HIGHLIGHT CME

The quality of diagnostic accuracy studies since the STARD statement

Has it improved?

N. Smidt, PhD; A.W.S. Rutjes, PhD; D.A.W.M. van der Windt, PhD; R.W.J.G. Ostelo, PhD;
P.M. Bossuyt, PhD; J.B. Reitsma, PhD; L.M. Bouter, PhD; and H.C.W. de Vet, PhD

Abstract—*Objective:* To assess whether the quality of reporting of diagnostic accuracy studies has improved since the publication of the Standards for the Reporting of Diagnostic Accuracy studies (STARD statement). *Methods:* The quality of reporting of diagnostic accuracy studies published in 12 medical journals in 2000 (pre-STARD) and 2004 (post-STARD) was evaluated by two reviewers independently. For each article, the number of reported STARD items was counted (range 0 to 25). Differences in completeness of reporting between articles published in 2000 and 2004 were analyzed, using multilevel analyses. *Results:* We included 124 articles published in 2000 and 141 articles published in 2004. Mean number of reported STARD items was 11.9 (range 3.5 to 19.5) in 2000 and 13.6 (range 4.0 to 21.0) in 2004, an increase of 1.81 items (95% CI: 0.61 to 3.01). Articles published in 2004 reported the following significantly more often: methods for calculating test reproducibility of the index test (16% vs 35%); distribution of the severity of disease and other diagnoses (23% vs 53%); estimates of variability of diagnostic accuracy between subgroups (39% vs 60%); and a flow diagram (2% vs 12%). *Conclusions:* The quality of reporting of diagnostic accuracy studies has improved slightly over time, without a more pronounced effect in journals that adopted the STARD statement. As there is still room for improvement, editors should mention the use of the STARD statement as a requirement in their guidelines for authors, and instruct reviewers to check the STARD items. Authors should include a flow diagram in their manuscript.

NEUROLOGY 2006;67:792–797

Many authors have emphasized the poor quality of research reports, which hampers an adequate judgment of the validity of a study.[1-3] Several groups have developed guidelines to improve the reporting of randomized controlled trials (CONSORT), diagnostic accuracy studies (STARD), observational studies (STROBE), systematic reviews of randomized controlled trials (QUOROM), and of observational studies (MOOSE).[4-11]

In January 2003, Standards for the Reporting of Diagnostic Accuracy studies (STARD statement) were published simultaneously in eight medical journals.[7,8] The STARD statement contains a checklist of 25 recommended items and encourages the use of a flow diagram to represent the design of the study and the flow of patients through the study.[7,8]

Many authors have evaluated the quality of reporting of diagnostic accuracy studies published before 2003 with the STARD checklist.[12-14] The quality of reporting varied from 6% for reporting the estimates of test reproducibility to 100% for discussion of the clinical applicability and research question.[12,14] Two studies found that around 40% of the studies reported on more than half of the STARD items.[12,14]

Our main objective is to examine whether the quality of reporting in diagnostic accuracy studies was improved since the publication of the STARD statement. Therefore we compared the quality of reporting of diagnostic accuracy studies published in journals with an impact factor of at least 4, before (in 2000)[12] and after (in 2004) the publication of the STARD statement. In addition, we compare the improvement in the quality of reporting in studies published in journals adopting the STARD statement vs non-adopting journals.

Methods. *Data sources.* One reviewer (N.S.) searched MEDLINE and used a validated strategy ([Sensitivity AND specificity.sh] OR [Specificit*.tw] OR [False negative.tw] OR [Accuracy.tw])[15] to identify articles on diagnostic accuracy published in six general medical journals (*Annals of Internal Medi-*

Editorial, see page 740

From the Institute for Research in Extramural Medicine (N.S., D.A.W.M.v.d.W., R.W.J.G.O., L.M.B., H.C.W.d.V.), VU University Medical Center, Amsterdam; and Department of Clinical Epidemiology & Biostatistics (A.W.S.R., P.M.B., J.B.R.), Academic Medical Center, University of Amsterdam, The Netherlands.

