Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD

Daniël A Korevaar,[1] W Annefloor van Enst,[2] René Spijker,[2] Patrick M M Bossuyt,[1] Lotty Hooft[2]

10.1136/eb-2013-101637

[1]Department of Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB), Academic Medical Centre (AMC), University of Amsterdam (UvA), Amsterdam, The Netherlands
[2]Dutch Cochrane Centre (DCC), Academic Medical Centre (AMC), University of Amsterdam (UvA), Amsterdam, The Netherlands

Correspondence to
Daniël A Korevaar
Department of Clinical Epidemiology, Biostatistics and Bioinformatics (KEBB), Academic Medical Centre (AMC), University of Amsterdam (UvA), Meibergdreef 9, Amsterdam 1105 AZ, The Netherlands;
d.a.korevaar@amc.uva.nl

Abstract

Background Poor reporting of diagnostic accuracy studies impedes an objective appraisal of the clinical performance of diagnostic tests. The Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement, first published in 2003, aims to improve the reporting quality of such studies.

Objective To investigate to which extent published diagnostic accuracy studies adhere to the 25-item STARD checklist, whether the reporting quality has improved after STARD's launch and whether there are any factors associated with adherence.

Study selection We performed a systematic review and searched MEDLINE, EMBASE and the Methodology Register of the Cochrane Library for studies that primarily aimed to examine the reporting quality of articles on diagnostic accuracy studies in humans by evaluating adherence to STARD. Study selection was performed in duplicate; data were extracted by one author and verified by the second author.

Findings We included 16 studies, analysing 1496 articles in total. Three studies investigated adherence in a general sample of diagnostic accuracy studies; the others did so in a specific field of research. The overall mean number of items reported varied from 9.1 to 14.3 between 13 evaluations that evaluated all 25 STARD items. Six studies quantitatively compared post-STARD with pre-STARD articles. Combining these results in a random-effects meta-analysis revealed a modest but significant increase in adherence after STARD's introduction (mean difference 1.41 items (95% CI 0.65 to 2.18)).

Conclusions The reporting quality of diagnostic accuracy studies was consistently moderate, at least through halfway the 2000s. Our results suggest a small improvement in the years after the introduction of STARD. Adherence to STARD should be further promoted among researchers, editors and peer reviewers.

Introduction

In 2003, the Standards for Reporting of Diagnostic Accuracy Studies (STARD) statement was published in 13 biomedical journals.[1 2] Diagnostic accuracy studies provide estimates of a test's ability to discriminate between patients with and without a predefined condition, by comparing the test results against a clinical reference standard. The STARD initiative was developed in response to accumulating evidence of poor methodological quality and poor reporting among test accuracy studies in the prior years.[3 4] The STARD checklist contains 25 items which invite authors and reviewers to verify that critical information about the study is included in the study report. In addition, a flow chart that specifies the number of included and excluded patients and characterises the flow of participants through the study is strongly recommended. Since its launch, the STARD checklist has been adopted by over 200 biomedical journals (http://www.stard-statement.org/).

Over the past 20 years, reporting guidelines have been developed and evaluated in many different fields of research. Although a modest increase in reporting quality is sometimes noticed in the years following the introduction of such guidelines,[5 6] improvements in adherence tend to be slow.[7] This makes it difficult to make statements about the impact of such guidelines. For STARD, there has been some controversy around its effect.[8] While one study noticed a small increase in reporting quality of diagnostic accuracy studies shortly after the introduction of STARD,[9] another study could not confirm this.[10]

Systematic reviews can provide more precise and more generalisable estimates of effect. A recently published systematic review evaluated adherence to several reporting guidelines in different fields of research, but STARD was not among the evaluated guidelines.[11] To fill this gap, we systematically reviewed all the studies that aimed to investigate diagnostic accuracy studies' adherence to the STARD checklist in any research field. Our main objective was to find out how diagnostic accuracy studies adhere to (specific items on) the STARD checklist. Our research questions were: (1) How is the current (or rather, most recent) quality of reporting of diagnostic accuracy studies? (2) Has the quality of reporting improved after the introduction of STARD? (3) How do diagnostic accuracy studies score on specific items on the checklist? (4) Are there any factors associated with adherence to the checklist?

Methods

Search and selection

The original protocol of this study can be obtained from the corresponding author. We performed a systematic review and searched MEDLINE and EMBASE, which, to our knowledge, provide the best sources for methodological reviews. To make sure that all relevant data were captured, we also searched the Methodology Register of the Cochrane Library, of which the content is sourced from MEDLINE and additional manual searches. We included studies that primarily aimed to examine the quality of reporting of articles of diagnostic accuracy studies in humans in any field of research, by evaluating their adherence to the STARD statement. Details on the search strategies are provided in Web only file 1. The final search was performed on 13 August 2013. The searches were performed without any restrictions for language, year of publication or study type. We excluded systematic reviews on the accuracy of a single test that had used the STARD checklist to score the quality of reporting in the included articles, as well as studies that investigated the influence of reporting quality on pooled estimates of test accuracy results. Such articles would be on a too specific topic to be able to make statements on the reporting quality of diagnostic accuracy studies in general. Studies focusing on reports about analytical rather than clinical performance were also excluded. Although the design of these two types of studies show many similarities, STARD was not designed for studies on analytical test performance and several items on the lists do not apply in this setting. We also excluded studies that evaluated less than 10 STARD items and studies that had not presented their results quantitatively (as a mean number of reported items or a score per individual item) because this would make an objective comparison between studies impossible.

