Special Reports

Quality of Reporting of Diagnostic Accuracy Studies[1]

Nynke Smidt, PhD
Anne W. S. Rutjes, MSc
Daniëlle A. W. M. van der Windt, PhD
Raymond W. J. G. Ostelo, PhD
Johannes B. Reitsma, MD, PhD
Patrick M. Bossuyt, PhD
Lex M. Bouter, PhD
Henrica C. W. de Vet, PhD

Published online before print
10.1148/radiol.2352040507
Radiology 2005; 235:347–353

Abbreviations:
MeSH = Medical Subject Headings
STARD = Standards for the Reporting of Diagnostic Accuracy

[1] From the Institute for Research in Extramural Medicine, VU University Medical Center, Van der Boechorststraat 7, 1081 BT Amsterdam, the Netherlands (N.S., D.A.W.M.v.d.W., R.W.J.G.O., L.M.B., H.C.W.d.V.); and Department of Clinical Epidemiology and Biostatistics, Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands (A.W.S.R., J.B.R., P.M.B.). Received March 12, 2004; revision requested May 21; revision received August 19; accepted October 1. Supported by grants from the Medical Sciences–Netherlands Organisation for Scientific Research (ZON-MW). **Address correspondence to** N.S. (e-mail: *n.smidt@vumc.nl*).

Authors stated no financial relationship to disclose.

Author contributions:
Guarantors of integrity of entire study, H.C.W.d.V., L.M.B.; study concepts and design, H.C.W.d.V., P.M.B., L.M.B., J.B.R.; literature research, N.S., A.W.S.R.; data acquisition, N.S., A.W.S.R.; data analysis/interpretation, N.S., H.C.W.d.V., A.W.S.R., R.W.J.G.O., D.A.W.M.v.d.W.; statistical analysis, N.S.; manuscript preparation and definition of intellectual content, N.S., H.C.W.d.V.; manuscript editing, N.S.; manuscript revision/review and final version approval, all authors

PURPOSE: To evaluate quality of reporting in diagnostic accuracy articles published in 2000 in journals with impact factor of at least 4 by using items of Standards for Reporting of Diagnostic Accuracy (STARD) statement published later in 2003.

MATERIALS AND METHODS: English-language articles on primary diagnostic accuracy studies in 2000 were identified with validated search strategy in MEDLINE. Articles published in journals with impact factor of 4 or higher that regularly publish articles on diagnostic accuracy were selected. Two independent reviewers evaluated quality of reporting by using STARD statement, which consists of 25 items and encourages use of a flow diagram. Total STARD score for each article was calculated by summing number of reported items. Subgroup analyses were performed for study design (case-control or cohort study) by using Student *t* tests for continuous outcomes and χ^2 tests for dichotomous outcomes.

RESULTS: Included were 124 articles published in 2000 in 12 journals: 33 case-control and 91 cohort studies. Only 41% of articles (51 of 124) reported on more than 50% of STARD items, while no articles reported on more than 80%. A flow chart was presented in two articles. Assessment of reporting on individual items of STARD statement revealed wide variation, with some items described in 11% of articles and others in 92%. Mean STARD score (0–25 points available) was 11.9 (range, 3.5–19.5). Mean difference in STARD score between cohort studies and case-control studies was 1.53 (95% confidence interval: 0.24, 2.82).

CONCLUSION: Quality of reporting in diagnostic accuracy articles published in 2000 is less than optimal, even in journals with high impact factor. Authors, editors, and reviewers should pay more attention to reporting by checking STARD statement items and including a flow diagram to represent study design and patient flow.

Supplemental material: *radiology.rsnajnls.org/cgi/content/full/2352040507/DC1*

Several systematic reviews have emphasized the poor quality of reporting in diagnostic accuracy studies (1–3). This poor reporting hampers an adequate judgment of both the internal and the external validity of a study. In 1995, Reid et al (4) evaluated the methodologic quality of 112 articles on diagnostic accuracy published in *Lancet, British Medical Journal, New England Journal of Medicine,* and *Journal of the American Medical Association (JAMA)* during the period of 1978–1993. On the basis of a set of seven methodologic standards, they concluded that the quality of the articles was poor. For example, only 8% of the articles included calculation of measures of diagnostic accuracy for relevant subgroups, and work-up bias was avoided in no more than 46% of the articles (4). The extent to which poor quality of reporting impeded the assessment of methodologic quality is unclear.

In 1999, Lijmer et al (1) demonstrated that case-control studies with healthy control subjects led to overestimation of diagnostic accuracy, compared with that in cohort studies. Furthermore, knowledge of the results of the index test and the use of clinical information about the study population when interpreting the reference standard resulted in an overestimation of diagnostic accuracy (1). Therefore, complete and accurate reporting is essential to judge the potential for bias and to assess the generalizability of results.