Supported by grant from The Netherlands Organisation for Health Research and Development (ZonMw).

Disclosure: The funding agency did not play a role in the design and conduct of the study, collection, management, analysis, and interpretation of the data, nor in the preparation, review, or approval of the manuscript. The authors report no conflicts of interest.

Received February 15, 2006. Accepted in final form June 28, 2006.

Address correspondence and reprint requests to Dr. Nynke Smidt, Department of Clinical Epidemiology & Biostatistics, Academic Medical Center, University of Amsterdam, PO Box 22700, 1100 DE Amsterdam, The Netherlands; e-mail: n.smidt@amc.uva.nl

cine, Archives of Internal Medicine, BMJ, JAMA, Lancet, New England Journal of Medicine) and six disease-specific or discipline-specific journals (*Archives of Neurology, Clinical Chemistry, Circulation, Gut, Neurology, Radiology*). The selection of these journals was based on the number of diagnostic accuracy studies published in 2000 and their impact factor (≥4).[12] The search was limited to studies focusing on human subjects and articles published in 2000 and 2004.

Study selection. Articles were included if 1) they were published in 2000 or 2004 in one of the 12 selected journals, 2) they concern diagnostic test research, 3) they were a primary study of diagnostic accuracy, in which the results of one or more tests were compared with the findings obtained with a reference standard, 4) they investigated a clinical population (no healthy volunteers or animals). Letters, editorials, abstracts, or technical briefs were excluded. Two reviewers (N.S., A.R.) independently assessed the title, abstract, and keywords of all potentially 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 25 items of the STARD statement were used to assess the quality of reporting.[7,8] For this assessment, the reviewers had to determine whether each item of the checklist was adequately described in the text. Reviewers were not expected to evaluate the likelihood of bias but only the quality of reporting.

The evaluation of the quality of reporting of diagnostic accuracy studies published in 2000 was carried out in spring 2003 (between March 2003 and May 2003). Between October 2004 and March 2005 the quality of reporting of studies published in 2004 was assessed. Two reviewers independently evaluated the included articles. Note that the reviewers were not blinded to the source (year of publication, journal, authors) of the articles. One reviewer (N.S.) assessed all articles and four other reviewers (A.R., H.V., D.W., and R.O.) each evaluated one fourth of all the articles published in 2000 and 2004. Disagreements between two reviewers were discussed and resolved in a 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 was calculated for 2000 and 2004. For each article, the total number of reported STARD items was counted (range 0 to 25), as indication of the quality of reporting. As six items (items 8, 9, 10, 11, 13, 24) concern the index test(s) as well as the reference standard, we counted the index test as ½ item and the reference standard as ½ item. The overall mean and SD of the total number of reported STARD items are presented.

Differences in reporting between studies published in 2000 and 2004 were analyzed for each item using logistic multilevel analyses, taking journal level effects into account. Using a linear multilevel analysis, differences in the number of reported STARD items between studies published in 2000 and 2004 were calculated. We also determined the effects of the use of the STARD statement in the editorial process of journals (adopters) on the quality of reporting of the individual items and on the total number of reported STARD items. In addition, the influence of the design (case control vs cohort) on the improvement in the quality of reporting was assessed.

p Values less than 0.05 were considered significant. Data entry using SPSS for Windows (Release 11.0.1, 2001) and statistical analysis using MLwiN (1.10, 2001) were done by N.S.

