The quality of reporting of diagnostic accuracy studies in pelvic floor ultrasound: A systematic review

A.T.M. Grob MSc\textsuperscript{1,2}, L.R. van der Vaart\textsuperscript{1}, dr. M.I.J. Withagen\textsuperscript{1}, prof. dr. C.H. van der Vaart\textsuperscript{1}

\textsuperscript{1}Department of Reproductive Medicine and Gynecology, University Medical Center, Utrecht, Netherlands
\textsuperscript{2}MIRA Institute for Biomedical Technology and Technical Medicine, University of Twente, Enschede, Netherlands
\textsuperscript{3}Medical student, Amsterdam Medical Center, Amsterdam, Netherlands

\textbf{Correspondence to:}

Anique T.M. Grob, Department of Reproductive Medicine and Gynecology, University Medical Center, PO box 85500, 3508 GA Utrecht, Netherlands; email: a.t.m.grob-3@umcutrecht.nl; phone: +31624935494

\textbf{Source of funding:} None

\textbf{Financial Disclosure:} The authors did not report any potential conflicts of interest.

\textbf{Short title:} STARD analysis on Pelvic Floor Ultrasound

\textbf{Keywords:} pelvic floor ultrasound, diagnostic accuracy study, STARD checklist, quality reporting
Introduction – In recent years, a large number of studies have been published on the clinical relevance of pelvic floor three-dimensional (3D) transperineal ultrasound. Several studies compare ultrasonography to other imaging modalities or clinical examination. The quality of reporting these studies is not known.

Objective/Purpose – To determine the compliance of diagnostic accuracy studies investigating pelvic floor 3D ultrasound, with Standards for Reporting of Diagnostic Accuracy (STARD) guidelines by means of a systematic review.

Method – This study reviewed published articles on pelvic floor 3D ultrasound identified by means of a systematic literature search that included the MEDLINE, Web of Science and SCOPUS databases. Prospective and retrospective studies that compared pelvic floor 3D ultrasound with other clinical and imaging diagnostics were included in the analysis. STARD compliance was assessed and quantified by two independent investigators, using 22 of the original 25 STARD checklist items. Items with the qualifier “if done” (items 13, 23 and 24) were excluded because they were not applicable to all papers. Each item was scored as reported (score = 1) or not reported (score = 0). We calculated observer variability, the total number of reported STARD items per article and summary scores for each item. We also statistically tested the difference in total score between STARD adopting and non-adopting journals, as well as the effect of year of publication.

Results – Forty studies published in 13 scientific journals were included in the analysis. The mean score (SD) of articles included was 16.0 (2.5) out of a maximum of 22 points. The lowest scores, below 55%, were found on quality of reporting on handling of indeterminate results or missing responses, adverse events and time interval between tests. Interobserver rating agreement of STARD items was substantial (ICC 0.77). The independent t-test showed no significant mean differences (±SD) in total score between the adopting 15.9 (± 2.6) and non-adopting 16.1 (± 2.5) journals. The mean STARD score for the period 2003-2009 (15.2 ± 2.5) was lower, but not statistically significant different as compared to the period 2010-2015 (16.6 ± 2.4).

Conclusion – The overall compliance with reporting guidelines of diagnostic accuracy studies of pelvic floor 3D transperineal ultrasound is relatively good, compared to other fields. However specific items should require more attention when reported.

Keywords: pelvic floor 3D ultrasound, diagnostic accuracy study, STARD checklist, quality reporting
Introduction

In everyday clinical practice history taking and physical examination are often followed by the execution of diagnostic tests. These tests aim to provide additional information on the nature and severity of the disease, and to reduce uncertainty about the diagnosis. In the end, the information gathered by diagnostic tests should improve outcome for the patient to an extent that would not have been reached without the test.

To improve awareness of the quality of reporting and to overcome over- or underestimating the outcome under study, a group of methodological researchers and editors developed the Standards for Reporting of Diagnostic Accuracy (STARD) statement [1]. The STARD checklist and the corresponding flowchart, which were published in 2003, are intended to support authors in reporting essential study elements of diagnostic research [2].

To evaluate the efficacy of an index test in comparison with the reference standard, studies on diagnostic accuracy are executed [9]–[15]. These reports help clinicians to decide if a diagnostic test is suitable for the disorder of interest. In order to be able to assess the value and to interpret the results of such studies, a clinician should be able to rely on accurate reporting of relevant study characteristics. In addition, a poorly reported article on diagnostic accuracy limits the possibility to identify potential bias and value the report.