The first checklist for reporting of diagnostic accuracy studies was published by Bruns et

al (5) in October 2000. In January 2003, guidelines for reporting studies of diagnostic accuracy (the Standards for the Reporting of Diagnostic Accuracy, or STARD) were published simultaneously in eight medical journals (*Radiology, American Journal of Clinical Pathology, Annals of Internal Medicine, British Medical Journal, Clinical Biochemistry, Clinical Chemistry, Clinical Chemistry of Laboratory Medicine, and Lancet*) (6,7). Similar guidelines for the reporting of randomized controlled trials (the Consolidated Standards for Reporting of Trials, or CONSORT), systematic reviews (the Quality of Reporting of Meta-analyses, or QUORUM), and observational studies (the Meta-analysis of Observational Studies in Epidemiology, or MOOSE) already exist (8–10).

After publication of the CONSORT statement, Moher et al (11) evaluated the quality of reports of 211 randomized controlled trials published in *British Medical Journal, JAMA, Lancet,* and the *New England Journal of Medicine* by using the CONSORT checklist. They concluded that the use of the CONSORT statement is associated with improvements in the quality of reports of randomized controlled trials (11). The presentation of a flow diagram was also associated with improved quality of reporting of randomized controlled trials (12).

Although Reid et al (4) had pointed out the poor quality of reporting in the 1990s, it is possible that the reporting has improved in more recent articles. Therefore, this study was designed to evaluate the quality of reporting in articles on diagnostic accuracy published in 2000 in journals with an impact factor of at least 4 by using the items of the STARD statement published later in 2003.

MATERIALS AND METHODS

Data Sources

One reviewer (N.S.) searched MEDLINE with a validated strategy to identify articles on diagnostic accuracy, as follows: "sensitivity AND specificity.sh" OR "specificit* .tw" OR "false negative.tw" OR "accuracy.tw" (where ".sh" indicates subject heading and ".tw" indicates text word) (13). The search was subsequently limited to publications in 2000, articles published in English, and studies focusing on human subjects. The journals were ranked according to the number of publications retrieved. From the top 50 in that ranking, those with an impact factor of 4 or higher were selected. Only articles published in these journals were included in the evaluation.

Study Selection

Articles were included if they reported on primary studies of diagnostic accuracy, in which the results of one or more tests were compared with the findings obtained with a reference standard in the same study population. Two reviewers (N.S., A.W.S.R.) independently assessed the title, abstract, and keywords of all eligible articles to determine whether they met the inclusion criteria. If there was any doubt, the full text of the article was retrieved and read by both reviewers. Disagreements were discussed and resolved in a consensus meeting.

Data Extraction

The STARD statement was used to assess the quality of reporting. The statement contains a list of 25 items and encourages the use of a flow diagram to represent the design of the study and the flow of patients through the study (6,7). For this assessment, the reviewers had to determine whether each item of the checklist was described adequately in the text. Note that the reviewers were not evaluating the likelihood of bias but only the quality of reporting. Two reviewers independently evaluated the quality of reporting in the included articles. One reviewer (N.S.) assessed all articles, and four other reviewers (A.W.S.R., H.C.W.d.V., D.A.W.M.v.d.W., R.W.J.G.O.) each evaluated a quarter of all the articles. Disagreements were discussed and resolved in a consensus meeting. If consensus could not be reached, a third reviewer made the final decision.

Statistical Analysis

For each item in the STARD statement, the total number of articles reporting the elements mentioned in that item is presented. A total STARD score for each article was calculated by summing the number of reported items (0–25 points available). Higher scores indicated better quality of reporting. Equal weights were applied to each of the items. Six items (items 8, 9, 10, 11, 13, and 24) concern the index tests, as well as the reference standard. Weights for these items were assigned to both the index test (0.5 point) and the reference standard (0.5 point) and evaluated separately. The overall mean and standard deviation of the total STARD scores are presented.

Subgroup analyses were performed to compare the quality of reporting among different journals and designs (case-control and cohort studies). Cohort studies are characterized by selection of subjects who underwent the index test, whereas in case-control studies, the subjects are selected on the basis of the results of the reference standard (14). Student *t* tests (independent samples) were used to calculate mean differences between the total STARD score of case-control and cohort studies. In addition, χ^2 tests were used to calculate differences between the number of articles reporting the items of the STARD statement in case-control and cohort studies. If the assumptions of the χ^2 tests were not met, the Fisher exact test was used. Differences in total STARD scores between the 12 journals were calculated by means of pairwise comparisons (Tukey honestly significant difference test). *P* values of less than .05 were considered to indicate a statistically significant difference. Statistical analysis was performed (N.S.) by using SPSS for Windows (release 11.0.1; SPSS, Chicago, Ill).

