Clinical Chemistry 49:1
7–18 (2003)

STARD Initiative

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

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] David Moher,[7]
Drummond Rennie,[8,9] Henrica C.W. de Vet,[10] and Jeroen G. Lijmer[1]

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 generalisability of the results.

A group of scientists and editors has developed the STARD (*Sta*ndards for *R*eporting of *D*iagnostic 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 healthcare.

Introduction

In studies of diagnostic accuracy, results from one or more tests are compared with the results obtained with the reference standard on the same subjects. Such accuracy studies are a vital step in the evaluation of new and existing diagnostic technologies *(1, 2)*.

Several factors threaten the internal and external validity of a study of diagnostic accuracy *(3–8)*. Some of these factors have to do with the design of such studies, others with the selection of patients, the execution of the tests or the analysis of the data. In a study involving several metaanalyses a number of design deficiencies were shown to be related to overly optimistic estimates of diagnostic accuracy *(9)*.

Exaggerated results from poorly designed studies can trigger premature adoption of diagnostic tests and can mislead physicians to incorrect decisions about the care for individual patients. Reviewers and other readers of diagnostic studies must therefore be aware of the potential for bias and a possible lack of applicability.

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 *(8)*. Furthermore, this review showed that information on key elements of design, conduct and analysis of diagnostic studies was often not reported.

To improve the quality of reporting of studies of

[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 p.m.bossuyt@amc.uva.nl.
Received September 15, 2002; accepted September 15, 2002.

Editor's Note: The STARD statement and checklist are published in the first 2003 issues of several journals, including *Annals of Internal Medicine (86)*, *BMJ, Clinical Chemistry (87)*, and *Radiology*. They are also available on several web sites, including that of *The Lancet* and http://www.consort-statement.org/stardstatement.htm (accessed October 31, 2002).

diagnostic accuracy the Standards for Reporting of Diagnostic Accuracy (STARD) initiative was started. The objective of the 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 and to judge the generalisability and applicability of the results. For this purpose, the STARD project group has developed a single-page checklist. Where possible, the decision to include items in the checklist was based on evidence linking these items to either bias, variability in results, or limitations of the applicability of results to other settings. The checklist can be used to verify that all essential elements are included in the report of a study.

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 first part of this document contains a summary of the design and terminology of diagnostic accuracy studies. The second part contains an item-by-item discussion with examples.

Studies of Diagnostic Accuracy

Studies of diagnostic accuracy have a common basic structure *(10)*. One or more tests are evaluated, with the purpose of detecting or predicting a target condition. The target condition can refer to a particular disease, a disease stage, a health status, or to any other identifiable condition within a patient, such as staging a disease already known to be present, or a health condition that should prompt clinical action, such as the initiation, modification or termination of treatment.

Here "test" refers to any method for obtaining additional information on a patient's health status. This includes laboratory tests, imaging tests, function tests, pathology, history and physical examination.

In a diagnostic accuracy study, the test under evaluation—referred to here as the index test—is applied to a series of subjects. The results obtained with the index test are compared with the results of the reference standard, obtained in the same subjects. In this framework, the reference standard is the best available method for establishing the presence or absence of the target condition. The reference standard can be a single test, or a combination of methods and techniques, including clinical follow-up of tested subjects.

The term accuracy refers to the amount of agreement between the results from the index test and those from the reference standard. Diagnostic accuracy can be expressed in a number of ways, including sensitivity – specificity pairs, likelihood ratios, diagnostic odds ratios, and areas under ROC curves *(11, 12)*.

study question, design and potential for bias

Early in the evaluation of a test, the author may simply want to know if the test is able to discriminate. The appropriate early question may be "do the test results in patients with the target condition differ from the results in healthy people." If preliminary studies answer this question affirmatively, the next study question is, "Are patients with specific test results more likely to have the target disorder than similar patients with other test results?" The usual study design to answer this is to apply the index test and the reference standard to a number of patients who are suspected of the target condition.

Some study designs are more prone to bias and have a more limited applicability than others. In this article, the term "bias" refers to difference between the observed measures of test performance and the true measures. No single design is guaranteed to be both feasible and able to provide valid, informative and relevant answers with optimal precision to all study questions. For each study, the reader must judge the relevance, the potential for bias and the limitations to applicability, making full and transparent reporting critical. For this reason, checklist items refer to the research question that prompted the study of diagnostic accuracy and ask for an explicit and complete description of the study design and results.

variability

Measures of test accuracy may vary from study to study. Variability may reflect differences in patient groups, differences in setting, differences in definition of the target condition and differences in test protocols or in criteria for test positivity *(13)*.

For example, bias may occur if a test is evaluated under circumstances that do not correspond to those of the research question. Examples are evaluating a screening test for early disease in patients with advanced stages of the disease, evaluating a physician's office test device in the specialty department of a university hospital.

The checklist contains a number of items to make sure that a study report contains a clear description of the inclusion criteria for patients, the testing protocols and the criteria for positivity, as well as an adequate account of subjects included in the study and their results. These items will enable readers to judge if the study results apply to their circumstances.

Items in the Checklist

The next section contains a point-by-point discussion of the items on the checklist. The order of the items corresponds to the sequence used in many publications of diagnostic accuracy studies. Specific requirements made by journals could lead to a different order.

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".

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.

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 pre-existing pulmonary diseases in studies of ventilation-perfusion scintigraphy *(24, 25)*.

Tests may perform differently in a primary care setting than in a secondary or tertiary care setting. Test performance may differ if the test is used for screening rather than for confirmation of diagnostic suspicion. The spectrum of the target disease as well as the range of other conditions that occur in patients suspected of the target disease can vary from setting to setting, depending on what referral mechanisms were in play *(26–28)*. For these reasons, the report should include a careful description of where patients were recruited and where the test and the reference standard were performed.