Ultrasound as an imaging and diagnostic tool has become more and more important in the field of urogynecology in the recent years. Its clinical ability and added value is used and described in many studies [3]–[8]. Since ultrasound imaging is a diagnostic instrument, information on the sensitivity and specificity of the images and retrieved parameters is crucial, to judge its clinical application.

To our knowledge, no evaluation has been performed to investigate STARD compliance in reporting diagnostic studies on 3D ultrasound of the pelvic floor. The objective of this systematic review was to determine the compliance of diagnostic accuracy studies of pelvic floor ultrasound with the STARD guideline.

Materials and Method

Data sources

To identify papers, a systematic literature search was performed. We searched MEDLINE (by using PubMed), Web of Science and SCOPUS, by using the MESH-search terms for the index test (ultrasound), and the anatomy being investigated (pelvic floor) as well as their synonyms. Additional search limits applied were “human”, “female”, “English” and “01-01-2003 to 12-31-2015” We performed our search in February 2016. We only included publications after 2003 since this was the first year that the guidelines were reported. An article was
excluded if children were studied, the study was a systematic review, the abstract or manuscript were missing or the technology (3D transperineal ultrasound) was incorrect.

Paper selection
We first screened titles and then abstracts and in a second step, read all full-text articles to evaluate all remaining potentially eligible studies for inclusion and exclusion criteria. Articles had to meet the following inclusion criteria: the study focus needs to be on one or more of the four major pelvic floor lines of interest (bladder neck mobility, genital hiatus, avulsion or prolapse), or related medical term.

Studies with a predictive design were excluded since these manuscript should be checked on quality of reporting following the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) guideline [16]. TRIPOD as a guideline is comparable to STARD, but with a focus on reporting full information on all aspects of a prediction modelling study, risk of bias and potential usefulness of the prediction model [16]. When the study only focused on reliability of one technique the study was excluded. Exclusion was also performed when the objective of a study was to define clinical diagnosis (e.g. “Is there an avulsion?”), rather than comparing ultrasound to another modality (e.g. physical examination/MRI) on a specific clinical measure (e.g. comparing sensitivity/specificity of avulsions based on ultrasound images compared to avulsions based on MRI).

Data extraction
We evaluated the quality of reporting by using the standard 25 questions STARD checklist. STARD is designed as a checklist to improve the completeness and transparency of reporting studies of diagnostic accuracy, to allow readers to assess the potential for bias in the study (internal validity) and to evaluate its generalisability (external validity). In this study, as did Walther in 2014, we changed the “design” of STARD in two ways. First we eliminated three items compared to the original checklist and second, we used a checklist meant for author publishing guidelines as a retrospective scoring instrument. We eliminated the same (if done) checklist items from scoring (13, 23 and 24) as did Walther and Wilczynski [17-18]. We did this given the same argument “If these items were not reported in the diagnostic accuracy papers evaluated, it would be impossible to determine whether this lack of reporting was because the item was not done or because it was not reported”. Our design differs from the study by Walther and co-workers, since they additionally eliminated item 9 from analysis since all papers scored a full 100% on this item, an exclusion that cannot be predefined. Additionally, our study differs from the study...
by Wilczynski and co-workers, because they only studied items “that have been empirically shown to have a potentially biasing effect on the results of diagnostic accuracy studies and those items that appear to account for variation between studies”. We did not exclude items based on an empirically shown potentially biasing effect, since variation in the items might well been found between different clinical domains.

The STARD checklist is known to have good reproducibility [17], [19]. The included articles were independently scored by the same two reviewers blinded for each other’s results. After scoring 12 papers, a consensus meeting was scheduled to make sure the perception of the STARD criteria context was aligned between reviewers, and to discuss potential discrepancies. After the consensus meeting both reviewers independently evaluated the remaining papers. Discrepancies in the analysis were resolved in consensus.

**Data and statistical analysis**

Equal weights were given to all items, each STARD item was scored as reported (score = 1) or not reported (score = 0). When multiple items were described within one STARD item (items 3,8,9,10,12,14,15,17,21,22) 1 point was granted when at least one of the items was described in the text. Items 8-11 and 20 on the checklist concern both the index test, as well as the reference standard. We defined that, since the outcome of a study can only be interpreted accurately when both the reference as well as the index test are described, to score 0 points if information on non or only for one of the tests was presented and 1 point when both tests were described. The total score for each article according to the STARD checklist was calculated by summing the awarded points for the 22 items included. Items were considered well reported when they were found in more than 80% of the papers and were poorly reported when the items was found in less than 50% of the papers.