RESULTS

Search and Selection

The search strategy resulted in the identification of 20 728 publications (Figure). All hits were grouped according to journal, and the number of publications for each journal was counted. Journals with an impact factor of at least 4 in the top 50 were *Annals of Internal Medicine, Archives of Internal Medicine, Archives of Neurology, British Medical Journal, Circulation, Clinical Chemistry, Gut, JAMA, Lancet, New England Journal of Medicine, Neurology,* and *Radiology* (Table 1). In these six general medical journals and six disease- or discipline-specific journals, the search strategy yielded 884 hits. On the basis of the title, abstract, and/or keywords, 219 articles were selected. As 46% (102 of 219) of the articles were published in *Radiology,* it was decided to limit the number of articles in this journal to 25 by selecting the first two articles published in this journal each month and the first three articles published in the December 2000 issue. The full text of the 142 selected articles was read by two independent reviewers. Subsequently, 18 articles were excluded because of a lack of reference standard ($n = 1$), no diagnostic research ($n = 13$), a letter to the editor instead of a full article ($n = 3$), and a mixture of human and animal research ($n = 1$). Finally, 124 articles fulfilled the selection criteria.

Article Characteristics

The 124 diagnostic articles consisted of 33 case-control studies and 91 cohort

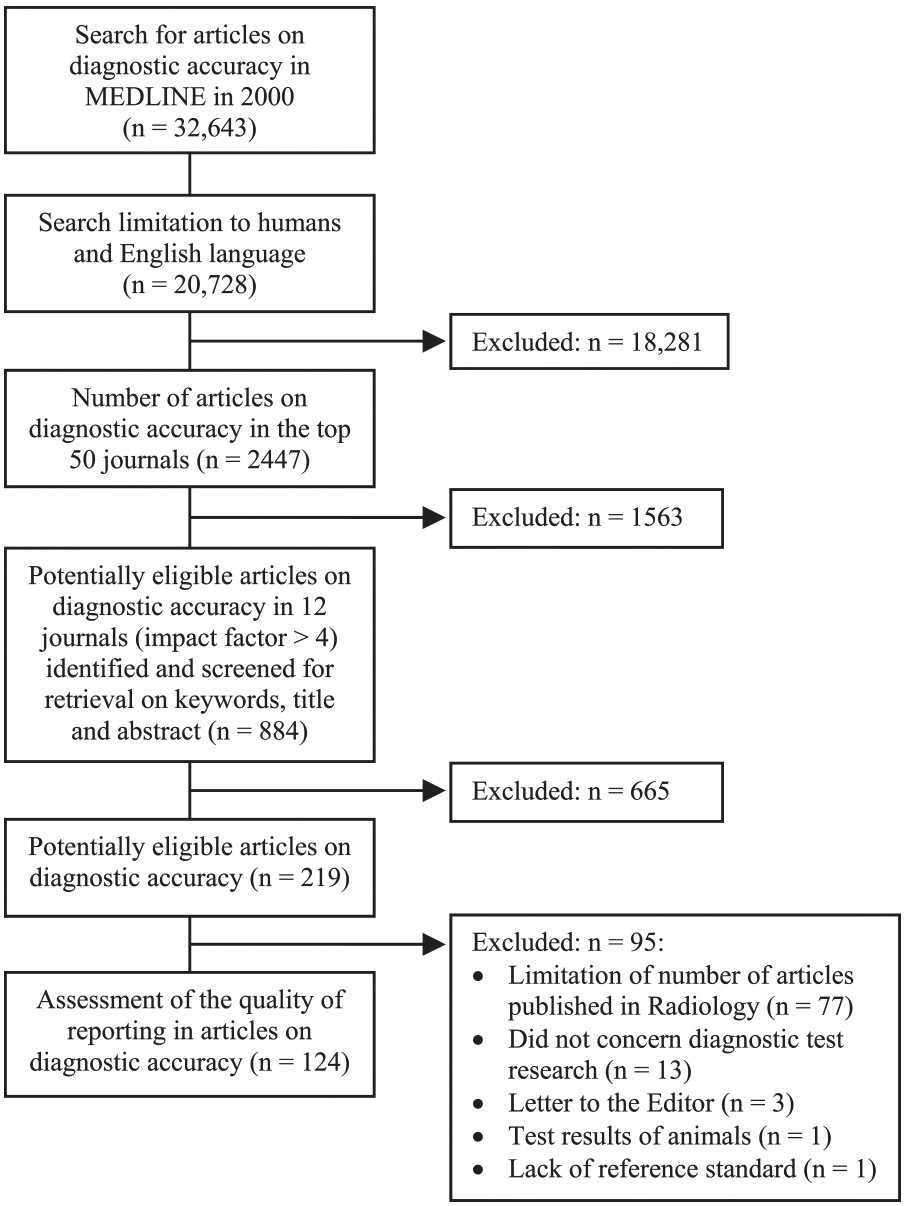

Diagram shows search and selection process of articles on diagnostic accuracy.

