (eg, 95% confidence intervals)	
	13	Methods for calculating test reproducibility, if done	
Results		Report	
Participants	14	When study was done, including beginning and ending dates of recruitment	
	15	Clinical and demographic characteristics of the study population (eg, age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers)	
	16	The number of participants satisfying the criteria for inclusion who 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 (eg, 95% confidence intervals)	
	22	How indeterminate results, missing responses, and outliers of index tests were handled	
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers, or centers, if done	
	24	Measures of test reproducibility, if done	
Discussion	25	Discuss the clinical applicability of the study findings	

STARD, Standards for Reporting of Diagnostic Accuracy.
* Insert the applicable manuscript page number.

We arranged the items under the usual headings of a medical research article, but this is not intended to dictate the order in which they have to appear within an article.

The guiding principle in the development of the STARD checklist was to select items that would help readers to judge the potential for bias in the study and to appraise the applicability of the findings. Two other general considerations shaped the content and format of the checklist. First, the STARD Group believes that one general checklist for studies of diagnostic accuracy, rather than different checklists for each field, is likely to be more widely disseminated and perhaps accepted by authors, peer reviewers, and journal editors. Although the evaluation of an imaging test differs from that of a test in the laboratory, we thought that these differences were more of degree than of kind. The second consideration was the development of a checklist specifically aimed at studies of diagnostic accuracy. We did not include general issues in the reporting of research findings, like the recommendations contained in the uniform requirements for manuscripts submitted to biomedical journals.[14]

Wherever possible, the STARD Group based the decision to include an item on evidence linking the item to biased estimates (internal validity) or to variation in measures of diagnostic accuracy (external validity). The evidence varied from narrative articles explaining theoretic principles and papers presenting results from statistical modeling to empiric evidence derived from diagnostic studies. For several items, the evidence is rather limited.