Two authors (DK and WvE) independently screened the titles and abstracts of the search results to identify potentially eligible studies. If at least one author identified an abstract as potentially eligible, the full text of the article was assessed by both authors. Disagreements were resolved through discussion, whenever possible. If agreement could not be reached, the case was discussed with a third author (LH). One author (DK) also reviewed reference lists of included studies for additional relevant papers.

Data collection

An extraction form was created before the literature search was performed, and piloted on three known eligible studies. After the pilot, the form was slightly modified. One author (DK) extracted relevant data from the included studies which were verified by the second author (WvE). Disagreements were resolved through discussion. If necessary, a third author (LH) made the final decision.

Of each included article, the first author, country, year of publication and journal were extracted. We also identified the inclusion and exclusion criteria, research field, primary aims, the number of studies included, which STARD items were evaluated and how they had been scored. In addition, we retrieved (descriptive) statistics regarding overall and item-specific STARD adherence, and adherence comparisons between articles published post-STARD versus those published pre-STARD. Any additional study characteristics mentioned to be associated with STARD adherence were extracted. We also extracted any statistics on inter-rater agreement in evaluating STARD items, and conclusions, interpretation and recommendations of the authors.

We assessed the quality of included studies by using the 11-item AMSTAR (Assessment of Multiple Systematic Reviews) tool.[12] As several items on this list do not apply to the studies included in our review, we omitted four items and only assessed items: item 1 (was an ‘a priori’ design provided?), item 2 (was there duplicate study selection and data extraction?), item 3 (was a comprehensive literature search performed?), item 4 (were inclusion and exclusion criteria provided?), item 5 (was a list of included and excluded studies provided?), item 6 (were the characteristics of included studies provided?) and item 9 (was the conflict of interest included?).

Analysis: overall adherence to STARD

We calculated κ statistics to assess inter-reviewer agreement for the two phases of study selection. For each included study, we calculated the overall STARD score, defined as the mean number of items reported by articles included in that study, and the proportion of articles adhering to each specific STARD item. For each STARD item, we calculated the median and range of these proportions.

Some studies also counted how often an item was partially reported. To be able to make comparisons between studies, we counted partially reported items as half in calculating proportions. Some STARD items pertain to the index test and the reference standard. Whenever these were analysed separately, half a point was allocated per reported item. If a study reported that an item on the STARD checklist was not applicable to all evaluated articles, that study was not included in our overall analysis for that specific item. If a study reported that a STARD item was applied to less than 100% of the evaluated articles, the score was calculated for the number of articles for which the item applied and the calculated proportions were adjusted.

Analysis: adherence to STARD before and after its launch

To obtain a summary estimate and the corresponding 95% CI of the difference in adherence before and after its launch, we used inverse variance random-effects meta-analysis.[13] Only studies specifically reporting pre-STARD and post-STARD results were included in this analysis. We explored statistical heterogeneity using the I^2 test.[14] We performed a subgroup analysis by separately analysing studies examining a general sample of diagnostic accuracy studies, rather than those investigating adherence in a specific field of research.

One included study only reported SDs for (equally sized) subgroups of STARD-adopting and non-adopting journals.[10] We calculated their overall SD by taking the square root of the pooled variances. SDs of one other study were obtained after contacting the authors.[15]

We used inverse variance random-effects meta-analysis to calculate summary ORs and 95% CIs for item-specific adherence in the pre-STARD versus post-STARD groups. Only studies specifically reporting the proportion of evaluated articles adhering to each individual item for the pre-STARD and post-STARD groups were included in this analysis.

Results

Search results and characteristics of included studies

Five hundred and eighteen studies were identified trough the search, of which 35 were deemed potentially eligible after screening titles and abstracts (figure 1). After studying the full texts, we were able to include 16 studies.[9–28] Reasons for exclusion of potentially eligible studies are provided in figure 1. No additional studies were identified through reference lists. Inter-reviewer

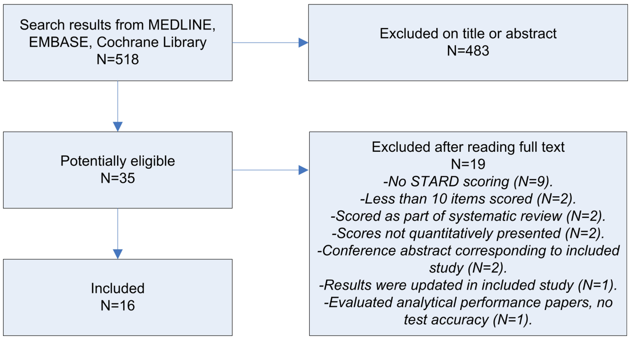

Search results from MEDLINE, EMBASE, Cochrane Library N=518
Excluded on title or abstract N=483
Potentially eligible N=35
Excluded after reading full text N=19
-No STARD scoring (N=9).
-Less than 10 items scored (N=2).
-Scored as part of systematic review (N=2).
-Scores not quantitatively presented (N=2).
-Conference abstract corresponding to included study (N=2).
-Results were updated in included study (N=1).
-Evaluated analytical performance papers, no test accuracy (N=1).
Included N=16

Figure 1 Flow chart for selection of studies.

agreement was substantial for the screening of titles and abstracts (κ=0.77 (95% CI 0.66 to 0.88)), and was perfect for the subsequent assessment of full-texts (κ=1.0).