Item 4. Describe 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?

An important element of the description is how eligible subjects were identified. Participant recruitment in diagnostic studies can start at different points *(10)*. Frequently, the study enrolls consecutive patients clinically suspected of the target condition because of presenting symptoms or referral by another healthcare professional. These patients then undergo the index test(s) as well as the reference standard.

Other designs are possible *(2)*. In some studies, patients are identified after having been subjected to the index test. Other studies start with patients in whom the reference standard established or excluded the presence of the target condition. These patients are then subjected to the index test. Still other studies include both patients already diagnosed with the target condition and participants in whom the condition was excluded. Other studies, often with retrospective data collection, include patients identified by searching hospital records to learn who received the reference standard, or the index test, or both *(29)*.

These alternative study designs are likely to influence the spectrum of disease in included patients, as well as the range and relative frequency of alternative conditions in patients without the target disorder.

In the example presented under Item 3, the reasons for attending the family planning clinic were not explicitly stated.

Item 5. Describe 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.

Example

Patients were prospectively enrolled during times the investigators or study associates were available. (30)

By definition, the targeted study population consists of all patients that satisfy the criteria for inclusion and are not disqualified by one or more of the exclusion criteria. The included patients (those whose findings comprise the study results) may be either a consecutive series of patients presenting at the study center, or a subselection. The subselection may or may not be truly random (e.g., by using a random numbers table).

It is important for readers to know the sampling scheme, since it may be helpful in judging the generalisability of the study findings.

Item 6. Describe data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?

Example

We reviewed the charts of 251 patients who underwent dual-detector spiral CT arthrography of the knee. The study population consisted of 50 consecutive patients who underwent spiral CT arthrography and subsequent arthroscopy at our institution but not prior arthroscopy in that knee. The other 201 patients included 12 who had undergone prior knee arthroscopy and subsequent arthroscopy, 69 who were referred by physicians outside of the institution, and 120 who did not undergo arthroscopy. (31)

If authors define the study question before they identify patients and collect data, they can target the collection of study data at the enrolled patients, using special case record forms or tailored data-entry forms. Prospective, dedicated data collection has many advantages: better data control, additional checks for data integrity and consistency, and a level of clinical detail appropriate to the problem *(32)*. As a result, there will be fewer missing or uninterpretable data items.

Alternatively, data collection can start after patients have undergone the index test and the reference standard. Retrospective data collection often relies on chart review. Studies with retrospective data collection may reflect routine clinical practice better than a prospective study, but also may fail to identify all eligible patients or to provide data of high quality *(29)*.

Item 7. Describe the reference standard and its rationale.

Example

The e-4 allele of the gene encoding apolipoprotein E (ApoE) is strongly associated with Alzheimer's disease, but its value in the diagnosis remains uncertain. (. . .) Using the pathological diagnosis of Alzheimer's disease as the standard, we compared the sensitivity and specificity of the clinical diagnosis of Alzheimer's disease, the ApoE genotype, and the clinical diagnosis and ApoE genotype determined sequentially. (33)

In studies of diagnostic accuracy, the reference standard is used to distinguish patients with the target condition from those without it. Some target conditions cannot be defined unambiguously. Depending on the study question, clinical relevance, management decisions or prognosis, or pathological diagnosis may define the target condition *(10)*.

When it is not possible to subject all patients to the reference standard for practical or ethical reasons, authors often use a composite reference standard. The components may reflect different definitions of the target condition or different strategies for diagnosing the target condition. One example comes from studies of using nuchal translucency in the first trimester of pregnancy as a marker for Down syndrome *(34)*. In several of those studies, positive test results were verified with karyotyping, whereas negative results were verified by awaiting delivery. Studies in which the decision to perform fetal karyotyping depended on the result of nuchal translucency measurement considerably overestimated the sensitivity of nuchal translucency *(34)*.

Authors should clearly define the reference standard and how the choice of the reference standard relates to the study question.

In the example, the authors use a neuropathological diagnosis after postmortem examination as the reference standard in patients referred to Alzheimer's disease centers for evaluations of dementia. Although pathological assessment is considered to be the gold standard of Alzheimer's disease diagnosis, the correlation of the clinical and pathological data is by no means perfect. Nor does every pathologist render the same diagnosis from a given set of tissue sections *(35)*.

Item 8. Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.

Example

Concentrations of prostate-specific antigen (PSA) were measured by the Tandem total PSA and free PSA monoclonal antibody-based assays (Hybritech).[ref] A new, time-resolved immunofluorometric assay, recently developed in our laboratory, was used to measure serum hK2 concentrations.[ref] Briefly, the hK2 assay uses a mouse monoclonal anti-hK2 capture antibody [coded G586, supplied by Hybritech (San Diego, CA) and raised against recombinant hK2], a biotinylated mouse monoclonal detection antibody (coded 8311; Diagnostic Systems Laboratories) and alkaline phosphatase-labeled streptavidin. We measured the alkaline phosphatase activity by adding the substrate diflunisal phosphate, incubating for 10 min, and then adding a Tb^{3+}-EDTA developing solution. The fluorescence was measured on a Cyberfluor 615 Immunoanalyzer (MDS Nordion). The hK2 assay has a detection limit of 0.006 μg/L and has <0.2% cross-reactivity to PSA. A full description of the method and its evaluation has been published elsewhere. [ref] (36)

Authors should describe the methods involved in the execution of index test and reference standard in sufficient detail to allow other researchers to replicate the study or to allow readers to judge the feasibility of the index test in their own setting. Differences in the execution of the index test and reference standard are a potential source of variation in diagnostic accuracy *(13, 24)*.

