

TOWARDS COMPLETE AND ACCURATE REPORTING OF STUDIES ON DIAGNOSTIC ACCURACY: THE STARD INITIATIVE

The STARD group

Test version, November 2001

This is a test version of the STARD checklist, the STARD flow diagram and the STARD statement. For evaluation purposes only.

The STARD group welcomes all comments, whether content or form, to improve the test version. Your comments and suggestions will be discussed among the members of STARD group during the final stage of preparing the first official version of the STARD checklist, flow diagram and statement. The official versions will be submitted for publication.

Addresses for comments and suggestions:

E-mail: stard@amc.uva.nl

Phone: +31-20-5666694

Fax: +31-20-6912683

Hans Reitsma MD PhD

Department of Clinical Epidemiology and Biostatistics

Academic Medical Center – University of Amsterdam

PO Box 22700, 1100 DE Amsterdam, The Netherlands

ABSTRACT

Objective - The objective of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative is to improve the quality of reporting of studies on diagnostic accuracy by using a checklist and flow diagram. Complete and accurate reporting allows the reader to detect the potential for bias in the study and to evaluate the generalisability of the results.

Methods - The STARD steering group started an extensive search of the literature to identify publications on the conduct and reporting of diagnostic studies. Potential items were extracted into a long list. This long list was shortened, where appropriate, during a two-day consensus meeting attended by researchers, editors, and members of professional organisations. Special attention was given to the development of a generic flow diagram for studies of diagnostic accuracy.

Results - The search on published guidelines regarding diagnostic research yielded 33 lists, which resulted in a long list of 75 potential items. During a consensus meeting, the long list was reduced to a 25-item checklist. A prototypical flow diagram was developed, which 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 transparent reporting. The STARD group anticipates that the use of the checklist in combination with the flow diagram will enhance the quality of reporting of studies on diagnostic accuracy.

INTRODUCTION

The world of diagnostic tests is highly dynamic. New tests are developed at a fast rate and the technology behind existing tests is continuously improved. Exaggerated and biased results from poorly designed and reported diagnostic studies can trigger premature dissemination and mislead physicians to 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 false test results but also limit health care 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 on diagnostic accuracy, the information from one or more tests under evaluation is compared with information from the reference standard as measured in the same series of subjects suspected of the condition of interest. The word test refers to any method for obtaining additional information on a patient's health status. It includes information from laboratory tests, imaging tests, function tests, pathology, history and physical examination. The condition of interest or target condition can refer to a particular disease or to any other identifiable condition within a patient that may prompt clinical action, like 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 one or even a combination of methods to establish the presence of the target condition, including laboratory tests, imaging tests, pathology, but also clinical follow-up of subjects. The term accuracy refers to the amount of agreement between the information from the test under evaluation 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 ROC curve.⁴⁻⁶

There are several potential threats to the internal and external validity of a study on diagnostic accuracy. A survey of studies on diagnostic accuracy published in four major medical journals between 1978 and 1993 revealed that the methodological quality was mediocre at best.⁷ Furthermore, it showed that information on key elements of design, conduct and analysis of diagnostic studies was often not reported.⁷ The absence of critical information about the design and conduct of diagnostic studies has been confirmed by authors of meta-analyses.^{8,9} Like any other

type of research, flaws in study design can lead to biased results. A report has shown 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 sub-standard reporting of diagnostic test evaluations. Urged by the findings that studies of poor quality overestimated diagnostic accuracy, the Group felt that the first step to correct these problems was to improve the quality of reporting of diagnostic studies. Following the successful CONSORT initiative¹¹⁻¹³, the Group aimed at the development of a comparable checklist of items that should be included in a report of 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 on diagnostic accuracy. Complete and accurate reporting allows the reader to detect the potential for bias in the study (internal validity) and to judge the generalisability and applicability of the results (external validity).

