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

PATRICK M. BOSSUYT,[1*] JOHANNES B. REITSMA,[1] DAVID E. BRUNS,[2,3]
CONSTANTINE A. GATSONIS,[4] PAUL P. GLASZIOU,[5] LES M. IRWIG,[6] JEROEN G. LIJMER,[1]
DAVID MOHER,[7] DRUMMOND RENNIE,[8,9] and HENRICA C.W. DE VET,[10] 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 to allow readers to assess the potential for bias in the study and to evaluate its generalisability.

Methods: The *Sta*ndards for *R*eporting of *D*iagnostic 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, and members of professional organisations shortened this list during a two-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 clinicians, researchers, reviewers, journals, and the public.

[1] Department of Clinical Epidemiology and Biostatistics, Academic Medical Center—University of Amsterdam, 1100 DE Amsterdam, The Netherlands.

[2] Department of Pathology, University of Virginia, Charlottesville, VA 22903.

[3] *Clinical Chemistry*, Washington, DC 20037.

[4] Centre for Statistical Sciences, Brown University, Providence, RI 02912.

[5] Centre for General Practice, University of Queensland, Herston QLD 4006, Australia.

[6] Department of Public Health & Community Medicine, University of Sydney, Sydney NSW 2006, Australia.

[7] Chalmers Research Group, Ottowa, Ontario, K1N 6M4 Canada.

[8] Institute for Health Policy Studies, University of California, San Francisco, San Francisco, CA 94118.

[9] *Journal of the American Medical Association*, Chicago, IL 60610.

[10] Institute for Research in Extramural Medicine, Free University, 1081 BT Amsterdam, The Netherlands.

*Address correspondence to this author at: Department of Clinical Epidemiology and Biostatistics, Academic Medical Center—University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands. Fax 31-20-6912683; e-mail Stard@amc.uva.nl.

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 healthcare costs by preventing unnecessary testing. Studies to determine the diagnostic accuracy of a test are a vital part in 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, 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 operator characteristic (ROC) curve *(4–6)*.

There are several potential threats to the internal and external validity of a study of diagnostic accuracy. A survey of studies of diagnostic accuracy published in four 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 metaanalyses *(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 to 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 felt that the first step to correct these problems was to improve the quality of reporting of diagnostic studies. Following the successful CONSORT 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 of 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 generalisability and applicability of the results (external validity).

Materials and Methods

The STARD steering committee (see appendix for membership and details) started with an extensive search to identify publications on the conduct and reporting of diagnostic studies. This search included the 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 two-day consensus meeting for invited experts from the following interest groups: researchers, editors, methodologists and professional organisations. The aim of the conference was 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 sessions and plenary sessions. Each small group focused on a group of related items of 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 group 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 Website 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 of steering 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 (e.g., 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 (see Fig. 1). We provide one prototypical flowchart that reflects the most commonly employed design in diagnostic research. Examples that

Table 1. 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.	
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 items 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.	
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.	

General example
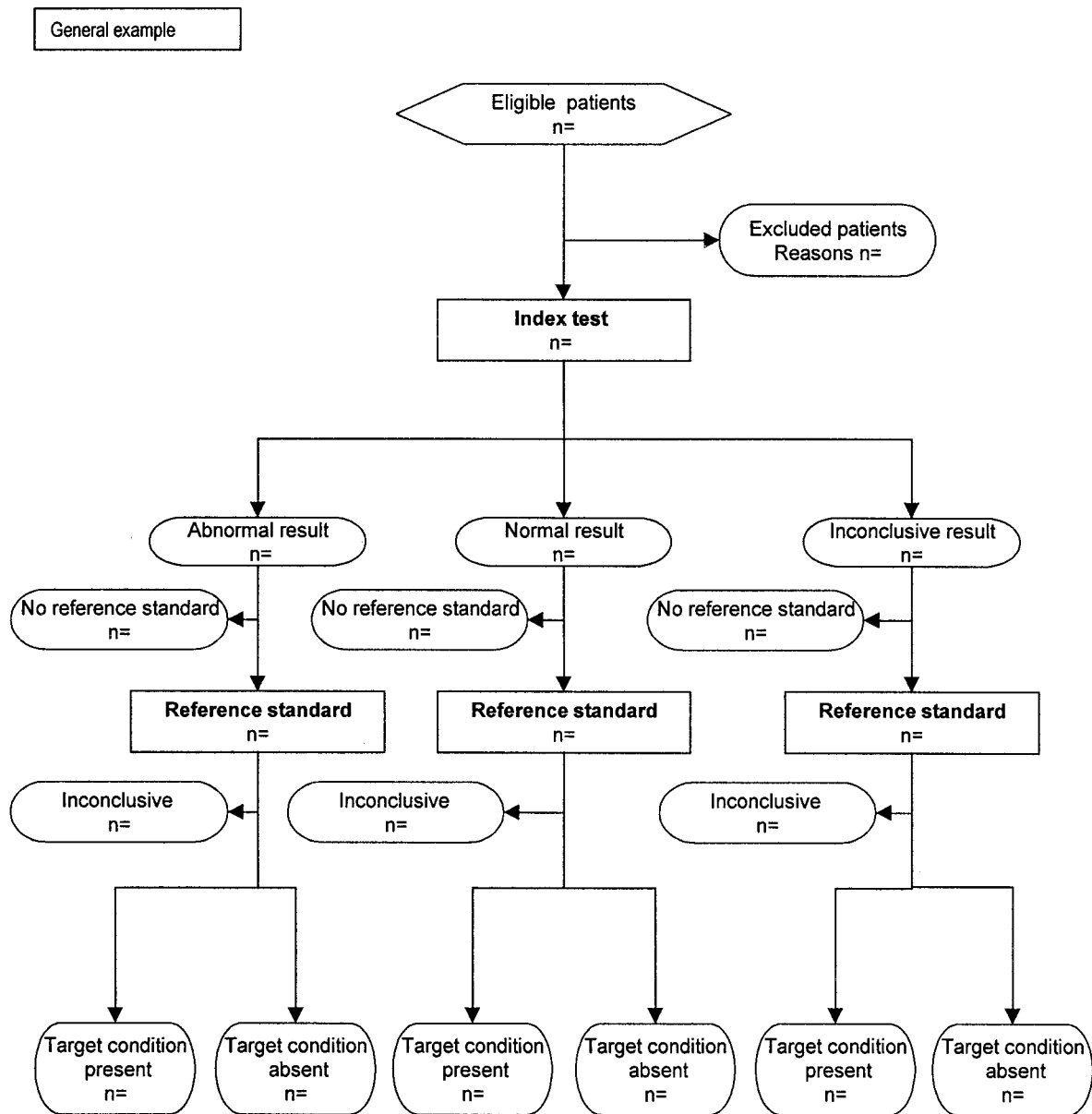
Fig. 1. Prototypical flow diagram of a diagnostic accuracy study.