The description should cover the full test protocol including the specification of materials and instruments together with their instructions for use, and specific measures (preparations) in participants (e.g., fasting before blood sample, the anatomic site of measurement). If no descriptions are available, details must be provided in the text. Between-study variability in measures of test accuracy due to differences in test protocol has been documented for a number of tests, including the use of hyperventilation before exercise electrocardiography and the use of tomography for exercise thallium scintigraphy *(23, 24)*.

Item 9. Describe definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.

Example

We chose three cutoff points of B type natriuretic peptide to achieve sensitivity values of at least 90%, 80%, and 70%. (37)

Test results can be truly dichotomous (e.g., present or absent), have multiple categories or be continuous. Readers need to know how the authors expressed results of the index test and reference standard.

If the authors defined several categories of results, readers need to know how and when they defined category boundaries and whether they defined them before the study, or after obtaining the results. In the latter case, there is an increased likelihood that the authors selected the cutoff value to maximize a particular test characteristic, which reduces the likelihood that another study will replicate the findings *(38, 39)*.

In the example, the authors are explicit about their selection of cut-offs for B type natriuretic peptide measurement in the diagnosis of left ventricular systolic dysfunction. They established these cut-offs *post hoc* to obtain pre-specified sensitivities.

Item 10. Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.

Example

Subjects were classified as to whether or not they were heavy drinkers based on their response to the Self Administered Alcohol Screening Test (SAAST)[ref] and the Khavari questionnaire on the amount of alcohol consumed during the past year.[ref] Both questionnaires were administered by a research associate. The research associate was trained by a Ph.D. psychologist who specializes in alcohol treatment. He discussed with her how to score questions and how to follow up ambiguous information, and he observed her administering more than 10 questionnaires. (40)

Variability in the manipulation, processing or reading of the index test or reference standard will affect measures of diagnostic accuracy *(41, 42)*. Many studies have shown reader variability, especially in the field of imaging *(43, 44)*. The amount of readers' training can help readers to judge whether similar results are attainable in their own settings, with possibly less experienced readers.

Professional background, expertise, and prior training to improve interpretation and to reduce interobserver variation all affect the quality of reading *(45, 46)*. Readers are more like to interpret results from (subjective) tests as abnormal in settings with higher prevalences of the target condition, a tendency known as context bias *(47)*.

The example describes the reference standard in a study of a model that uses results of commonly performed laboratory tests to identify men who are heavy drinkers.

ITEM 11. DESCRIBE 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.

Example

All images were interpreted on the computer workstation by two radiologists (J.Y., R.K.H.) independently, and subsequently a consensus reading was performed. The radiologists were blinded to the patient's history, including whether the patient had been recruited for screening or for symptoms, and to results of standard colonoscopy and histologic analysis. (14)

Knowledge of the results of the reference standard can influence the reading of the index test, and vice versa. Such knowledge is likely to increase the agreement between results of the index test and those of the reference standard, leading to inflated measures of diagnostic accuracy. The distortion of measures of diagnostic accuracy caused by knowledge of the result of the reference standard while interpreting the index test is known as test review bias *(23)*. Knowing the result of the index test while interpreting the reference standard has been named diagnostic review bias *(23)*. The observation that interpretations become more accurate by providing additional clinical information to interpreters is known as clinical review bias *(6, 48, 49)*.

Withholding information from the readers of the test is known as blinding or masking. Readers can be masked for the results of other tests or even for all information related to the patient.

Blinding of readers of tests is important. In a meta-regression analysis of a wide range of tests, test review bias produced a moderate exaggeration of measures of diagnostic accuracy *(9)*. Individual studies have shown a substantial effect of inappropriate masking *(24)*.

The example shows how the readers of CT colonography for colorectal polyp and cancer detection were blinded to additional clinical information as well as to the results of colonoscopy, the reference standard.

ITEM 12. DESCRIBE METHODS FOR CALCULATING OR COMPARING MEASURES OF DIAGNOSTIC ACCURACY, AND THE STATISTICAL METHODS USED TO QUANTIFY UNCERTAINTY (E.G., 95% CONFIDENCE INTERVALS).

Example

The statistical significance of the differences in sensitivities between magnetic resonance angiography (MRA) and duplex were assessed by means of the McNemar test. (50)

Several measures of diagnostic accuracy exist *(12)*. Authors should report in sufficient detail the methods used in calculating the measures that they considered appropriate.

Estimates of diagnostic accuracy are subject to chance variation, with larger studies usually resulting in more precise estimates. Authors should therefore quantify the amount of statistical uncertainty around the observed value *(51)*. Articles that describe methods for calculating the precision around frequently used measures of diagnostic accuracy are readily available *(12)*.

Alternatively, statistical techniques can be used to test more specific hypotheses, such as the superiority of one test over another, or the hypothesis that a specific measure of diagnostic accuracy surpasses a pre-specified value.

In the example, the authors used McNemar test statistic to reject the null hypothesis that magnetic resonance angiography had the same sensitivity as duplex sonography for diagnosing renovascular disease.

ITEM 13. DESCRIBE METHODS FOR CALCULATING TEST REPRODUCIBILITY, IF DONE.

Example

Interobserver variability in the interpretation of conventional angiography and magnetic resonance angiography (MRA) were computed using the κ statistic including 95% confidence intervals. (50)

The index test and the reference standard are seldom perfect. Their reproducibility varies, and limited reproducibility adversely affects diagnostic accuracy *(52)*.

Observer variability can arise with imaging tests when the reader must summarize visual observations in a statement about the presence of disease. It also arises during classification, when the reader must use the data to place patients into diagnostic categories *(41)*. Instrument variability concerns the amount of variation that arises during the operation of devices or systems, such as automated laboratory measurements. Other terms for this form of variation include imprecision, analytic methodological variation, or analytical noise (error). Poor reproducibility adversely affects diagnostic accuracy. If possible, authors should evaluate the reproducibility of the test methods used in their study and report their procedure to do so.