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, and the Cochrane Research and Methods database. In addition, the steering committee members examined reference lists of retrieved articles, searched personal files, and contacted other experts in field of diagnostic research. All relevant publications were reviewed and extracted into a long list of potential checklist items.

Subsequently, the STARD steering committee prepared 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 long list of potential items, where appropriate, and to discuss the optimal format and phrasing of the checklist.

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 long 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. This version was then discussed the next day among all attendees of the consensus meeting, and additional changes were made. There was a final round of comments from the members of the STARD working group by electronic mail.

The conference version was then field-tested by potential users and additional comments were collected. This version was made available on the Internet at the CONSORT Website (www.consort-statement.org), together with a call for comments. All comments were discussed among the STARD group to arrive at the checklist presented in this paper.

RESULTS

The search on published guidelines for diagnostic research yielded 33 lists. Based on these published guidelines and from input of steering and working group members, a long list of 75 items was assembled. During the consensus meeting on September 16 and 17, 2000, the long list was reduced into a 25-item checklist. Major revisions were made with respect to the phrasing and format of the checklist during the conference. The STARD group received several valuable comments and remarks during the various stages of evaluation after the conference. This resulted in the version of the STARD checklist as presented in table 1.

Table 1. STARD checklist of items to improve the reporting of studies on diagnostic accuracy. *Test version, November 2001. For evaluation purposes only.*

Section and Topic	Item #	Describe	Reported on page #
TITLE/ABSTRACT/KEYWORDS	1	The article as a study on diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity')	
INTRODUCTION	2	The research question(s) such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	
METHODS			
<i>Participants</i>	3	The study population: the inclusion and exclusion criteria, setting(s) and location(s) where the data were collected	
	4	Participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index test(s) or the reference standard?	
	5	Participant sampling: was this a consecutive series of patients defined by selection criteria in (3) and (4)? If not specify how patients were further selected.	
	6	Data collection: were the participants identified and data collected before the index test(s) and reference standards were performed (prospective study) or after (retrospective study)?	
<i>Reference standard</i>	7	The reference standard and its rationale	
<i>Test methods</i>	8	Technical specification of material and methods involved including how and when measurements were taken, and/or cite references for index test(s) and reference standard	
	9	Definition and rationale for the units, cutoffs and/or categories of the results of the index test(s) and the reference standard	
	10	The number, training and expertise of the persons (a) executing and (b) reading the index test(s) and the reference standard	
	11	Whether or not the reader(s) of the index test(s) and reference standard were blind (masked) to the results of the other test(s) and describe any information available to them	
<i>Statistical methods</i>	12	Methods for calculating measures of diagnostic accuracy or making comparisons, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals)	
	13	Methods for calculating test reproducibility, if done	
RESULTS			
<i>Participants</i>	14	When study was done, including beginning and ending dates of recruitment	
	15	Clinical and demographic characteristics (e.g. age, sex, spectrum of presenting symptom(s), comorbidity, current treatment(s), recruitment center)	
	16	How many participants satisfying the criteria for inclusion did or did not undergo the index test and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended)	
<i>Reference standard</i>	17	Time interval and any treatment administered between index and reference standard	
	18	Distribution of severity of disease (define criteria) in those with the target condition; describe other diagnoses in participants without the target condition	
<i>Test results</i>	19	A cross tabulation of the results of the index test(s) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard	
	20	Indeterminate results, missing responses and outliers of index test(s) stratified by reference standard result and how they were handled	
	21	Adverse events of index test(s) and reference standard	
<i>Estimation</i>	22	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals)	
	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	The clinical applicability of the study findings	

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 figure 1). We provide one prototypical flowchart that reflects the most commonly employed design in diagnostic research.