studies, including five reporting on population screening. Most articles (75%, 93 of 124) were published in disease- or discipline-specific medical journals, such as *Radiology, Neurology, Clinical Chemistry, Archives of Neurology, Archives of Internal Medicine,* and *Circulation*. Case-control studies were more often published in disease- or discipline-specific journals (30%, 28 of 93) than in general medical journals (16%, five of 31).

Quality of Reporting in Diagnostic Articles

Interrater agreement on the items of the STARD statement was good (overall agreement, 81.3%; κ statistic, 0.62). In six articles, disagreements between two reviewers could not be resolved, and the decision was made by one of the other reviewers. Most disagreements were caused by poor reporting of the design or doubts about the identity of the index and/or reference test. The time needed to perform the quality assessment was approximately 1 hour for each article.

Overall, the items of the STARD statement were poorly reported. The mean STARD score of the 124 articles was 11.9 (standard deviation, 3.3). Only 41% (51 of 124) of the articles reported more than 50% of the items (STARD score ≥ 12.5), and none of the them reported more than 80% (STARD score ≥ 20). A flow chart was reported in only two articles (2%). The quality of the reporting of the items of the STARD statement for each article separately is presented in the online Appendix E1 (*radiology.rsnajnls.org/cgi/content/full/2352040507/DC1*; for further information, contact N.S. at *n.smidt@vumc.nl*).

STARD Statement

The overall quality of the reporting of the items of the STARD statement in the articles is presented in Table 2. There is a broad variation in the quality of the reporting of these items (11%–92%). Poorly (<20%) reported items were *(a)* identification of the article as a study of diagnostic accuracy (item 1), *(b)* methods used for calculating or comparing measures of diagnostic accuracy (item 12), *(c)* methods used for calculating test reproducibility (item 13), *(d)* adverse events from performing the test(s) (item 20), and *(e)* estimates of test reproducibility of the reference standard (item 24b). The best reported item was discussion of the clinical applicability of the study findings (item 25). For each section (title, abstract, and keywords; introduction; methods; results; and discussion) of the STARD statement, the most remarkable findings are discussed as follows.

Title, abstract, and keywords (item 1).— To identify articles on diagnostic accuracy (item 1), keywords such as *sensitivity and specificity* or *diagnostic accuracy* would improve and simplify the search and the selection of articles on diagnostic accuracy. Only four of the 12 journals (*Circulation, Gut, Neurology,* and *Radiology*) presented keywords in the article itself. No more than two (3%) of the 71 articles published in these journals used the keywords *sensitivity and specificity* or *diagnostic accuracy*. Furthermore, less than 3% (three of 124) of all articles mentioned the words *diagnostic accuracy* in the title, and only 9% (11 of 124) mentioned them in the abstract.

The STARD statement recommends the use of the Medical Subject Headings (MeSH) term *sensitivity and specificity*. In this search, 686 (78%) of the 884 articles were identified by this MeSH term. However, only 100 of the 686 articles actually concerned a diagnostic accuracy study (positive predictive value, 15%). Nevertheless, the sensitivity of this search term was high, with 81% (100 of 124) of the included articles being identified correctly in MEDLINE.

Introduction (item 2).—In 90% of all articles (112 of 124), the research question became clear after reading the abstract and introduction (item 2). However, information regarding the index tests, the reference standard, and the target condition was scattered throughout the text. Only 32% of the articles (40 of 124) mentioned the index test, the reference standard, and the target condition in their research question. In many articles, the reference standard was lacking in the formulation of the research question (64%, 79 of 124).

Methods (items 3–13).—Only 28% of all articles (35 of 124) reported the inclusion and exclusion criteria, the setting, and the location where the data were collected (item 3). This low percentage was mainly due to the absence of exclusion criteria (69 of the 124 articles [56%]). The inclusion criteria were relatively well reported (108 of 124; 87%), but only 56% of the articles (70 of 124) reported how patients were selected (item 5). A consecutive series of patients was apparently included in 36% of the studies [45 of 124]). The reference standard and its rationale were reported clearly in 57% of the articles (item 7). In 40% of the articles (50 of 124), only the reference standard was reported, while in four articles (3%), the identity of the reference standard remained unclear. Information concerning the index test was better reported than that for the reference standard (items 8–13 and 24). In particular, information regarding the number and training of the persons executing and evaluating the reference test(s) and the blinding of the readers to the tests was reported poorly (items 10 and 11).

