Diagnostic and prognostic aspects of tubal patency testing
Coppus, S.F.P.J.

Citation for published version (APA):
Coppus, S. F. P. J. (2012). Diagnostic and prognostic aspects of tubal patency testing
Quality of reporting of test accuracy studies in reproductive medicine: impact of the standards for reporting of diagnostic accuracy (STARD)-initiative.

Sjors Coppus
Fulco van der Veen
Patrick Bossuyt
Ben Willem Mol

Fertility and Sterility 2006, 86: 1321-1329

The content of this article was presented as a Prize Paper Candidate session oral presentation at the 61st ASRM Annual Meeting, 2005, Montreal, Canada.
Abstract

BACKGROUND: To evaluate the extent to which test accuracy studies published in two leading reproductive medicine journals in the years 1999 and 2004 adhered to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) initiative parameters, and to explore whether the introduction of the STARD statement has led to an improved quality of reporting.

METHODS: Structured literature search. Articles that reported on the diagnostic performance of a test in comparison with a reference standard were eligible for inclusion. For each article we scored how well the twenty-five items of the STARD checklist were reported. These items deal with the study question, study participants, study design, test methods, reference standard, statistical methods, reporting of results and conclusions. We calculated the total number of reported STARD items per article, summary scores for each STARD item and the average number of reported STARD items per publication year.

RESULTS: We found 24 studies reporting on test accuracy in reproductive medicine in 1999 and 27 studies in 2004. The mean number of reported STARD items for articles published in 1999 was 12.1 ± 3.3 (range 6.5-20) and 12.4 ± 3.2 (range 7-17.5) in 2004, after publication of the STARD statement. Overall, less than half of the studies reported adequately on 50% or more of the STARD items. The reporting of individual items showed a wide variation. There was no significant improvement in mean number of reported items for the articles published after the introduction of the STARD statement.

CONCLUSIONS: Authors of test accuracy studies in the two leading fertility journals poorly report the design, conduct, methodology and statistical analysis of their study. Strict adherence to the STARD guidelines should be encouraged.
Introduction

The number of available diagnostic tests is growing exponentially. Evaluation of their accuracy in ruling in or ruling out disease is increasingly important as these data can help us identify which tests are useful to perform.

A problem in test accuracy studies is that rigorous methodological standards for research on diagnostic tests have been slower to develop than standards for studies of therapeutic interventions. This could be explained by the fact that test evaluation research is less straightforward than the evaluation of therapy, where randomized controlled trials have become the de facto standard method for evaluating effectiveness. In the last 10 years, more and more empirical evidence has been gained on design features of test accuracy studies associated with bias and lack of applicability. Several factors can threaten the internal and external validation of accuracy studies. One study showed that the methodological quality of studies of diagnostic accuracy, as published in four major medical journals between 1978 and 1993, was mediocre at best. Information on key elements of design, conduct and analysis was often not reported (1).

In 1999, Lijmer et al. (2) showed that many reports of diagnostic tests contain methodological shortcomings that can result in an overestimation of diagnostic test accuracy. A systematic review showed that a number of issues in the design and conduct of diagnostic accuracy studies can lead to bias or variation, with the best-documented effects found for demographic features, disease prevalence and severity, partial verification bias, clinical review bias, and observer and instrument variation (3).

In reproductive medicine, tests are ordered to establish the presence or absence of disease or to predict clinical outcome such as ovarian response or pregnancy. In diagnostic accuracy studies the results of a test are compared to those of a reference standard, the best available method to detect the presence of absence of disease. These results are expressed then expressed in accuracy statistics, such as sensitivity and specificity. In studies on the prognostic performance of tests, test results are compared with a clinically relevant endpoint and statistics are calculated to express the predictive capacity of the test.
To interpret the results of clinical studies on testing correctly, readers must understand the design, conduct, and data analysis and must be able to judge the internal validity and generalizability of the results. This goal can only be achieved through complete transparency of reporting in the articles. In recent years, a growing number of guidelines have been published to support authors of medical articles to clearly report the methodology used. For randomised clinical trials, the CONSORT statement (Consolidated Standards of Reporting Trials) was published in 1996 (4) and in a revised form in 2001 (5). For meta-analyses QUOROM (Quality of Reporting of Meta-analyses) and MOOSE (Meta-analysis of Observational Studies in Epidemiology) are available (6, 7).