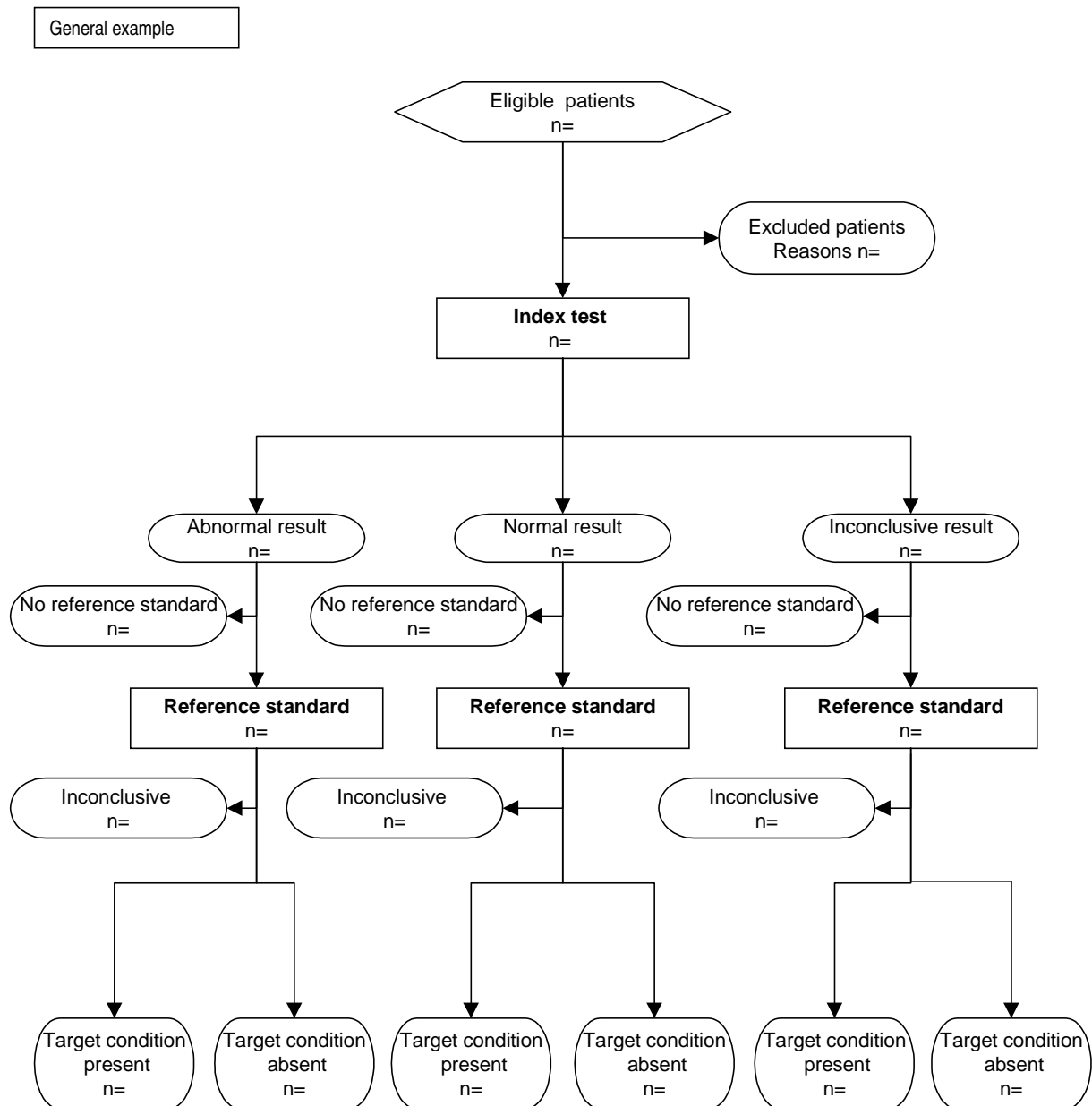


Figure 1. Prototypical flow diagram of a study on diagnostic accuracy.

Test version, November 2001, For evaluation purposes only.