Only 37% of the articles (46 of 124) clearly reported whether the results of the reference standard and clinical information about the study population were given to the readers of the index test (item 11a). In most articles (62%, 77 of 124), information regarding the revelation of clinical information about the study population to the readers of the index test was lacking. If it was reported clearly that the index test was performed before the reference test, we assume that the readers of the index test had been blinded to the results of the reference test. Information regarding the revelation of the results of the index test, other tests, or clinical information about the study population to the readers of the reference standard was reported in only 18% of the articles (23 of 124) (item 11b).

The methods for calculating measures of diagnostic accuracy, such as sensitivity, specificity, likelihood ratios, diagnostic odds ratios, and receiver operating characteristic curves, were reported in 65% of the articles (81 of 124). Only 14% of the articles (17 of 124) adequately reported the statistical methods used to calculate measures of diagnostic accuracy, particularly with regard to the quantification of estimates of the diagnostic accuracy (eg, 95% confidence limits, item 12). Methods used to study the reproducibility of the index test and the reference standard were reported poorly, by only 16% (20 of 124) and 5% (six of 124) of the articles, respectively (item 13). Six articles (5%) referred to previous research on the reproducibility of the test(s).

Results (items 14–24).—Clinical and demographic characteristics, such as age and sex of the study population and the spectrum of the symptoms at presentation, were reported clearly in 52% of the articles (65 of 124, item 15). Less frequently reported clinical characteristics were co-morbidity (20 of 124, 16%) and current treatments (33 of 124, 27%).

Eighty-three percent of the articles (103 of 124) reported the number of par-

TABLE 1
Top 50 Journals That Frequently Publish Articles on Diagnostic Accuracy

Ranking	Journal Name	Number of Hits	Impact Factor in 2000*
1	Radiology	298	4.1
2	J Clin Microbiol	241	3.5
3	Am J Roentgenol	127	1.9
4	Am J Cardiology	124	2.8
5	Cancer	114	3.6
6	Neurology	102	4.8
7	Crit Care Med	93	3.8
8	Clin Chem	84	4.3
9	J Urol	84	2.9
10	Circulation	83	10.9
11	Lancet	81	10.2
12	Chest	71	2.5
13	Obstetrics and Gynecology	60	2.1
14	BMJ	59	5.3
15	New England Journal of Medicine	44	29.5
16	Br J Radiol	38	1.0
17	Pediatrics	38	3.7
18	Clin Radiol	38	0.9
19	Archives of Internal Medicine	35	6.1
20	Annals of Emergency Medicine	34	2.2
21	Scandinavian Journal of Gastroenterology	32	1.8
22	Archives of Neurology	31	4.4
23	J Clin Epidemiol	31	2.1
24	Radiotherapy and Oncology	30	2.5
25	Ann Internal Medicine	29	9.8
26	Arch Pathol Lab Med	29	1.4
27	Archives of Pathology and Laboratory Medicine	29	1.4
28	Annals of Oncology	26	3.2
29	Acad Radiol	26	0.9
30	Oncology	24	2.6
31	Gut	24	5.4
32	Archives of Disease in Childhood	23	1.9
33	Annals of Rheumatic Diseases	22	2.4
34	Arch Phys Med Rehab	22	1.4
35	Archives of Surgery	21	2.6
36	Ophthalmology	21	3.0
37	European Journal of Cancer	21	2.7
38	Medical Journal of Australia	18	1.9
39	Cardiology	15	0.7
40	Australian and New Zealand Journal of Surgery	15	0.6
41	British Journal of Surgery	14	2.9
42	Scand J Clin Lab Invest	14	1.1
43	JAMA	14	15.4
44	Am Fam Physician	14	0.9
46	Archives of Dermatology	14	3.3
47	British Journal of Ophthalmology	14	1.9
48	Br J General Pract	11	1.6
49	Am J Phys Med Rehab	8	0.9
50	Baillieres Best Pract Res Clin Obst Gyn	7	0.9