DOI: 10.1092/8EXCCM6YR1THUBAF

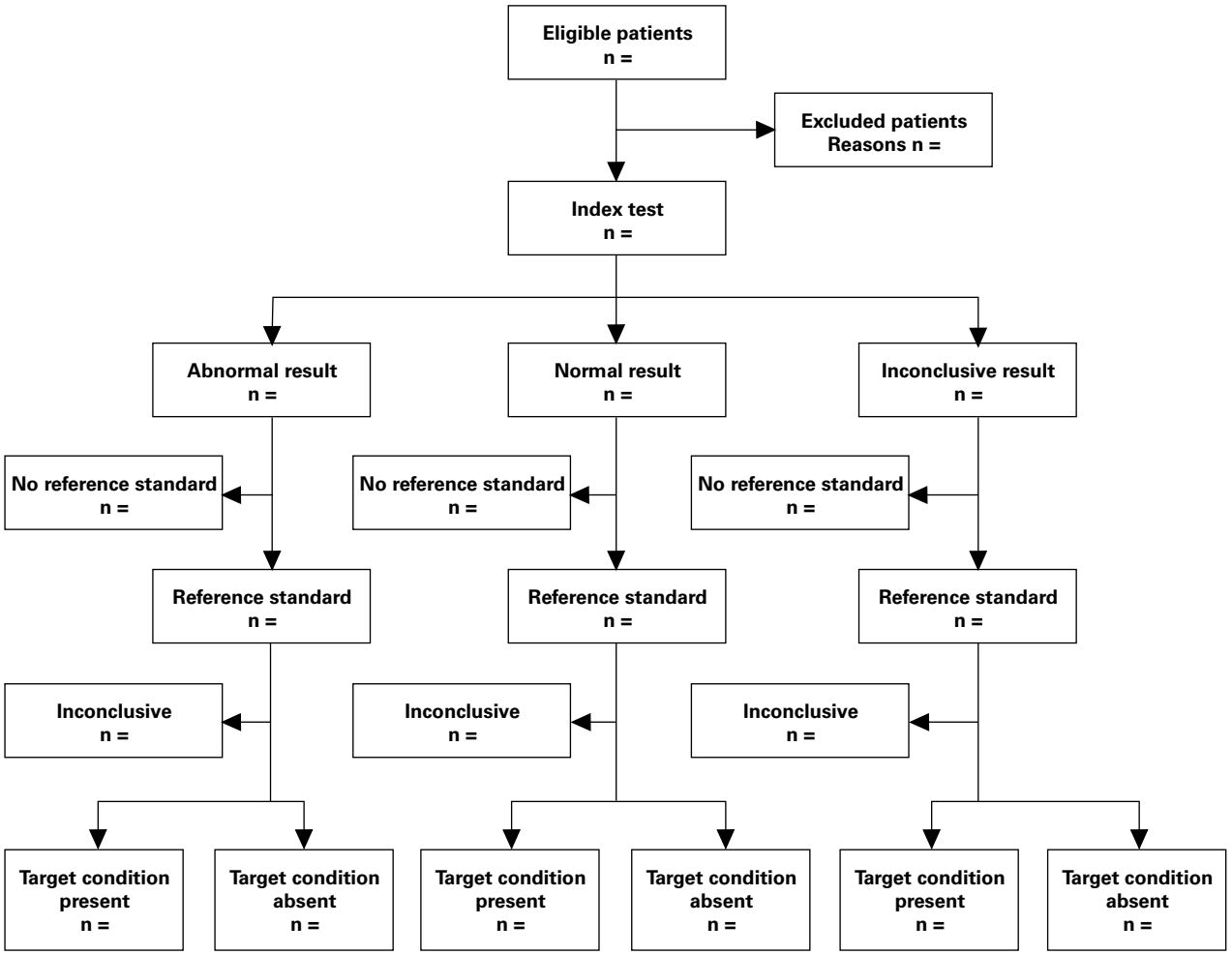

Figure 1 Prototypical flow diagram of a study on diagnostic accuracy.

A separate background document explains the meaning and rationale of each item and briefly summarizes the type and amount of evidence.[15] This background document should enhance the use, understanding, and dissemination of the STARD checklist.

The STARD Group put considerable effort into the development of a flow diagram for diagnostic studies. A flow diagram has the potential to communicate vital information about the design of a study and the flow of participants in a transparent manner.[16] A comparable flow diagram has become an essential element in the CONSORT standards for reporting of randomized trials. The flow diagram could be even more essential in diagnostic studies, given the variety of designs used in diagnostic research. Flow diagrams in the reports of diagnostic accuracy studies indicate the process of sampling and selecting participants (external validity), the flow of participants in relation to the timing and outcomes of tests as a transparent method, the number of subjects who fail to receive either the index test and/or the reference standard (potential for verification bias[17-19]), and the number of patients at each stage of the study, thus providing the correct denominator for proportions (internal consistency).

The STARD Group plans to measure the impact of the statement on the quality of published reports on diagnostic accuracy using a before-and-after evaluation.[13] Updates of STARD will be provided when new evidence on sources of bias or variability becomes available. We welcome any comments, whether on content or form, to improve the current version.

Address reprint requests to Drs Bossuyt and Reitsma: Dept of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, the Netherlands.

** For a list of members of the STARD Steering Committee and the STARD Group, see Appendix 1.*

Supported in part by the Dutch Health Care Insurance Board, Amstelveen, the Netherlands; the International Federation of Clinical Chemistry, Milano, Italy; the Medical Research Council's Health Services Research Collaboration, Bristol, England; and the Academic Medical Center, Amsterdam, the Netherlands.

DOI: 10.1092/8EXCCM6YR1THUBAF

Appendix 1

Members of the STARD Steering Committee

Patrick Bossuyt, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands
David Bruns, *Clinical Chemistry*, Charlottesville, NC
Constantine Gatsonis, Brown University, Center for Statistical Sciences, Providence, RI
Paul Glasziou, Mayne Medical School, Department of Social & Preventive Medicine, Herston, Australia
Les Irwig, University of Sydney, Department of Public Health & Community Medicine, Sydney, Australia
Jeroen Lijmer, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands
David Moher, Chalmers Research Group, Ottawa, Canada
Drummond Rennie, *Journal of the American Medical Association*, Jacksonville, FL
Riekie de Vet, Free University, Institute for Research in Extramural Medicine, Amsterdam, the Netherlands