reflect other designs are on the STARD Web site (see www.consort-statement.org.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 flowchart can help authors in describing essential elements of the design and conduct of the study, the execution of tests, and the results.

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 imaging tests differs from that of tests in the laboratory, we felt 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 recommenda-

tions 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 theoretical principles and papers presenting results from statistical modelling to empirical evidence derived from diagnostic studies. For several items, the evidence is rather limited.

A separate background document explains the meaning and rationale of each item and briefly summarises 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 employed 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, the number of subjects who fail to receive either the index test and/or the reference standard [potential for verification bias; Refs. *(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.

Members of the STARD steering committee

Patrick Bossuyt
Academic Medical Center, Dept. of Clinical Epidemiology, Amsterdam, The Netherlands

Constantine Gatsonis
Brown University, Centre for Statistical Sciences Providence, United States of America
Les Irwig
University of Sydney, Dept. of Public Health & Community Medicine, Sydney, Australia
David Moher
Chalmers Research Group, Ottowa, Ontario, Canada
Riekie de Vet
Free University, Institute for Research in Extramural Medicine, Amsterdam, The Netherlands
David Bruns
Clinical Chemistry, Charlottesville, United States of America
Paul Glasziou
Mayne Medical School, Dept. of Social & Preventive Medicine, Herston, Australia
Jeroen Lijmer
Academic Medical Center, Dept. of Clinical Epidemiology, Amsterdam, The Netherlands
Drummond Rennie
Journal of the American Medical Association, Chicago, United States of America

Members of the STARD group

Doug Altman, Institute of Health Sciences, Centre for Statistics in Medicine (Oxford, United Kingdom); Stuart Barton, *British Medical Journal*, BMA House (London, United Kingdom); Colin Begg, Memorial Sloan-Kettering Cancer Center, Department 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, US FDA, Center for Devices and Radiological Health (Rockville, MD); Frank Davidoff, *Annals of Internal Medicine* (Philadelphia, PA); Jon Deeks, Institute of Health Sciences, Centre for Statistics in Medicine (Old Road, United Kingdom); Paul Dieppe, Department Social Medicine, University of Bristol (Bristol, United Kingdom); Kenneth Fleming, John Radcliffe Hospital, (Oxford, United Kingdom); 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 Epidemiology & Biostatistics (Montreal, Canada); Richard Horton, *The Lancet*, (London, United Kingdom); Myriam Hunink, Erasmus Medical Center, Department of Epidemiology & Biostatistics (Rotterdam, The Netherlands); Jos Kleijnen, NHS Centre for Reviews and Dissemination (York, United Kingdom); Andre Knottnerus, Maastricht University, Netherlands School of Primary Care Research (Maastricht, The Netherlands); Erik Magid, Amager Hospital, Department of Clinical Bio-

chemistry (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, United Kingdom); Anthony Proto, *Radiology* Editorial Office (Richmond, VA);Hans Reitsma, Academic Medical Center, Department of Clinical Epidemiology (Amsterdam, The Netherlands); David Sackett, Trout Centre (Ontario, 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).

References

1. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587–94.
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;27:245–54.
4. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981;94:557–92.
5. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. The selection of diagnostic tests. In: Sackett D, editor. Clinical epidemiology, 2nd ed. Boston/Toronto/London: Little, Brown and Company; 1991: 47–57.
6. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283–98.
7. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645–51.
8. Nelemans PJ, Leiner T, de Vet HCW, van Engelshoven JMA. Peripheral arterial disease: Meta-analysis of the diagnostic performance of MR angiography. Radiology 2000;217:105–14.
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;3:361–9.
10. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–6.
11. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637–9.
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;285:1987–91.
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;285:1992–5.
14. International Committee of Medical Journal Editors. Uniform Requirements for manuscripts submitted to biomedical journals. JAMA. 1997;277:927–34. Also available at: ACP Online, http://www.acponline.org.
15. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD Statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003;49:7–18.
16. Egger M, Jüni P, Barlett C. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996–9.
17. Knottnerus JA. The effects of disease verification and referral on the relationship between symptoms and diseases. Med Decis Making 1987;7:139–48.
18. Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. Med Decis Making 1987;7:115–9.
19. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411–23.