* According to *www.jcrweb.com.*

TABLE 2
Report of Items of the STARD Statement in 124 Articles Retrieved

Category and Item No.	All Articles (n = 124)*	Cohort Studies (n = 91)†	Case-Control Studies (n = 33)‡
Title, abstract, and keywords			
1. Identification of the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity").	13 (10)	9 (10)	4 (12)
Introduction			
2. Statement of research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	112 (90)	83 (91)	29 (88)
Methods			
3. Study population: Inclusion and exclusion criteria, setting, and locations where data were collected.	35 (28)	28 (31)	7 (21)
4. Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had undergone the index tests or the reference standard?	103 (83)	76 (84)	27 (82)
5. Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.	70 (56)	58 (64)	12 (36)
6. Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	99 (80)	76 (84)	23 (70)
7. The reference standard and its rationale.	70 (56)	51 (56)	19 (58)
8. Technical specifications of material and methods involved, including how and when measurements were taken, and/or citation of references for			
(a) index tests and	115 (93)	83 (91)	32 (97)
(b) reference standard.	83 (67)	62 (68)	21 (64)
9. Definition of and rationale for the units, cutoffs, and/or categories of the results of the			
(a) index tests and the	103 (83)	77 (85)	26 (79)
(b) reference standard.	75 (60)	60 (66)	15 (45)
10. The number, training, and expertise of the persons executing and evaluating the			
(a) index tests and the	51 (41)	42 (46)	9 (27)
(b) reference standard.	32 (26)	29 (32)	3 (9)
11. Whether the readers of the			
(a) index tests and	46 (37)	31 (34)	15 (45)
(b) reference standard	23 (18)	18 (20)	5 (15)
were blind (masked) to the results of the other test and description of any other clinical information available to the readers.			
12. Methods for calculating or comparing measures of diagnostic accuracy and statistical methods used to quantify uncertainty (eg, 95% confidence intervals).	17 (14)	13 (14)	4 (12)
13. Methods for calculating test reproducibility, if done			
(a) for the index test and	20 (16)	14 (15)	6 (18)
(b) for the reference standard.	6 (5)	6 (6)	0 (0)
Results			
14. When study was performed, including beginning and end dates of recruitment.	60 (48)	48 (53)	12 (36)
15. Clinical and demographic characteristics of the study population (at least information on age, sex, and spectrum of presenting symptoms).	65 (52)	46 (50)	19 (58)
16. Number of participants satisfying the criteria for inclusion who did or did not undergo index tests and/or reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended).	75 (60)	58 (64)	17 (52)
17. Time interval between index tests and reference standard and any treatment administered in between.	33 (27)	28 (31)	5 (15)
18. Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.	28 (22)	18 (20)	10 (30)
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.	104 (84)	76 (84)	28 (85)
20. Any adverse events from performing the index tests or the reference standard.	21 (17)	19 (21)	2 (6)
21. Estimates of diagnostic accuracy and measures of statistical uncertainty (eg, 95% confidence intervals).	40 (32)	32 (35)	8 (24)
22. How indeterminate results, missing data, and outliers of the index tests were handled.	73 (59)	60 (66)	13 (39)
23. Estimates of variability of diagnostic accuracy between subgroups of participants, readers, or centers, if done.	48 (39)	36 (40)	12 (36)
24. Estimates of test reproducibility, if done			
(a) for the index test and	40 (32)	25 (27)	15 (45)
(b) for the reference standard.	8 (6)	7 (8)	1 (3)
Discussion			
25. Discussion of the clinical applicability of the study findings.	114 (92)	82 (90)	32 (97)

Note.—Data are number of articles. Numbers in parentheses are percentages.
* Mean STARD score, 11.9 ± 3.3. Range, 3.5–19.5.
† Mean STARD score, 12.4 ± 3.0. Range, 3.5–19.5.
‡ Mean STARD score, 10.8 ± 3.7. Range, 4.5–19.0.

TABLE 3
Quality of Reporting of Articles in 12 High-Impact Journals

Journal Name	Impact Factor*	No. of Articles (n = 124)	Cohort Study (n = 91)	Case-Control Study (n = 33)	Mean STARD Score ± SD†
New England Journal of Medicine	29.5	7	7	0	14.3 ± 2.7
JAMA	15.4	4	4	0	15.5 ± 2.3
Circulation	10.9	13	11	2	10.3 ± 3.6
Lancet	10.2	9	7	2	12.4 ± 3.5
Annals of Internal Medicine	9.8	3	2	1	13.2 ± 1.3
Archives of Internal Medicine	6.1	6	4	2	11.3 ± 3.6
Gut	5.4	13	11	2	12.7 ± 3.1
British Medical Journal	5.3	2	2	0	9.8 ± 2.5
Neurology	4.8	20	8	12	10.8 ± 3.6
Archives of Neurology	4.4	7	4	3	12.3 ± 2.7
Clinical Chemistry	4.3	15	9	6	10.0 ± 3.0
Radiology	4.1	25	22	3	13.2 ± 2.3

Note.—Data are number of articles, unless specified otherwise.
* In 2000, according to *www.jcrweb.com.*
† SD = standard deviation. Each item was given equal weight (0–25 points available).

ticipants who met the inclusion criteria and those who did or did not undergo the index test and reference standard. Seventy-five (60%) articles explained why participants failed to undergo one or more of the tests (item 16). In 43 of the 75 articles, however, none of the participants failed to undergo the index test or reference standard. A flow diagram, describing the design of the study and the number of participants, was presented in only two articles (2%).