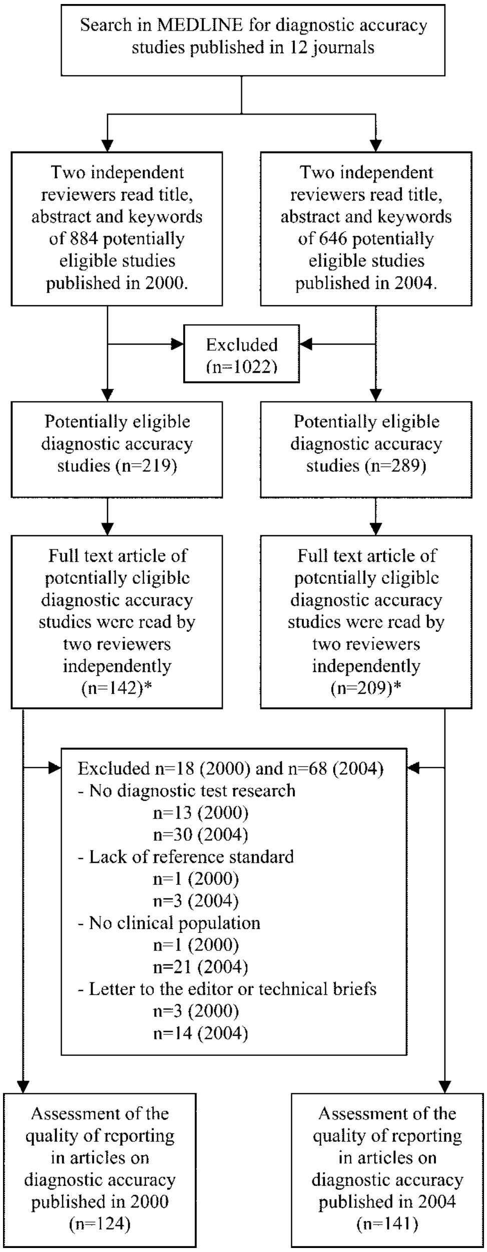
Search in MEDLINE for diagnostic accuracy studies published in 12 journals
Two independent reviewers read title, abstract and keywords of 884 potentially eligible studies published in 2000.
Two independent reviewers read title, abstract and keywords of 646 potentially eligible studies published in 2004.
Excluded (n=1022)
Potentially eligible diagnostic accuracy studies (n=219)
Potentially eligible diagnostic accuracy studies (n=289)
Full text article of potentially eligible diagnostic accuracy studies were read by two reviewers independently (n=142)*
Full text article of potentially eligible diagnostic accuracy studies were read by two reviewers independently (n=209)*
Excluded n=18 (2000) and n=68 (2004)
- No diagnostic test research
n=13 (2000)
n=30 (2004)
- Lack of reference standard
n=1 (2000)
n=3 (2004)
- No clinical population
n=1 (2000)
n=21 (2004)
- Letter to the editor or technical briefs
n=3 (2000)
n=14 (2004)
Assessment of the quality of reporting in articles on diagnostic accuracy published in 2000 (n=124)
Assessment of the quality of reporting in articles on diagnostic accuracy published in 2004 (n=141)

*Figure 1. Flow diagram of search and selection process of diagnostic accuracy articles published in 2000 and 2004. *Due to large number of articles published in Radiology (n = 102 in 2000 and n = 108 in 2004), the assessment of the full text articles published in this journal was limited to one fourth of the total number of potentially eligible publications in Radiology (n = 25 in 2000 and n = 28 in 2004).*

Results. *Search and selection.* Figure 1 presents the search and selection process of diagnostic accuracy studies published in 2000 and 2004 in the journals at issue. In these 12 journals, the search strategy identified 884 hits in 2000 and 646 hits in 2004. Based on the title, abstract, and keywords, a total of 508 articles were independently selected by two reviewers (N.S., A.R.). As a large number of articles were published in *Radiology*, we decided to limit the number of articles in this journal to one fourth of the total number of potentially eligible publications in *Radiology*, which resulted in 25 articles for 2000 and 27 for 2004. All potentially eligible articles published in *Radiology* were ranked according their publication date. Subsequently, at least two articles were selected randomly from each month. Two independent reviewers read the full text of the 351 articles and excluded 86 articles for the following reasons: no diagnostic test research (n = 43), lack of reference standard (n = 4), no clinical population (n = 22), no full text article, but a letter to the Editor or technical brief (n = 17). In total, 124 articles published in 2000 and 141 articles published in 2004 were included.