Members of the STARD Group

Doug Altman, Institute of Health Sciences, Centre for Statistics in Medicine, Oxford, England
Stuart Barton, *British Medical Journal*, BMA House, London, England
Colin Begg, Memorial Sloan-Kettering Cancer Center, Department of Epidemiology & Biostatistics, New York, NY
William Black, Dartmouth Hitchcock Medical Center, Department of Radiology, Lebanon, NH
Harry Büller, Academic Medical Center, Department of Vascular Medicine, Amsterdam, the Netherlands
Gregory Campbell, Center for Devices and Radiological Health, US Food and Drug Administration, Rockville, MD
Frank Davidoff, *Annals of Internal Medicine*, Philadelphia, PA
Jon Deeks, Institute of Health Sciences, Centre for Statistics in Medicine, Oxford, England
Paul Dieppe, Department of Social Medicine, University of Bristol, Bristol, England
Kenneth Fleming, John Radcliffe Hospital, Oxford, England
Rijk van Ginkel, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands
Afina Glas, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands
Gordon Guyatt, McMaster University, Clinical Epidemiology and Biostatistics, Hamilton, Canada
James Hanley, McGill University, Department of Epidemiology & Biostatistics, Montreal, Canada
Richard Horton, *The Lancet*, London, England
Myriam Hunink, Erasmus Medical Center, Department of Epidemiology & Biostatistics, Rotterdam, the Netherlands
Jos Kleijnen, NHS Centre for Reviews and Dissemination, York, England
Andre Knottnerus, Maastricht University, Netherlands School of Primary Care Research, Maastricht, the Netherlands
Erik Magid, Amager Hospital, Department of Clinical Biochemistry, Copenhagen, Denmark
Barbara McNeil, Harvard Medical School, Department of Health Care Policy, Boston, MA
Matthew McQueen, Hamilton Civic Hospitals, Department of Laboratory Medicine, Hamilton, Canada
Andrew Onderdonk, Channing Laboratory, Boston, MA
John Overbeke, *Nederlands Tijdschrift voor Geneeskunde*, Amsterdam, the Netherlands
Christopher Price, St Bartholomew's-Royal London School of Medicine and Dentistry, London, England
Anthony Proto, *Radiology* Editorial Office, Richmond, VA
Hans Reitsma, Academic Medical Center, Department of Clinical Epidemiology, Amsterdam, the Netherlands
David Sackett, Trout Research and Education Centre, Irish Lake, Canada
Gerard Sanders, Academic Medical Center, Department of Clinical Chemistry, Amsterdam, the Netherlands
Harold Sox, *Annals of Internal Medicine*, Philadelphia, PA
Sharon Straus, Mt Sinai Hospital, Toronto, Canada
Stephan Walter, McMaster University, Clinical Epidemiology and Biostatistics, Hamilton, Canada

STARD, Standards for Reporting of Diagnostic Accuracy.

Acknowledgment: This initiative was supported by a large number of people around the globe who commented on earlier versions.

References

1. Guyatt GH, Tugwell PX, Feeny DH, et al. A framework for clinical evaluation of diagnostic technologies. *CMAJ*. 1986;134:587-594.
2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making*. 1991;11:88-94.
3. Kent DL, Larson EB. Disease, level of impact, and quality of research methods: three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. *Invest Radiol*. 1992;27:245-254.
4. Griner PF, Mayewski RJ, Mushlin AI, et al. Selection and interpretation of diagnostic tests and procedures: principles and applications. *Ann Intern Med*. 1981;94:557-592.
5. Sackett DL, Haynes RB, Guyatt GH, et al. The selection of diagnostic tests. In: Sackett D, ed. *Clinical Epidemiology*. 2nd ed. Boston, MA: Little, Brown; 1991:47-57.
6. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978;8:283-298.
7. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research: getting better but still not good. *JAMA*. 1995;274:645-651.
8. Nelemans PJ, Leiner T, de Vet HCW, et al. Peripheral arterial disease: meta-analysis of the diagnostic performance of MR angiography. *Radiology*. 2000;217:105-114.
9. Devries SO, Hunink MGM, Polak JF. Summary receiver operating characteristic curves as a technique for meta-analysis of the diagnostic performance of duplex ultrasonography in peripheral arterial disease. *Acad Radiol*. 1996;3:361-369.
10. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*. 1999;282:1061-1066.
11. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA*. 1996;276:637-639.
12. Moher D, Schulz KF, Altman D. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *JAMA*. 2001;285:1987-1991.
13. Moher D, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA*. 2001;285:1992-1995.
14. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *JAMA*. 1997;277:927-934. Also available at: ACP Online, http://www.acponline.org. Accessed November 12, 2002.
15. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Clin Chem*. 2003;49:7-18.
16. Egger M, Jüni, Barlett C. Value of flow diagrams in reports of randomized controlled trials. *JAMA*. 2001;285:1996-1999.
17. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. *Med Decis Making*. 1987;7:139-148.
18. Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. *Med Decis Making*. 1987;7:115-119.
19. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411-423.

DOI: 10.1092/8EXCCM6YR1THUBAF

HOLOGIC®

First and Only FDA Cleared
Digital Cytology System

Genius™ Cervical AI

Genius™ Review Station

Genius™ Digital Imager

Empower Your Genius With Ours

Make a Greater Impact on Cervical Cancer
with the Advanced Technology of the
Genius™ Digital Diagnostics System

Click or Scan
to discover more

genius™
DIGITAL DIAGNOSTICS