For diagnostic studies, a group of scientists and editors developed the STARD (Standards for Reporting of Diagnostic Accuracy) statement to improve the reporting quality of studies on diagnostic accuracy. This STARD statement contains a 25-item checklist and a prototypical flow chart that helps authors to write a proper report of a diagnostic accuracy study and enables readers to assess the potential for bias in a study and to evaluate the generalizability of the results. STARD was primarily developed for authors to ensure completeness of reporting, not for reviewers to determine the quality of the study methodology.

In January 2003, this statement was first published simultaneously in eight medical journals, (American Journal of Clinical Pathology, Annals of Internal Medicine, British Medical Journal, Clinical Chemistry, Clinical Biochemistry, Clinical Chemistry and Laboratory Medicine, the Lancet and Radiology) together with an explanation and elaboration document published in Clinical Chemistry (8, 9). Since then, it has been published or promoted in several other biomedical journals, including Academic Radiology, British Journal of Obstetrics and Gynaecology, Clinical Radiology, Gynecological Oncology, JAMA, Annals of Clinical Microbiology and others (10-15).

In reproductive medicine, neither Human Reproduction nor Fertility and Sterility published STARD or promoted this statement with an editorial. In the instructions for authors of these journals on how to prepare a paper, only Human Reproduction lists the STARD statement as compulsory when writing an article on test accuracy.

Since it is not known how reports on diagnostic and prognostic accuracy in reproductive medicine comply with the STARD-checklist, and to explore whether
Quality of reporting of diagnostic studies

the introduction of the STARD-statement has improved the quality of report, we conducted a systematic literature survey in which the quality of reporting of these studies published in *Fertility and Sterility* and *Human Reproduction* was assessed for the years 1999 (pre-STARD) and 2004 (post-STARD).

**Materials and Methods**

*Study selection and data extraction*

We performed a systematic search in all issues of *Fertility and Sterility* and *Human Reproduction* published in 1999 (pre-STARD) and in 2004 (post-STARD) for articles reporting on the diagnostic or prognostic accuracy of a test. A search was run in Pubmed on journal name and publication year. This search was then limited to articles that dealt with human research and had an abstract. Hereafter we excluded all (randomized) clinical trials, meta-analyses, letters, editorials, practice guidelines and reviews with the search limit options of the search system.

We screened the remaining articles for eligibility based on title and abstract. Articles were eligible for inclusion if they reported on primary studies on diagnostic or prognostic accuracy. Diagnostic accuracy studies were defined as the assessment of a new test in comparison with a reference standard. Prognostic accuracy studies were defined as studies in which a test was performed to predict clinical outcomes in ART such as ovarian hyperstimulation syndrome, ovarian response, or pregnancy. A flowchart of the study selection process is shown in Fig. 1.

All selected articles were closely examined on the extent to which they reported the STARD criteria. A standardized score form for the quality of reporting of diagnostic accuracy studies that had been developed and tested at the department of Clinical Epidemiology and Biostatistics from the Academic Medical Centre, University of Amsterdam was used for this purpose (16). A single trained reviewer scored all articles, with a secondary reviewer (BWJM) checking a random sample of 20% to ensure accuracy in interpretation of the articles. Inter-rater agreement between these two reviewers was expressed as overall agreement percentage and more formally tested with the kappa statistic.
Chapter 2

Pubmed search for articles published in Fertility and Sterility and Human Reproduction in 1999 and 2004 (n = 2566)