Article characteristics. The percentage cohort studies and case control studies published in 2000 and 2004 were

Table 1 *Number and characteristics of diagnostic accuracy articles published in 2000 and 2004 in 12 medical journals with high impact factor (>4)*

	Diagnostic accuracy studies published in 2000 (n = 124)					Diagnostic accuracy studies published in 2004 (n = 141)				
Journal	Impact factor*	No. of articles	Flow diagram included	Cohort (n = 91)	Case control (n = 33)	Impact factor†	No. of articles	Flow diagram included	Cohort (n = 96)	Case control (n = 45)
Adopter‡		78	1	54	24		95	14	61	34
Journal of the American Medical Association	15.4	4	—	4	—	24.8	9	2	8	1
Lancet	10.2	9	—	7	2	21.7	5	2	2	3
Annals of Internal Medicine	9.8	3	—	2	1	13.1	6	5	6	—
British Medical Journal	5.3	2	—	2	—	7.0	3	1	3	—
Neurology	4.8	20	—	8	12	6.0	21	3	9	12
Clinical Chemistry	4.3	15	—	9	6	6.5	24	—	6	18
Radiology	4.1	25	1	22	3	5.1	27	1	27	—
Nonadopter		46	1	37	9		46	3	35	11
New England Journal of Medicine	29.5	7	—	7	—	38.6	3	1	3	—
Circulation	10.9	13	1	11	2	12.6	25	1	19	6
Archives of Internal Medicine	6.1	6	—	4	2	7.5	4	—	4	—
Gut	5.4	13	—	11	2	6.6	7	1	6	1
Archives of Neurology	4.4	7	—	4	3	4.8	7	—	3	4

* Impact factor in 2000 according to www.jcrweb.com.

† Impact factor in 2004 according to www.jcrweb.com.

‡ Adopter of the Standards for the Reporting of Diagnostic Accuracy studies (STARD statement) before January 1, 2004.

similar (table 1). Seventy-three percent (91/124) of the diagnostic articles published in 2000 were cohort studies, including five reporting on population screening. From the 124 studies that were published in 2004, 96 (68%) were cohort studies with 14 studies concerning screening in the general population.

Most diagnostic accuracy articles had been published in disease-specific or discipline-specific journals, such as *Circulation, Clinical Chemistry, Neurology,* and *Radiology*.

In most studies, the diagnostic accuracy of one or more imaging tests (56%) or laboratory tests (35%) were examined. In less than 10% of the studies, the diagnostic value of history taking, questionnaires, or physical examination was investigated.

Journal characteristics. After the publication of the STARD statement in January 2003, the *Annals of Internal Medicine, BMJ, Clinical Chemistry, JAMA, The Lancet, Neurology*, and *Radiology* mentioned the STARD statement in their instructions for authors. These journals were considered as adopting journals. Although all adopting journals advise their authors to follow the STARD guidelines, there was a broad variation in the clearness and strictness in the language of their expectations. For example, *Neurology* requires authors to submit a checklist (for review purposes) and a flow diagram (for publication if the article is accepted),[16] whereas the *Lancet* only states that studies of diagnostic accuracy should be reported according to STARD guidelines.

The other journals did not mention the STARD statement in their instructions for authors and were therefore considered as non-adopting journals. In the summer of 2004, *Gut* joined the BMJ Publishing Group and subsequently changed their guidelines for authors and adopted the STARD statement. As the statement had not been used in the editorial process of articles published in 2004 (personal communication), *Gut* was considered to be a non-adopting journal for the current analysis.