Agreement between reviewers, as a measure of subjectivity of the assessment, was calculated by using Cohen k statistics. For calculating the Cohen k statistics, only the remaining papers after the consensus meeting were included. According to Landis and Koch [20], a k value of 0.41-0.60 indicates moderate agreement between the reviewers; a k value of 0.61-0.80 substantial agreement and a k value of 0.81-1.00, almost perfect agreement.

Normally distributed data are reported as means +/- standard deviations (SD), and percentages are reported with 95% confidence intervals (CI). To test for differences in total scores between STARD adopting and non-adopting journals, we performed an independent samples t-test. To determine whether a journal was STARD adopting, we contacted all journals included in this review and asked them whether the journal was STARD adopting, and
if so the year of implementation. The online author guidelines per journal were checked in July 2016 in case of non-responders. To check for the correlation between year of publication and total STARD score the Pearson correlation coefficient is calculated. In addition, an independent t-test is performed to compare the STARD score for the first period of the study (2003-2009) and second period (2010-2015). Statistical analyses were performed with statistical software (SPSS Statistics 20).

Results

The systematic literature search yielded 1808 eligible papers. After excluding manuscript based on the criteria “duplicate”, “children”, “review”, “missing abstract” and “missing article”, and scoring the abstracts on the conditions under study (bladder neck mobility, avulsion, prolapse and genital hiatus) 1552 papers were excluded, leaving 256 papers for further analyses. Thirty-three studies did not report on 3D/4D transperineal ultrasound and 27 studies were prognostic (TRIPOD scoring list) instead of diagnostic studies, leaving 223 papers. Of these 223 papers 39 studies were observer reliability studies, and 117 studies did not compare ultrasound to another diagnostic method (e.g. POP-Q, MRI), leaving 40 papers based on transperineal ultrasound imaging compared to a gold standard (STARD) [21-60]. Figure 1 shows a flowchart representing the inclusion of papers.
Overall, the STARD score of the included papers ranges from 11 to 21 with a mean of 16.0 (SD = 2.5). The reporting of each item is presented in Table 1. The best reported item was item 15, which refers to reporting demographic characteristics of the study population. All papers fulfilled the requirements for scoring on this item. Other well reported items (>80%) were item 1, 2, 4, 6, 7, 8, 9, 12, 19 and 25. Three items (17, 20 and 22) were especially poorly reported (<50%). None of the papers mentioned how indeterminate results were handled and only two papers describes the occurrence of an adverse event. Only 35% of all papers mentioned the time interval.
between tests, 52.5% reports on distribution of severity of disease and 55% describes whether the study population was a consecutive study or how patient were further selected.

The 40 papers were published in 14 different medical journals, of whom four journals (14 papers) advised the authors to use the STARD checklist in their author guidelines (Table 2). The independent t-test showed no significant mean differences (±SD) in total score between the adopting 15.9 (± 2.6) and non-adopting 16.1 (± 2.5) journals. Over time the mean STARD score per year increased (Pearson correlation coefficient 0.28, p=0.08) shown in Figure 2. The mean STARD score for the period 2003-2009 (15.2 ± 2.5) was lower as compared to the period 2010-2015 (16.6 ± 2.4).

Overall agreement of the reviewers in scoring the STARD items was 90.5% The k value was 0.77 (95% CI: 0.72-0.80), indicating substantial agreement between the reviewers. The items with the highest disagreement were items 5 (consecutive patient sampling), 17 (time interval reporting), 18 (distribution of severity of disease) and 19 (reporting of a the distribution of test results) with ICC values of 0.51; 0.66; 0.50 and 0.33 respectively.
The STARD checklist strongly recommends the use of a flow diagram to illustrate the key elements of the study design and the patient flow though the study setup, such a diagram was only provided in one paper.

Discussion

In this study we assessed the quality of reporting of pelvic floor 3D ultrasound. The results of our study indicate that the quality of diagnostics accuracy reports is fair but not optimal. The mean score of 16.2 out of 22 points indicates that there is room for further improvement. However, within the timeframe we selected for our review we could demonstrate that the adherence to the STARD criteria is improving. Interestingly, papers in journals who explicitly state in their guidelines to authors that diagnostic study reports should adhere to the STARD criteria, do not perform better as compared to journals who do not mention the STARD criteria.