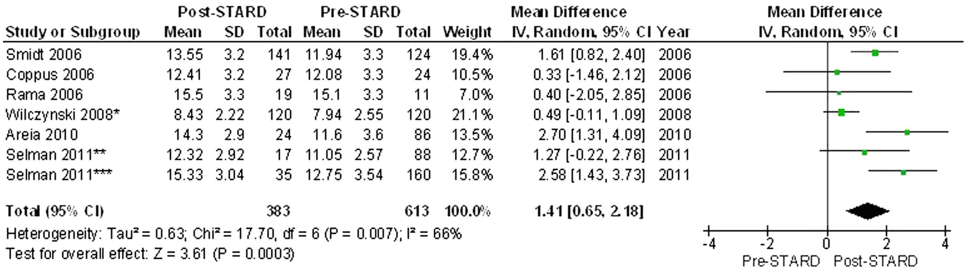
The characteristics of the included studies are provided in table 1. Three studies investigated adherence to STARD in a general sample of articles on diagnostic accuracy studies, the other 13 had performed so in a specific field of research. None of the included studies had evaluated a recent sample of articles: one study evaluated articles published through 2010, one study through 2008, two studies through 2007 and four studies through 2006. All other studies included only articles published before 2006. Twelve studies included articles published before and after STARD's launch. One study investigated only articles published pre-STARD and three studies investigated only articles published post-STARD.

The number of evaluated articles varied markedly between the included studies, with a median of 55 (range 16–300). Most of the studies (n=13) evaluated all 25 STARD items. However, among three of these, one item was found not applicable to all included articles. The other three studies had evaluated 24, 22 and 13 items of the 25 items, respectively. κ Values for overall inter-rater agreement on the STARD-items were reported by nine studies: moderate agreement (κ=0.41–0.6) was achieved in one study, substantial agreement (κ=0.61–0.8) in six studies and almost perfect agreement (κ=0.81–1.0) in two other studies.[29] An overall percentage agreement was reported by seven studies; this varied between 81% and 95%. Four studies did not report on inter-rater agreement.

An a priori study design was provided by only one included study. Seven studies performed the complete study selection in duplicate, while three did so in part. Eleven studies evaluated the reporting quality of all the included studies in duplicate, and three did so for a part of the included studies. All the included studies provided comprehensive data on the literature searches and the inclusion and exclusion criteria. Although more than half (n=9) of the studies provided a list of included studies, only two provided a list of excluded studies. Characteristics of included studies were provided, to some extent, by all studies: all gave information on the research field in which included articles were performed and 12 studies gave information on the type of tests used. Only three studies gave information on the included studies' design.

Overall adherence to STARD

The overall mean STARD score varied from 9.1 to 14.3 for the 13 studies that had evaluated all 25 STARD items, with a median of 12.8 items (table 1). Fifteen (94%) of the included studies concluded that the adherence to STARD was poor, medium, suboptimal or needed improvement. One study used more conservative language and concluded that adherence of included articles was highly variable. Seven studies evaluating all 25 items only reported post-STARD results or reported pre-STARD and post-STARD results separately. The overall mean number of items reported in these post-STARD results varied from 12.0 to 15.5, with a median of 13.6. Most of the included studies



	Post-STARD			Pre-STARD				Mean Difference	
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI	Year
Smidt 2006	13.55	3.2	141	11.94	3.3	124	19.4%	1.61 [0.82, 2.40]	2006
Coppus 2006	12.41	3.2	27	12.08	3.3	24	10.5%	0.33 [-1.46, 2.12]	2006
Rama 2006	15.5	3.3	19	15.1	3.3	11	7.0%	0.40 [-2.05, 2.85]	2006
Wilczynski 2008*	8.43	2.22	120	7.94	2.55	120	21.1%	0.49 [-0.11, 1.09]	2008
Areia 2010	14.3	2.9	24	11.6	3.6	86	13.5%	2.70 [1.31, 4.09]	2010
Selman 2011**	12.32	2.92	17	11.05	2.57	88	12.7%	1.27 [-0.22, 2.76]	2011
Selman 2011***	15.33	3.04	35	12.75	3.54	160	15.8%	2.58 [1.43, 3.73]	2011
Total (95% CI)			**383**			**613**	**100.0%**	**1.41 [0.65, 2.18]**	

Heterogeneity: Tau² = 0.63; Chi² = 17.70, df = 6 (P = 0.007); I² = 66%
Test for overall effect: Z = 3.61 (P = 0.0003)

Figure 2 Forest plot for studies included in meta-analysis comparing adherence post-Standards for Reporting of Diagnostic Accuracy Studies (STARD) and pre-STARD. *Wilczynski[10] evaluated only 13 STARD items; the other studies evaluated 25 STARD items. **Results of the studies on obstetrics. ***Results of the studies on gynaecology.