Quality of reporting. Reviewing procedure. The inter-reviewer agreement on the items of the STARD statement was good (overall agreement 81%, Kappa statistics 0.62). In 15 articles (6%), disagreements between two reviewers could not be resolved and the decision was made by one of the other reviewers. Doubts about identity of the index and reference test and poor reporting of the design caused most disagreements. The average time needed to complete the assessment of the quality of reporting of one article was 47 (range 23 to 83) minutes.

Individual STARD items. The quality of reporting of the individual items of the STARD statement is presented in table 2. There was large variation in the quality of reporting between individual items, varying from 6% (item 24b and item 13b) to 98% (item 25).

The results of the logistic multilevel analysis showed that seven items were significantly more often reported in studies published in 2004, including item 4 (OR = 4.2 [95% CI: 1.0 to 17.4]), item 5 (OR = 2.75 [95% CI: 1.0 to 7.4]), item 9a (OR = 3.0 [95% CI: 1.3 to 7.0]), item 13a (OR = 2.8 [95% CI: 1.5 to 5.0]), item 18a (OR = 3.8 [95% CI: 2.2 to 6.5]), item 23 (OR = 2.3 [95% CI: 1.4 to 3.8]), and item 25 (OR = 4.0 [95% CI: 1.1 to 15.0]).

None of the individual items showed a significant reduction in the quality of reporting between studies published in 2000 and 2004.

Total number of reported STARD items. The mean total number of reported STARD items for articles published was 11.8 (SD 3.3) in 2000 and 13.6 (SD 3.2) in 2004 (maximum is 25). The results of the linear multilevel analysis confirmed that the quality of reporting of diagnostic accuracy studies improved over time. Studies published in 2004 reported on average 1.8 items (95% CI: 0.6 to 3.0) more

Table 2 *Reporting of individual items of the Standards for the Reporting of Diagnostic Accuracy studies (STARD statement) in articles on diagnostic accuracy published in 2000 and 2004*

Item		Articles published in 2000 (n = 124), n (%)	Articles published in 2004 (n = 141), n (%)	OR (95% CI)*
Title/abstract/keywords				
1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading "sensitivity and specificity")	13 (10.5)	26 (18)	1.9 (0.9, 4.0)
Introduction				
2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups	112 (90)	136 (96.5)	Not possible
Methods				
3	The study population: The inclusion and exclusion criteria, setting and locations where data were collected	35 (28)	30 (21)	0.7 (0.3, 1.5)
4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	103 (83)	130 (92)	4.2 (1.0, 17.4)
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 (56.5)	108 (77)	2.8 (1.0, 7.4)
6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	99 (80)	119 (84)	1.8 (0.6, 5.6)
7	The rationale of the reference standard	70 (56.5)	64 (45)	0.7 (0.4, 1.1)
8	Technical specifications of materials and methods involved including how and when measurements were taken, or cite references for			
	a) index tests and	115 (92.5)	137 (97)	Not possible
	b) reference standard	83 (67)	101 (72)	1.3 (0.7, 2.8)
9	Definition of and rationale for the units, cutoffs, or categories of the results of the			
	a) index tests and the	103 (83)	132 (94)	3.0 (1.3, 7.0)
	b) reference standard	75 (60.5)	102 (72)	1.7 (0.95, 3.1)
10	The number, training, and expertise of the persons executing and reading the			
	a) index tests and the	51 (41)	72 (51)	2.7 (0.8, 8.4)
	b) reference standard	32 (26)	46 (33)	1.6 (0.7, 3.6)
11	Whether or not the readers of the			
	a) index tests and	46 (37)	55 (39)	1.2 (0.5, 2.6)†
	b) reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers	23 (18.5)	39 (28)	1.7 (0.8, 3.5)
12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g., 95% CI)	17 (14)	28 (20)	1.6 (0.6, 3.9)
13	Methods for calculating test reproducibility, if done			
	a) for the index test	20 (16)	49 (35)	2.8 (1.5, 5.0)
	b) for the reference standard	6 (5)	9 (6)	1.3 (0.5, 3.9)
Results				
14	When study was performed, including beginning and end dates of recruitment	60 (48)	89 (63)	2.1 (0.9, 4.9)
15	Clinical and demographic characteristics of the study population (at least information on age, sex, spectrum of presenting symptoms)	65 (52)	84 (60)	1.4 (0.8, 2.4)
16	The number of participants satisfying the criteria for inclusion who did or did not undergo the index tests and the reference standard; describe why participants failed to undergo either test (a flow diagram is strongly recommended)	75 (60.5)	83 (59)	0.9 (0.4, 2.1)
17	Time interval between the index tests and the reference standard, and any treatment administered in between	33 (27)	35 (25)	0.9 (0.5, 1.7)
18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition	28 (23)	74 (52.5)	3.8 (2.2, 6.5)
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)	124 (88)	1.5 (0.7, 3.3)
20	Any adverse events from performing the index tests or the reference standard	21 (17)	16 (11)	0.6 (0.3, 1.3)
21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% CI)	40 (32)	57 (40)	1.2 (0.5, 2.9)
22	How indeterminate results, missing data, and outliers of the index tests were handled	73 (59)	80 (57)	0.9 (0.5, 1.6)
23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers, or centers, if done	48 (39)	84 (60)	2.3 (1.4, 3.8)
24	Estimates of test reproducibility, if done			
	a) index test	40 (32)	62 (44)	1.6 (0.98, 2.7)
	b) reference standard	8 (6.5)	8 (6)	0.9 (0.3, 2.4)
Discussion				
25	Discuss the clinical applicability of the study findings	114 (92)	138 (98)	4.0 (1.1, 15.0)