To the best of our knowledge this is the first time that the reporting of quality of pelvic floor 3D ultrasound as a diagnostic tool has been studied. When we compare our findings to other papers on the quality of reporting diagnostic accuracy in other fields of medicine, we found that our average score of 16.0 out of 22 points (73%) was relatively high. In the paper by Zafar and co-workers a different scoring scale was used, but with a mean score of 19.8 out of a 50 points (40%) their relative score is poorer. The same accounts for the study performed by Areia and co-workers, who used the original, 25-points scale. They found a mean score of 12.2 out of 25 points (49%) [12], [14]. Several explanations for this difference can be given. First the other studies were performed in 2010 (including published articles from 1998 to 2008) and in 2008 (including published articles from 1995 to 2006) and as we also have shown, the quality of reporting in the early period was poorer as compared to the later period. The second explanation could be that we scored to liberal or other authors to strict. Since the STARD criteria leave room for interpretation this potential bias cannot be ruled out. Since our interobserver reliability was good we believe our results accurately represent the current status of reporting diagnostic studies in pelvic floor ultrasound. Finally, it is not always obvious, based on title or abstract, that the paper is a diagnostic study and therefore need to be checked for the STARD criteria. This is supported by the fact that journals who specifically state that STARD criteria needed to be used, did not perform better that other journals who did not mention STARD in the author guidelines.

When comparing the scores on individual items to the studies by Paranjothy (2007), Areia (2010) and Zafar 2008 we found that the items 17, 20 and 22 were consistently poorly reported (<50%).[10], [12], [14]. Especially items
22 “Report how indeterminate results, missing responses and outliers of the index tests were handled”, is one of crucial information since it is related to the potential risk of selection bias. Authors should inform their readers on this in a consistent way. Another potential contribution to bias is the lack of reporting the distribution of severity of disease in those with the target condition (item 18). This is only described in 52.5% of the reports, disabling readers to determine whether study results are defined for all severities of disease. These items are thus clearly underreported in current papers and should particularly be paid attention to by researches in this area when they submit their work for peer review.

To reduce our own observer bias in our evaluation of the quality of reporting, each paper was independently evaluated by two reviewers. The interrater reproducibility indicated substantial agreement. This is in line with the results of Smidt and co-workers [19], who investigated the interrater reproducibility of the STARD checklist for evaluating studies of diagnostic accuracy. However, certain items were more reliably scored than others. Items, that in itself contained multiple questions were found more difficult to score with a 0 or 1. These items (e.g. 3 and 17) needed more discussion to reach consensus between the reviewers.

Our analysis was based on the original STARD checklist that was published in 2003. However, last year, the STARD group published an updated STARD checklist [61] and most STARD adopting journals now recommend to use the updated version. We deliberately have chosen to score all papers based on the 2003 STARD guidelines. As we believe that scoring a paper based on a checklist that was not available at the moment of writing the original study is not valid.

Our study had some limitations. The first relates to the scoring method of granting 0 or 1 point per item. When multiple items were described within one STARD item a paper was granted 1 point when at least one of the items was described in the text. Following the scoring as described by Zafar and co-working (score 2 = completely reported; score 1 = partly reported and score 0 is not reported) might have provided more detail, however this approach reduces the observer reliability [14]. A second limitation of our study was that the exclusion of three items represents a deviation from the original checklist. Nevertheless, we believe the exclusion was justified, because the “if done” items 13, 23 and 24 were not applicable to all studies. A final limitation was our choice to limit our search to the English language report, although we believe that inclusion of studies published in other languages would not alter our conclusion.
In conclusion, our study demonstrates that the overall compliance with reporting guidelines of studies addressing diagnostic accuracy of pelvic floor 3D ultrasound has improved over time and is relatively good compared to other fields of medicine. However specific items should require more attention when reported.
References


[10] Paranjothy B, Shunmugam M, Azuara-Blanco A. The quality of reporting of diagnostic accuracy studies in


[20] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;


Oversand SH, Stan IK, Shek KL, Dietz HP. The association between different measures of pelvic floor muscle function and female pelvic organ prolapse. *Int Urogynecol J* 2015; 26: 1777-1781.


Table 1 The reporting of the STARD items.