Table 1 Characteristics of included studies

First author	Country	Year	Journal	Research field	Number of articles included	Timeframe	Number of STARD items evaluated	Mean STARD score (% items evaluated)	Authors' conclusions on quality of reporting
Areia[16]	Portugal	2010	*Endoscopy*	Endoscopy	110	1998–2008	25	12.9 (52)	'Recent publications in diagnostic endoscopy achieve only medium quality'
Coppus[17]	The Netherlands	2006	*Fertility and Sterility*	Reproductive medicine	51	1999 vs 2004	25	12.3 (49)	'The quality of reporting in articles on test accuracy in reproductive medicine is poor to mediocre'
Fontela[18]	Canada	2009	*PlosONE*	Commercial tests for tuberculosis, HIV, malaria	90	2004–2006	25	13.6 (54)	'Diagnostic studies on tuberculosis, malaria and HIV commercial tests [...] were often poorly reported'
Freeman[19]	UK	2009	*European Journal of Obstetrics & Gynecology and Reproductive Biology*	Non-invase prenatal diagnostic tests for Rhesus D genotyping	27	1996–2006	25	9.1 (36)	'Articles have consistent weaknesses in their reporting'
Gómez Sáez[20]	Spain	2009	*Medicina Clinica*	Any research field, 4 Spanish journals	58	2004–2007	25	12.0 (48)	'Despite efforts by different groups of research to achieve higher methodological quality in the diagnostics field, on average, they follow less than half of the items proposed by STARD'
Johnson[21]	UK	2007	*Ophthalmology*	Optical coherence tomography (OCT) in glaucoma.	30	2001–2006	25*	13.2 (53)	'Quality of reporting of the diagnostic accuracy of OCT in glaucoma is suboptimal.'
Lumbreras[22]	Spain	2006	*Gaseta Sanitària*	Genetic-molecular research.	44	2002–2005	24	9.8 (41)	'The articles on genetic-molecular diagnostic tests (...) fail to satisfy most of the quality requirements assembled in the STARD proposal'
Paranjothy[23]	UK	2007	*Journal of Glaucoma*	Scanning laser polarimetry (SLP) for diagnosing glaucoma.	20	1997–2000 vs 2004–2005	25*	13.5 (54)	'The quality of reporting of diagnostic accuracy tests for glaucoma with SLP is suboptimal'
Rama[24]	UK	2006	*Clinical Orthopaedics and Related Research*	Orthopedics.	37	2002–2004	25	14.2 (57)	'Current standards of reporting of diagnostic accuracy studies in orthopaedic journals are suboptimal.'
Selman[15]	UK	2011	*BMC Women's Health*	Obstetrics and gynaecology.	300	1977–2007	25	12.5 (50)	'The reporting of included studies in this review overall was poor.'
Shunmugam[25]	UK	2006	*Investigative Ophthalmology & Visual Science*	Heidelberg retina tomography (HRT) for glaucoma detection.	29	1995–2004	25*	14.3 (57)	'The quality of reporting of diagnostic accuracy tests for glaucoma with HRT is suboptimal.'
Siddiqui[26]	UK	2005	*British Journal of Ophthalmology*	Ophthalmology.	16	2002	25	11.6 (47)	'The current standards of reporting of diagnostic accuracy tests are highly variable.'
Smidt[9]	The Netherlands	2006	*Neurology*	Six general and six disease/discipline-specific journals.	265	2000 vs 2004	25	12.8 (51)	'After publication of STARD, the quality of reporting of diagnostic accuracy studies has slightly improved. There is still room for improvement.'
Wilczynski[10]	Canada	2008	*Radiology*	Twelve journals on radiology, internal medicine or general medicine.	240	2001–2002 vs 2004–2005	13	8.2 (63)	'We found low rates of adherence to the STARD checklist items.'
Zafar[27]	UK	2008	*Clinical and Experimental Ophthalmology*	Diabetic retinopathy (DR) screening.	76	1995–2006	25	9.9 (40)	'The quality of diagnostic accuracy reports in DR screening is suboptimal.'
Zintzaras[28]	Greece	2012	*BMC Musculoskeletal Disorders*	Anti-CCP2 for the diagnosis of rheumatoid arthritis.	103	2003–2010	22	14.0 (64)	'The overall reporting quality was relatively good but needs further improvement.'

*One of the 25 evaluated STARD-items was not applicable to all the articles included in this study.
BMC, British Medical Council; DR, diabetic retinopathy; STARD, Standards for Reporting of Diagnostic Accuracy Studies.

recommended the use of STARD as a guideline to improve the quality of reporting of diagnostic accuracy studies, and no study discouraged it.

The medians and ranges of the proportions of adherence to individual STARD-items reported by included studies are provided in table 2. There was a large between-study variation in adherence to specific items. Overall, only 12 items had a median proportion exceeding 50%; only three items had a median proportion above 75%. When only evaluating post-STARD results, these median proportions were slightly better: 15 items exceeding 50% and 6 items exceeding 75%. Six items (8, 9, 10, 11, 13 and 24) concern the index test as well as the reference standard. Reporting of the index test was better than reporting of the reference standard for all of these items.

Several studies reported on factors potentially associated with quality of reporting. One study found that adherence to STARD was significantly better for cohort studies compared with case–control studies,[9] but another study could not confirm this.[24] Other factors reported to be significantly associated with higher STARD scores were sample size (higher scores among larger studies[15]) and research field (obstetric studies scored better than gynaecological studies,[15] and tuberculosis and malaria studies scored better than HIV