Excluded: n = 439

Search limited to human research with available abstract (n = 2127)

Excluded: n = 670:
• Trial n = 452
• Editorial n = 4
• Letter n = 57
• Meta-analysis n = 16
• Practice guideline n = 14
• Review n = 127

Potentially eligible articles identified and screened for retrieval based on title and abstract (n = 1457)

Excluded: n = 1397

Potentially eligible articles (n = 60)

Excluded: n = 9
No diagnostic or prognostic test evaluation n = 7
Correspondence n = 2

Assessment of the quality of reporting on diagnostic and prognostic test accuracy studies in reproductive medicine (n = 51)

FIGURE 1. Flowchart showing search and selection process of articles reporting on diagnostic or prognostic test accuracy.

Statistical analysis

The 25-items of the STARD list can be divided into items addressing formulation of the study question, study participants, study design, test methods, reference standard, statistical methods, report of results and conclusions (Table 1). For each item of the STARD statement, the total number of articles reporting all the...
elements needed for that item was summed. Equal weights were applied to each item. The total number of reported STARD items, was also calculated for each article by summing the number of reported items (0-25 points possible). A higher number of items indicates better reporting and visa versa.

Six items (items 8, 9, 10, 11, 13 and 24) concern both the index test as well as the reference test. If only the index test or only the reference standard was reported adequately, this item was scored as 0.5 point. The overall mean, range, and standard deviation of the total number of reported STARD items were calculated. Testing for significant improvement of the total number of reported STARD items 1 year after the introduction of this statement was performed with an unpaired two-tailed t-test for independent samples, with the significance level set at \( P<0.05 \).

**Results**

A total number of 2127 articles reporting on human research with an available abstract published in 1999 and 2004 in *Fertility and Sterility* and *Human Reproduction* were identified. After limiting the search, as described in the method section, the remaining 1457 articles were screened for eligibility. Of these, 51 (3.5%) were marked as prognostic or diagnostic accuracy studies (17-67). Twenty-eight of these were published in *Fertility and Sterility* and 23 in *Human Reproduction*. Agreement between the two reviewers was good with an overall agreement percentage of 91.5%. The kappa-statistic had a value of 0.82, indicating a very good agreement beyond change.

The mean number of reported STARD items in the 24 accuracy reports published in 1999 was 12.1 (SD 3.3) with a range of 6.5 to 20 on a scale from 0 to 25 points. In 2004, the 27 included articles reported a mean of 12.4 (SD 3.2) items with a range of 7 to 17.5 (Fig. 2). The mean difference was a nonsignificant 0.3 points (\( P=0.7 \)). No subgroup analysis was performed. In both years, slightly less than half of studies (11/24 and 13/27) reported on more than 50% of the items (≥12.5 items). The maximum number of reported items in a single article in 1999 was 20 of the 25 items versus 17.5 of the 25 items in 2004.

The reporting of individual items showed a wide variation (0% to 96%), which is shown in Table 1. Although we interpreted item 1 more broadly (allowing
“diagnostic” to be read as “prognostic”), articles about diagnostic and prognostic accuracy were indexed poorly, with only 29% and 19% of articles clearly indexing the study as an article reporting on diagnostic or prognostic accuracy in the way that is recommended by the STARD committee. The aim of the study was stated in more than 80% of articles, with an adequate description of the test under study, the reference test, and target condition or outcome of interest.

The methods section lacked a detailed definition of the study population in which the tests were carried out in 30% to 40% of studies. Of the 16 items in the method section -item 8, 9, 10, 11 and 13 counted twice, for index test and reference test separately- only three of these were reported in more than 75% of the articles in 1999, a number that increased to six in 2004. For 1999 these items were the description of participant recruitment, of data collection and adequate description of the index test (items 4, 6 and 8a). In 2004, these items were participant recruitment, participant sampling, data collection, description of the index test, and definition of and rationale for units and cutoffs for the index test and reference test.
standard (items 4, 5, 6, 8a, 9a and 9b). The reference standard and its rationale or the outcome of interest was only reported in 52% of the studies.