<table>
<thead>
<tr>
<th>Item</th>
<th>Topic</th>
<th>Reported (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identify the article as a study of diagnostic accuracy (recommended MeSH heading 'sensitivity and specificity')</td>
<td>39 (97.5)</td>
</tr>
<tr>
<td>2</td>
<td>State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.</td>
<td>39 (97.5)</td>
</tr>
<tr>
<td>3</td>
<td>Describe the study population: the inclusion and exclusion criteria, setting and locations where the data were collected</td>
<td>30 (75)</td>
</tr>
<tr>
<td>4</td>
<td>Describe participant recruitment: was recruitment: based on presenting symptoms, results from previous test, or the fact that the participants had received the index tests or the reference standard</td>
<td>36 (90)</td>
</tr>
<tr>
<td>5</td>
<td>Describe participant sampling: was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.</td>
<td>22 (55)</td>
</tr>
<tr>
<td>6</td>
<td>Describe data collection: was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?</td>
<td>39 (97.5)</td>
</tr>
<tr>
<td>7</td>
<td>Describe the reference standard and its rationale</td>
<td>38 (95)</td>
</tr>
<tr>
<td>8</td>
<td>Describe technical specification of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard</td>
<td>39 (97.5)</td>
</tr>
<tr>
<td>9</td>
<td>Describe definition and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard</td>
<td>35 (87.5)</td>
</tr>
<tr>
<td>10</td>
<td>Describe the number, training and expertise of the persons executing and reading the index tests and the reference standard</td>
<td>26 (66.7)</td>
</tr>
<tr>
<td>11</td>
<td>Describe whether or not the readers of the index tests and the reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers</td>
<td>28 (70)</td>
</tr>
<tr>
<td>12</td>
<td>Describe methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% CI)</td>
<td>34 (85)</td>
</tr>
<tr>
<td>14</td>
<td>Report when study was done, including beginning and ending dates of recruitment</td>
<td>28 (70)</td>
</tr>
</tbody>
</table>
Report clinical and demographic characteristics of the study population (e.g., age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers) 40 (100)

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) 25 (62.5)

Report time interval from the index test to the reference standard, and any treatment administered between 14 (35)

Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target standard 21 (52.5)

Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard 37 (92.5)

Report any adverse events from performing the index test or the reference standard 2 (5)

Report estimated diagnostic accuracy and measures of statistical uncertainty (e.g., 95% CI) 29 (72.5)

Report how indeterminate results, missing responses and outliers of the index tests were handled 0 (0)

Discuss the clinical applicability of the study findings 39 (97.5)

Table 2 Medical journals included - representing number of papers included in STARD analysis, compliance to STARD in author guidelines and Impact Factor 2015 (* retrieved from author guidelines, ¯ retrieved from editorial board response)

<table>
<thead>
<tr>
<th>Journal</th>
<th>Frequency (%)</th>
<th>STARD adopting (year)</th>
<th>Impact Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>American Journal of Obstetrics &amp; Gynecology</td>
<td>1 (2.5)</td>
<td>Yes (2004)*</td>
<td>4.681</td>
</tr>
<tr>
<td>Australian and New Zealand Journal of Obstetrics and Gynaecology</td>
<td>3 (7.5)</td>
<td>No*</td>
<td>1.738</td>
</tr>
<tr>
<td>BJOG: An International Journal of Obstetrics and Gynaecology</td>
<td>2 (5.0)</td>
<td>Yes* (n.a.)</td>
<td>4.039</td>
</tr>
<tr>
<td>Brazilian Journal of Physical Therapy</td>
<td>1 (2.5)</td>
<td>No*</td>
<td>0.979</td>
</tr>
<tr>
<td>British Journal of Radiology</td>
<td>1 (2.5)</td>
<td>No*</td>
<td>1.840</td>
</tr>
<tr>
<td>European Journal of Obstetrics &amp; Gynecology and Reproductive Biology</td>
<td>1 (2.5)</td>
<td>No*</td>
<td>1.662</td>
</tr>
<tr>
<td>Female Pelvic Medicine &amp; Reconstructive Surgery</td>
<td>1 (2.5)</td>
<td>No*</td>
<td>1.331</td>
</tr>
<tr>
<td>International Journal of Colorectal Disease</td>
<td>3 (7.5)</td>
<td>No*</td>
<td>2.383</td>
</tr>
<tr>
<td>International Urogynecology Journal</td>
<td>13 (32.5)</td>
<td>No*</td>
<td>1.834</td>
</tr>
<tr>
<td>Medical Ultrasonography</td>
<td>1 (2.5)</td>
<td>No*</td>
<td>1.167</td>
</tr>
<tr>
<td>Neurology and Urodynamics</td>
<td>2 (5.0)</td>
<td>No*</td>
<td>3.128</td>
</tr>
<tr>
<td>Obstetrics &amp; Gynecology</td>
<td>2 (5.0)</td>
<td>Yes (2004)*</td>
<td>5.656</td>
</tr>
<tr>
<td>Ultrasound in Obstetrics &amp; Gynecology</td>
<td>9 (22.5)</td>
<td>Yes (2011)*</td>
<td>4.254</td>
</tr>
</tbody>
</table>