Table 2 Proportions of adherence to individual STARD items

	Overall			Post-STARD results only		
STARD item	Studies evaluating item (n)	Median of proportions (%)	Range (%)	Studies evaluating item (n)	Median of proportions (%)	Range (%)
25. Clinical applicability of findings	14	98	41–100	5	98	84–99
4. Participant recruitment	16	85	55–100	7	93	60–98
2. Research questions/aims	14	84	24–100	5	88	76–96
8. Technique of	16	73	31–98	7	74	40–97
a. Index test	5	92	49–95	4	84	58–97
b. Reference standard	5	63	13–86	4	55	23–72
15. Characteristics of study population	16	73	42–90	7	70	60–93
7. Reference standard and rationale	16	70	28–98	7	76	45–98
9. Units/cut-offs/categories for	16	70	0–98	7	83	63–85
a. Index test	5	84	68–95	4	91	71–94
b. Reference standard	5	73	55–76	4	75	56–80
3. Study population	16	68	23–92	7	63	21–88
6. Data collection	16	68	21–100	7	83	43–95
19. Cross tabulation of results	15	65	2–99	6	66	28–99
18. Distribution of severity of disease	16	62	0–97	7	52	11–98
21. Estimates of diagnostic accuracy	15	56	12–97	6	56	22–97
12. Methods for statistics used	15	49	8–90	6	49	11–90
14. Dates of study	16	47	6–73	7	73	42–81
1. Study identified as test accuracy study	13	40	8–100	5	24	18–99
5. Participant sampling	16	40	12–89	7	64	31–89
23. Estimates of variability of accuracy	15	37	0–100	6	39	0–100
17. Time interval between tests	15	34	0–77	6	38	25–74
11. Blinding of results of	16	29	14–54	7	33	16–55
a. Index test	5	43	33–72	4	50	26–67
b. Reference test	5	23	12–48	4	25	15–48
22. How uninterpretable results were handled	15	28	8–62	6	25	8–57
10. Persons executing	16	26	2–73	7	20	2–42
a. Index test	5	33	7–46	4	26	4–51
b. Reference standard	5	20	0–35	4	14	0–33
16. Eligible patients not undergoing either test	16	24	5–78	7	53	13–70
a. Flow diagram	12	5	0–16	4	8	0–22
13. Methods for test reproducibility for:	15	16	0–88	6	18	0–88
a. Index test	4	20	12–53	3	35	6–48
b. Reference standard	4	7	0–12	3	4	0–6
24. Estimates of test reproducibility, for:	15	8	0–96	6	8	0–96
a. Index test	4	20	13–38	3	22	6–44
b. Reference standard	4	3	0–8	3	0	0–6
20. Adverse events	12	7	0–33	6	11	1–18

studies[18]). Factors that did not show a significant difference were geographical area,[15] level of evidence[24] and pooled sensitivity and specificity,[28] but these findings were not replicated in a subsequent study.

Adherence to STARD before and after its launch

Of the 12 studies that had included articles published before and after the publication of STARD, 6 reported results for the pre-STARD and post-STARD groups. These were included in the meta-analysis. Combining these studies in a meta-analysis showed that significantly more items were reported post-STARD, with an estimate difference of 1.41 items (95% CI 0.65 to 2.18). However, the great majority of the 383 post-STARD articles included in this analysis were published in the 2 years after introduction of STARD (2004 and 2005, n=349); only 34 articles were published after 2005. As expected, I^2 test showed evidence of substantial statistical heterogeneity (66%). Subgroup analysis of the two studies that reported on a general sample of diagnostic accuracy studies[9 10] showed a non-significant increase in the number of reported STARD-items (difference of 1.02 items (95% CI −0.08 to 2.12), I^2=80%).

Six other studies have reported some form of analysis of STARD adherence over time. One of these noticed an upward trend in the number of items reported pre-STARD and post-STARD.[23] Four others could not confirm this: two studies reported that introduction of STARD did not seem to have improved the quality of reporting of articles included in their analysis,[21 22] one study observed no improvement of quality of reporting over time[27] and one study noticed a (non-significant) decline in adherence after STARD publication.[20]

The pre-STARD versus post-STARD meta-analyses for individual items are reported in Web only file 2. Six items were significantly more reported after the publication of STARD: item 4 (describes participant recruitment), item 5 (describes participant sampling), item 6 (describes data collection), item 14 (reports dates of study), item 15 (reports characteristics of study population) and item 23 (reports estimates of variability of accuracy). Although still rare, the number of studies reporting a flow diagram also increased significantly. None of the STARD items showed a significant decrease in frequency of reporting.

Discussion

In this systematic review, we evaluated adherence to STARD. We were able to include 16 studies, together evaluating 1496 articles on diagnostic accuracy studies. The overall quality of reporting in these articles, published both in general and in disease-specific journals, was moderate, at least through halfway the 2000s, confirming the necessity of the introduction of STARD. Results of overall adherence were consistent among all included studies, and varied from 9.1 to 14.3 items being reported, of the 25 items on the checklist. Several factors were reported to be associated with STARD adherence by individual studies, but none of these associations was confirmed by a second study.

Although modest, there seemed to be an improvement in reporting quality (1.41 items (95% CI 0.65 to 2.18)) in the first years after STARD's publication in 2003 compared with articles published pre-STARD. Even though the CI is wide, this improvement is significant. The fact that the quality of the seven analyses included in this meta-analysis was acceptable, and that all of them showed an increase in reported items (three of them significant), increases our confidence in the estimates of effect.