For quantitative assays, it is useful to report imprecision as the coefficient of variation at two or more specified mean values near clinical decision points as obtained by repeating the test over a specified number of days. Within-run coefficients of variation are appropriate if all patient samples were analyzed in a single run.

In the example, the authors used the kappa statistic to express interobserver variability for conventional angiography and MRA in the detection of renovascular disease.

ITEM 14. REPORT WHEN STUDY WAS DONE, INCLUDING BEGINNING AND ENDING DATES OF RECRUITMENT.

Example

We retrospectively screened all blood cultures from patients on an oncology ward at New England Medical Center, a 300-bed tertiary care university-affiliated hospital, between August 1994 and June 1996. (53)

The technology behind many tests advances continuously, leading to improvements in diagnostic accuracy. There may be a considerable gap between the dates of the study and the publication date of the study report. Readers will therefore want to know the dates during

which a study was conducted. This information may also provide an indication about the rate of recruitment.

ITEM 15. REPORT CLINICAL AND DEMOGRAPHIC CHARACTERISTICS OF THE STUDY POPULATION (E.G., AGE, SEX, SPECTRUM OF PRESENTING SYMPTOMS, COMORBIDITY, CURRENT TREATMENTS, RECRUITMENT CENTERS).

Example

Demographic, clinical and x-ray angiographic characteristics of the 109 study patients. (20)

CHARACTERISTIC	VALUE
Female sex—no. (%)	34 (31)
Age—yr	
Mean ± SD	59 ± 10
Range	27–75
Chest pain—no. (%)	86 (79)
Prior myocardial infarction—no. (%)	26 (24)
History of systemic hypertension—no. (%)	54 (50)
Current or prior cigarette smoking—no. (%)	58 (53)
Cholesterol > 200 mg/dl—no. (%)	67 (61)
Diabetes—no. (%)	19 (17)
Family history of premature coronary disease*—no. (%)	43 (39)
Findings on x-ray angiography—no. (%)	
One-vessel disease	31 (28)
Two-vessel disease	20 (18)
Three-vessel disease	13 (12)

*** A family history was defined as a history of myocardial infarction or angina in a first-degree relative before the age of 65.**

An adequate description of the demographic and clinical characteristics of the participants allows the reader to judge the applicability of the study findings to another population. Most authors present the demographic and clinical characteristics of the study group in a table.

ITEM 16. REPORT 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).

Example 1

During the course of the study, 272 patients with suspected deep vein thrombosis (DVT) were referred to the participating centers. Of these, 28 were excluded from the study for the following reasons: previous DVT (21), contrast allergy (1), renal failure (1), and unwillingness to provide consent (5). Of the remaining 244 patients, 25 were excluded from the analysis because of inadequate or failed venography and 5 were excluded because of inadequate or failed impedance plethysmography. (54)

Example 2 (see Figure 1)

The study report should present the number of participants that were assessed for eligibility, if available. This number is a useful indicator of how closely the targeted study population resembles the patient population.

The flow diagram provides the exact number of patients at each stage of the study and thus the correct denominator for calculating rates and proportions. It also shows the number of subjects who failed to receive either the index test and/or the reference standard.

Measures of diagnostic accuracy will be biased if the result of the index test influences the decision to order the reference standard test *(56–63)*. The terms used to describe this effect, include (partial) verification bias, work-up bias, (primary) selection bias, sequential ordering bias, and verification bias (the most general term). Verification bias occurs in up to 26% of diagnostic studies and is especially common when the reference standard is an invasive procedure *(60)*.

We strongly recommend the use of a flow diagram to illustrate the design of the study and provide the exact number of participants at each stage of the study. A flow diagram can communicate transparently the key elements of a study design. A flow diagram has been a helpful addition to reports of randomized clinical trials *(64)*.

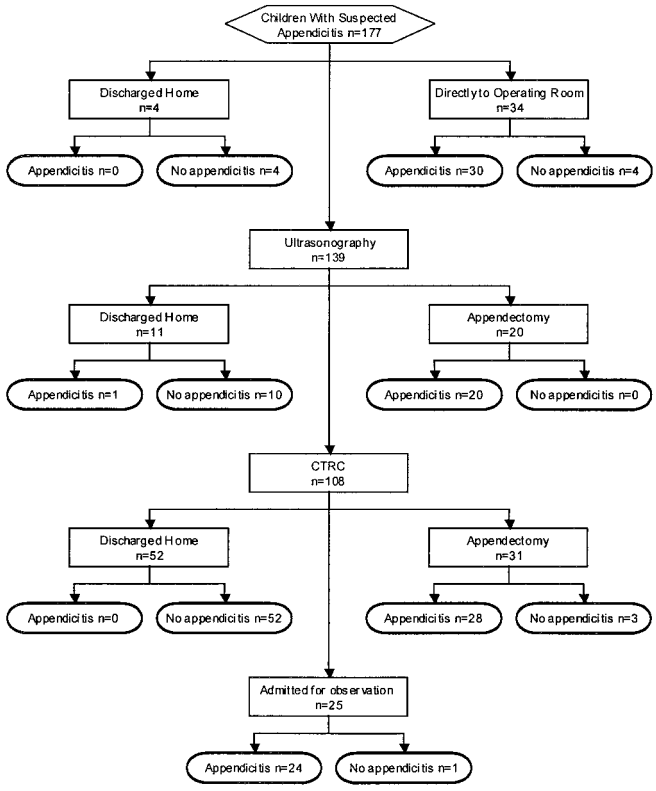

Fig. 1. Example of a flow diagram of a diagnostic accuracy study *(55)*.