Information about the time interval between the index test and the reference standard and about the treatment administered between the tests was given in 33 (27%) articles (item 17). Twenty-two of these 33 articles did not report on the treatment between the tests, but the time interval between the tests was so small that treatment could not have affected the results of the second test.

Although 109 of 124 articles reported estimates of diagnostic accuracy (eg, sensitivity and specificity), 29 of these gave no information about the number of true-positive, true-negative, false-positive, and false-negative findings. Thirty-two percent of the articles (40 of 124) reported statistical uncertainty (ie, 95% confidence intervals) for the measures of diagnostic accuracy (item 21).

Discussion (item 25).—Most articles (114 of 124, 92%) discussed the clinical applicability of the study findings. In addition to scoring the items of the STARD statement, the reviewers were asked to compose a 2 × 2 table for each article. This was possible for 73% of the articles (91 of 124). However, true-positive and true-negative findings often had to be deduced from the results of sensitivity and specificity, which implied that the number of indeterminate or missing results had to be ignored in the reconstruction of the 2 × 2 table.

Subgroup Analysis

Results of subgroup analyses showed that the quality of reporting for case control studies was not as good as that for cohort studies (Table 2). The mean STARD score ± standard deviation was 12.4 ± 3.0 for the 91 cohort studies and 10.8 ± 3.7 for the case-control studies. The mean difference in STARD score between cohort studies and case-control studies was 1.5 (95% confidence interval: 0.2, 2.8). Large differences (≥15%) in the quality of reporting between cohort and case-control studies were found for the following items: *(a)* participant sampling (item 5); *(b)* definition of and rationale for the units, cutoffs, and/or categories of the results of the reference standard (item 9b); *(c)* the number, training, and expertise of the persons executing and evaluating the tests (items 10a and 10b); *(d)* recruitment period (item 14); *(e)* time interval between the index tests and the reference standard and any treatment administered between the tests (item 17); *(f)* adverse events of the tests (item 20); *(g)* how indeterminate results, missing data, and outliers of the index tests were handled (item 22); and *(h)* estimates of reproducibility of the index test (item 24a). Statistically significant differences ($P < .05$) between case-control and cohort studies were found for the following items: participant sampling (item 5); number, training, and expertise of the persons executing and evaluating the reference standard (item 10b); and the handling of indeterminate results (item 22). Only 27% of the case-control studies (nine of 33) adequately reported on at least 50% of the items, while 46% of the cohort studies (42 of 91) reported on more than 50% of the items.

Mean STARD score and standard deviations are presented for each journal in Table 3. The mean STARD score varied from 9.8 in the *British Medical Journal* to 15.5 in *JAMA*. However, none of the pairwise comparisons were statistically significant.

DISCUSSION

The results of this study indicate that the quality of reporting in articles on diagnostic accuracy published in 2000 is disappointingly poor, even in journals with a high impact factor. Only 41% of the articles adequately reported on at least 50% of the items, and none of the articles provided information on more than 80% of the STARD items. The mean STARD score (out of 25 available points) of the 124 articles was 11.9 ± 3.3. The advantage of using an overall score is its simplicity, but an overview of specific items that are poorly reported—and therefore need improvement—is, in our opinion, more important. Therefore, we elaborated in detail on these individual items.

First, we strongly recommend the use of a flow diagram, because for most of the articles, the reviewers had to spend a considerable amount of time identifying the index test and the reference standard, the sequences of performing these tests, and the number of patients who underwent each test. Second, accurate identification of articles on diagnostic accuracy in the literature is important, and therefore, the

use of uniform terms (MeSH headings) in keywords, titles, or abstracts is important. Just as clinical trials are labeled as a specific type of publication in MEDLINE (PubMed), studies on diagnostic accuracy should also be labeled as a specific type of publication. The STARD group proposed systematic use of the MeSH term *sensitivity and specificity,* because this is indicative of a study on diagnostic accuracy and is a term that has been used frequently in the past. Moons and Harrell (15) suggested use of the term *posttest probability,* because studies on diagnostic accuracy do not necessarily have to determine sensitivity and specificity. However, posttest probability is not yet registered as a MeSH term. We recommend the use of *diagnostic accuracy* as publication type, and *posttest probability* should be included as a new MeSH term, in addition to *sensitivity and specificity*.