Our study has several potential limitations. Most of the studies evaluated articles on diagnostic accuracy studies published before 2006; none evaluated articles published after 2010. Therefore, we cannot comment on how diagnostic accuracy studies currently adhere to STARD. Most of the included studies reported a substantial inter-rater agreement on individual items, with marked differences between studies in reported frequencies of adherence to specific items (table 2). There was also considerable heterogeneity in our meta-analysis comparing pre-STARD and post-STARD adherence. It is likely that this can, at least partially, be explained by between-study differences in scoring for specific items. For example, while some studies indicated that for item 3, at least the inclusion and exclusion criteria had to be reported, others only considered this item as fully reported when the setting and locations were also described. Only seven studies specifically reported how often an item was judged not to be applicable to the evaluated articles, while the others did not. Therefore, we were not always able to do a mathematical correction for non-applicable items. It is difficult to say whether between-study differences in scores of specific items were caused by a great diversity in adherence in the respective research fields, by heterogeneity in methods of scoring or both. We would have liked to compare the differences in compliance between STARD-adopting and non-adopting journals, and between high-impact and low-impact journals, but were unable to do so, because this information was almost never available in the included studies.

Although the overall quality of reporting was moderate, several items scored relatively good, with a median proportion of 70% or higher: item 2 (research questions/aims), item 4 (participant recruitment), item 7 (reference standard), item 8 (technique of index test and reference standard), item 9 (units/cut-offs/categories of tests), item 15 (study group characteristics) and item 25 (clinical applicability of findings). Worrisome is the fact that more than half of the 25 STARD items had median proportions of adherence under 50%. Especially, the reporting of study methods and results was suboptimal.

Seven items scored remarkably poor, with a median proportion of 30% or lower: item 10 (persons executing the tests), item 11 (blinding of readers), item 13 (methods for calculating test reproducibility), item 16 (the number of eligible patients not undergoing either test), item 20 (adverse events), item 22 (handling of missing results) and item 24 (estimates of test reproducibility). This is particularly alarming because several of these items can be related to biased results. If no or incomplete information on such items is reported, the potential for bias cannot be determined. Review bias, which can result when readers of a test have knowledge of the outcome of other tests or additional clinical information (item 11),[3] and verification bias, which occurs when a patient is only tested by the reference standard in case of a positive index test (item 16),[30] are likely to give inflated estimates of diagnostic accuracy. Limited test reproducibility (items 13 and 24), an effect of instrumental and/or observer variability, and not including missing responses or outliers (item 22), can also introduce biased or imprecise accuracy estimates.[2] Interestingly, for all the six items that apply to the index test and reference standard, adherence was better for the index test. Since accuracy estimates of an index test completely depend on the reference standard, authors should be encouraged to provide all the relevant information of both tests. Finally, flow charts were rarely reported, both pre-STARD and post-STARD. Since these highly facilitate a reader's assessment of study design, their use should be further promoted.

Owing to a constant increase in technological and scientific innovations, the number of available diagnostic tests has been growing exponentially over the past decades. Diagnostic tests are indispensable in patient management since many clinical decisions depend on their results. Implementation and proper usage of a test in any given clinical setting should be based on a thorough consideration of its costs, safety and clinical performance and utility. High-quality diagnostic accuracy studies are crucial in this consideration. Compared with other forms of research, diagnostic accuracy studies are probably more sensitive to bias.[3 31] The STARD checklist facilitates a complete and transparent reporting of diagnostic accuracy studies and, consequently, allows readers (clinicians, editors, reviewers, policy makers, etc) to identify sources of bias that may influence the clinical value and generalisability of a test. While reviews of diagnostic studies often struggle with high heterogeneity, complete and transparent reporting would facilitate an identification of potential sources of heterogeneity.

Although we have presented evidence that the quality of reporting of diagnostic accuracy studies is slowly increasing, it seems that there is still significant room for improvement. A recent study showed that adherence to guidelines is also suboptimal in other fields of research.[11] Although the scientific community seems to become more and more aware of the importance of transparent reporting, further enforcement of reporting guidelines among researchers, editors and peer reviewers is a necessity. We strongly recommend authors of diagnostic accuracy studies to take STARD into account from the stage of designing the study and onwards. This way, the items can easily be incorporated in the final article. In addition, this may lead to an increased awareness among authors about potential sources of bias, which allows them to take preventive measures and, consequently, also increases the methodological quality of their study. In addition, we recommend that an evaluation of adherence to STARD should be performed on a more recent cohort of diagnostic accuracy studies. A systematic review has recently shown that, after the introduction of the CONSORT (Consolidated Standards of Reporting Trials) statement, adopting journals had a larger increase in reporting quality of randomised controlled trials than non-adopting journals.[7] Such information may be useful in the effort to convince journal editors of the necessity of adopting reporting guidelines. Future evaluations can compare reporting quality of diagnostic accuracy studies between STARD-adopting and non-adopting journals. This way, an estimation of the impact of adopting STARD on reporting quality can be made.

Competing interests None.

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/eb-2013-101637).