* Difference between the quality of reporting of the individual items reported in articles published in 2000 and 2004, adjusted for journal level and estimated with logistic multilevel analysis; OR above 1.0 signifies that an item was more frequently reported in 2004.

† The improvement in quality of reporting item 11A was significantly larger in cohort studies than in case control studies (OR: 0.3 [95% CI; 0.1 to 1.0]).

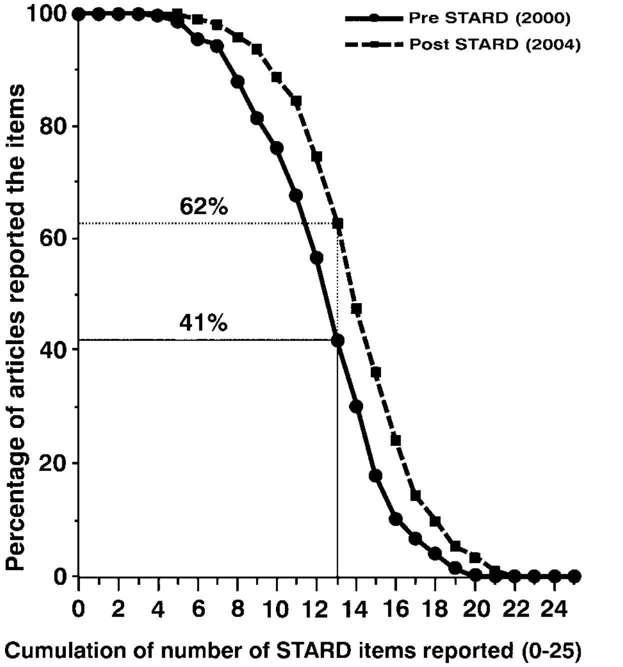

Figure 2. The percentage of articles published in 2000 and 2004 that present at least the indicated number of Standards for the Reporting of Diagnostic Accuracy studies (STARD statement) items. The straight line represents the articles published in 2000 and the dotted line represents the articles published in 2004.

than studies published in 2000. In 2004, 62% of the articles reported on more than half of the 25 items, vs 41% in 2000 (figure 2). None of the articles published in 2000 reported more than 20 of the 25 items, whereas 3 (2%) did so in 2004.