ITEM 17. REPORT TIME INTERVAL FROM THE INDEX TESTS TO THE REFERENCE STANDARD, AND ANY TREATMENT ADMINISTERED BETWEEN THEM.

Example

Patients were scheduled to undergo CT colonography before conventional colonoscopy, both of which were performed on the same day. (14)

In epidemiological terms, studies of diagnostic accuracy are cross-sectional. The results of the index test and reference standard are performed on the same patients at the same time *(10)*. When a delay occurs between doing the index test and the reference standard the condition of the patient may change, leading to worsening or improvement of the target condition or the alternative conditions.

Similar concerns apply if treatment is started after doing the index test but before doing the reference standard.

ITEM 18. REPORT DISTRIBUTION OF SEVERITY OF DISEASE (DEFINE CRITERIA) IN THOSE WITH THE TARGET CONDITION; OTHER DIAGNOSES IN PARTICIPANTS WITHOUT THE TARGET CONDITION.

Demographic and clinical features of the study population can affect measures of diagnostic accuracy. This variability is known as spectrum bias *(56)*. The spectrum effect includes the severity of the target condition, demographic features, and comorbidity. All of these elements have caused variability in measures of test accuracy, but most notable examples involved differences in the severity of the target condition *(65–70)*.

Many target conditions are not pure dichotomous states but cover a continuum, ranging from minute pathological changes to advanced clinical disease. Test sensitivity is often higher in studies with a higher proportion of patients with more advanced stages of the target condition *(56* On the other hand, in the presence of comorbid conditions, false-positive or false-negative test results may occur more often *(25, 56, 71)*.

Accordingly, it is of important to describe the severity of disease in the study group.

ITEM 19. REPORT 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 REPORT THE DISTRIBUTION OF THE TEST RESULTS BY THE RESULTS OF THE REFERENCE STANDARD.

Example 1

Distribution of cytologic outcomes within each histologic type of thyroid carcinoma. (non-diagn: nondiagnostic; Fol neopl. Follicular neoplasia; PAP: papillary carcinoma; FOL: follicular carcinoma; MED medullary carcinoma; ANAPL anaplastic carcinoma. (72)

	Nondiagn	Normal	Atypia	Fol. neopl	Suspect	Malignant
PAP	12	30	5	17	18	18
FOL	18	31	3	40	5	3
MED	15	15	4	11	28	27
ANAPL	18	12	5	5	13	47
Total	14	28	4	23	14	17

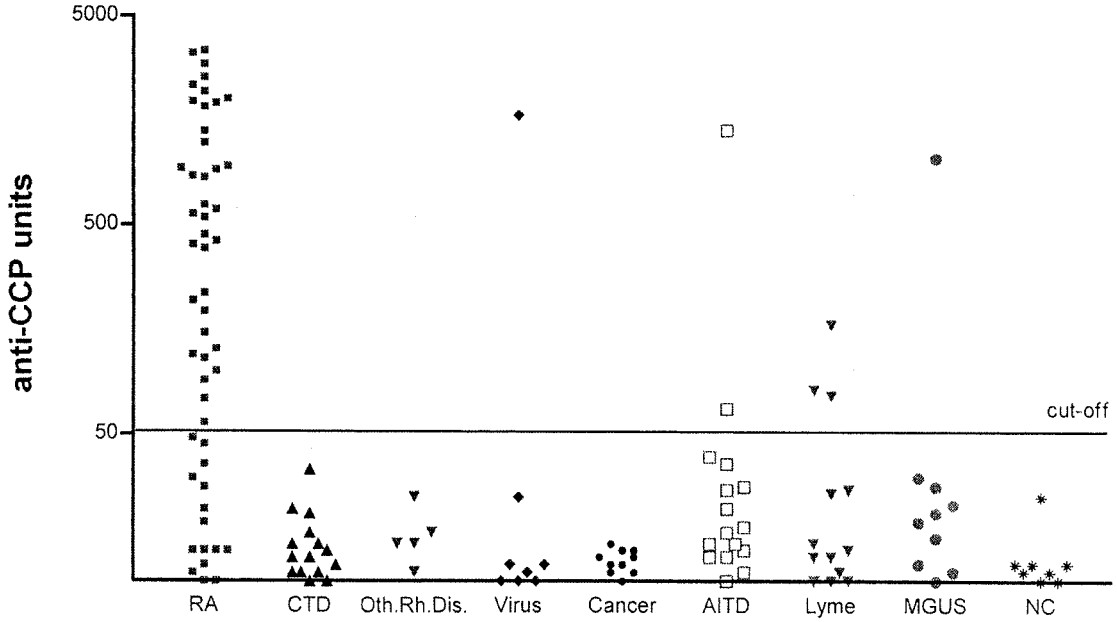

Fig. 2. Distribution on a log scale of the test results according to anti-CCP units for the different groups of patients.

A cutoff value set at 50 units guarantees a good specificity because all but seven of the non-RA patients have an antibody concentration below the threshold. *CTD*, connective tissue disease; *Oth. Rh. Dis.*, other rheumatoid diseases; *NC*, healthy controls *(73)*.

Example 2 (see Figure 2)
Scientists want to verify important results, and so reanalysis is an important aspect of the scientific method. To facilitate this process, authors should present results in the form of absolute numbers. Cross tabulations of test results in categories and graphs of distributions of continuous results are essential to allow scientific colleagues to (re)calculate measures of diagnostic accuracy or to perform alternative analyses, including metaanalysis. Authors should report all test results, including indeterminate test results on the index test and reference test.

One example, with a few categories of test results, is taken from a study of fine-needle aspiration cytology in histologically proven thyroid carcinoma; the second example shows the distribution of the concentration of anti-citruline antibodies in patients with the target condition (rheumatoid arthritis) and in patients with several alternative diagnoses.