References

1. **Bossuyt PM**, Reitsma JB, Bruns DE, *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. Standards for Reporting of Diagnostic Accuracy. *Clin Chem* 2003;**49**:1–6.
2. **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.
3. **Lijmer JG**, Mol BW, Heisterkamp S, *et al.* Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA* 1999;**282**:1061–6.
4. **Reid MC**, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. *JAMA* 1995;**274**:645–51.
5. **Plint AC**, Moher D, Morrison A, *et al.* Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust* 2006;**185**:263–7.
6. **Moher D**, Jones A, Lepage L. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *JAMA* 2001;**285**:1992–5.
7. **Turner L**, Shamseer L, Altman DG, *et al.* Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst Rev* 2012;**11**:MR000030.
8. **Bossuyt PM.** STARD statement: still room for improvement in the reporting of diagnostic accuracy studies. *Radiology* 2008;**248**:713–14.
9. **Smidt N**, Rutjes AW, van der Windt DA, *et al.* The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology* 2006;**67**:792–7.
10. **Wilczynski NL.** Quality of reporting of diagnostic accuracy studies: no change since STARD statement publication—before-and-after study. *Radiology* 2008;**248**:817–23.
11. **Samaan Z**, Mbuagbaw L, Kosa D, *et al.* A systematic scoping review of adherence to reporting guidelines in health care literature. *J Multidiscip Healthc* 2013;**6**:169–88.
12. **Shea BJ**, Grimshaw JM, Wells GA, *et al.* Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. *BMC Med Res Methodol* 2007;**7**:10.
13. **DerSimonian R**, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177–88.
14. **Higgins JP**, Thompson SG, Deeks JJ, *et al.* Measuring inconsistency in meta-analyses. *BMJ* 2003;**327**:557–60.
15. **Selman TJ**, Morris RK, Zamora J, *et al.* The quality of reporting of primary test accuracy studies in obstetrics and gynaecology: application of the STARD criteria. *BMC Womens Health* 2011;**11**:8.
16. **Areia M**, Soares M, Dinis-Ribeiro M. Quality reporting of endoscopic diagnostic studies in gastrointestinal journals: where do we stand on the use of the STARD and CONSORT statements? *Endoscopy* 2010;**42**:138–47.

17. **Coppus SF,** van d V, Bossuyt PM, *et al*. Quality of reporting of test accuracy studies in reproductive medicine: impact of the Standards for Reporting of Diagnostic Accuracy (STARD) initiative. *Fertil Steril* 2006;**86**:1321–9.
18. **Fontela PS,** Pant PN, Schiller I, *et al*. Quality and reporting of diagnostic accuracy studies in TB, HIV and malaria: evaluation using QUADAS and STARD standards. *PLoS ONE* 2009;**4**:e7753.
19. **Freeman K,** Szczepura A, Osipenko L. Non-invasive fetal RHD genotyping tests: a systematic review of the quality of reporting of diagnostic accuracy in published studies. *Eur J Obstet Gynecol Reprod Biol* 2009;**142**:91–8.
20. **Gomez SN,** Hernandez-Aguado I, Lumbreras B. [Observacional study: evaluation of the diagnostic research methodology in Spain after STARD publication]. *Med Clin (Barc)* 2009;**133**:302–10.
21. **Johnson ZK,** Siddiqui MA, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies of optical coherence tomography in glaucoma. *Ophthalmology* 2007;**114**:1607–12.
22. **Lumbreras B,** Jarrin I, Hernandez AI. Evaluation of the research methodology in genetic, molecular and proteomic tests. *Gac Sanit* 2006;**20**:368–73.
23. **Paranjothy B,** Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using scanning laser polarimetry. *J Glaucoma* 2007;**16**:670–5.
24. **Rama KR,** Poovali S, Apsingi S. Quality of reporting of orthopaedic diagnostic accuracy studies is suboptimal. *Clin Orthop Relat Res* 2006;**447**:237–46.
25. **Shunmugam M,** Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in glaucoma using the Heidelberg retina tomograph. *Invest Ophthalmol Vis Sci* 2006;**47**:2317–23.
26. **Siddiqui MA,** Azuara-Blanco A, Burr J. The quality of reporting of diagnostic accuracy studies published in ophthalmic journals. *Br J Ophthalmol* 2005;**89**:261–5.
27. **Zafar A,** Khan GI, Siddiqui MA. The quality of reporting of diagnostic accuracy studies in diabetic retinopathy screening: a systematic review. *Clin Experiment Ophthalmol* 2008;**36**:537–42.
28. **Zintzaras E,** Papathanasiou AA, Ziogas DC, *et al*. The reporting quality of studies investigating the diagnostic accuracy of anti-CCP antibody in rheumatoid arthritis and its impact on diagnostic estimates. *BMC Musculoskelet Disord* 2012;**13**:113.
29. **Bland JM,** Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;**1**:307–10.
30. **Reitsma JB,** Moons KG, Bossuyt PM, *et al*. Systematic reviews of studies quantifying the accuracy of diagnostic tests and markers. *Clin Chem* 2012;**58**:1534–45.
31. **Ochodo EA,** Bossuyt PM. Reporting the accuracy of diagnostic tests: the STARD initiative 10 years on. *Clin Chem* 2013;**59**:917–19.

Web only file 1: Search strategy.
Literature searches were performed on August 13th, 2013.

Ovid SP Embase Classic+Embase 1947 to Present

1	stard.ti,ab.	610
2	(reporting adj3 standard$ adj3 diagnos$).ti,ab.	199
3	(quality adj4 report$ adj6 diagnos$).ti,ab.	174
4	(reporting adj3 standard$ adj3 accuracy).ti,ab.	165
5	(quality adj4 report$ adj6 accuracy).ti,ab.	96
6	or/1-5	855
7	"review"/	1988330
8	review$.ti,ab.	1459791
9	MEDLINE.tw.	64633
10	exp systematic review/ or systematic review.tw.	83327
11	meta-analysis/	74817
12	(search* adj12 (literature or database?)).ti,ab.	80642
13	or/7-12	2855260
14	6 and 13	313