Flow diagram. Flow diagrams were sparse. Only 2 of the studies (2%) published in 2000 and 17 (12%) of those published in 2004 included a flow diagram. Most flow diagrams were published in journals that had adopted the STARD statement (table 1). Studies that included a flow diagram were associated with higher quality of reporting (mean difference 1.7 items [95% CI: 0.2 to 3.2]).

Adopting journals vs non-adopting journals. In 2000, the mean number of items reported in studies published in adopting and non-adopting journals was 11.9 (SD 3.2) and 12.0 (SD 3.4). In 2004, these numbers were 13.5 (SD 3.5) and 13.7 (SD 2.3). No significant differences in improvement were observed between adopting and non-adopting journals (mean difference = −0.06 [95% CI: −2.5 to 2.4]). For none of the individual items, significant differences were found in improvement between adopting and non-adopting journals.

Cohort studies vs case control studies. Between 2000 and 2004, the mean number of items reported in cohort studies had changed from 12.4 (SD 3.0) to 14.4 (SD 2.8) vs 10.8 (SD 3.7) and 11.8 (SD 3.2) for case control studies, a nonsignificant difference in improvement (1.1 items [95% CI: −0.6 to 2.8]) although quality of reporting was higher in cohort studies (mean difference = 1.7 items [95% CI: 0.9 to 2.5]).

Except for item 11a (OR = 0.3 [95% CI: 0.1 to 1.0]), no significant differences were found in improvement of reporting quality for individual items between cohort and case control studies.

Discussion. After publication of the STARD statement, the quality of reporting of diagnostic accuracy studies has slightly improved. It is unclear whether this small improvement is caused by the publication of the STARD statement as we did not find any differences in the reporting of the items between adopting and non-adopting journals. This could alternatively be attributed to an increasing awareness of authors, reviewers, and editors of the importance of the quality of reporting of research articles.

One obvious explanation for the small improvement in quality of reporting is the timing of the post-STARD evaluation. Is it probably too early to expect an improvement in the quality of reporting of studies published in 2004, as these studies were published only 1 to 2 years after the publication of the STARD statement? Yet some items of the STARD statement, such as recruitment period (item 14), measures of statistical uncertainty for the estimates of diagnostic accuracy (item 21), and presentation of a flow diagram could easily have been included in the manuscript. An improvement in the quality of reporting of these items suggests that authors, reviewers, and editors have used the STARD statement. Other items, such as reasons why participants failed to undergo one of the tests (item 16) and methods and calculation of test reproducibility (items 13 and 24), are more difficult to satisfy retrospectively, as these items concern the design of the study. Improvement in the quality of reporting of these items would take more time.

In contrast, the quality of reporting of randomized controlled trials (RCTs) published 2 years after the publication of the CONSORT statement had improved and the use of the CONSORT statement in the editorial process was associated with improvements in the quality of reporting.[17] Some authors suggested that knowledge of design principles of RCTs and the effects of bias are better known now compared to diagnostic accuracy studies and are relatively simple and straightforward for readers to appraise.[18]

Part of the improvement in the quality of reporting can also be attributed to measurement error. The assessment of the 2004 articles took place 2 years after the 2000 articles. To be sure that the reviewers used the same criteria in the same way, we have carried out a reproducibility study.[19] Although the overall reproducibility of the assessment of the quality of reporting using the STARD checklist was found to be good, substantial disagreements were found for some items, so any small improvement in the quality of reporting of these items should be interpreted with caution.[19] The presentation of a flow diagram, including the design of the study and the flow of patients through the study, would be helpful in improving the quality of reporting, as it explicitly clarifies items that caused most confusion among reviewers.[19]

The absence of a significant difference between studies published in adopting and non-adopting journals warrants discussion. As the STARD statement is available to everyone, some authors and reviewers may have used the STARD statement for their manuscripts irrespective of the guidelines of the journal of submission. In addition, the absence of a difference could be explained by the way the STARD statement was used within the editorial process. We found a large variation in formulations in the guidelines for authors in the adopting journals, with regard to the clearness and strictness of the use of the STARD statement. In journals with strict and clear guidelines, one would expect better quality of reporting. We could not analyze this effect on the quality of reporting, as the numbers of articles published in each journal were too small for such a comparison.