<table>
<thead>
<tr>
<th>Category and item no.</th>
<th>1999</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title, abstract and keywords</strong></td>
<td>(n=24)</td>
<td>(n=27)</td>
</tr>
<tr>
<td>1. Identify the article as a study of diagnostic accuracy (recommend MESH heading &quot;sensitivity and specificity&quot;)</td>
<td>7 (29)</td>
<td>5 (19)*</td>
</tr>
<tr>
<td><strong>Introduction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. State the research question or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.</td>
<td>21 (88)</td>
<td>22 (81)</td>
</tr>
<tr>
<td><strong>Methods</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Describe the study population: the inclusion and exclusion criteria, setting, and locations where data were collected.</td>
<td>17 (71)</td>
<td>17 (63)</td>
</tr>
<tr>
<td>4. Describe participant recruitment: was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had undergone the index test(s) or the reference standard?</td>
<td>20 (83)</td>
<td>26 (96)</td>
</tr>
<tr>
<td>5. Describe participant sampling: was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.</td>
<td>15 (63)</td>
<td>21 (78)</td>
</tr>
<tr>
<td>6. Describe data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?</td>
<td>19 (79)</td>
<td>23 (85)</td>
</tr>
<tr>
<td>7. Describe the reference standard and its rationale.</td>
<td>10 (42)</td>
<td>14 (52)</td>
</tr>
<tr>
<td>8. Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. For the index test</td>
<td>23 (96)</td>
<td>24 (89)</td>
</tr>
<tr>
<td>b. For the reference standard</td>
<td>16 (67)</td>
<td>16 (59)</td>
</tr>
<tr>
<td>9. Describe definition of and rationale for the units, cut-offs and/or categories of the results of the index test and the reference standard.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. For the index test</td>
<td>16 (67)</td>
<td>25 (93)</td>
</tr>
<tr>
<td>b. For the reference standard</td>
<td>16 (67)</td>
<td>21 (78)</td>
</tr>
<tr>
<td>10. Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. For the index test</td>
<td>6 (25)</td>
<td>5 (19)</td>
</tr>
<tr>
<td>b. For the reference standard</td>
<td>2 (8)</td>
<td>2 (7)</td>
</tr>
<tr>
<td>11. Describe whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. For the index test</td>
<td>10 (42)</td>
<td>7 (26)</td>
</tr>
<tr>
<td>b. For the reference standard</td>
<td>4 (17)</td>
<td>6 (22)</td>
</tr>
<tr>
<td>12. Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. For the index test</td>
<td>14 (62)</td>
<td>13 (48)</td>
</tr>
<tr>
<td>b. For the reference standard</td>
<td>5 (27)</td>
<td>1 (4)</td>
</tr>
</tbody>
</table>
Results

14 Report when study was done, including beginning and ending dates of recruitment.
15 Report clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).
16 Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).
17 Report time interval from the index tests to the reference standard, and any treatment administered between them.
18 Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.
19 Report a cross tabulation of the results (including indeterminate and missing results) by the results of the reference standard; for continuous results report the distribution of the test results by the results of the reference standard.
20 Report any adverse events from performing the index tests or the reference standard.
21 Report measures of diagnostic accuracy and measures of statistical uncertainty (e.g., 95% confidence intervals).
22 Report how indeterminate results, missing responses, and outliers of the index tests were handled.
23 Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.
24. Estimates of test reproducibility, if done
   a. For the index test
   b. For the reference standard

Discussion

25. Discussion of the clinical applicability of the study findings.

* Number in parentheses expresses percentage
¶ Mean STARD score, 12.1 ± 3.3; range 6.5-20
¶¶ Mean STARD score, 12.4 ± 3.2; range 7-17.5