ITEM 20. REPORT ANY ADVERSE EVENTS FROM PERFORMING THE INDEX TESTS OR THE REFERENCE STANDARD.

Example
A mean period of 15 min was sufficient for full investigation of the uterine cavity. The mean tolerance on a pain scale range of 0 to 10 was 1. However, sonohysterography has not been tolerated once (indication of pain to 10). The patient had pelvic pains that were regressive with phosphoglucinol and after 20 min of rest in decubitus position. Only one complication was recorded during the subsequent 3 days, an endometritis in a patient with unbalanced diabetes. Ampicilen antibiotic treatment was efficient permitting a complete recovery. (74)

Not all tests are safe. Measuring and reporting of adverse events in studies of diagnostic accuracy can provide additional information about the clinical usefulness of a particular test. The requirement to report adverse events applies equally to diagnostic research and research on treatments *(75)*.

It can also be important to learn about the invasiveness and risks of the reference standard used. For example, if in the evaluation of the positive results of Hemocult screening, colonscopy, sigmoidscopy and double contrast barium enema were to be used, one might expect complications (perforation or hemorrhage) once in 300–900 subjects investigated *(76)*.

The example comes from the first part of the results section of a study of sonohysterography for the diagnosis of intrauterine abnormalities, with histopathology and clinical outcome as the composite reference standard.

ITEM 21. REPORT ESTIMATES OF DIAGNOSTIC ACCURACY AND MEASURES OF STATISTICAL UNCERTAINTY (E.G., 95% CONFIDENCE INTERVALS).

Example
ROC plots comparing CDTect values with %CDT values for men and women independently are given in Fig. 2. (. . .) The areas under the curves (with 95% confidence intervals) were 0.88 (0.85– 0.91) and 0.89 (0.86–0.92) for men (P = 0.67) and 0.72 (0.68–0.76) and 0.76 (0.72–0.81) for women (P = 0.26), respectively. (77)

The final aim of a study of diagnostic accuracy is to produce an expression of how well the test results corresponded with the presence or absence of the target condition, as established by the reference standard. The values presented in the report should be taken as estimates. Due to chance variations in the patients submitted to the tests and other factors, the results are likely to differ over replications of the study in the same study population *(51)*. The reporting of precision will show the reader the range of likely values around an estimate of diagnostic accuracy.

Many journals require or strongly encourage the use of confidence intervals as measures of precision. A 95% confidence interval is conventional. Only 50% of the reports of diagnostic evaluations published in 1996 or 1997 in the British Medical Journal reported precision for the estimates of diagnostic accuracy *(78)*.

ITEM 22. REPORT HOW INDETERMINATE RESULTS, MISSING RESPONSES AND OUTLIERS OF THE INDEX TESTS WERE HANDLED.

Uninterpretable, indeterminate and intermediate test results pose a problem in the assessment of a diagnostic test *(71, 79, 80)*. By itself, the frequency of these test results is an important indicator of the overall usefulness of the test. Furthermore, ignoring such test results can produce biased estimates of diagnostic accuracy if these results occur more frequently in patients with the target condition than in those without it, or vice versa.

Uninterpretable, indeterminate and intermediate test results have many causes *(79)*. An test result may fail technically or from an insufficient sample, such as the absence of cells in a needle biopsy from a tumor (uninterpretable result) *(45, 81, 82)*. A test result may be invalidated by a concomitant medical condition or therapy that affects the test, e.g., the effect of beta-adrenergic blockers on heart rate response during an exercise test (indeterminant result) *(24)*.

The occurrence of uninterpretable, indeterminate and intermediate test results varies from test to test, but frequencies up to 40% have been reported *(79)*. Intermediate test results (not clearly positive or negative) may have diagnostic value, as in the case of ventilation perfusion scans that are neither normal nor high probability for pulmonary embolism(s) *(83)*. The incorporation of such test results into clinical decisions making varies *(80)*.

The Table under Item 19 rightfully includes nondiagnostic test results.

ITEM 23. REPORT ESTIMATES OF VARIABILITY OF DIAGNOSTIC ACCURACY BETWEEN SUBGROUPS OF PARTICIPANTS, READERS OR CENTERS, IF DONE.

Example

For detection of hemodynamically significant main renal artery stenosis, sensitivity and specificity were 90% (. . .) for magnetic resonance angiography (. . .) When patients with fibromuscular dysplasia were excluded from the analysis, the sensitivity of magnetic resonance angiography increased to 97%, with a negative predictive value of 98%. (50)

Since variability is the rule rather than the exception, researchers should explore possible sources of heterogeneity in results, within the limits of the available sample size. The best practice is to plan subgroup analyses before the start of the study *(84)*.

In the example above, the authors report separate estimates for patients with fibromuscular dysplasia. They did not specify whether they planned this subgroup analysis before the data collection.

ITEM 24. REPORT ESTIMATES OF TEST REPRODUCIBILITY, IF DONE.

Example

The interobserver variability in the grading of stenotic renal artery lesions (grades 1 to 4) with conventional angiography and MRA were identical, with a κ *value of 0.77 and 95% confidence intervals ranging from 0.67 to 0.86. For the detection of hemodynamically significant lesions, interobserver variability was 0.87 (0.78 to 0.95) for MRA and 0.88 (0.79 to 0.97) for conventional angiography. (50)*

We recommend that authors report all measures of test reproducibility that they performed during the study (see Item 13). For quantitative analytical methods, report the coefficient of variation (CV) at concentrations that are relevant to the study, state those concentrations and the number of determinations (for within-run CV, if relevant) or the number of days of testing (for day-to-day, total CV), or both.