The STARD statement consists of 25 individual items. Failure to report some items withholds information from the reader with regard to applicability, but does not necessarily invalidate the evidence. Poor reporting of other items, such as the blinding of the readers of the tests (item 11), description of the criteria for the tests (items 8 and 9), description of the study population (item 15), and the number of included patients that underwent both test(s) (item 16), may reflect biased results.[20,21] Our study shows that these items were poorly reported. This does not necessarily mean that bias is present, but that the likelihood of bias cannot be determined.

In general, the quality of reporting of cohort studies was better than in case control studies. It should be emphasized that, in theory, case control studies are also able to satisfy all individual items.

In our study, we selected journals that frequently publish studies on diagnostic accuracy and had an impact factor of at least 4. The quality of reporting of diagnostic accuracy studies in journals with lower impact factors showed similar results as our pre-STARD evaluation.[12,14] Therefore, the results of our study can be generalized to journals with lower impact factors. However, we expect that an improvement in the quality of reporting may become apparent first in journals with higher impact factors, as they put higher demands on manuscripts than journals with lower impact factors.

References

1. Chan AW, Altman DG. Epidemiology and reporting of randomized trials published in PubMed journals. Lancet 2005;365:1159–1162.
2. Honest H, Khan KS. Reporting of measures of accuracy in systematic reviews of diagnostic literature. BMC Health Services Research 2002;2:1–4.
3. Pocock SJ, Collier TJ, Dandreo KJ, et al. Issues in the reporting of epidemiological studies: a survey of recent practice. BMJ 2004;329:883.
4. Begg CB, Cho MK, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. JAMA 1996;276:637–639.
5. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT Statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. Ann Intern Med 2001;134:657–662.
6. Altman DG, Schultz KF, Moher D, for the CONSORT Group. The revised CONSORT Statement for reporting randomized trials: explanation and elaboration. Ann Intern Med 2001;134:663–694.
7. 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.
8. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy; the STARD initiative. Clin Chem 2003;49:1–6.
9. Moher D, Cook DJ, Eastwood S, 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. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 2000;283:2008–2012.
11. Von Elm E, Egger M. The scandal of poor epidemiological research. BMJ 2004;329:868–869.
12. Siddiqui MAR, Azuara-Blanco A, Burr J. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. Br J Ophthalmol 2005;89:261–265.
13. Stengel D, Bauwens K, Rademacher G, Mutze S, Ekkernkamp A. Association between compliance with methodological standards of diagnostic research and reported test accuracy: meta-analysis of focused assessment of US for trauma. Radiology 2005;236:102–111.
14. Smidt N, Rutjes AWS, Van der Windt AWM, et al. Quality of reporting of diagnostic accuracy studies. Radiology 2005;235:347–353.
15. Devillé 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.
16. Holloway RG. Improving the flow of diagnostic information. The importance of STARD for authors and readers. Neurology 2003;61:600–601.
17. 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.
18. Reeves BC. Evidence about evidence. Br J Ophthalmol 2005;89:253–254.
19. Smidt N, Rutjes AWS, Van der Windt DAWM, et al. Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. BMC Medical Research Methodology 2006;6:12.
20. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–1066.
21. Whiting P, Rutjes AWS, Reitsma JB, Glas AS, Bossuyt PMM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy. Ann Intern Med 2004;140:189–202.