Information concerning the index test was generally better reported than that for the reference standard (items 8-13 and 24). Notably information regarding the number and training of the persons executing and evaluating the index test and reference test, and blinding of readers to the tests was reported in a minority of reports. In 2004, methods for calculating measures of diagnostic or prognostic accuracy, such as sensitivity, specificity, likelihood ratios, predictive values and ROC curves, were reported in only three articles. These were also the only articles that reported a statistical method used to quantify uncertainty, in all cases 95% confidence intervals. Indices of accuracy and the corresponding confidence intervals were reported in another three articles. Reporting indices of accuracy was done more often in articles in 1999. However, because the STARD statement invites authors to report confidence intervals as well, no more than 2 articles fulfilled item 12 (data not shown).
Quality of reporting of diagnostic studies

The study population in which tests were executed was reported in 71% and 63% of articles. Of the 11 items in the result section, study period, clinical and demographic characteristics, and time interval between tests were reported with a frequency of more than 60%. A flow diagram, which is strongly encouraged by the STARD steering committee, was presented in only 3 of the 51 articles.

Discussion

The results of this study indicate that the quality of reporting in articles on test accuracy in reproductive medicine is poor to mediocre. For both publication years examined, more than 50% of articles report less than 50% of STARD items. The mean number of reported STARD items was 12.1 for the articles published in 1999 and 12.4 for articles published in 2004. If we compare our results to data on quality of reporting from other fields of medicine, the standards of reporting in reproductive medicine are quite comparable to these other specialities. Articles published in the year 2000 in journals with impact factor of four or higher that regularly publish articles on diagnostic accuracy, reported 11.9 STARD items on average, with only 41% of articles reporting on more than 50% of STARD items (16). In ophthalmology, only 44% of the 16 diagnostic accuracy studies published in 2002 in the five leading ophthalmic journals reported more than half of the STARD items (68).

Description of study population was reported in 60% to 70% of the articles. Although this may seem a reasonable score, in our opinion this percentage is worrisome. An absent or incomplete description of the study population affects the generalizability of the study. Empirical evidence exists that this may be accompanied by overestimated test accuracy (2).

The indexing of the articles as diagnostic studies fulfilled the criteria of the STARD list in only a minority of reports. This is of concern due to the following reason: with the number of diagnostic meta-analyses growing, electronic databases have become indispensable tools to identify studies. To facilitate retrieval of their diagnostic study, authors should identify their studies as such. STARD recommends the use of the term “diagnostic accuracy” in the title or abstract of a report that compares the results of index test(s) with that of a reference standard. The MESH
heading “Sensitivity and Specificity” is also recommended. Several reports however, have shown that a search with this term does not identify all accuracy studies, while incorrectly identifying many articles as such. “Sensitivity and Specificity” may therefore be a less suitable set of keywords (69-70). Using the term “diagnostic accuracy” in title or abstract would enable the reader to easily and quickly identify a test accuracy study as such.

In the present study, we only valued the quality of reporting (defined as the number of reported STARD items), not the methodological quality or the degree of bias of the studies. If a study reported that the evaluation of the index and reference test was not blinded and this was clearly stated as such in the report, item 11 was considered to be well reported. Yet this may indicate a possible methodological shortcoming, depending on the amount of subjectivity associated with reading both tests. Good reporting though allows the reader-clinician to judge the potentials for bias in a study. The Quality Assessment of Studies of Diagnostic Accuracy Included in Systematic Reviews (QUADAS) tool, which contains items concerning bias in diagnostic accuracy studies, is a checklist that can be used to judge the potential for bias in a diagnostic accuracy study, and has been developed to be used in systematic reviews of diagnostic test accuracy (71). Some of the items listed in the QUADAS tool, were also selected by the US Department of Health and Human Services in their evidence-based practice report on systems to rate the strength of scientific evidence (www.ahrq.gov/clinic/tp/strengthtp.htm).