DISCUSSION

The purpose of the STARD checklist is to improve the quality of the reporting of diagnostic studies. The choice of items in the STARD checklist, including the flowchart, is to assure that information on essential elements of the design, the conduct, the execution of tests and results are presented and unambiguously described. We arranged the items under the familiar headings (introduction, methods, results, discussion) of a medical research article. However, the emphasis of STARD is to assure that the various items are all reported, not to dictate the order or place within the article itself. To emphasise this point, we added the column '*reported on page #*' to the checklist.

General considerations in the development of the STARD checklist

The guiding principle in the development of the STARD checklist was the selection of 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 believed that one general checklist for studies of diagnostic accuracy, rather than different checklists for each field, would have a better chance of being widely accepted. Although the evaluation of an imaging test differs from that of a lab test, the STARD group felt that these differences were more in degree than in kind. The second consideration was to develop a checklist specifically aimed at studies of diagnostic accuracy. This meant that general issues in the reporting of research findings, like the recommendations contained in the Uniform Requirements for Manuscripts submitted to Biomedical Journals¹⁴, were excluded from the STARD checklist.

Rationale for inclusion of items and the importance of the background document

The decision to include items was based on evidence that linked these items to biased estimates (internal validity) or to variation in measures of diagnostic accuracy (external validity). The evidence varied from narrative articles explaining theoretical principles to papers presenting results from statistical modelling to empirical evidence derived from real-life diagnostic studies. This heterogeneity in evidence was one of the reasons to prepare a background document. In the background document the meaning and rationale of each item will be explained together with a summary of the type and amount of evidence. The existence of such a background document could enhance the use, understanding and dissemination of the STARD checklist. Therefore, we aim to coincide the publication of the STARD checklist with that of the STARD background document. This approach was also motivated by the experience of the CONSORT group where the explanation and explanatory

document was published at time of the revised CONSORT checklist.¹⁵ The members of the CONSORT group felt that part of the critique on the first checklist could have been prevented by the presence of the explanatory document.

Flow diagram

The STARD group put considerable efforts into the development of flow diagrams 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 The flow diagram is an essential element in the CONSORT standards for reporting of randomized trials, but its role could even be more vital in diagnostic studies given the variety of designs employed in diagnostic research. Flow diagrams in reports of diagnostic accuracy studies could provide insight into:

- the sampling and selection process of participants (external validity)
- the flow of participants in relation to the timing and outcomes of tests, in particular the number of subjects who fail to receive either the index test and/or the reference standard (potential of verification bias^{17,18})
- the number of patients at each stage of the study which identifies the correct denominator for proportions the number of patients at each stage of the study which identifies the correct denominator for rates and proportions (internal consistency)

In this statement, we provide one archetypal flow diagram general, but more examples, reflecting other designs, can be found at the CONSORT Web site at www.consort-statement.org

Evaluation and revision

We plan to measure the impact of the statement on the quality of published reports on diagnostic accuracy using a before-and-after evaluation.¹³ The checklist will be regularly updated and revised when new evidence comes available or based on comments from users of the checklist. Therefore, we welcome any comment, whether content or form, to improve the current version.

APPENDIX

Members of the STARD steering committee

Patrick Bossuyt

Academic Medical Center, Dep of Clinical Epidemiology,
Amsterdam, The Netherlands

Constantine Gatsonis

Brown University, Centre for Statistical Sciences
Providence, United States of America

Les Irwig

University of Sydney, Dep of Public Health &
Community Medicine, Sydney, Australia

David Moher

Thomas C. Chalmers Centre for Sys. Reviews
Ottawa, Ontario, Canada

Riekje 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, Dep. of Social &
Preventive Medicine, Herston, Australia

Jeroen Lijmer

Academic Medical Center, Dep of Clinical
Epidemiology, Amsterdam, The Netherlands

Drummond Rennie

Journal of the American Medical Association,
Jacksonville, United States of America

Members of the STARD working group

Doug Altman

Institute of Health Sciences, Centre for Statistics in
Medicine, Oxford, United Kingdom