ITEM 25. DISCUSS THE CLINICAL APPLICABILITY OF THE STUDY FINDINGS.

Example

Although several studies on assays for brain natriuretic peptide in select patient groups have been published, these are the first data on the performance characteristics of an assay for NT-proBNP in a large generalisable series of randomly selected adults with validated diagnoses of heart failure and with a comparator normative population randomly selected from the same populations as the cases. (. . .) These data suggest that, in clinical practice, the assay would have three practical uses: screening patients with existing clinical labels of heart failure (70 of the 103 patients so categorised in this study had heart failure ruled out on NT-proBNP testing); triaging patients presenting with symptoms suggestive of heart failure (shortness of breath, lethargy) for echocardiography; and screening patients at high risk of heart failure. We suspect the assay would perform well in these settings, but the first indication was not formally tested in this study, and the third indication was tested in only 134 patients. (85)

Because of the variability in tests characteristics due to differences in design, patients and procedures, the findings from one particular study may not be applicable to the decision problem of interest to the readers *(13)*.

In addition to a discussion about the potential methodological shortcomings of the study and a general interpretation of the results in the context of current evidence, we recommend that authors point out the differences between the context of the study and other settings and patient groups in which the test is likely to be used.

Comments

We are aware that studies of diagnostic accuracy are not the only type of studies to evaluate diagnostic tests. A wide range of other designs is used, including randomized clinical trials *(2)*.

The methodology for designing and conducting studies of diagnostic accuracy is still maturing. Our understanding of the sources of variability and the potential for bias is growing. As a result, we expect to update the STARD checklist periodically.

Diagnostic tests are an essential part of medicine. Complete and informative reporting can only lead to better decisions in healthcare.

References

1. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134:587–94.
2. Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:19–38.
3. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994;271:389–91.
4. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271:703–7.
5. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85–91.
6. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411–23.
7. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA 1984;252:2418–22.
8. 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.
9. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282:1061–6.
10. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:39–59.
11. 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.
12. Habbema JDF, Eijkemans R, Krijnen P, Knottnerus JA. Analysis of

data on the accuracy of diagnostic tests. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:117–44.

13. Irwig LM, Bossuyt PM, Glasziou PP, Gatsonis C, Lijmer JG. Designing studies to ensure that estimates of test accuracy will travel. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:95–116.
14. Yee J, Akerkar GA, Hung RK, Steinauer-Gebauer AM, Wall SD, McQuaid KR. Colorectal neoplasia: performance characteristics of CT colonography for detection in 300 patients. Radiology 2001; 219:685–92.
15. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1:447–58.
16. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309:1286–91.
17. McKibbon KA, Walker-Dilks CJ. Beyond ACP Journal Club: how to harness MEDLINE for diagnostic problems. ACP J Club 1994; 121(Suppl 2):A10–2.
18. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53:65–9.
19. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. Medinfo 2001;10: 390–3.
20. Kim WY, Danias PG, Stuber M, Flamm SD, Plein S, Nagel E, et al. Coronary magnetic resonance angiography for the detection of coronary stenoses. N Engl J Med 2001;345:1863–9.
21. World Medical Association Declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. JAMA 1997;277:925–6.
22. Newhall WJ, Johnson RE, DeLisle S, Fine D, Hadgu A, Matsuda B, et al. Head-to-head evaluation of five chlamydia tests relative to a quality-assured culture standard. J Clin Microbiol 1999;37: 681–5.
23. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic problems of exercise testing for coronary artery disease: groups, analysis and bias. Am J Cardiol 1980;46:807–12.
24. Detrano R, Gianrossi R, Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. Prog Cardiovasc Dis 1989;32:173–206.
25. Stein PD, Gottschalk A, Henry JW, Shivkumar K. Stratification of patients according to prior cardiopulmonary disease and probability assessment based on the number of mismatched segmental equivalent perfusion defects. Approaches to strengthen the diagnostic value of ventilation/perfusion lung scans in acute pulmonary embolism. Chest 1993;104:1461–7.
26. Knottnerus JA, Knipschild PG, Sturmans F. Symptoms and selection bias: the influence of selection towards specialist care on the relationship between symptoms and diagnoses. Theor Med 1989; 10:67–81.
27. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992;45: 1143–54.
28. Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. Scand J Prim Health Care 1993;11:241–6.
29. van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. J Clin Epidemiol 1995;48:417–22.
30. Kline JA, Israel EG, Michelson EA, O'Neil BJ, Plewa MC, Portelli DC. Diagnostic accuracy of a bedside D-dimer assay and alveolar dead-space measurement for rapid exclusion of pulmonary embolism: a multicenter study. JAMA 2001;285:761–8.
31. Vande Berg BC, Lecouvet FE, Poilvache P, Dubuc JE, Bedat B, Maldague B, et al. Dual-detector spiral CT arthrography of the knee: accuracy for detection of meniscal abnormalities and unstable meniscal tears. Radiology 2000;216:851–7.
32. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 1996;25:435–42.
33. Mayeux R, Saunders AM, Shea S, Mirra S, Evans D, Roses AD, et al. Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. N Engl J Med 1998; 338:506–11.
34. Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. Obstet Gynecol 1999;94:864–9.
35. Corley DE, Kirtland SH, Winterbauer RH, Hammar SP, Dail DH, Bauermeister DE, et al. Reproducibility of the histologic diagnosis of pneumonia among a panel of four pathologists: analysis of a gold standard. Chest 1997;112:458–65.
36. Magklara A, Scorilas A, Catalona WJ, Diamandis EP. The combination of human glandular kallikrein and free prostate-specific antigen (PSA) enhances discrimination between prostate cancer and benign prostatic hyperplasia in patients with moderately increased total PSA. Clin Chem 1999;45:1960–6.
37. Smith H, Pickering RM, Struthers A, Simpson I, Mant D. Biochemical diagnosis of ventricular dysfunction in elderly patients in general practice: observational study. BMJ 2000;320:906–8.
38. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
39. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15: 361–87.
40. Hartz AJ, Guse C, Kajdacsy-Balla A. Identification of heavy drinkers using a combination of laboratory tests. J Clin Epidemiol 1997; 50:1357–68.
41. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). J Clin Epidemiol 1992;45: 567–80.
42. Brealey S, Scally AJ, Thomas NB. Review article: methodological standards in radiographer plain film reading performance studies. Br J Radiol 2002;75:107–13.
43. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–9.
44. Brealey S, Scally AJ. Bias in plain film reading performance studies. Br J Radiol 2001;74:307–16.
45. Ronco G, Montanari G, Aimone V, Parisio F, Segnan N, Valle A, et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. Cytopathology 1996;7:151–8.
46. Cohen MB, Rodgers RP, Hales MS, Gonzales JM, Ljung BM, Beckstead JH, et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. Arch Pathol Lab Med 1987;111:518–20.
47. Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. JAMA 1996;276:1752–5.
48. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. AJR Am J Roentgenol 1981;137:1055–8.
49. Berbaum KS, Franken EA Jr, Dorfman DD, Barloon T, Ell SR, Lu CH, et al. Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986;21:532–9.