Confidence intervals (CIs), which are important to judge the reliability of the estimates of diagnostic accuracy, were reported in only eight studies. It is noteworthy that, although in studies of effects CIs are more or less obligatory; these are much less being reported in studies of diagnostic accuracy. Like measures of effect in therapeutic trials, estimates of diagnostic accuracy are also subject to chance variation, and authors should quantify the amount of uncertainty around their observed values. Methods for calculating the precision around frequently used measures of diagnostic accuracy are readily available in literature (72, 73). One other review revealed that CIs were reported in only 50% of 16 diagnostic evaluation reports published in the BMJ in 1996 and 1997 (74). In our study confidence intervals were reported in only 16% of studies.
Quality of reporting of diagnostic studies

In this study, we evaluated whether or not information on each of the 25 STARD items was reported. Although we have referred to this evaluation in this manuscript as the “quality of reporting”, the latter may be defined by an additional number of features, such as clarity of the language used for reporting, absence of ambiguity, consistency between abstract, methods, results and discussion. Such a more detailed analysis was beyond the scope of our paper.

It has been shown that the introduction of a guideline on how to report randomized controlled trials has led to improvement in the quality of reporting of these studies. Articles in journals that had adopted CONSORT showed greater improvement in quality of reporting than did articles in a journal that did not advocate its use (75). A recent study provided additional evidence to suggest a role for checklists. Articles published in Clinical Chemistry - a journal that adopted a preliminary checklist for the report of diagnostic accuracy studies as early as 1996 - showed a greater improvement in reporting from 1996 to 2002 than articles published in Clinical Chemistry and Laboratory Medicine, a journal that did not use a similar checklist before (76).

Our study was unable to detect a significant improvement of the quality of reporting between 1999 and 2004 in a “before-and-after STARD” evaluation. A possible explanation may be that the time involved between submitting an article to a journal and the final acceptance for publication (the so-called “peer-review time”) did not allow authors to incorporate the STARD-guidelines into their articles. Another possibility might be that authors, reviewers and editors are less familiar with the current STARD checklist than with others such as CONSORT.

Possible limitations of our study are the relatively small number of studies included, and the fact that we applied the STARD list also to prognostic accuracy studies. As far as the first limitation is concerned, we decided to include studies published from January 2004 onward, specifically with the aim of reducing the effect of the above mentioned peer review time. This obviously limited the number of eligible articles. With respect to the second limitation, we believe that it is justified to apply the STARD-criteria to prognostic accuracy studies as well, as in reproductive medicine these studies have a striking similarity with diagnostic test accuracy studies. For the evaluation of the prognostic accuracy of tests, transparent reporting of study...
execution, participants, flow of participants throughout the study, statistical uncertainty, definition of outcome and rationale, reproducibility of test results, and so on, is as essential as it is with diagnostic test accuracy studies. In conclusion, this study shows that the quality of reporting on diagnostic and prognostic accuracy studies in reproductive medicine is comparable to other fields of medicine but leaves ample room for improvement. Authors, reviewers, editors and readers should be aware of the STARD criteria for reporting on diagnostic/prognostic accuracy studies. We welcome stricter adherence to this guideline.

Acknowledgements
The authors thank Nynke Smidt, Ph.D., from the Institute for Research in Extramural Medicine, VU University Medical Center Amsterdam and Anne W.S. Rutjes, Ph.D., from the department of Clinical Epidemiology and Biostatistics of the Academic Medical Centre, University of Amsterdam, for generously providing the STARD scoring form.
References


55. Van Rooij IAJ, de Jong E, Broekmans FJM, Looman CWN, Habbema JDF, te Velde ER. High follicle-stimulating hormone levels should not necessarily lead to the exclusion of subfertile patients from treatment. Fertil Steril 2004;81:1478-1485.