Colin Begg

Memorial Sloan-Kettering Cancer Center, Dep
Epidemiology & Biostatistics, New York, United States
of America

Harry Büller

Academic Medical Center, Dep of Vascular
Medicine, Amsterdam, The Netherlands

Frank Davidoff

Annals of Internal Medicine
Philadelphia, United States of America

Paul Dieppe

Dept Social Medicine
University of Bristol, Bristol, United Kingdom

Rijk van Ginkel

Academic Medical Center, Dep of Clinical
Epidemiology, Amsterdam, The Netherlands

Gordon Guyatt

McMaster University, Clinical Epidemiology and
Biostatistics, Hamilton, Canada

Richard Horton

The Lancet,
London, United Kingdom

Stuart Barton

British Medical Journal,
BMA House, London, United Kingdom

William Black

Dartmouth Hitchcock Medical Center, Dep of
Radiology, United States of America

Gregory Campbell

US FDA, Center for Devices and Radiological Health
Rockville, United States of America

Jon Deeks

Institute of Health Sciences, Centre for Statistics in
Medicine, Old Road, United Kingdom

Kenneth Fleming

John Radcliffe Hospital, Oxford, United Kingdom

Afina Glas

Academic Medical Center, Dep of Clinical
Epidemiology, Amsterdam, The Netherlands

James Hanley

McGill University, Dep Epidemiology & Biostatistics,
Montreal, Canada

Myriam Hunink

Erasmus Medical Center, Dep Epidemiology &
Biostatistics, Rotterdam, The Netherlands

Jos Kleijnen

NHS Centre for Reviews and Dissemination
York, United Kingdom

Erik Magid

Amager Hospital, Dep Clinical Biochemistry
Copenhagen, Denmark

Matthew McQueen

Hamilton Civic Hospitals, Dep of Laboratory Medicine
Hamilton, Canada

John Overbeke

Nederlands Tijdschrift voor Geneeskunde,
Amsterdam, The Netherlands

Anthony Proto

Radiology Editorial Office, Richmond, United States
of America

David Sackett

Trout centre,
Ontario, Canada

Harold Sox

Dartmouth Hitchcock Medical Center
Dep of Medicine, Lebanon, United States of America

Stephan Walter

McMaster University, Clin Epidemiology and
Biostatistics, Hamilton, Canada

Andre Knottnerus

Maastricht University, Netherlands School of Primary
Care Research, Maastricht, The Netherlands

Barbara McNeil

Harvard Medical School, Dep of Health Care Policy,
Boston, United States of America

Andrew Onderdonk

Channing Laboratory, Boston, United States of
America

Christopher Price

St Bartholemew' s - Royal London School of Medicine
and Dentistry, London, United Kingdom

Hans Reitsma

Academic Medical Center, Dep of Clinical
Epidemiology, Amsterdam, The Netherlands

Gerard Sanders

Academic Medical Center, Dep of Clinical Chemistry
Amsterdam, The Netherlands

Sharon Straus

Mt. Sinai Hospital, Toronto, 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;134:587-594.
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, Greenland P. 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, Tugwell P. The selection of diagnostic tests. In: Sackett D, editor. *Clinical Epidemiology*. 2nd ed. Boston/Toronto/London: Little, Brown and Company; 1991. p. 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, van Engelshoven JMA. 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, Bossel 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;285:1987-1991.
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. Available at: ACP Online, <http://www.acponline.org>.
15. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, Gøtzsche PC, Lang T. The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Ann Intern Med* 2001;134:663-694.
16. Egger M, Jüni, Barlett C. Value of flow diagrams in reports of randomized controlled trials. *JAMA* 2001;285:1996-1999.
17. Panzer RJ, Suchman AL, Griner PF. Workup bias in prediction research. *Med Decis Making* 1987;7:115-119.
18. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987;6:411-423.