Ovid SP Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present

1	stard.ti,ab.	453
2	(reporting adj3 standard$ adj3 diagnos$).ti,ab.	166
3	(quality adj4 report$ adj6 diagnos$).ti,ab.	140
4	(reporting adj3 standard$ adj3 accuracy).ti,ab.	143
5	(quality adj4 report$ adj6 accuracy).ti,ab.	80
6	or/1-5	654
7	review$.ti,ab.	1250160
8	review.pt.	1893377
9	MEDLINE.tw.	62288
10	systematic review.tw.	44661
11	meta-analysis.pt.	49854
12	(search* adj12 (literature or database?)).ti,ab.	72043
13	or/7-12	2537849
14	13 and 6	257

Cochrane Library (methods register)

1	stard:ti,ab	55
2	(reporting near/6 diagnos*):ti,ab	484
3	(quality near/8 diagnos*):ti,ab	591
4	(reporting near/8 accuracy):ti,ab	116
5	(quality near/8 accuracy):ti,ab	90
6	#1 or #2 or #3 or #4 or #5	246

Total number of search results after deduplication: 518

Web only file 2: Table of pooled results of adherence per individual STARD item comparing studies published pre- and post-STARD.

		Pre-STARD		**Post-STARD**		**Overall**
	Studies	**Evaluated articles**	**Item reported**	**Evaluated articles**	**Item reported**	**Effect estimate**
STARD item	**n**	**n**	**n (%)**	**n**	**n (%)**	**OR (95%CI)**
1. Study identified as test accuracy study	4	396	108 (27.3%)	220	64 (29.1%)	1.80 [0.63, 5.08]
2. Research questions/aims	4	396	339 (85.6%)	220	204 (92.7%)	1.52 [0.70, 3.31]
3. Study population	5	516	290 (56.2%)	340	147 (43.2%)	0.94 [0.63, 1.40]
4. Participant recruitment	5	516	395 (76.6%)	340	316 (92.9%)	2.89 [1.41, 5.92]
5. Participant sampling	5	516	234 (45.3%)	340	223 (65.6%)	1.89 [1.33, 2.69]
6. Data collection	5	516	359 (69.6%)	340	266 (78.2%)	1.44 [1.02, 2.03]
7. Reference standard and rationale	5	516	325 (63.0%)	340	315 (92.6%)	1.72 [0.79, 3.74]
8. Technique of:						
8a. Index test	5	516	355 (68.8%)	340	286 (84.1%)	1.32 [0.64, 2.72]
8b. Reference standard	5	546	194 (35.5%)	340	190 (55.9%)	1.28 [0.65, 2.53]
9. Units/cut-offs/categories for:						
9a. Index test	5	516	437 (84.7%)	340	289 (85.0%)	1.37 [0.43, 4.34]
9b. Reference standard	4	428	276 (64.5%)	324	218 (67.3%)	1.37 [0.99, 1.89]
10. Persons exectuting:						
10a. Index test	5	516	113 (21.9%)	340	120 (35.3%)	1.21 [0.86, 1.71]
10b. Reference standard	5	516	57 (11.0%)	340	72 (21.2%)	1.22 [0.82, 1.83]
11. Blinding of results of:						
11a. Index test	5	516	307 (59.5%)	340	171 (50.3%)	0.93 [0.58, 1.47]
11b. Reference test	5	516	112 (21.7%)	340	110 (32.4%)	1.36 [0.87, 2.13]
12. Methods for statistics used	4	396	123 (31.1%)	220	58 (26.4%)	1.49 [0.54, 4.09]
13. Methods for test reproducibility for:						
13a. Index test	3	308	56 (18.2%)	203	64 (31.5%)	1.01 [0.28, 3.62]
13b. Reference standard	3	308	11 (3.6%)	203	10 (4.9%)	0.55 [0.07, 4.65]
14. Dates of study	5	516	332 (64.3%)	340	241 (70.9%)	1.69 [1.23, 2.33]
15. Characteristics of study population	5	516	293 (56.8%)	340	250 (73.5%)	1.85 [1.00, 3.43]
16. Eligible patients not undergoing either test	5	516	296 (57.4%)	340	198 (58.2%)	1.02 [0.76, 1.38]
16a. Flow diagram	*4*	*492*	*15* (3.0%)	*313*	*45 (14.4%)*	4.79 [1.39, 16.50]
17. Time interval between tests	4	396	127 (32.1%)	220	75 (34.1%)	1.38 [0.69, 2.76]
18. Distribution of severity of disease	5	516	315 (61.0%)	340	215 (63.2%)	2.18 [0.92, 5.15]
19. Cross tabulation of results	4	396	235 (59.3%)	220	159 (72.3%)	0.90 [0.53, 1.53]
20. Adverse events	3	236	35 (14.8%)	185	20 (10.8%)	0.74 [0.40, 1.37]
21. Estimates of diagnostic accuracy	4	396	177 (44.7%)	220	96 (43.6%)	1.46 [0.67, 3.18]
22. How uninterpretable results were handled	4	396	194 (49.0%)	220	113 (51.4%)	1.10 [0.75, 1.61]
23. Estimates of variability of accuracy	4	396	133 (33.6%)	220	119 (54.1%)	2.62 [1.57, 4.37]
24. Estimates of test reproducibility, for:						
24a. Index test	3	308	71 (23.1%)	203	70 (34.5%)	0.84 [0.31, 2.26]
24b. Reference standard	2	148	8 (5.4%)	168	8 (4.8%)	0.87 [0.32, 2.40]
25. Clinical applicability of findings	4	396	384 (97.0%)	220	215 (97.7%)	1.31 [0.16, 10.57]