The STARD statement focuses on the quality of reporting, not the methodologic quality of a diagnostic study. For example, if the authors stated that the reviewers of the reference standard were not blinded to the results of the index test, we considered item 11 to be well reported, even though this indicates a potential methodologic shortcoming. We believe that there is a positive association between the methodologic quality of a study and the quality of reporting. It is easier to report on a well-performed study than on a study that was poorly designed or in which a large number of protocol deviations occurred. Moreover, in the latter case, the authors may be less inclined to report in detail what happened. Increased attention to the quality of reporting and strict requirements for reporting in journals might, in the long term, thus also improve the methodologic quality of diagnostic research.

Lijmer et al (1) showed that various methodologic characteristics of a diagnostic study might influence the results of diagnostic accuracy. Their analysis was hampered by the poor reporting in many studies. Improved reporting may lead to better estimation of the influence of methodologic characteristics on diagnostic accuracy. Moreover, better estimates of biases or sources of variation within diagnostic studies can be made if all STARD items are reported. The STARD guidelines are not the first to focus on the reporting of studies. CONSORT, QUOROM, and MOOSE have emphasized the importance of better reporting of other study designs (8–10).

The quality of reporting in articles on diagnostic accuracy is of great importance for assessing the generalizability of the results. It is also essential for the detection of methodologic flaws, the recalculation of sensitivity and specificity, repetition of the study, and application of the results in clinical practice. Fortunately, a number of journals have already changed their instructions to authors and require authors to complete the STARD checklist and to include a flow diagram that represents the design of the study and the flow of patients.

Our study has a few limitations. First, the identification of studies of diagnostic accuracy is difficult. We searched MEDLINE by using a validated search strategy to identify all studies on diagnostic accuracy published in 2000. However, the search strategy has a sensitivity of 80.0% and a specificity of 97.3% (13). Therefore, we may have missed studies on diagnostic accuracy that were not identified with our search strategy.

Second, the generalizability of the results of this study may be questioned. We evaluated the quality of reporting of studies on diagnostic accuracy published in 2000 in 12 journals. For this purpose, journals were selected if they occurred in the top-50 ranking of journals that frequently publish articles on diagnostic accuracy and if they had an impact factor of at least 4. However, it remains unclear whether results would be similar for journals that only rarely publish diagnostic accuracy studies or for journals with an impact factor of less than 4.

Furthermore, as almost 50% of all identified articles on diagnostic accuracy were published in *Radiology,* we decided to limit the number of articles published in *Radiology* to 25. As the quality of reporting could have been improved during the year, we selected the first two articles of each month and the first three articles published in the December 2000 issue. In our opinion, the quality of reporting of those articles not selected for the review will be similar to the selected articles.

We strongly recommend that authors, editors, and reviewers use the STARD statement for preparing, writing, and reviewing articles on diagnostic accuracy. We also stress that special attention should be paid to the identification of the article as a work pertaining to diagnostic accuracy and that a flow diagram should be included to represent the design of the study and the flow of patients. Hopefully this will lead to an improvement in the quality of reporting in the near future.

References

1. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282:1061–1066.
2. Scholten RJ, Deville WL, Opstelten W, Bijl D, van der Plas CG, Bouter LM. The accuracy of physical diagnostic tests for assessing meniscal lesions of the knee: a meta-analysis. J Fam Pract 2001; 50:938–944.
3. Devillé WL, Van der Windt DA, Daferagi A, Bezemer PD, Bouter LM. The test of Lasegue: systematic review of the accuracy in diagnosing herniated discs. Spine 2000; 25:1140–1147.
4. 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.
5. Bruns DE, Huth EJ, Magid E, Young DS. Towards a checklist for reporting of studies of diagnostic accuracy of medical tests. Clin Chem 2000; 46:893–895.
6. Bossuyt PM, Reitsma JB, Bruns DE, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clin Chem 2003; 49:7–18.
7. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Radiology 2003; 226: 24–28.
8. Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. BMJ 1996; 313:570–571.
9. Moher D, Cook DJ, Eastwood S, et al, for the QUOROM group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet 1999; 354:1896–1900.
10. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis of observational studies in epidemiology: a proposal for reporting. JAMA 2000; 283:2008–2012.
11. Moher D, Jones A, Lepage L, for the CONSORT group. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. JAMA 2001; 285:1992–1995.
12. Egger M, Jüni P, Bartlett C. Value of flow diagrams in reports of randomised controlled trials. JAMA 2001; 285:1996–1999.
13. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000; 53:65–69.
14. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003; 56:1118–1128.
15. Moons KG, Harrell FE. Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. Acad Radiol 2003; 10:670–671.