50. Leung DA, Hoffmann U, Pfammatter T, Hany TF, Rainoni L, Hilfiker P, et al. Magnetic resonance angiography versus duplex sonography for diagnosing renovascular disease. Hypertension 1999;33: 726–31.

51. Lang TA, Secic M. Generalizing from a sample to a population: reporting estimates and confidence intervals. In: Lang TA, Secic M, eds. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. Philadelphia: American College of Physicians, 1997:55–63.

52. Quinn MF. Relation of observer agreement to accuracy according to a two-receiver signal detection model of diagnosis. Med Decis Making 1989;9:196–206.

53. DesJardin JA, Falagas ME, Ruthazer R, Griffith J, Wawrose D, Schenkein D, et al. Clinical utility of blood cultures drawn from indwelling central venous catheters in hospitalized patients with cancer. Ann Intern Med 1999;131:641–7.

54. Wells PS, Brill-Edwards P, Stevens P, Panju A, Patel A, Douketis J, et al. A novel and rapid whole-blood assay for D-dimer in patients with clinically suspected deep vein thrombosis. Circulation 1995; 91:2184–7.

55. Garcia Pena BM, Mandl KD, Kraus SJ, Fischer AC, Fleisher GR, Lund DP, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. JAMA 1999;282:1041–6.

56. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299:926–30.

57. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin Epidemiol 1992;45:581–6.

58. Zhou XH. Effect of verification bias on positive and negative predictive values. Stat Med 1994;13:1737–45.

59. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39:207–15.

60. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. Invest Radiol 1985;20:751–6.

61. Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. J Clin Epidemiol 1996;49: 735–42.

62. Diamond GA, Rozanski A, Forrester JS, Morris D, Pollock BH, Staniloff HM, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. J Chronic Dis 1986;39:343–55.

63. Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. Med Decis Making 1992;12:22–31.

64. Egger M, Juni P, Bartlett C. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996–9.

65. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986; 104:66–73.

66. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992;117:135–40.

67. O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. Neurology 1996;47:140–4.

68. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–7.

69. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64–71.

70. Harris JM Jr. The hazards of bedside Bayes. JAMA 1981;246: 2602–5.

71. Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA 1982;248:2467–70.

72. Giard RW, Hermans J. Use and accuracy of fine-needle aspiration cytology in histologically proven thyroid carcinoma: an audit using a national nathology database. Cancer 2000;90:330–4.

73. Bizzaro N, Mazzanti G, Tonutti E, Villalta D, Tozzoli R. Diagnostic accuracy of the anti-citrulline antibody assay for rheumatoid arthritis. Clin Chem 2001;47:1089–93.

74. Bonnamy L, Marret H, Perrotin F, Body G, Berger C, Lansac J. Sonohysterography: a prospective survey of results and complications in 81 patients. Eur J Obstet Gynecol Reprod Biol 2002;102: 42–7.

75. Ioannidis JP, Lau J. Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas. JAMA 2001;285: 437–43.

76. Towler BP, Irwig L, Glasziou P, Weller D, Kewenter J. Screening for colorectal cancer using the faecal occult blood test, hemoccult. Cochrane Database Syst Rev 2000:CD001216.

77. Anton RF, Dominick C, Bigelow M, Westby C. Comparison of Bio-Rad %CDT TIA and CDTect as laboratory markers of heavy alcohol use and their relationships with γ-glutamyltransferase. Clin Chem 2001;47:1769–75.

78. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. BMJ 1999;318:1322–3.

79. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. J Chronic Dis 1986; 39:575–84.

80. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making 1987;7:107–14.

81. Pisano ED, Fajardo LL, Tsimikas J, Sneige N, Frable WJ, Gatsonis CA, et al. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: The Radiologic Diagnostic Oncology Group 5 Study. The RDOG5 investigators. Cancer 1998;82:679–88.

82. Giard RW, Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer 1992;69:2104–10.

83. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. JAMA 1990; 263:2753–9.

84. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78–84.

85. Hobbs FD, Davis RC, Roalfe AK, Hare R, Davies MK, Kenkre JE. Reliability of N-terminal pro-brain natriuretic peptide assay in diagnosis of heart failure: cohort study in representative and high risk community populations. BMJ 2002;324:1498–500.

86. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Ann Intern Med 2003;138:40–4.

87. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Clin Chem 2003;49:1–6.